Microbiome and its role in shaping planetary health

Microbiome and Its Role in Shaping Planetary Health

Ankit Singh Tanwar

Cover Page Story

The cover picture represents two microbial ecosystems existing, on one hand it represents microbial cultures grown in a petri dish (lab ecosystem) on the other hand it represents microbial diversity on planet Earth (global ecosystem). It signifies how using lab grown microbial cultures, humanity can understand global microbial diversity of our planet.

Cover Photo by

Daniel Olah on Unsplash

ISBN: 978 94 6469 775 9

Microbiome and Its Role in Shaping Planetary Health

DISSERTATION

To obtain the degree of Doctor at Maastricht University on the authority of the Rector Magnificus

Prof. Dr. Pamela Habibović,

in accordance with the decision of the Board of Deans, to be defended in public on Wednesday the 24th of January 2024 at 10:00 hrs

By

Ankit Singh Tanwar

Born March 12th, 1994, in Vishakhapatnam, India

Supervisor:

Prof. Dr. Angela Brand

Co-Supervisor:

Prof. Dr. Kapaettu Satyamoorthy, Shri Dharmasthala Manjunatheshwara (SDM) University, Dharwad, India

Assessment Committee:

Prof. Dr. Kasia Czabanowska, Maastricht University (Chair)

Prof. Dr. Frank Bier, University of Potsdam, Germany

Dr. Harald Peter, Fraunhofer Institute for Cell Therapy and Immunology (IZI-BB), Germany

Dr. Lars M.T. Eijssen, Maastricht University, The Netherlands

Table of Contents

Abbreviations

AMR – Antimicrobial resistance

ARA – Average Relative Abundance

ARGs – Antibiotic Resistance Genes

AST – Antimicrobial Susceptibility Testing

CARD – Comprehensive Antibiotic Resistance Database

CDI – *Clostridioides difficile* infections

CICs – Cancer-initiating cells

COGs – Cluster of Orthologues

CTD – Comparative Toxicogenomics Database

DCA – Deoxycholic acid

DestinE – Destination Earth

DFU – Diabetic Foot Ulcer

EC – Enzyme Commission

ECDC – European Centre for Disease Prevention and Control

EHRs – Electronic Health Records

EMT – Epithelial-mesenchymal transition

EOSC – European Open Science Cloud

ETBF – Enterotoxigenic *Bacteroides fragilis*

FMT – Fecal Microbiota Transplantation

GBIT – Genome-based Information and Technologies

GDC – Genomic Data Commons

GHDx – Global Health Data Exchange

GOHDCC – Global Open Health Data Cooperatives Cloud

HDCC – Health Data Cooperatives Cloud

HDCs – Health Data Cooperatives

HPV – Human Papillomavirus

ICGC – International Cancer Genome Consortium

IHR – International Health Regulations

IoT – Internet of Things

LDA – Linear Discriminant Analysis

LEfSe – Linear discriminant analysis Effect Size

MIC – Minimal Inhibitory Concentration

MMPs – Matrix Metalloproteinases

NGS – Next Generation Sequencing

NIH – National Institute of Health

OSCC – Oral Squamous Cell Carcinoma

OSDC – Open Science Data Cloud

OSF – Oral Submucous Fibrosis

OTU – Operational Taxonomic Unit

PCA – Principal Component Analysis

PGHD – Patient-generated health data

SARS-CoV-2 – Severe Acute Respiratory Syndrome Coronavirus 2

SCFAs – Short-chain fatty acids

SSRIs – Selective serotonin reuptake inhibitors

TIMPs – Tissue inhibitors of metalloproteinases

VFDB – Virulence Factor Database

WGS – Whole Genome Sequencing

WHO – World Health Organization

Chapter

1

Introduction

The microbiome refers to the complex community of microorganisms that inhabit a particular environment, such as the human body. These microorganisms, including bacteria, fungi, and viruses, can have a significant impact on the health and well-being of their hosts (Methé et al., 2012; Shreiner et al., 2015). Within the microbiome, the virome, which comprises the collection of viruses, has also been identified as playing an important role in human health (Wylie et al., 2012).

There are several different types of microbiomes found in the human body, including the gut, oral, and skin microbiomes (Li et al., 2012). Each of these microbiomes plays a unique role in maintaining host health. For example, the gut microbiome aids in the digestion of food, synthesis of vitamins, and protection against pathogenic bacteria. The oral microbiome helps to prevent tooth decay and gum disease, whereas the skin microbiome helps to protect against infection and maintain skin integrity (Huttenhower et al., 2012).

Recent studies have shown that the virome may play a role in a variety of diseases and conditions. For example, human papillomavirus (HPV) and hepatitis B and C have been linked to certain types of cancer (Broecker and Moelling, 2021). Additionally, viruses have also been associated with neurological disorders such as multiple sclerosis and Alzheimer's disease (Wouk et al., 2021).

This growing body of evidence suggests that the microbiome and virome may play a role in the development and progression of a wide range of diseases and conditions. However, more research is needed to fully understand the role of these microorganisms in human health. Further studies are required to uncover the underlying mechanisms and interactions between microorganisms and host, and how it may impact the health outcomes. Thus, the microbiome and virome are important area of study and have the potential to lead to new diagnostic and therapeutic approaches for various diseases.

Human Microbiome

The human microbiome is a collection of microbes, their genetic material, and the substances they produce that live on or inside the human body. The combination of microorganisms within our bodies is unique to each individual, just like fingerprints, and even distinctive to each body site (Mousa et al., 2022). These microorganisms, including bacteria, viruses, fungi, and protozoa, are found in various parts of the body such as the skin, mouth, gut, and respiratory tract. Microbial diversity and abundance are influenced by a variety of variables, such as nutrition, host genetics, diseases, medications, and lifestyle (Shreiner et al., 2015). These elements come together to play a crucial role in maintaining a balanced and diverse healthy microbiota.

Evidence suggests that microbial dysbiosis is associated with several health conditions such as obesity, diabetes, inflammatory bowel disease (IBD), metabolic and mental disorders. Some studies finding suggests that changes in the microbiome are linked with conditions such as depression and anxiety (Doelman et al., 2021; Mousa et al., 2022; Ogunrinola et al., 2020). Numerous studies have linked microbiome dysbiosis to a specific disease, but the exact mechanisms behind this relationship are not well understood. It is thought that changes in the types and abundance of microorganisms may alter their interactions and the substances they produce, leading to changes in the host's metabolism and other bodily functions (Yadav et al., 2018).

Many factors can influence the relationship between the microbiome and the host, including environmental, epigenetic, and genetic factors. One example can be overuse of antibiotics, use of antibiotics can disrupt the microbiome balance, leading to the overgrowth of certain microorganisms and suppression of others. This can lead to the development of conditions such as *Clostridioides difficile* infection (CDI), a serious and potentially life-threatening condition that can occur after antibiotic treatment (Dilnessa et al., 2022).

It is well known that antibiotics, particularly broad-spectrum ones, can change the composition of the gut microbiota. However, research has also shown that other non-antibiotic substances, such as oral steroids, antidepressants, vitamin D supplements and platelet aggregation inhibitors can also impact the composition and diversity of the gut microbiota in certain populations (Vila et al., 2020; Weersma et al., 2020). For example, the prevalence of *Haemophilus parainfluenzae*, a bacterium that is more prevalent in people with irritable bowel syndrome, has been connected to the usage of benzodiazepines (Vila et al., 2020). *Bifidobacterium dentium* is preferentially more prevalent when using proton pump inhibitors. While tricyclic antidepressants increase the abundance of *Clostridium leptum*, Selective serotonin reuptake inhibitors (SSRIs) antidepressants enhance the abundance of *Eubacterium ramulus* (Vila et al., 2020).

Diet, microbiota, and host interactions are complex and multifaceted. The composition of the microbiome changes when the diet is altered. Due to differences in how food is metabolised, each person's microbiome reacts to diet differently and the composition of the microbiome influences the host's metabolic capacities (Johnson et al., 2019). The microbiome is increasingly thought to play a role in how diet affects the development and progression of some disorders, including celiac disease and inflammatory bowel syndrome (Glassner et al., 2020; Krishnareddy, 2019). Similar to diet, the composition of the microbiome is influenced by dietary supplements such as vitamins and minerals (Zimmer et al., 2012).

Exposure to stress can alter the makeup of the gut microbiota, which can increase the likelihood of developing certain diseases. Research has demonstrated that social stress can decrease the number of anti-inflammatory microbes, such as para *Bacteroides* taxa, in the gut, leading to a decrease in anti-inflammatory substances produced by the microbiota, such as short-chain fatty acids (SCFAs). This decrease in anti-inflammatory substances can contribute to an increase in inflammation (Maltz et al., 2019, 2018).

The role of host genetics in shaping the microbiota is not fully understood (Gaulke and Sharpton, 2018). However, it has been observed that individuals living in low-income countries tend to have a more diverse microbiome compared to those in Western countries. Interestingly, research has shown that immigrants from Southeast Asia who move to the United States experience a 15% loss in microbiome diversity immediately. This reduction in microbiome diversity has been linked to an increased risk of certain conditions, such as obesity and cardiovascular disorders. It is believed that this loss of diversity may be due to a combination of factors, such as a shift to a high-calorie Western diet, changes in drinking water, and the use of drugs and antibiotics (Vangay et al., 2018).

The changes in the microbiome structure that often occur in older individuals have been attributed to a variety of factors, such as changes in lifestyle and diet, reduced mobility and intestinal function, decreased immune

function, altered gut morphology, recurrent infections and increased use of medications and drugs (Kim and Benayoun, 2020). In the gut microbiomes, age-related patterns of association have been found. In three separate studies involving a total of 9,000 individuals, researchers identified a trend of healthy ageing marked by a reduction of key gut microbial taxa, specifically *Bacteroides*. Healthier individuals had increasingly distinct microbiome compositions in comparison to other study participants (Lau et al., 2021; Wilmanski et al., 2021).

The microbiome is greatly impacted by various aspects of an individual's lifestyle, such as their level of physical activity, smoking habits, drug use, and the environment in which they live. These factors can all play a role in determining microbiome diversity (Mousa et al., 2022). For example, exercise has been shown to increase microbiome diversity and can also affect the communication between the gut and brain (Dalton et al., 2019; Kim et al., 2018). Physically active people typically have more diverse microbiomes, which are frequently distinguished by an abundance of healthy bacteria including *Roseburia hominis, Faecalibacterium prausnitzii* and *Akkermansia muciniphila* (Bressa et al., 2017).

The composition of the microbiome is altered by the environment, which therefore influences disease susceptibility (Littleford-Colquhoun et al., 2019). Westernization, for example, is thought to reduce microbiome diversity and raise the risk of illnesses like obesity and infections (Winglee et al., 2017). Additionally, industrialization appears to be associated with higher levels of antibiotic resistance in the microbiome, as compared to the pre-antibiotic era. The industrialization era is also associated with a higher rate of horizontal gene transfer, which could enable the acquisition of new capabilities and an increased ability to adapt to changing conditions (Groussin et al., 2021; Lau et al., 2021).

Recent breakthroughs in microbiome research have expanded our knowledge of how the microbiome influences both the development and prevention of various diseases. While the specific mechanisms behind these effects are not yet fully understood, further research is needed to fully understand the role of the microbiome in shaping human health (Mousa et al., 2022).

Cancer and Microbiome

Cancer is a complex and devastating disease that affects millions of people worldwide. It is caused by the uncontrolled growth and spread of abnormal cells in the body, leading to the formation of tumors and other complications. According to the World Health Organization (WHO), cancer is one of the leading causes of death globally, accounting for an estimated 10 million deaths in 2020 (WHO, 2022). Cancer can occur in any part of the body and can take many different forms, each with its unique set of symptoms, causes, and treatments. Some of the most common types of cancer include lung cancer, breast cancer, prostate cancer, colon cancer, and skin cancer.

The causes of cancer can be genetic, environmental, or a combination of both. Some risk factors for cancer include smoking, exposure to radiation, a poor diet, lack of exercise, and certain infections such as human papillomavirus (HPV) and hepatitis B and C (American Cancer Society, 2022). The treatment of cancer often

involves a combination of different therapies, including surgery, radiation therapy, chemotherapy, targeted therapy, and immunotherapy. The choice of treatment depends on the type and stage of cancer, as well as the patient's overall health and personal preferences. Advances in medical research and technology have improved the prognosis for many people with cancer, with early detection and appropriate treatment leading to better outcomes (National Cancer Institute, 2021).

Each year, around 880,000 patients worldwide are diagnosed with head and neck cancer, resulting in approximately 440,000 deaths. In 2020, head and neck cancer ranked as the eighth most common tumor globally, with oral squamous cell carcinoma (OSCC) being the most prevalent malignant tumor in the oral and maxillofacial area. OSCC differs from other types of head and neck cancers in terms of epidemiology, clinical characteristics, and treatment therapy. OSCC can arise from various parts of the oral cavity, such as the alveolar ridge, buccal mucosa, floor of the mouth, upper jaw, and tongue. The risk factors for OSCC include tobacco use, alcohol consumption, and exposure to the human papillomavirus (HPV). While there is a high correlation between head and neck cancer and HPV infection, only around 25% of OSCC patients are HPV positive (Zou et al., 2022).

The global incidence of OSCC is 3.90 per 100,000, while the global mortality rate for oral cancer is 1.94 per 100,000. The 5-year overall survival rate for OSCC linked to carcinogens is no more than 60%. Treatment of OSCC typically involves surgery to remove the cancerous tissue, followed by radiation therapy or chemotherapy to kill any remaining cancer cells. The choice of treatment depends on the stage of the cancer, as well as the patient's overall health and personal preferences (Zou et al., 2022; Dolens et al., 2021).

Epidemiological studies consistently report increased risks of cancers in men and women with periodontal disease or tooth loss, conditions caused by oral bacteria. It is well established that oral bacteria are critical to the development of oral diseases and are linked in a number of studies to the risk of oral and gastrointestinal cancers, with the most consistent increased risks noted in studies of oral and oesophageal cancers, followed by evidence for pancreatic and gastric cancer (Ahn et al., 2012).

Establishing the association of the oral microbiome with cancer may lead to significant advances in understanding of cancer aetiology, potentially opening a new research paradigm for these diseases. Multi-disciplinary collaborations in epidemiology, microbiology, genetics, immunology, and bioinformatics are needed to broaden our understanding of the relationship of oral bacteria to cancer risk (Ahn et al., 2012).

Eubiosis is a state of balance within the microbial ecosystem of the human body, while dysbiosis is characterized by a lack of diversity and an overabundance of harmful bacteria. Microbes can play a role in the development of cancer through various mechanisms such as the production of toxins, changes in metabolic compounds, hormonal imbalances, chronic inflammation, changes in the immune system, and genetic damage and mutations. These microbes can also become a part of the environment around tumors in the respiratory and digestive tracts and can impact the growth of cancerous cells. Dysbiosis in the gut can not only affect the levels of nutrients and other compounds in the body but can also produce toxins that can influence the

progression of cancer. Additionally, the defensive mechanisms of certain bacteria can lead to genetic mutations that contribute to the formation of tumors (Parida and Sharma, 2021).

Helicobacter pylori, a type of bacteria that colonizes the stomach, is believed to be linked to around 60% of stomach cancer cases. It has also been associated with other types of cancer such as laryngeal carcinoma, hepatobiliary cancer, prostate and colon cancer (Li et al., 2020; Parida and Sharma, 2021). *Chlamydia trachomatis*, a gram-negative bacterium, has been linked to cervical cancer, and *Salmonella typhi* has been associated with gallbladder cancer. *Fusobacterium nucleatum* and enterotoxigenic *Bacteroides fragilis* (ETBF) have been linked to colorectal cancer (Kordahi et al., 2021). Additionally, a subtype of *Escherichia coli* that produces a compound called colibactin has been shown to increase the risk of tumor formation in the intestines in mice model. Both colibactin and another toxin called cytolethal distending toxin (CDT) can cause double-stranded DNA damage in mammalian cells (Rajagopala et al., 2017; Schwabe and Jobin, 2013).

Several bacteria have been identified as potential biomarkers for colorectal cancer, including *Fusobacterium nucleatum, Streptococcus gallolyticus*, Grp B2 *E. coli, Enterococcus faecalis* and Enterotoxigenic *Bacteroides fragilis*. *F. nucleatum*, which is associated with colon cancer, has been found to activate a signalling pathway called NF-kB, which is a key regulator of cancer-associated inflammation. It also directly suppresses the ability of the immune system to kill tumor cells by binding to a receptor called T cell immunoglobulin and ITIM domain (TIGIT), which is expressed on some T cells and natural killer cells (Rajagopala et al., 2017). Furthermore, *F. nucleatum* produces a bacterial cell surface adhesion component called FadA, which binds to host E-cadherin and leads to the activation of β-catenin, leading to cancer development (Cullin et al., 2021; Garrett, 2015). BFT (*B. fragilis*) cleaves E-cadherin, disrupt catenin-cadherin complexes and activate the β-catenin-cMyc hyperproliferative pathway to affect colonic epithelial cells (Cullin et al., 2021). Additionally, members of the nucleotide-binding oligomerization domain-like receptor (NLR) family may also play a role in the development of colorectal cancer (Garrett, 2015; Schwabe and Jobin, 2013).

Pancreatic cancer has been linked to certain changes in the composition of gut bacteria, specifically an increase in *Bacteroidetes* and a decrease in *Firmicutes*. A higher presence of certain bacteria in the oral cavity, such as *Enterobacteriaceae*, *Lachnospiraceae G7*, *Bacteroidaceae*, or *Staphylococcaceae*, has also been associated with an increased risk of pancreatic cancer (Wong-Rolle et al., 2021). Similarly, patients with liver cancer and cirrhosis have been found to have a higher ratio of *Bacteroides* to *Prevotella*, along with increased levels of *Erysipelotrichaceae* and decreased levels of *Leuconostocaceae* and *Fusobacterium* in their gut microbiome.

Pathogens and Infections

Bacterial pathogens are microorganisms that can cause disease in humans, animals, and plants. These bacteria can be transmitted through various means, including through food, water, contact with infected individuals, and insect bites. These pathogens can enter the body through various routes, including the respiratory, gastrointestinal, and genitourinary tracts (Doron and Gorbach, 2008). Some common examples of bacterial pathogens include:

- *Salmonella*: This group of bacteria causes food poisoning, with symptoms such as abdominal cramps, diarrhea, and fever. *Salmonella* is commonly transmitted through contaminated food (Doron and Gorbach, 2008).
- *Escherichia coli* (*E. coli*): A type of bacteria that lives in the intestinal tract of humans and animals. Some strains of *E. coli* can cause diarrhoea, urinary tract infections, and respiratory illness. *E. coli* infections can be transmitted through contaminated food or water, as well as through contact with infected individuals (Fleckenstein and Kuhlmann, 2019).
- *Staphylococcus aureus* (*S. aureus*): This type of bacteria is found on the skin and in the nasal passages of humans and animals. It can cause a range of infections, including skin infections, pneumonia, and toxic shock syndrome. *S. aureus* can be transmitted through skin-to-skin contact or through contact with contaminated surfaces (Cheung et al., 2021).
- *Streptococcus pneumoniae* (*S. pneumoniae*): This bacterium is a common cause of pneumonia, as well as other respiratory infections such as bronchitis and sinusitis. It can also cause meningitis and sepsis. *S. pneumoniae* is transmitted through respiratory droplets or close contact with infected individuals (Feldman and Anderson, 2020).
- *Mycobacterium tuberculosis* (*M. tuberculosis*): This bacterium is the cause of tuberculosis, a serious and potentially deadly lung infection. It is transmitted through the air when an infected individual coughs or sneezes (Roberts and Buikstra, 2019).

Infections caused by bacterial pathogens can range from mild to severe and can be treated with antibiotics. However, overuse and misuse of antibiotics has led to the development of antibiotic-resistant strains of bacteria, making it more difficult to treat infections (Lim et al., 2016). To prevent the spread of bacterial infections, it is important to practice good hygiene. Vaccines are also available for some bacterial pathogens, such as *Streptococcus pneumoniae* and *Haemophilus influenzae* type B (CDC, 2022b).

It is also important to note that not all bacteria are harmful and many are beneficial for the host, such as gut bacteria that aid in digestion.

Antimicrobial Resistance

Antimicrobial resistance (AMR) is the ability of microorganisms to resist the effects of antimicrobial drugs, such as antibiotics, antivirals and antifungals. This can occur naturally or because of the overuse and misuse of these drugs. When microorganisms are repeatedly exposed to antimicrobial drugs, they can develop mechanisms to evade the drugs' effects (CDC, 2021). This can happen through genetic mutations, the transfer of resistance genes between microorganisms, or the expression of previously dormant genes. Once a microorganism becomes resistant to a drug, it can pass on its resistance to other microorganisms through genetic transfer (Blair et al., 2015).

AMR is a major public health concern as it can lead to the spread of infections that are difficult to treat, resulting in prolonged illness, disability, and death. It also increases the cost of healthcare and can lead to the development of new, more virulent strains of microorganisms (Woolhouse et al., 2016).

Antibiotic resistance, a subcategory of AMR, is particularly concerning as it can lead to the spread of antibiotic-resistant bacteria. These bacteria can cause infections that are difficult or impossible to treat, leading to increased morbidity and mortality.

The World Health Organization (WHO) (World Health Organization, 2021) has identified a number of key areas for action to combat AMR, including:

1. Surveillance: monitoring the spread of resistant microorganisms and tracking changes in resistance patterns.
2. Infection prevention and control: implementing measures to reduce the spread of infections, such as hand hygiene and the use of personal protective equipment.
3. Rational use of antimicrobials: promoting the appropriate use of antimicrobial drugs in human and animal health to slow the development of resistance.
4. Research and development: investing in research to develop new antimicrobial drugs, diagnostic tools, and vaccination strategies.
5. International cooperation: working with other countries and international organizations to coordinate efforts to combat AMR.

In conclusion, antimicrobial resistance is a growing public health concern that requires a multifaceted approach to combat. This includes surveillance, infection prevention and control, rational use of antimicrobials, research and development, and international cooperation. It is also important for individuals to take responsibility for their own health and to use antimicrobial drugs appropriately.

Emerging Pathogens

SARS-CoV-2 is the novel coronavirus responsible for the ongoing pandemic of COVID-19, which began in Wuhan, China in December 2019 and has since spread to become a global public health crisis. The virus is a member of the coronavirus family, which also includes the viruses responsible for SARS (severe acute respiratory syndrome) and MERS (Middle East respiratory syndrome) (Geanta et al., 2022; World Health Organization, 2019).

SARS-CoV-2 is primarily spread through respiratory droplets from an infected person, and can cause a range of symptoms, from mild to severe. The most common symptoms include fever, cough, and difficulty breathing, and the virus can lead to severe respiratory illness and even death, particularly in older adults and individuals with underlying health conditions (CDC, 2022a; World Health Organization, 2019).

The World Health Organization (WHO) declared COVID-19 a pandemic on March 11, 2020, due to the rapid spread of the virus and the large number of confirmed cases and deaths reported globally. As of January 2022, there have been over 100 million confirmed cases and more than 2 million deaths worldwide. To control the spread of the virus, many countries have implemented measures such as social distancing, quarantine, and travel restrictions, as well as widespread testing and contact tracing (World Health Organization, 2022b).

Vaccines have also been developed and distributed globally, with several shown to be highly effective in preventing severe illness and death (World Health Organization, 2022a).

However, the pandemic has also led to significant economic and social impacts, including widespread job loss and disruption to education and other essential services. Additionally, there have been concerns about the disproportionate impact of the virus on marginalized communities and the potential long-term effects of the pandemic on mental health and well-being.

Application of Omics

Microbial genomics is the study of the genomic content of microorganisms, including bacteria, archaea, and viruses. It involves the sequencing, assembly, and analysis of microbial genomes, as well as the identification and functional annotation of their genes. Microbial genomics has revolutionized our understanding of the diversity and evolution of microorganisms, as well as their roles in various ecological and biomedical contexts (Aguiar-Pulido et al., 2016).

One major application of microbial genomics is in the field of biotechnology, where it has been used to identify and exploit microbial genes with industrial or medicinal value. For example, genomic studies have led to the discovery of enzymes that are used in the production of biofuels, detergents, and industrial chemicals (Lorenz and Eck, 2005). In the pharmaceutical industry, microbial genomics has been used to identify new drug targets and to develop antibiotics and other therapies (Bashir et al., 2014).

Microbial genomics has also had a significant impact on our understanding of the human microbiome, which refers to the collective genomes of the microorganisms that inhabit the human body. These microorganisms play important roles in maintaining human health and are involved in a variety of physiological processes, including digestion, immunity, and metabolism (Peterson et al., 2009). The study of the human microbiome has led to the development of probiotics, which are live microorganisms that are consumed to promote health, and to the identification of potential connections between the microbiome and diseases such as obesity, diabetes, and cancer (Chopra et al., 2020).

Another important application of microbial genomics is in the field of environmental microbiology, where it has been used to study the diversity and function of microorganisms in various ecosystems. This includes the study of microorganisms in the soil, water, and air, as well as in extreme environments such as polar regions and deep-sea vents (Aguiar-Pulido et al., 2016). Microbial genomics has helped us to better understand the role of microorganisms in nutrient cycling and the impact of human activities on microbial communities.

Microbial genomics has also had an impact on our understanding of the evolution of microorganisms, as it has allowed us to study the evolutionary relationships between different species and trace their evolutionary history. This has led to the identification of new species and the discovery of horizontal gene transfer, which is the process by which microorganisms acquire genes from other species (Aguiar-Pulido et al., 2016).

Overall, microbial genomics has had a significant impact on our understanding of microorganisms and their roles in various ecological and biomedical contexts. It has led to numerous technological and medical advances and will continue to be an important area of study in the future.

Global Cloud Platforms

The use of cloud computing in healthcare has grown rapidly in recent years, as healthcare organizations seek to improve patient care and reduce costs. One specific application of cloud computing in healthcare is the use of global healthcare data cloud platforms. These platforms allow for the storage, sharing, and analysis of healthcare data on a global scale, enabling improved research and collaboration among healthcare organizations (Tanwar et al., 2021).

One example of a global healthcare data cloud platform is the Global Health Data Exchange (GHDx, 2019). Developed by the Institute for Health Metrics and Evaluation (IHME) at the University of Washington, the GHDx is a platform for sharing and analyzing global health data. The platform currently contains over 2 billion data points from more than 190 countries, including data on mortality, cause of death, and risk factors for disease. The GHDx allows researchers and policymakers to access and analyze this data to inform public health policy and improve global health outcomes.

Another example of a global healthcare data cloud platform is the International Cancer Genome Consortium (ICGC) Data Portal (Zhang et al., 2019). The ICGC is a global effort to collect and share genomic data on cancer, intending to improve the understanding of the disease and develop new treatments. The ICGC Data Portal provides access to genomic and clinical data on cancer patients from institutions around the world. This data is available to researchers in order to support cancer research and improve patient care.

A third example of a global healthcare data cloud platform is the National Institutes of Health (NIH) Genomic Data Commons (GDC, 2016). The GDC is a platform for storing, sharing, and analyzing cancer genomics data. The platform currently contains data from over 17,000 cancer patients and allows researchers to access and analyze this data to improve their understanding of cancer and develop new treatments.

All these platforms are examples of how healthcare organizations are leveraging cloud computing to improve patient care and reduce costs by sharing and analyzing data on a global scale. While these platforms are designed to support research and collaboration in specific areas of healthcare, they also demonstrate the potential of global healthcare data cloud platforms to improve global health outcomes (Tanwar et al., 2021).

Aim of the thesis

The thesis aims to:

I. Describe the importance of human microbiome-associated health data to understand microbial diversity and its role in shaping human health.
II. Explain the importance of genomic surveillance to identify emerging pathogens and predict future pandemics.

III. Illustrate the significance of a global health data cloud platform to unifying research by collaborating on a global scale.

Outline of the thesis

The dissertation has nine chapters, starting with a general introduction, followed by seven chapters underlining the key areas of research. The final chapter provides a general discussion on the outcomes of mentioned seven chapters followed by impact paragraph to draw conclusions for future research and final summary of the dissertation. In total, seven scientific articles published in peer-reviewed impact factor journals form the core body of this dissertation. The graphical summary and major checkpoints of this dissertation are highlighted in **figure 1**.

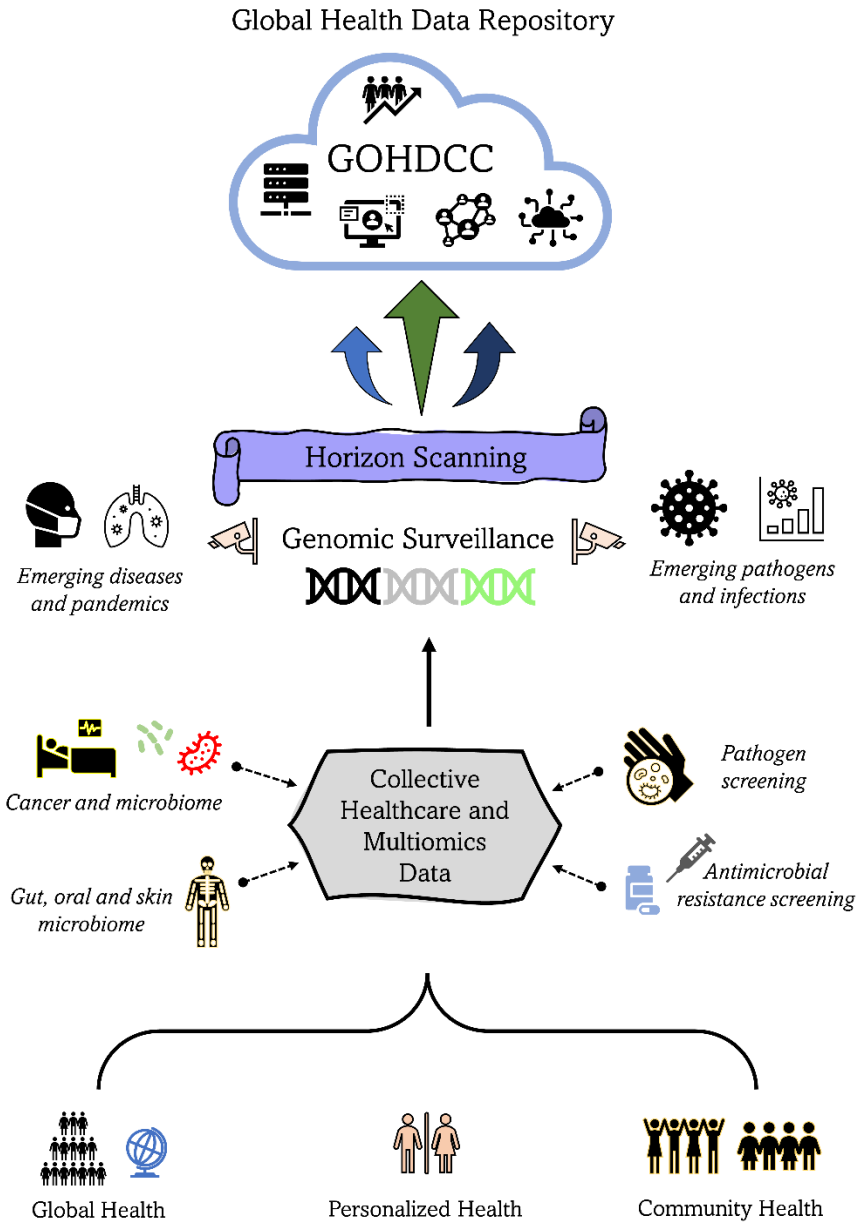


Figure 1: Highlighting major levels and checkpoints of the thesis. Illustrate applications of human microbiome, health data and genomic surveillance.

Chapter 2: "Human microbiome: Microbes shaping human health"

In this chapter the association between *Ayurveda prakriti's* and microbial diversity for 272 healthy individuals has been drawn. Gut and oral microbiomes of healthy individual were utilized to study the relationship between *prakriti* type and microbial composition. Microbial composition plays a vital role in shaping human health and considered as a reliable indicator of health.

Chapter 3: "Cancer insights: A pathway and network analysis to understand oral cancer"

In this chapter complex networks and molecular mechanisms involved in pathogenesis of oral submucous fibrosis (OSF) and oral squamous cell carcinoma (OSCC) are described. A pathway analysis to identify signature genes that affect epithelial-mesenchymal transition (EMT) that play a crucial role in the advancement of oral cancer are highlighted in this study.

Chapter 4: "Emerging pathogens: Comparative genome analysis of *Clostridia* species"

This chapter discuss about the pathogenic potential of *Clostridia* strains when the genomes were compared to a known pathogen. It highlights the application of comparative genome analysis and how it can help to identify emerging pathogens. This study describes the virulence factors encoded in bacterial genome and how genomic analysis can help one to differentiate non-pathogens and pathogenic strains.

Chapter 5: "Antimicrobial resistance and virulence: Genome comparison of *Staphylococcus aureus* strains"

In this chapter, four strains of *Staphylococcus aureus* (*S. aureus*) isolated from diabetic foot ulcer patients were compared on genome scale. Minor genomic differences provide evidence for change in phenotypic characteristics. Four *S. aureus* strains were compared to distinguish them based on their antimicrobial properties and virulence capabilities. This work, which used whole-genome sequence analysis, was focused on finding potential virulence factors, genes that cause antibiotic resistance, mobile genetic elements, biofilm-forming ability, and sporulation factors that contribute to the pathogenicity of bacterial strains.

Chapter 6: "Microbial genomics: A comprehensive guide for microbial genome research"

In this chapter our aim was to emphasize the widely recognized and extensively utilized tools and references for different aspects of microbial research, including genome assembly and annotation, profiling antibiotic genes, identifying virulence factors, and studying drug interactions. Furthermore, we explored the recommended methodologies in computer-based research on microbial genomes, current developments in the analysis of microbial genomic data, the integration of multi-omics data, the proper utilization of machine learning algorithms, and the availability of open-source bioinformatics resources for genome data analytics.

Chapter 7: "Genomic surveillance: Linking omics data for pandemic preparedness"

This chapter highlights the recent COVID-19 pandemic as a planetary health concern and necessary steps to be taken to prepare for future pandemics. This chapter describe the importance of omics data and how high-

throughput technologies are changing the face and pace of science. Multiomics technologies played a very crucial role in COVID-19 pandemic by providing huge amount of genomic information in such a short time. This chapter also discuss on employing omics systems to monitor new pathogens and conduct genomic surveillance.

Chapter 8: "Global health data cloud: Laying new directions for collaborative science"

To facilitate international research and development, we suggest the Global Open Health Data Cooperatives Cloud (GOHDCC), a platform for exchanging health data on a worldwide scale. This platform is advantageous for all stakeholders involved in the healthcare system since it is citizen-led and jointly governed. The concept emphasises on the importance of big data management by integrating cloud computing to manage enormous amount of data. Additionally, it introduces the Open Science Data Cloud (OSDC) and European Open Science Cloud (EOSC), two current cloud-based health data systems.

References:

Aguiar-Pulido V, Huang W, Suarez-Ulloa V, et al. Metagenomics, metatranscriptomics, and metabolomics approaches for microbiome analysis. Evol Bioinforma 2016;12:5–16.

Ahn J, Chen CY and Hayes RB. Oral microbiome and oral and gastrointestinal cancer risk. Cancer Causes Control 2012;23(3):399–404.

AMC. Cancer Types | Cancer Resources | American Cancer Society. 2022. Available from: https://www.cancer.org/cancer.html [Last accessed: 3/13/2023].

Bashir Y, Pradeep Singh S and Kumar Konwar B. Metagenomics: an application based perspective. Chinese J Biol 2014;2014:1–7.

Blair JMA, Webber MA, Baylay AJ, et al. Molecular mechanisms of antibiotic resistance. Nat Rev Microbiol 2015;13(1):42–51.

Bressa C, Bailén-Andrino M, Pérez-Santiago J, et al. Differences in gut microbiota profile between women with active lifestyle and sedentary women. PLoS One 2017;12(2):e0171352.

Broecker F and Moelling K. The roles of the virome in cancer. Microorganisms 2021;9(12).

CDC. About Antibiotic Resistance | CDC. 2021. Available from: https://www.cdc.gov/drugresistance/about.html [Last accessed: 2/7/2023].

CDC. Symptoms of COVID-19 | CDC. 2022a. Available from: https://www.cdc.gov/coronavirus/2019-ncov/symptoms-testing/symptoms.html [Last accessed: 2/7/2023].

CDC. Vaccines and Immunizations. Centers Dis Control Prev 2022.

Cheung GYC, Bae JS and Otto M. Pathogenicity and virulence of *Staphylococcus aureus*. Virulence 2021;12(1):547–569.

Chopra RS, Chopra C and Sharma NR, (eds). Metagenomics: techniques, applications, challenges and opportunities. 1st ed. Springer Singapore; 2020.

Cullin N, Azevedo Antunes C, Straussman R, et al. Microbiome and cancer. Cancer Cell 2021;39(10):1317–1341.

Dalton A, Mermier C and Zuhl M. Exercise influence on the microbiome–gut–brain axis. 2019;10(5):555–568.

Dilnessa T, Getaneh A, Hailu W, et al. Prevalence and antimicrobial resistance pattern of *Clostridium difficile* among hospitalized diarrheal patients: a systematic review and meta-analysis. PLoS One 2022;17(1):e0262597.

Doelman A, Tigchelaar S, McConeghy B, et al. Characterization of the gut microbiome in a porcine model of thoracic spinal cord injury. BMC Genomics 2021;22(1):1–17.

Dolens E da S, Dourado MR, Almangush A, et al. The Impact of Histopathological Features on the Prognosis of Oral Squamous Cell Carcinoma: A Comprehensive Review and Meta-Analysis. Front Oncol 2021;11:1–13.

Doron S and Gorbach SL. Bacterial infections: overview. In: International Encyclopedia of Public Health Elsevier; 2008; pp. 273–282.

Elinav E, Garrett WS, Trinchieri G, et al. The cancer microbiome. Nat Rev Cancer 2019;19(7):371–376.

Feldman C and Anderson R. Recent advances in the epidemiology and prevention of *Streptococcus pneumoniae* infections. F1000Research 2020;9.

Fleckenstein JM and Kuhlmann FM. Enterotoxigenic *Escherichia coli* infections. Curr Infect Dis Rep 2019;21(3):1–9.

Garrett WS. Cancer and the microbiota. Science 2015;348(6230):80–86.

Gaulke CA and Sharpton TJ. The influence of ethnicity and geography on human gut microbiome composition. Nat Med 2018;24(10):1495–1496.

GDC. Home | NCI Genomic Data Commons. 2016. Available from: https://gdc.cancer.gov/ [Last accessed: 2/7/2023].

Geanta M, Tanwar AS, Lehrach H, et al. Horizon scanning: rise of planetary health genomics and digital twins for pandemic preparedness. Omi A J Integr Biol 2022;26(2):93–100.

GHDx. Global Health Data Exchange | GHDx. 2019. Available from: https://ghdx.healthdata.org/ [Last accessed: 2/7/2023].

Glassner KL, Abraham BP and Quigley EMM. The microbiome and inflammatory bowel disease. J Allergy Clin Immunol 2020;145(1):16–27.

Groussin M, Poyet M, Sistiaga A, et al. Elevated rates of horizontal gene transfer in the industrialized human microbiome. Cell 2021;184(8):2053-2067.e18.

Huttenhower C, Gevers D, Knight R, et al. Structure, function and diversity of the healthy human microbiome. Nature 2012;486(7402):207–214.

Johnson AJ, Vangay P, Al-Ghalith GA, et al. Daily sampling reveals personalized diet-microbiome associations in humans. Cell Host Microbe 2019;25(6):789-802.e5.

Kim M and Benayoun BA. The microbiome: an emerging key player in aging and longevity. Transl Med Aging 2020;4:103–116.

Kim N, Yun M, Oh YJ, et al. Mind-Altering with the gut: modulation of the gut-brain axis with probiotics. J Microbiol 2018 563 2018;56(3):172–182.

Kordahi MC, Stanaway IB, Avril M, et al. Genomic and functional characterization of a mucosal symbiont involved in early-stage colorectal cancer. Cell Host Microbe 2021;29(10):1589-1598.e6.

Krishnareddy S. The microbiome in celiac disease. Gastroenterol Clin North Am 2019;48(1):115–126.

Lau AWY, Tan LTH, Ab Mutalib NS, et al. The chemistry of gut microbiome in health and diseases. Prog Microbes Mol Biol 2021;4(1).

Li K, Bihan M, Yooseph S, et al. Analyses of the microbial diversity across the human microbiome. PLoS One 2012;7(6):e32118.

Li L, Tan J, Liu L, et al. Association between *H. Pylori* infection and health outcomes: an umbrella review of systematic reviews and meta-analyses. BMJ Open 2020;10(1):e031951.

Lim C, Takahashi E, Hongsuwan M, et al. Epidemiology and burden of multidrug-resistant bacterial infection in a developing country. Elife 2016;5.

Littleford-Colquhoun BL, Weyrich LS, Kent N, et al. City life alters the gut microbiome and stable isotope profiling of the eastern water dragon (*Intellagama Lesueurii*). Mol Ecol 2019;28(20):4592–4607.

Lorenz P and Eck J. Metagenomics and industrial applications. Nat Rev Microbiol 2005;3(6):510–516.

Maltz RM, Keirsey J, Kim SC, et al. Prolonged restraint stressor exposure in outbred cd-1 mice impacts microbiota, colonic inflammation, and short chain fatty acids. PLoS One 2018;13(5):e0196961.

Maltz RM, Keirsey J, Kim SC, et al. Social stress affects colonic inflammation, the gut microbiome, and short-chain fatty acid levels and receptors. J Pediatr Gastroenterol Nutr 2019;68(4):533–540.

Methé BA, Nelson KE, Pop M, et al. A framework for human microbiome research. Nature 2012;486(7402):215–221.

Mousa WK, Chehadeh F and Husband S. Recent advances in understanding the structure and function of the human microbiome. Front Microbiol 2022;13:111.

NCI. Treatment for Cancer - NCI. 2021. Available from: https://www.cancer.gov/about-cancer/treatment [Last accessed: 3/13/2023].

Ogunrinola GA, Oyewale JO, Oshamika OO, et al. The human microbiome and its impacts on health. Int J Microbiol 2020.

Parida S and Sharma D. The microbiome and cancer: creating friendly neighborhoods and removing the foes within. Cancer Res 2021;81(4):790–800.

Peterson J, Garges S, Giovanni M, et al. The NIH human microbiome project. Genome Res 2009;19(12):2317–2323.

Rajagopala S V., Vashee S, Oldfield LM, et al. The human microbiome and cancer. Cancer Prev Res 2017;10(4):226–234.

Roberts CA and Buikstra JE. Bacterial Infections. In: Ortner's identification of pathological conditions in human skeletal remains Academic Press; 2019; pp. 321–349.

Schwabe RF and Jobin C. The microbiome and cancer. Nat Rev Cancer 2013;13(11):800–812.

Shreiner AB, Kao JY and Young VB. The gut microbiome in health and in disease. Curr Opin Gastroenterol 2015;31(1):69–75.

Tanwar AS, Evangelatos N, Venne J, et al. Global open health data cooperatives cloud in an era of covid-19 and planetary health. Omi A J Integr Biol 2021;25(3):169–175.

Vangay P, Johnson AJ, Ward TL, et al. US immigration westernizes the human gut microbiome. Cell 2018;175(4):962-972.e10.

Vila AV, Collij V, Sanna S, et al. Impact of commonly used drugs on the composition and metabolic function of the gut microbiota. Nat Commun 2020;11(1):1–11.

Weersma RK, Zhernakova A and Fu J. Interaction between drugs and the gut microbiome. Gut 2020;69(8):1510–1519.

WHO. Cancer. 2022. Available from: https://www.who.int/news-room/fact-sheets/detail/cancer [Last accessed: 3/13/2023].

Wilmanski T, Diener C, Rappaport N, et al. Gut microbiome pattern reflects healthy ageing and predicts survival in humans. Nat Metab 2021;3(2):274–286.

Winglee K, Howard AG, Sha W, et al. Recent urbanization in China is correlated with a westernized microbiome encoding increased virulence and antibiotic resistance genes. Microbiome 2017;5(1):1–13.

Wong-Rolle A, Wei HK, Zhao C, et al. Unexpected guests in the tumor microenvironment: microbiome in cancer. Protein Cell 2021;12(5):426–435.

Woolhouse M, Waugh C, Perry MR, et al. Global disease burden due to antibiotic resistance - state of the evidence. J Glob Health 2016;6(1):1–5.

World Health Organization. Coronavirus. 2019. Available from: https://www.who.int/health-topics/coronavirus#tab=tab_1 [Last accessed: 2/7/2023].

World Health Organization. COVID-19 Vaccines. 2022a. Available from: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/covid-19-vaccines [Last accessed: 2/7/2023].

World Health Organization. Strategy and planning. 2022b. Available from: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/strategies-and-plans [Last accessed: 2/7/2023].

World Health Organization. WHO strategic priorities on antimicrobial resistance: preserving antimicrobials for today and tomorrow. World Heal Organ 2021;1–12.

Wouk J, Rechenchoski DZ, Rodrigues BCD, et al. Viral infections and their relationship to neurological disorders. Arch Virol 2021;166(3):733–753.

Wylie KM, Weinstock GM and Storch GA. Emerging view of the human virome. Transl Res 2012;160(4):283–290; doi: 10.1016/j.trsl.2012.03.006.

Yadav M, Verma MK and Chauhan NS. A review of metabolic potential of human gut microbiome in human nutrition. Arch Microbiol 2018;200(2):203–217.

Zhang J, Bajari R, Andric D, et al. The international cancer genome consortium data portal. Nat Biotechnol 2019;37(4):367–369.

Zimmer J, Lange B, Frick JS, et al. A vegan or vegetarian diet substantially alters the human colonic faecal microbiota. Eur J Clin Nutr 2012;66(1):53–60.

Zou Z, Li B, Wen S, et al. The Current Landscape of Oral Squamous Cell Carcinoma: A Comprehensive Analysis from ClinicalTrials.gov. Cancer Control 2022;29:1–14.

Chapter

2

Human Microbiome: Microbes Shaping Human Health

Shalini, Tirumalapura Vijayanna, Apoorva Jnana, Sitaram Jaideep Sriranjini, **Ankit Singh Tanwar**, Angela Brand, Thokur Sreepathy Murali, Kapaettu Satyamoorthy, and G. G. Gangadharan. "Exploring the signature gut and oral microbiome in individuals of specific *Ayurveda prakriti*." *Journal of Biosciences* 46, no. 3 (2021): 54.

DOI: 10.1007/s12038-021-00182-2; **IF 2.8 (2021)**

Exploring the Signature Gut and Oral Microbiome in Individuals of Specific *Ayurveda Prakriti*

Abstract

Diagnosis and treatment of various diseases in *Ayurveda*, the Indian system of medicine, relies on '*prakriti*' phenotyping of individuals into predominantly three constitutions, *kapha*, *pitta* and *vata*. Recent studies propose that microbiome play an integral role in precision medicine. A study of the relationship between *prakriti* – the basis of personalized medicine in *Ayurveda* and that of gut microbiome, and possible biomarker of an individual's health, would vastly improve precision therapy. Towards this, we analyzed bacterial metagenomes from buccal (oral microbiome) and fecal (gut microbiome) samples of 272 healthy individuals of various predominant *prakriti*s. Major bacterial genera from gut microbiome included *Prevotella, Bacteroides* and *Dialister* while oral microbiome included *Streptococcus, Neisseria, Veilonella, Haemophilus, Porphyromonas* and *Prevotella*. Though the core microbiome was shared across all individuals, we found *prakriti* specific signatures such as preferential presence of *Paraprevotella* and Christensenellaceae in *vata* individuals. A comparison of core gut microbiome of each *prakriti* with a database of 'healthy' microbes identified microbes unique to each *prakriti* with functional roles similar to the physiological characteristics of various *prakriti*s as described in *Ayurveda*. Our findings provide evidence to *Ayurvedic* interventions based on *prakriti* phenotyping and possible microbial biomarkers that can stratify the heterogenous population and aid in precision therapy.

Keywords

Ayurveda; gut microbiome; oral microbiome; *prakriti*; precision medicine

1. Introduction

Ayurveda, the Indian system of medicine, has an established and unique approach towards the diagnosis and treatment of various diseases. *Doshas* are the biological, functional units described in *Ayurveda* for understanding of both *prakriti* and *vikruti* in an individual. The three doshas are *Vata* (kinetic) – representing the movements in the body, *Pitta* (metabolic) – representing metabolism and transformation in the body, and *Kapha* (potential) – representing the growth and maintenance in the body (Prasher *et al*. 2016). The body–mind constitution of an individual, termed as '*prakriti*' plays a pivotal role in the management of diseases and selection of formulations and different dosage forms of treatment in *Ayurveda*. Though *prakriti* is claimed as a genetic determinant influenced by the nature of male and female gametes (*Shukra shonita*), it is additionally influenced by several factors such as *Rutu* (season), *Matu Ahara Vihara* (maternal diet and lifestyle), *Kala-Garbhashaya* (age of parents and female reproductive system), *Matruja* (maternal factors), *Pitruja* (paternal factors), *Aatmaja* (actions of soul), *Sattvaja* (psychological factors), *Saatmyaja* (congenious factors) and *Kaalaja* (time of conception and the seasonal influence) (Sharma and Dash 2009). In *Ayurveda* system of medicine, an individual is classified into one of the seven *prakriti* types based on the constitution which is determined at the stage of conception. Further, *dosha* (bio humors), *dhatu* (body tissues) and the *mala* (metabolic waste) form the core of the body (Murthy 2009), and the combination of the three *doshas* (*vata*,

pitta and *kapha*) help in determining the *prakriti* of an individual. The proportions of *vata*, *pitta* and *kapha* govern the functional attributes such as kinetic energy, metabolism and potential energy and any disturbance to this equilibrium can lead to *vikruti* or disease condition. Hence, it is of significance to understand the basic constitutional type and the need for personalized treatment for disease management. Predominance of a single *dosha prakriti* is seen in only about 10% of the individuals and genome analysis performed on these individuals clearly indicate distinct genome wide differences providing genetic basis for the *prakriti* phenotyping in *Ayurveda* (Aggarwal *et al*. 2010; Rotti *et al*. 2014). In addition, striking differences at the molecular level, in biochemical and hematological parameters, and genome wide gene expression were discovered in few studies conducted on the single dosha *prakriti* types (Govindaraj *et al*. 2015; Prasher *et al*. 2008, 2016; Rotti *et al*. 2015).

The human gut microbiome hosts abundant, highly diverse and metabolically active microorganisms known to influence physiology and metabolism of an individual. It is linked with energy metabolism, modulation of immune system and inflammation (Nicholson *et al*. 2012). Advances in sequencing technologies have facilitated in understanding and appreciating the significant diversity and functional interrelationships between microbial communities that determine the health status of an individual. The structure and function of these microbial communities play an indispensable role in maintaining the homeostasis and any perturbations to this stable community structure can lead to disease states (Liu *et al*. 2012). Studies clearly indicate that the gut microbiome of individuals can be used as reliable predictors of health status with gut microbiome of healthy individuals having a dominance of bacterial phyla such as Bacteroidetes, Firmicutes along with lesser abundance of Proteobacteria, Actinobacteria, Verrucomicrobia and Fusobacteria irrespective of the diet or population considered (Durack and Lynch 2019; Gupta *et al*. 2020).

Hence, we designed to explore the relationship between *prakriti*, the basis of personalized medicine in *Ayurveda* and that of gut microbiome, which is increasingly being contemplated as a reliable biomarker to assess health of an individual. Towards this objective, we have analyzed the bacterial metagenomes from saliva and stool samples representing oral and gut microbiome respectively of 272 healthy individuals and their relationship with *prakriti* so as to correlate microbial diversity patterns with the *prakriti* identity.

2. Materials and methods

2.1 *Recruitment of subjects*

Screening and recruitment of the subjects were done in Institutions within the campus of Ramaiah Indic Specialty *Ayurveda* Restoration hospital and two *Ayurveda* colleges in Bangalore. A total of 2000 volunteers were screened and of these, 272 healthy volunteers with all the biochemical and haematological parameters within normal limits were included in the study. The study protocol was approved by the ethics committee of M.S. Ramaiah Medical College and Hospitals and informed consent was obtained from all the participants. The tests included complete blood count, lipid profile test, liver function test, serum creatinine, blood urea nitrogen, serum TSH, fasting blood sugar and urine routine to ensure their values are within defined, physiological limits. Subjects of both gender in the ratio of 102:170 (Male:Female) were recruited and the

participants had an average age of 21 years, height of 163.51 cm, weight of 70.04 kg and BMI of 22.30 kg/m^2. We excluded subjects with history of smoking, alcohol consumption, any form of drug addictions, diabetes, hypertension and other chronic diseases from the study. In addition, subjects who had taken antibiotics 6 months prior to sample collection were also excluded from the study.

2.2 *Assessment of Prakriti*

The *prakriti* of the volunteers were ascertained by 3 different modes: (a) TNMC questionnaire designed on the basis of literature in *Ayurveda* texts comprising 37 objective questions related to the person's physical characteristics, psychological make-up and physiological habits (Bhalerao *et al*. 2012), (b) assessment by a senior *Ayurveda* physician using classical method of interview and physical examination to assess physical, physiological and psychological characteristics as described in *Ayurveda* literature and (c) assessment using 'Ayusoft' (http://ayusoft.cdac.in), a software developed based on information from *Ayurveda* literature. For the first 50 subjects, *prakriti* was assessed by the senior physician, TNMC and the software to check for concordance and remaining volunteers were assessed using TNMC and Ayusoft (table 1).

Table 1: Classification of prakriti types in the current study.

Sl. No.	*Prakriti* type	Definition
1	*Vata*	More than 70% of characteristics of *Vata* and other two *doshas* making 30% together)
2	*Pitta*	More than 70% of characteristics of *Pitta* and other two *doshas* making 30% together
3	*Kapha*	More than 70% of characteristics of *Kapha* and other two *doshas* making 30% together

2.3 *Library preparation and metagenomic sequencing*

The stool and saliva samples of 272 recruited volunteers were collected as per standard protocols (Qiagen, 2010, 2016). Samples were stored at 4°C until processed after which they were stored at -20°C. This was chosen to be the most feasible method available which would not alter the concentration and composition of the DNA (Ribeiro *et al*. 2018). The bacterial DNA from the samples were isolated using the Qiagen kits (cat No. 51304 and 51504). The extracted DNA samples were stored in -80°C till further processing. Sequencing libraries were prepared as per the Illumina MiSeq Metagenomics workflow. Briefly, variable V3 and V4 regions of the 16S rRNA gene were amplified with 16S rRNA universal primers fused with Illumina adapters (Illumina 2013; Klindworth *et al*. 2013). Primers were 5' TCGTCGGCAGCGTCAGATGTGTA-TAAGAGACAGCCTACGGGNGGCWGCAG and 5'GTCTCGTGGGCTCGGAGATGTGTATAAGAGA-CAGGACTACHVGGGTATCTAATCC. PCR was carried out in 25 µl volume containing 10.5 µl of microbial DNA (12.5 ng), 1 µl of each primer (5 µM), and 12.5 µl of 2X KAPA Hifi HotStart Ready Mix and cycling conditions were 95°C for 3 min, 25 cycles of 95°C for 30 s, 55°C for 30 s, 72°C for 30 s, and 72°C for 5 min.

The resulting PCR products were run on an Agilent 2100 Bioanalyzer and then purified using 0.9X Agencourt AMPure XP beads. The index PCRs were carried out in 50 µl reactions containing 5 µl of purified DNA, 5 µl each of Nextera XT Index Primer 1 and 2, 25 µl of 2x KAPA Hifi HotStart Ready Mix and 10 µl of nuclease-free water. The PCR conditions were as follows: 95°C for 3 min, 8 cycles of 95°C for 30 s, 55°C for 30 s, 72°C for 30 s, and 72°C for 5 min. The libraries were then cleaned up using 1X AMPure XP beads and the size of amplicons were verified on Agilent 2100 Bioanalyzer. Equal Volume of normalized 2 nM libraries were pooled, denatured and diluted to 4 pM before loading onto the MiSeq flow cell and sequenced on Illumina MiSeq platform using a 2 x 300 bp paired end protocol.

2.4 *Raw data processing and taxonomy analysis*

Raw sequence data was analyzed and processed separately for fecal (n = 232) and buccal (n = 256) samples using Quantitative Insights into Microbial Ecology (QIIME 2) software version 2020.2.0 (Bolyen *et al*.2019). Paired-end raw sequencing data for a total of 488 samples were demultiplexed and sequences with a quality score (q-score of ≥28) were selected for further processing. Sequences were trimmed (440 base for buccal samples; 439 bases for fecal samples) and selected for OTU (operational taxonomic unit) clustering using Deblur (Amir *et al*. 2017). Using Deblur, sequences were clustered into OTUs by aligning sequences locally using SortMeRNA (Kopylova *et al*.2012) against Greengenes (DeSantis *et al*. 2006) 16SrRNA database (version 13_8) and filtered at 88% sequence similarity. Chimeric sequences were identified using *de novo* (vsearch) method (Edgar *et al*. 2011; Rognes *et al*. 2016). After identification and removal of chimeric reads, OTUs were further clustered at 97% sequence similarity using closed reference clustering workflow (Caporaso *et al*. 2010). To explore the taxonomic composition in all samples, taxonomy was assigned to all the sequences identified as OTUs/features after closed-reference clustering using q2-feature-classifier plugin in QIIME 2.

2.5 *Microbiome diversity analysis*

Microbiome diversity analysis was performed with OTUs that were classified up to genera level. Analysis at species level was performed for identification of *Prakriti* specific taxa. Alpha diversity (Chao1 index) was calculated using MicrobiomeAnalyst (Chong *et al*.2020) to estimate species richness and diversity. Beta diversity was assessed by principal coordinate analysis on weighted UniFrac distances via q2-diversity plugin of QIIME 2. Relative abundance (reads corresponding to an OTU/total number of reads of the sample) and average relative abundance (ARA; relative abundance in all samples/total number of samples) was calculated to assess the contribution of each OTU at phylum and genera level towards the gut and oral microbiome. Shared and unique bacterial genera were analyzed with a web-based tool InteractiVenn (Heberle *et al*. 2015). BugBase web server was used to predict the functional traits of the microbiome from each *prakriti* types based on OTU abundance data using default parameters (Ward *et al*. 2017). Kruskal Wallis test was used to find whether the functional traits differed significantly across different *prakriti* types.

2.6 *Detection of microbial enterotypes*

Microbial enterotypes were determined via 'between class analysis' which performs a principal component analysis with partitioning around medoids clustering using R package 'ade4'. It involves identification of inherent clusters in the data based on a distance matrix of Jensen-Shannon divergence indices. Optimum number of clusters was determined by Calinski-Harabasz index and validated by Silhouette index (Arumugam *et al*. 2011). Spearman's correlation test was performed to assess the correlation of the major drivers of the clusters.

2.7 *Analysis of 'healthy' core gut microbiome*

Core microbiome was defined as OTUs present in at least 50% of the samples in each *prakriti* (*kapha*, *pitta*, *vata*) irrespective of their abundance. Data from GutFeeling KB (GFKB https://hive.biochemistry.gwu.edu/gfkb), a healthy human reference microbiome database, was used as a 'healthy control' dataset and compared with microbes identified in the current study. Genera from each *prakriti*'s core microbiome that were absent in GFKB were checked for their presence in other *prakriti*s and the resulting unique genera for each *prakriti* was analyzed based on its relative abundance (King *et al*. 2019).

2.8 Identification of *prakriti* specific signatures

Prakriti specific signatures were identified using LEfSe (Segata *et al*. 2011) which identifies biomarkers based on linear discriminant analysis (LDA). LefSe was performed with OTUs that were classified to genera level using default parameters (effect size 2) (Goecks *et al*. 2010). Strict parameters (multiclass analysis all against all) were used for identification of species-specific signatures. The male and female datasets of gut and oral microbiome respectively were processed independently. Significant biomarkers were annotated to their species level via manual BLAST (highest scoring hits were retained). Only non-redundant OTUs with highest LDA scores in each *prakriti* were considered for analysis.

3. Results

3.1 *Microbiome of healthy individuals belonging to different prakriti*

In this study, we obtained a total of 32,750,999 and 36,023,626 16S rRNA sequence reads from fecal and buccal samples respectively. After quality filtering using QIIME, 8,127,407 reads from stool samples and 9,092,545 reads from buccal samples were taken up for identification of OTUs and further analysis. A total of 836 features (trim length of 439) representing 232 samples from gut microbiome were selected after closed reference OTU clustering at 97% identity while 589 features (trim length of 440) representing 255 samples were selected for oral microbiome analysis. Further, p-sampling depth was chosen as 2500 for gut microbiome samples and at 3000 for oral microbiome samples based on a rarefaction analysis. The final dataset included 1,448,987 sequences from 209 stool samples and 1,873,939 sequences from 200 buccal samples for overall analysis. The gut microbiome dataset included samples from 46, 52 and 35 females and 35, 12 and 29 males of *kapha*, *pitta* and *vata prakriti*s respectively. The oral microbiome dataset included samples from 41, 48 and 34 females and 36, 10 and 31 males belonging to *kapha*, *pitta* and *vata prakriti*s respectively.

3.2 *Microbial diversity patterns*

3.2.1 *Gut microbiome:* A total of 109 OTUs representing 12 different phyla were identified in gut microbiome from a total of 209 samples from healthy individuals. Irrespective of the *prakriti* or gender, Bacteroidetes and Firmicutes were the dominant phyla and accounted for more than 90% of the gut microbiome, while Proteobacteria and Actinobacteria contributed <10% to the overall microbiome (figure 1a). Chao1 diversity index was used to calculate the diversity of gut microbiome based on age, gender and BMI values (figure 1b). We found that the microbiome was more diverse in aged individuals and male samples were more heterogeneous than female samples but with lower diversity index. With increase in BMI values, we found lower diversity values. When the samples were grouped based on *prakriti*s, a higher median species diversity was observed in individuals belonging to *vata prakriti* (figure 1c).

Major bacterial genera from the gut microbiome included *Prevotella, Bacteroides* and *Dialister* (figure 1d). *Sutterella* was less abundant in *vata prakriti* individuals when compared to individuals from other *prakriti*s. We found 13 OTUs out of the 109 OTUs to be highly prevalent across our samples (presence in more than 90% of samples) with only one OTU (*Roseburia*) shared across all the 209 samples studied. Only two OTUs (*Prevotella* and *Bacteroides*) had ARA >10%, while several OTUs with high prevalence across all samples showed low ARA of less than 5% indicating enormous diversity of the gut microbiome. Only 34 of the 109 OTUs were present in more than 50% of the samples while 38 OTUs were present in less than 10% of samples. Out of the 109 OTUs considered, 91 OTUs or 83% of the OTUs were shared among all the three *prakriti*s (figure 1e). Only 10 OTUs were unique to any of the three *prakriti* types but these OTUs had very low prevalence. A Principal Component Analysis of gut microbiome samples performed using Bray-Curtis measure of distances showed that the samples could be grouped into two major clusters based on the presence of two major genera – *Prevotella* and *Bacteroides*; but no clusters were seen based on *prakriti*s (figure 1f). Similarly, when we performed a between-class analysis based on Jensen Shannon Divergence distances to decipher gut enterotypes (Arumugam *et al*. 2011), two enterotypes based on abundance of *Prevotella* and *Bacteroides* were observed (figure 1g). We found a negative correlation between relative abundance of *Prevotella* and *Bacteroides* in the gut microbiome samples ($P < 0.001$; $r = 0.7671$) wherein samples which showed higher abundance of *Prevotella* had lower abundance of *Bacteroides* and vice versa (figure 1h).

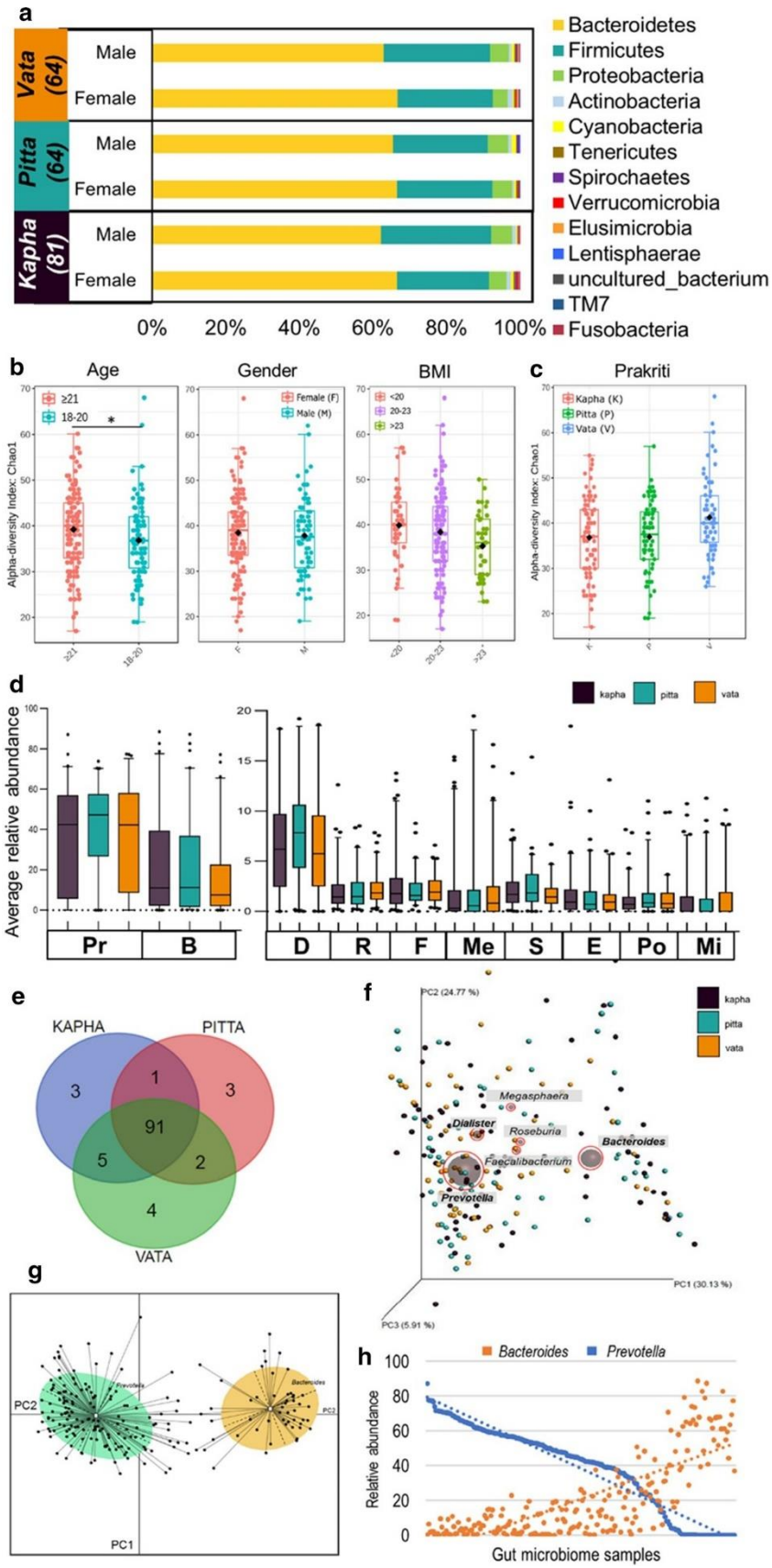


Figure 1: Gut microbiome: **(a)** Phylum plot showing average relative abundance (ARA) of major phyla in different cohorts. **(b)** Alpha diversity values for microbiome based on age, gender, BMI. **(c)** Alpha diversity values based on prakriti. **(d)** Box plot for the 10 most abundant OTUs (Pr – *Prevotella*, B – *Bacteroides*, D – *Dialister*, R – *Roseburia*, F – *Faecalibacterium*, Me – *Megasphaera*, S – *Sutterella*, E – Enterobacteriaceae,

Po – *Porphyromonas*, Mi – *Mitsuokella*) as determined by ARA. Boxes represent the interquartile range (IQR) and the line represents the median. Whiskers denote the lowest and highest values within 1.5 x IQR. Circles represent outliers beyond the whiskers. **(e)** Venn diagram showing unique and shared genera between the *prakriti* types. **(f)** PCoA with weighted unifrac distances. Each dot represents a sample while grey circles represent major OTUs. Sizes of the grey circles represent their ARA. **(g)** Clustering of gut microbial taxa into enterotypes based on between-class analysis using Jensen- Shannon distance. X axis shows cluster number and Y axis shows silhouette width (a measure of cluster separation) (Wu *et al*. 2011). Each dot represents a sample. **(h)** Relative abundance of *Prevotella* and *Bacteroides* in different samples. Dots represent relative abundance in each sample and dotted lines represent the trendline. The two species showed an inverse relationship with reference to their relative abundance in different samples **(h)**.

3.2.2 *Oral microbiome:* With respect to buccal samples, a total of 132 OTUs representing 12 different phyla were associated with oral microbiome from a total of 200 samples screened from healthy individuals. The dominant phyla were Firmicutes and Proteobacteria, while Bacteroidetes, Actinobacteria and Fusobacteria were also found to contribute to the overall diversity in the oral microbiome (figure 2a). Similar to the gut microbiome, the oral microbiome was significantly more diverse (Chao1) in aged individuals ($P < 0.05$, T test) and the diversity index of the oral microbiome in males was slightly lower than that of females. Higher diversity values were found in individuals with lower BMI values as was observed in gut microbiome (figure 2b). The median species diversity was found to be slightly higher in samples from individuals belonging to *pitta prakriti* among the three *prakriti* types (figure 2c).

Major bacterial genera in the oral microbiome included *Streptococcus, Neisseria, Veilonella, Haemophilus, Porphyromonas* and *Prevotella* which were consistently present across all *prakriti* types (figure 2d). We found 28 OTUs to be highly prevalent with presence in 180 or more samples (90% prevalence); of these, 9 OTUs were found in all the 200 samples studied. Only four OTUs (*Streptococcus, Neisseria, Veilonella* and *Haemophilus*) had ARA >10% across all samples while *Prevotella* and *Porphyromonas* had ARA >5% when all samples from different *prakriti*s were pooled together; *Bulleidia* which was present in all the samples had an ARA value of only 0.89 across all samples. Only 41 of the 132 OTUs were present in more than 50% of the samples while 62 OTUs were present in less than 10% of the samples. Out of 132 OTUs considered for the oral microbiome, 99 OTUs (75%) were shared among the three *prakriti*s, while only 21 OTUs were unique to either one of the three *prakriti*s. Of these 21 OTUs, 8 were unique to *kapha prakriti*,5 to *pitta* and 8 to *vata prakriti* (figure 2e). But none of these OTUs were found in more than 4 samples, indicating that the core microbiome was indeed truly shared among all *prakriti*s. However, certain species were found to be preferentially associated with specific *prakriti*s. *Actinobacillus* was found with slightly higher ARA (>3%) in males with *pitta prakriti* but was relatively less abundant in all other *prakriti* types. Similarly, TM7-3 was found more abundantly in *vata* females and *Capnocytophaga* was found more abundantly in female samples belonging to *kapha* and *pitta prakriti*s. A Principal Component Analysis with weighted unifrac distance measures could not cluster the samples into *prakriti* types based on the abundance of major genera (figure 2f).

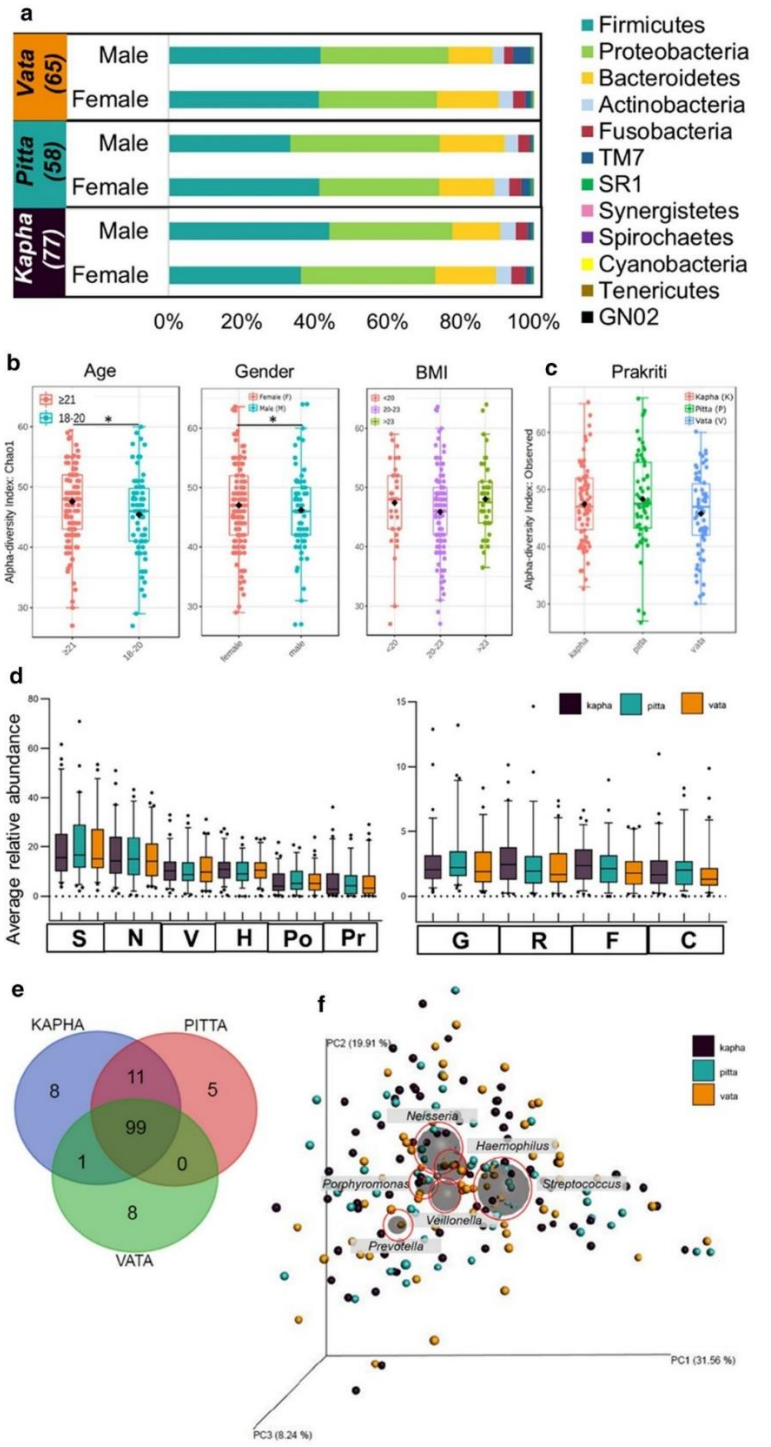


Figure 2: Oral microbiome: **(a)** Phylum plot showing average relative abundance (ARA) of major phyla in different cohorts. **(b)** Alpha diversity values for microbiome based on age, gender, BMI. **(c)** Alpha diversity values based on *prakriti*. **(d)** Box plot for the 10 most abundant OTUs (S – *Streptococcus*, N – *Neisseria*, V – *Veillonella*, H – *Haemophilus*, Po – *Porphyromonas*, Pr – *Prevotella*, G – *Granulicatella*, R – *Rothia*, F – *Fusobacterium*, C – *Capnocytophaga*) as determined by ARA. Boxes represent the interquartile range (IQR) and the line represents the median. Whiskers denote the lowest and highest values within 1.5 x IQR. Circles represent outliers beyond the whiskers. **(e)** Venn diagram showing unique and shared genera between the prakriti types. **(f)** PCoA with weighted unifrac distances. Each dot represents a sample while grey circles represent major OTUs. Sizes of the grey circles represent their ARA.

3.3 *Characterization of the healthy gut microbiome*

A comparison of core gut microbiome (present in >50% of the samples in each cohort) from our study with GFKB database of microbiome of healthy individuals was performed and OTUs missing from this list were analyzed for possible correlation with specific *prakriti* phenotype. We found *Butyricoccus* to be dominantly present in *kapha prakriti*, *Turicibacter* in *pitta prakriti*, and Para *Prevotella*, Christensenellaceae, *Mitsuokella*, S24-7 and Barnesiellaceae enriched in *vata prakriti* based on this classification (figure 3).

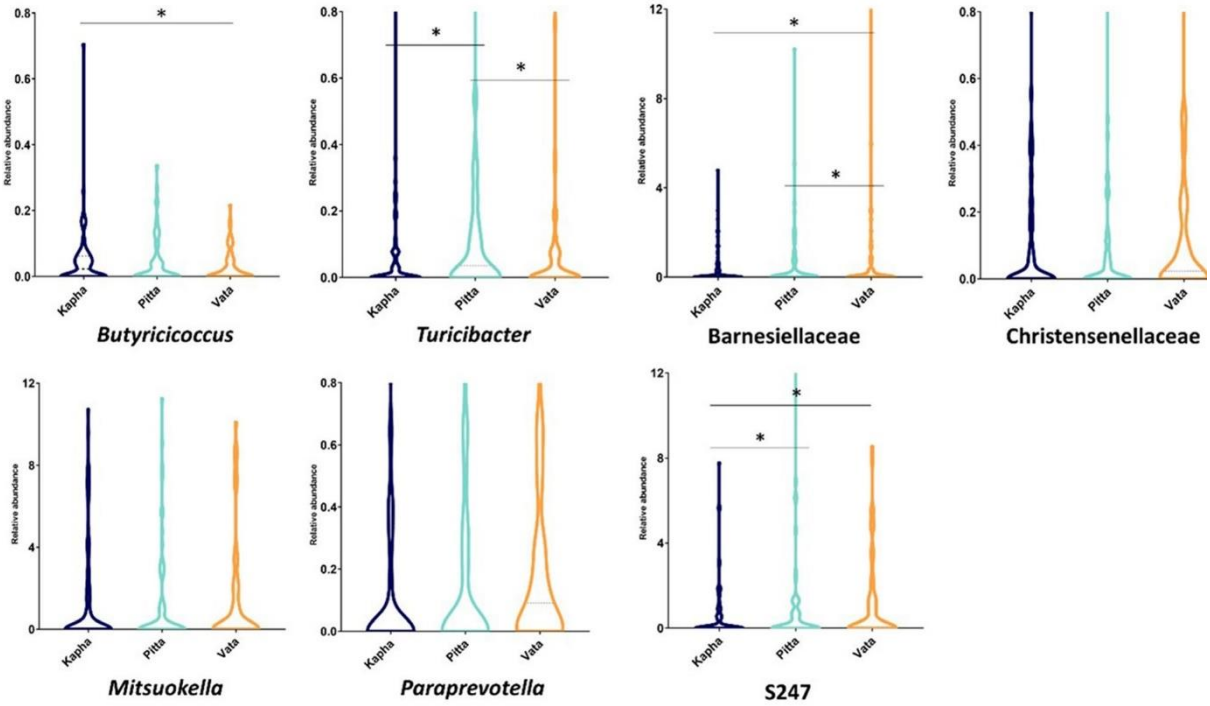


Figure 3: Comparison of 'core' gut microbiome of each *prakriti* with GFKB microbiome database of healthy individuals: *Butyricicoccus* was dominantly present in *kapha prakriti*, *Turicibacter* in *pitta prakriti*, and *Paraprevotella*, Christensenellaceae, *Mitsuokella*, S24-7 and Barnesiellaceae were enriched in *vata prakriti* as measured by one-way ANOVA followed by a post-hoc analysis with Tukey's test. *P <0.05

3.4 *Microbial features specific to prakriti*

Analysis of the OTUs from gut and oral microbiome using BugBase showed that facultative anaerobes were higher in individuals belonging to *kapha prakriti* (in both gut and oral microbiome). Interestingly, there was discordance in the pattern of potentially pathogenic microbes between gut and oral microbiome. While in the gut, potentially pathogenic organisms were found to be higher in *kapha prakriti* and lowest in *pitta prakriti*, in the oral microbiome, potential pathogens were higher in *vata prakriti* and least abundant in *pitta prakriti* (figure 4a,b).

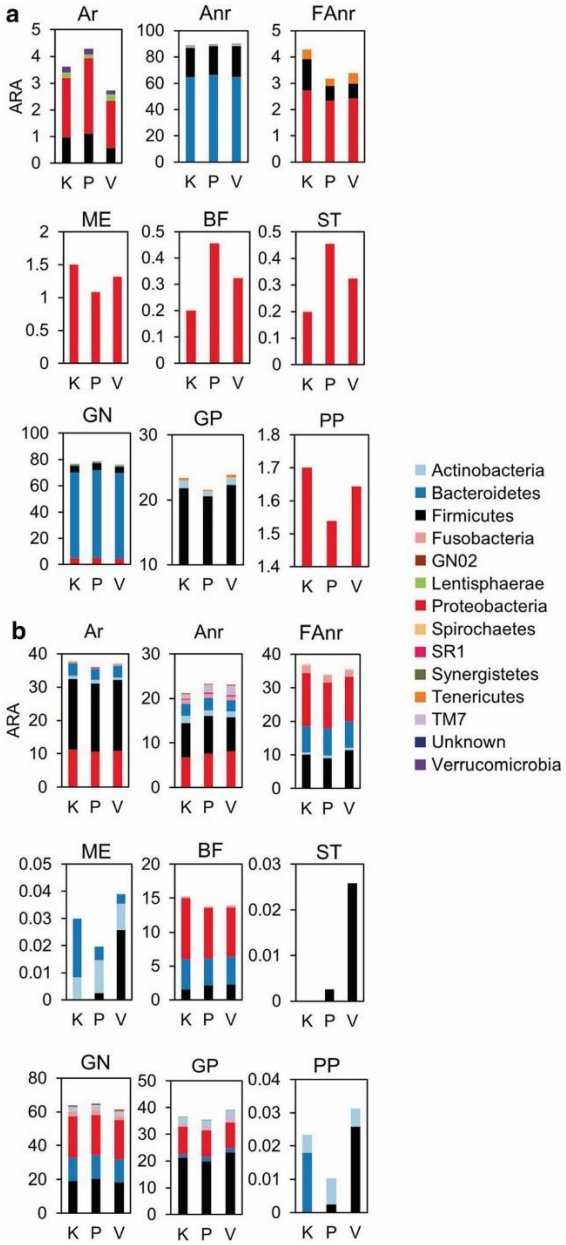


Figure 4: Classification of microbiome based on Bugbase: (a) Contribution of different groups of bacteria from gut microbiome to different categories based on Bugbase analysis. (b) Contribution of different groups of bacteria from oral microbiome to different categories based on Bugbase analysis; K – *Kapha*, P – *Pitta*, V – *Vata*, ARA – Average Relative Abundance, Ar – Aerobic, Anr – Anaerobic, FAnr – Facultative anaerobic, ME – Mobile elements, BF – Biofilm former, ST – Stress tolerance, GN – Gram negative, GP – Gram positive, PP – Potential pathogen.

3.4.1 *Gut microbiome:* A linear discriminant effect size (LEfSe) analysis showed that more bacterial groups were preferentially associated with individuals of *vata prakriti* (figure 5). Further to this, we identified signature species associated with specific gender in each *prakriti*. This showed several species such as *Prevotella copri* (*vata* females) and *Blautia wexlerae* (*vata* males) to be preferentially associated with specific *prakriti* phenotypes (table 2). In addition, we found more bacterial species to be specifically associated with female subjects.

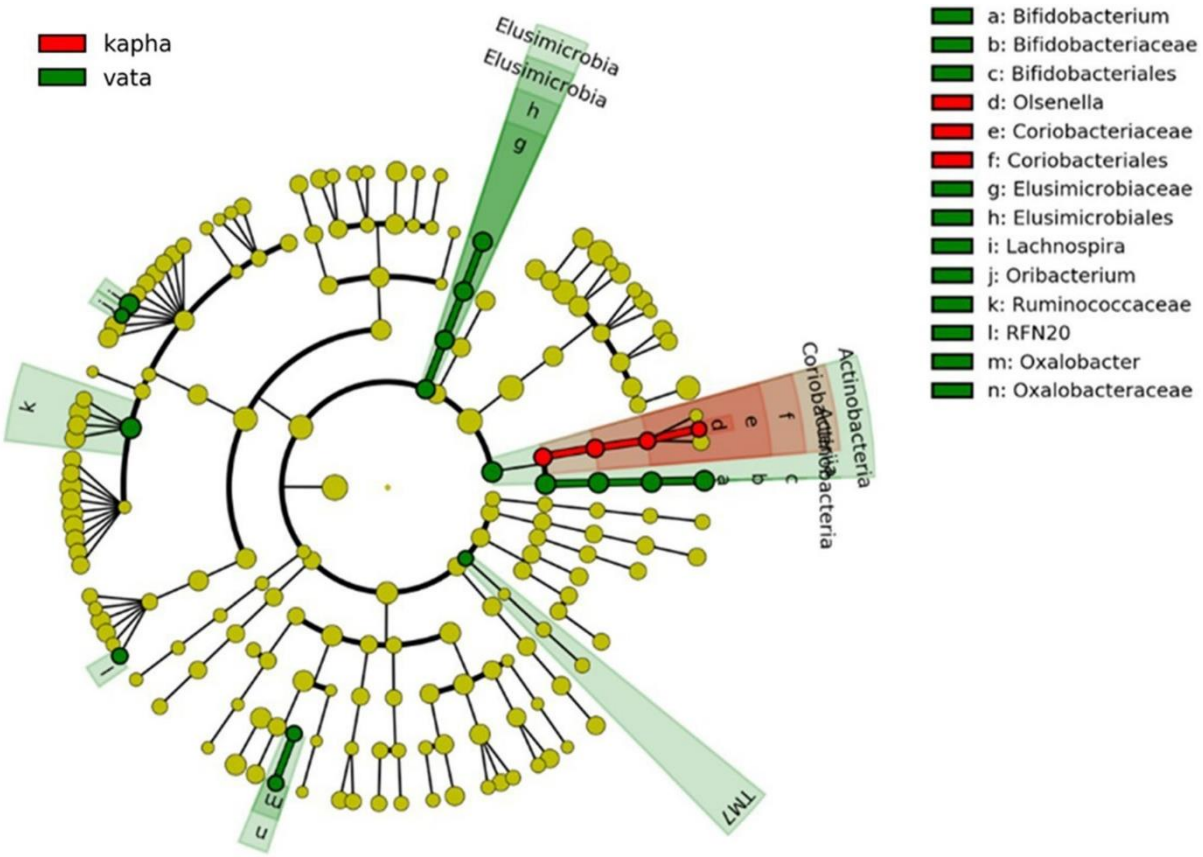


Figure 5: LEfSe analysis of gut microbiome: Cladogram showing OTUs that are significantly different among the three *prakritis* based on LEfSe analysis. The taxonomic levels are represented by rings with phyla at the outermost ring and genera at the innermost ring. Each circle is a member within that level. Coloured circles denote significant enrichment in their respective *prakritis* ($p<0.05$, LDA > 2).

Table 2: Bacterial species specific to prakriti and gender in gut microbiome (LDA score > 2.5) and their physiological relevance

Prakriti	Gender	Greengenes Taxonomy ID	Species	Physiological relevance in human gut	Reference
Vata	Female	187324	*Bacteroides caccae*	Anti-inflammatory, microbial disease target in inflammatory bowel disorder	Hiippala *et al.* 2020; Wei *et al.* 2001
Vata	Female	235262	*Bifidobacterium adolescentis*	Common commensal among adults	Duranti *et al.* 2016
Vata	Female	545061	*Prevotella copri*	Involved in complex fiber breakdown, provides tolerance to inflammation	De Filippis *et al.* 2019
Vata	Female	41229	*Sutterella megalosphaeroides*	Novel species of *Sutterella* (commensal of human gut), depletion leads to onset of Celiac disease	Hiippala *et al.* 2016; Sakamoto *et al.* 2018
Vata	Female	358781	*Ruminococcus bromii*	Primary degrader of resistant starch (common commensal of human gut)	Crost *et al.* 2018
Vata	Female	175751	*Kineothrix alysoides*	Saccharolytic butyrate producer (a key product for gut health)	Haas and Blanchard 2017
Vata	Female	266210	*Megasphaera elsdenii*	Lactate metabolizer, lowers rumen acidosis	Chen *et al.* 2019
Vata	Female	149335	*Mitsuokella jalaludinii*	Phytase producer, strictly anaerobic bacterium	Lan *et al.* 2002
Vata	Female	581048	*Elusimicrobium minutum*	Commonly associated with insect gut	Herlemann *et al.* 2009
Vata	Female	276149	*Parabacteroides gordonii*	Novel species of *Parabacteroides* reported to be negatively correlated with BMI	Wang *et al.* 2019b
Pitta	Female	246930	*Dialister hominis*	Pro-inflammatory, potential marker of spondylarthritis	Tito *et al.* 2017
Pitta	Female	341460	*Haemophilus parainfluenzae*	Common commensal of the oral microflora	Yang *et al.* 2018
Pitta	Female	368490	*Turicibacter sanguinis*	Neuromodulatory properties	Fung *et al.* 2019
Pitta	Female	4300127	*Mitsuokella multacida*	Butyrate producer (common commensal of human gut)	Chakravarthy *et al.* 2018
Pitta	Female	551822	*Intestinibacter bartlettii*	Implicated in etiology of neurodevelopmental disorders	Bojović *et al.* 2020
Kapha	Female	336012	*Bacteroides uniformis*	Beta glucuronidase glycoside hydrolase producer (common commensal of human gut)	Pellock *et al.* 2018
Kapha	Female	910353	*Streptococcus salivarius*	A prolific colonizer of human oropharyngeal tract, produces bacteriocins that offers protection from pathogens	Patras *et al.* 2015
Kapha	Female	4310221	*Millionella massiliensis*	Novel species reported in human gut	Mailhe *et al.* 2017
Kapha	Female	165118	*Paraprevotella clara*	Positively correlated with sedentary lifestyle, Succinic and acetic acid producer	Bressa *et al.* 2017
Kapha	Female	196769	*Bacteroides cellulosilyticus*	Extensively involved in carbohydrate utilization	McNulty *et al.* 2013
Pitta	Male	184287	*Ruminococcus bicirculans*	Non cellulolytic (common commensal of human gut), negatively correlated with mean arterial blood pressure and positively correlated with weight loss diets	Salonen *et al.* 2014; Wegmann *et al.* 2014
Pitta	Male	215269	*Gemmiger formicilis*	Associated with immune related colitis	Chaput *et al.* 2017
Vata	Male	192746	*Blautia wexlerae*	Anti-inflammatory, found significantly lower in obese children	Benítez-Páez *et al.* 2020
Vata	Male	842596	*Coprococcus* spp.	Butyrate producer, positively associated with quality of life indicators and depleted in depression	Duncan *et al.* 2002; Valles-colomer *et al.* 2019

3.4.2 *Oral microbiome:* A LEfSe analysis of the oral microbiome at the genera level did not show any significant differentially abundant features (data not shown). However, at the species level, several species of *Prevotella* and *Fusobacterium nucleatum* were found to be preferentially associated with specific *prakriti*s. We found more number of bacterial species that were differentially abundant to be associated with male subjects compared to females (table 3).

Table 3: Bacterial species specific to prakriti and gender in oral microbiome (LDA score > 2.5) and their physiological relevance

Prakriti	Gender	Greengenes Taxonomy ID	Species	Physiological relevance in human gut	Reference
Vata	Female	4459993	*Porphyromonas catoniae*	Prominently found in healthy caries-free adults	Soares-Castro *et al.* 2019
Kapha	Female	239506	*Prevotella histicola*	Commonly a human gut commensal with disease suppressing properties (arthritis, multiple sclerosis)	Maeda and Takeda 2019; Mangalam *et al.* 2017
Pitta	Male	4456252	*Megasphaera stantonii/ cerevisiae*	*M. stantonii* isolated from chicken caecum, *M. cerevisiae* is commonly found in nonpasteurized low-alcohol beer (Note: *Megasphaera* species are a major periodontal pathogen)	Maki and Looft 2018
Pitta	Male	4307464	*Brachymonas denitrificans*	Aerobic, chemo-organotroph	Laviad *et al.* 2015
Pitta	Male	4332410	*Prevotella nigrescens*	Found in subgingival plaque and is mainly necessary for periodontal health. However, it is reportedly found even in diseased conditions in a highly virulent manner	Szafrański *et al.* 2015
Pitta	Male	851923	*Prevotella oralis*	Major commensal of oral cavity	Karatay *et al.* 2015
Kapha	Male	4391542	*Fusobacterium nucleatum*	Reported as a gastric cancer specific bacterial signature/Common opportunistic commensal of oral microbiome	Brennan and Garrett 2019; Hsieh *et al.* 2018

4. Discussion

In the current study, an attempt was made to understand the influence of Ayurvedic *prakriti* phenotypes on the diversity of the gut/oral microbiome in healthy individuals. With this aim, healthy individuals belonging to three different *prakriti* phenotypes were studied for their oral and gut microbiome. Volunteers from different ethnicity and race residing in a single geographical area were considered to minimize location-based heterogeneity in the results (Mobeen *et al*. 2019).

We refrained from including subjects with habits like smoking and alcohol consumption as it is known to alter the gut microbiome. Subjects who had taken antibiotics 6 months prior to sample collection were excluded as it is also known to influence the gut microbiome. Antibiotics cause large disturbances in the microbiome composition and species-species interactions and thus disrupt the community structure (Modi *et al*. 2014).

The composition of the gut microbiome, though relatively simple at birth, undergoes series of changes in its composition and functions by the early age of 3 to 5 years, due to the influence of various genetic and environmental factors such as diet, lifestyle, age, geography, mode of delivery, infection, infant feeding modality (formula versus breastfed), maternal diet, diseases and medication (Rodríguez *et al*. 2015; Schmidt *et al*. 2018; Wen and Duffy 2017). It has been reported that the gut microbiota of infants delivered through caesarean section did not resemble their mother's gut profile (Arboleya *et al*. 2018; Backhed *et al*. 2015). Exercise and high fiber diet, such as fruits, vegetables, legumes, and whole-wheat grain products have also been shown to increase the microbial diversity (Clarke *et al*. 2014; Flint *et al*. 2012). In fact, diet directly regulates the gut microbial ecosystem and has a profound effect on the colonization of the gut thereby shaping the gut microbial ecosystem in the early stages of life (Rodríguez *et al*. 2015). Carbohydrates, proteins and fats are the macronutrients known to influence the gut microbial system (Rowland *et al*. 2018). For instance, *Bacteroides* have been shown to be predominant in the gut associated with western diet while, *Prevotella* enterotype was found to be associated with the plant based polysaccharides and non-western diet

(Gorvitovskaia *et al*. 2016). In another study, feces samples of people who consume coffee was shown to have increased presence of *Bifidobacterium* (Jaquet *et al*. 2009).

The fecal microbiome has been shown to be influenced by the type of diet consumed by the healthy subjects on previous few days. The animal-based diet is known to alter the microbiome which may be due to fecal bile acid profiles and the growth of microorganisms capable of triggering inflammatory bowel disease. These results suggest that dysbiosis of gut microbiome can be caused by the high fat diet which may lead to influence several diseases (Wen and Duffy 2017).

Similar to earlier studies on gut microbiome from Indian population (Arumugam *et al*. 2011; Chaudhari *et al*. 2020; Chauhan *et al*. 2018), we found Bacteroidetes and Firmicutes phyla to be the major component of the gut microbiome irrespective of the *prakriti* type or gender (figure 1a) suggesting their importance in maintaining the general wellbeing of the individual. However, in spite of this general similarity, certain subtle differences were discerned with respect to the composition of the gut microbiome between healthy individuals.

Interestingly, the major factors such as age, diet, lifestyle, stress and environment that can influence and cause alterations in the three *doshas* in an individual (Lakhotia 2014) are also known to affect the microbiome composition and function. This suggests a possible link between *prakriti* constitution and the microbiome assemblage and how subtle physiological or lifestyle changes can lead to disequilibrium and diseased state in healthy individuals. In our study, we found that the overall species diversity was significantly higher in aged individuals for both gut and oral microbiome (figure 1b, 2b). Other studies from India have reported *Prevotella, Bacteroides* and *Dialister* to be the major bacteria genera associated with gut microbiome (Chaudhari *et al*. 2019, 2020). We found these three genera as major contributors to the gut microbiome in all *prakriti* types (figure 1d) irrespective of the gender. Reports suggest *Prevotella* to be a major gut microbe associated with plant-rich diet since it plays an important role in metabolism of plant-based products (Chaudhari *et al*. 2020; Chen *et al*. 2017). Earlier studies on gut microbiome of western African population with diet rich in carbohydrates and fibres were also shown to be enriched with *Prevotella* (Bhute *et al*. 2016; De Filippo *et al*. 2010). Similarly, *Bacteroides* is reported to be associated with a non-vegetarian diet (Chaudhari *et al*. 2020). We found a negative correlation between relative abundance of *Prevotella* and *Bacteroides* suggesting enrichment of bacterial species based on the diet of the individuals. Das *et al*. (2018) in their study based on fecal microbiome of 84 healthy individuals from three Indian communities concluded that diet can significantly influence the gut microbiome. Interestingly, they found *Prevotella* to be more abundant in gut microbiome of individuals with non-vegetarian diet. However, Nishijima *et al*. (2016) compared the gut microbiome from healthy individuals from 12 different countries and found *Bacteroides* to be the most abundant genus reported from gut microbiome from different countries including China, Spain, Denmark and concluded that diet might not be the only factor that determines gut microbial diversity. We did not discern a clear clustering of samples based on gut microbiome composition of the three *prakriti* phenotypes. In a similar study, Mobeen *et al*. (2020) also did not observe a distinct clustering specific to the three *prakriti* types. This could be explained by the fact that the core microbiome of these *prakriti* phenotypes shared most of the bacterial species while very few unique species were associated with each type. Studies also indicate that gut microbiome of dizygotic twins is

much less similar compared to that of monozygotic twins indicating the importance of genetic factors in determining the gut microbiome (Goodrich *et al*. 2016a). Tschop *et al*. (2009) suggested that the core microbiome was shared among individuals albeit with differences in abundance of different organisms. In fact, Arumugam *et al*. (2011) opined that the presence of a few dominant species along with several low abundance species might contribute to homeostasis of the gut microbiome in healthy individuals. It could be argued that the less abundant species that form the core microbiome might play a crucial role in determining the functional attributes characteristic of each *prakriti* phenotype in healthy individuals and their perturbations due to internal and external factors might lead to diseased state.

It is an interesting and powerful hypothesis that despite being healthy, *Ayurveda* can discern individuals based on their constitution type and offer prognostic value in terms of predicting disease susceptibility. One could then wonder if this would have a scientific basis in terms of differences in the corresponding microbial assemblage. With this in mind, we compared the core microbiome from the three *prakriti* phenotypes (present in >50% of the samples in each cohort) and compared it with GFKB database which lists the microbiome typical of healthy individuals. GFKB database provides a baseline microbiome data from healthy individuals consisting of more than 150 organisms belonging to 59 bacterial genera (King *et al*. 2019). The OTU elements forming the core microbiome of each *prakriti*, that were not included in the GFKB database were taken further ahead to identify the unique OTUs characteristic of each *prakriti* (figure 3) and their functional roles were assessed from literature. *Butyricicoccus* which was enriched in *kapha prakriti* is a widely established commensal butyrate, short chain fatty acid (SCFA) producer that is known to suppress inflammatory bowel disorder (Wang *et al*. 2019a) and ulcerative colitis (Devriese *et al*. 2017). It is considered as a probiotic that can protect intestinal barriers from potentially harmful microbes (Ma *et al*. 2020). This correlates well with the *kapha* phenotype which relates to strength, stability (Sharma 2016) and functionally to disease resistance (Dey and Pahwa 2014). *Turicibacter* found in *pitta prakriti* individuals has been linked to host genetics (Kemis *et al*. 2019), host immunity and is associated with inflammation and cancer (Goodrich *et al*. 2016a, b). *Turicibacter sanguinis*, a common gut commensal, is involved in signalling intestinal cells to release serotonin, thereby playing a crucial role in altering immune and metabolic conditions (Fung *et al*. 2019). This correlates with the phenotype of *pitta prakriti*, which is associated with digestion, metabolism and transformation (processes involving energy exchange) (Sharma 2016). In a balanced state, these individuals are capable of quick metabolism of toxic substances and a good amount of disease resistance; while in an imbalanced state, *pitta* individuals are more prone to metabolism and digestion disorders such as coronary disease, ulcer, cancer of stomach, inflammation of lymph system and others (Dey and Pahwa 2014; Mishra *et al*. 2001).

Vata prakriti individuals had a higher number of bacteria absent in the GFKB database such as *Paraprevotella, Mitsuokella*, Barnesiellaceae, Christensenellaceae and S24-7 (figure 3). *Paraprevotella* is usually found to be negatively correlated with BMI, percentage body fat, adiposity index and estimated visceral fat, which fits the description of *vata* phenotype. A study by Bressa *et al*. (2017) found *Paraprevotella* to be significantly higher in active women. Christensenellaceae is reported to have 'inverse relationship' with BMI (Waters and Ley 2019) and aligns well with the *vata* phenotype (lower BMI). *Mitsuokella* is an anti-inflammatory,

polysaccharide degrading, short chain fatty acid producer. This genus has been found to be uniquely enriched in Indian gut populations as opposed to US and Chinese populations (Jain *et al*. 2018) and in individuals who consume a plant-based diet (Jayasudha *et al*. 2018; Shankar *et al*. 2017). S24-7, unique to *vata* phenotype, is reported to offer protection from type 2 diabetes (Hansen *et al*. 2015), while few other studies have reported it as an opportunistic pathogen (Ormerod *et al*. 2016) and to increase with onset of arthritis (Liu *et al*. 2016; Rogier *et al*. 2017). Though widely reported from animal studies (Lagkouvardos *et al*. 2019), their relative significance in human gut microbiome remains largely unexplored in spite of their wide prevalence. *Ayurveda* texts attribute rheumatoid arthritis to an accumulation of *ama* in the joints and *vata* imbalance (Gupta *et al*. 2015). Barnesiellaceae has been reported as a significant biomarker for high fiber diet (Ong *et al*. 2018), while Bressa *et al*. (2017) correlated it with sedentary lifestyle. The presence of these 'unique' genera in specific *prakriti*s can provide these constitutions with adaptable genomes which when perturbed by external factors can shift the balance from healthy to diseased state in individuals.

Oral microbiome is considered to be extremely diverse with more than 600 bacterial species occupying different niches with varying abundances in the oral cavity (Dewhirst *et al*. 2010). The overall diversity of oral microbiome is considered to be second only to the gut microbiome (Verma *et al*. 2018). Increasing evidence suggest that oral microbiome not only plays a major factor in infections leading to periodontitis and tonsillitis but also other systemic diseases such as diabetes, stroke, etc. (Genco *et al*. 2005; Joshipura *et al*. 2003). Phyla Firmicutes, Proteobacteria and Bacteroidetes have been reported to be major contributors to the oral microbiome (Dewhirst *et al*. 2010) as has been reported here (figure 2a). In our study, we also found TM7 as a component of oral microbiome in males of *vata prakriti*. Initially reported from peat bogs, TM7 consists of unculturable group of bacterial organisms reported to occur with high prevalence but low abundance in oral microbiome (Brinig *et al*. 2003; Podar *et al*. 2007; Rheims *et al*. 1996). The presence of division TM7 in the oral microbiome of apparently healthy individuals is interesting considering the fact that TM7 members are considered to be associated with sub-gingival plaques leading to periodontitis (Liu *et al*. 2012). Chaudhari *et al*. (2020) found significant difference in the alpha diversity estimates including Chao1 in the skin microbiome across different age groups. We also observed higher diversity in slightly higher aged individuals with reference to the oral microbiome but no significant difference was observed with reference to gender or BMI (figure 2b).

The major bacterial genera associated with oral microbiome in our study were found to be *Streptococcus* and *Neisseria* which agrees with earlier studies carried out on Indian population (Chaudhari *et al*. 2019). Chaudhari *et al*. (2020) reported that five genera namely *Neisseria, Streptococcus, Prevotella, Porphyromonas,* and *Haemophilus* contributed to more than two-third of the oral microbiome from 54 healthy individuals belonging to six joint families. In our study, we found members of the genus *Streptococcus* to be the most abundant species in the oral cavity (figure 2d). In addition to the 5 dominant genera, we found *Veilonella* which was consistently present across all *prakriti*s to be another major contributor to the oral microbiome. Both *Haemophilus* and *Veilonella* are reported to occupy distinct habitats in the mouth (Welch *et al*. 2016). Though *Mycoplasma* is considered to be extremely common in human saliva, we found this species in very few

samples. However, the presence of *Porphyromonas* and *Fusobacterium*, two common genera implicated in periodontitis, with high prevalence and abundance in healthy samples was intriguing.

Using BugBase, we found that anaerobes contributed more than 80% of the gut microbiome while in case of oral microbiome, aerobes, anaerobes and facultative anaerobes were found to contribute more or less equally to the overall diversity (figure 4a,b). In both oral and gut microbiome, majority of the organisms were found to be Gram negative in nature. Interestingly, a larger proportion of microbes identified from oral microbiome were reported to be biofilm formers. Though, some species were identified as potential pathogens in both oral and gut microbiome, the proportion of potentially pathogenic organisms were least abundant in individuals belonging to *pitta prakriti* in both cases (figure 4a,b).

A linear discriminant effect size analysis showed preferential association of several bacterial species in different *prakriti*s and gender (tables 2 and 3). *Fusobacterium nucleatum* which is reported as a biomarker for gastric cancer (Hsieh *et al*. 2018) was preferentially associated with oral microbiome of healthy *kapha* males. Further studies in this regard would provide valuable information on utilizing these microbes as prognostic and diagnostic biomarkers for specific disease states.

Recent studies propose that microbiome might play an integral role in precision medicine approach as prognostic, diagnostic and therapeutic biomarkers considering their role in disease pathogenesis and response to treatment. Gut microbiome is considered to be a better choice for personalized medicine due to their subtle heterogeneity among different individuals and their ability to respond to therapeutic interventions (Kashyap *et al*. 2017). While reduction in microbial diversity and their concomitant functional roles in elderly individuals is associated with higher chronic intake of drugs (Ticinesi *et al*. 2017), enrichment of certain bacteria such as *Akkermansia, Bifidobacterium* and Christensenellaceae have shown to be associated with increased life expectancy (Biagi *et al*. 2016). The *prakriti* concept of *Ayurveda* system places great emphasis on differences in functional attributes (*prakritis*) among individuals which can be used for personalized treatment. Similarly, functional omics which display much clear variations when perturbed are known to succinctly reflect these changes than overall alterations in microbial community structure (Heintz- Buschart and Wilmes 2018). However, the subtle differences in the microbiome in healthy individuals of different *prakriti* types clearly indicate that sudden variations in microbiome can also provide us valuable clues on predisposition to disease state. Hence, it becomes imperative to build information related to the baseline levels of the microbiome from healthy individuals so that any changes to these 'normal' microbiome levels can provide us with clues that can help in predicting the disease outcome and tailoring the treatment modalities to specific needs by suitable alterations to gut microbiome.

In our current study, a large number of bacterial genera were shared among all *prakriti* types. It is possible that the assessment of biochemical parameters alone does not suffice to classify an individual as healthy. In addition, we have sampled from a cosmopolitan city (Bangalore) which has individuals of multi-ethnic backgrounds with wide variation in diet, lifestyle and other factors that can influence the gut microbiome. To establish a better correlation that can capture subtle differences among the *prakriti* types, several factors such

as lifestyle, diet, drugs, age and stress need to be accounted for as these can influence the microbial diversity even in healthy individuals leading to heterogeneous data within the various *prakriti* types thus making it difficult to discern specific patterns.

Further studies to explore the association of *prakriti* with gut microbiome should involve sampling of a broad population of healthy individuals over time and account for all factors that can influence the microbial diversity patterns such as age, gender, geography, food habits and cultural traditions, in order to discover features of gut microbiome that are unique to different geographical areas/lifestyles and aid discovery of statistically enriched biomarkers for each *prakriti*. Further, studies of the microbial diversity patterns while accounting for parameters mentioned can enable exploration of how westernization (mainly in terms of diet and lifestyle) of the Indian population might have influenced shifts in microbial landscape – changes that potentially mediate the suite of pathophysiological states correlated with Westernization.

Acknowledgements

GG would like to thank Gokula Education Foundation for support, Science and Engineering Research Board, New Delhi, for funding, Dr Aswin Sai Narain Sesha- sayee, NCBS, for sequencing, and Ramaiah Medical College, Ramaiah College of Management, Ramaiah Nursing College, Ramaiah Pharmacy College, Ramakrishna *Ayurveda* Medical College and Indian Institute of *Ayurveda* Medicine for logistic support. KS thanks Science and Engineering Board (SERB), Department of Science and Technology (DST), TIFAC-CORE, India, and MAHE for the support and encouragement. AJ thanks MAHE for Dr. TMA Pai PhD Scholarship.

References:

Aggarwal S, Negi S, Jha P, Singh PK, Stobdan T, Pasha MAQ, Ghosh S, Agrawal A, et al. 2010 EGLN1 involvement in high-altitude adaptation revealed through genetic analysis of extreme constitution types defined in Ayurveda. Proc. Natl. Acad. Sci. U.S.A. 107 18961–18966.

Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Xu ZZ, Kightley EP, Thompson LR, et al. 2017 Deblur rapidly resolves single-nucleotide community sequence patterns. mSystems 2 e00191-16.

Arboleya S, Suárez M, Fernández N, Mantecón L, Solís G, Gueimonde M and de Los Reyes-Gavilan CG 2018 C-section and the neonatal gut microbiome acquisition: consequences for future health. Ann. Nutr. Metab. 73 17–23.

Arumugam M, Raes J, Pelletier E, Paslier DL, Yamada T, Mende DR, Fernandes GR, Tap J, et al. 2011 Enterotypes of the human gut microbiome. Nature 473 174–180.

Bäckhed F, Roswall J, Peng Y, Feng Q, Jia H, Kovatcheva- Datchary P, Li Y, Xia Y, et al. 2015 Dynamics and stabilization of the human gut microbiome during the first year of life. Cell Host Microbe 17 690–703.

Benítez-Páez A, del Pugar EMG, López-Almela I, Moya- Pérez Á, Codoñer-Franch P and Sanz Y 2020 Depletion of *Blautia* species in the microbiota of obese children relates to intestinal inflammation and metabolic phenotype worsening. mSystems 5 e00857-19.

Bhalerao S, Deshpande T and Thatte U 2012 *Prakriti* (Ayurvedic concept of constitution) and variations in platelet aggregation. BMC Complement. Altern. Med. 12 248.

Bhute S, Pande P, Shetty SA, Shelar R, Mane S, Kumbhare SV, Gawali A, Makhani H, et al. 2016 Molecular characterization and meta-analysis of gut microbial communities illustrate enrichment of *Prevotella* and *Megasphaera* in Indian subjects. Front. Microbiol. 7 660.

Biagi E, Franceschi C, Rampelli S, Severgnini M, Ostan R, Turroni S, Consolandi C, Quercia S, et al. 2016 Gut microbiota and extreme longevity. Current Biology 26 1480–1485.

Bojović K, Ignjatović D, Bajić SS, Milutinović DV, Tomić M, Golić N and Tolinački M 2020 Gut microbiota dysbiosis associated with altered production of short chain fatty acids in children with neurodevelopmental disorders. Front. Cell. Infect. Microbiol. 10 223.

Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, et al. 2019 Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat. Biotechnol. 37 852–857.

Brennan CA and Garrett WS 2019 *Fusobacterium nucleatum*— symbiont, opportunist and oncobacterium. Nat. Rev. Microbiol. 17 156–166.

Bressa C, Bailén-Andrino M, Pérez-Santiago J, González- Soltero R, Pérez M, Montalvo-Lominchar MG, Maté- Muñoz JL, Domínguez R, et al. 2017 Differences in gut microbiota profile between women with active lifestyle and sedentary women. PLoS One 12 e0171352.

Brinig MM, Lepp PW, Ouverney CC, Armitage GC and Relman DA 2003 Prevalence of bacteria of division TM7 in human subgingival plaque and their association with disease. Appl. Environ. Microbiol. 69 1687–1694.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, et al. 2010 QIIME allows analysis of high-throughput community sequencing data. Nat. Methods 7 335–336.

Chakravarthy SK, Jayasudha R, Ranjith K, Pinna NK, Mande SS, Sharma S, Garg P, Murthy SI, et al. 2018 Alterations in the gut bacterial microbiome in fungal Keratitis patients. PLoS One 13 e0199640.

Chaput N, Lepage P, Coutzac C, Soularue E, Le Roux K, Monot C, Boselli L, Routier E, et al. 2017 Baseline gut microbiota predicts clinical response and colitis in metastatic melanoma patients treated with ipilimumab. Ann. Oncol. 28 1368–1379.

Chaudhari D, Dhotre D, Agarwal D, Gondhali A, Nagarkar A, Lad V, Patil U, Juvekar S, et al. 2019 Understanding the association between the human gut, oral and skin microbiome and the Ayurvedic concept of *prakriti*. J. Biosci. 44 1–8.

Chaudhari DS, Dhotre DP, Agarwal DM, Gaike AH, Bhalerao D, Jadhav P, Mongad D, Lubree H, et al. 2020 Gut, oral and skin microbiome of Indian patrilineal families reveal perceptible association with age. Sci. Rep. 10 1–13.

Chauhan NS, Pandey R, Mondal AK, Gupta S, Verma MK, Jain S, Ahmed V, Patil R, et al. 2018 Western Indian rural gut microbial diversity in extreme *Prakriti* endo-phenotypes reveals signature microbes. Front. Microbiol. 9 118.

Chen L, Shen Y, Wang C, Ding L, Zhao F, Wang M, Fu J and Wang H 2019 *Megasphaera elsdenii* lactate degradation pattern shifts in rumen acidosis models. Front. Microbiol. 10 162.

Chen T, Long W, Zhang C, Liu S, Zhao L and Hamaker BR 2017 Fiber-utilizing capacity varies in *Prevotella*-versus *Bacteroides*-dominated gut microbiota. Sci. Rep. 7 1–7.

Chong J, Liu P, Zhou G and Xia J 2020 Using MicrobiomeAnalyst for comprehensive statistical, functional, and meta-analysis of microbiome data. Nat. Protoc. 15 799–821.

Clarke SF, Murphy EF, O'Sullivan O, Lucey AJ, Humphreys M, Hogan A, Hayes P, O'Reilly M, et al. 2014 Exercise and associated dietary extremes impact on gut microbial diversity. Gut 63 1913–1920.

Crost EH, Le Gall G, Laverde-Gomez JA, Mukhopadhya I, Flint HJ and Juge N 2018 Mechanistic insights into the cross-feeding of *Ruminococcus gnavus* and *Ruminococcus bromii* on host and dietary carbohydrates. Front. Microbiol. 9 2558.

Das B, Ghosh TS, Kedia S, Rampal R, Saxena S, Bag S, Mitra R, Dayal M, et al. 2018 Analysis of the gut microbiome of rural and urban healthy indians living in sea level and high altitude areas. Sci. Rep. 8 1–15.

De Filippis F, Pasolli E, Tett A, Tarallo S, Naccarati A, De Angelis M, Neviani E, Cocolin L, et al. 2019 Distinct genetic and functional traits of human intestinal *Prevotella copri* strains are associated with different habitual diets. Cell Host Microbe 25 444–453.

De Filippo C, Cavalieri D, Di Paola M, Ramazzotti M, Poullet JP, Massart S, Collini S, Pieraccini G, et al. 2010 Impact of diet in shaping gut microbiota revealed by a comparative study in children from Europe and rural Africa. Proc. Natl. Acad. Sci. U.S.A. 107 14691–14696.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, et al. 2006 Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microbiol. 72 5069–5072.

Devriese S, Eeckhaut V, Geirnaert A, Van den Bossche L, Hindryckx P, Van de Wiele T, Van Immerseel F, et al. 2017 Reduced mucosa-associated *Butyricicoccus* activity in patients with ulcerative colitis correlates with aberrant claudin-1 expression. J. Crohn's Colitis 11 229–236.

Dewhirst FE, Chen T, Izard J, Paster BJ, Tanner ACR, Yu WH, Lakshmanan A and Wade WG 2010 The Human oral microbiome. J. Bacteriol. 192 5002–5017.

Dey S and Pahwa P 2014 Prakriti and its associations with metabolism, chronic diseases, and genotypes: Possibilities of new born screening and a lifetime of personalized prevention. J. Ayurveda Integr. Med. 5 15–24.

Duncan SH, Barcenilla A, Stewart CS, Pryde SE and Flint HJ 2002 Acetate utilization and butyryl coenzyme A (CoA): acetate-CoA transferase in butyrate-producing bacteria from the human large intestine. Appl. Environ. Microbiol. 68 5186–5190.

Durack J and Lynch SV 2019 The gut microbiome: Relationships with disease and opportunities for therapy. J. Exp. Med. 216 20–40.

Duranti S, Milani C, Lugli GA, Mancabelli L, Turroni F, Ferrario C, Magnifesta M, Viappiani A, et al. 2016 Evaluation of genetic diversity among strains of the human gut commensal *Bifidobacterium adolescentis*. Sci. Rep. 6 1–10.

Edgar RC, Haas BJ, Clemente JC, Quince C and Knight R 2011 UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27 2194–2200.

Flint HJ, Scott KP, Louis P and Duncan SH 2012 The role of the gut microbiota in nutrition and health. Nat. Rev. Gastroenterol. Hepatol. 9 577–589.

Fung TC, Vuong HE, Luna CDG, Pronovost GN, Aleksandrova AA, Riley NG, Vavilina A, McGinn J, et al. 2019 Intestinal serotonin and fluoxetine exposure modulate bacterial colonization in the gut. Nat. Microbiol. 4 2064–2073.

Genco RJ, Grossi SG, Ho A, Nishimura F and Murayama Y 2005 A proposed model linking inflammation to obesity, diabetes, and periodontal infections. J. Periodontol. 76 2075–2084.

Goecks J, Nekrutenko A, Taylor J and Galaxy team 2010 Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol. 11 1–13.

Goodrich JK, Davenport ER, Beaumont M, Jackson MA, Knight R, Ober C, Spector TD, Bell JT, et al. 2016a Genetic determinants of the gut microbiome in UK twins. Cell Host Microbe 19 731–743.

Goodrich JK, Davenport ER, Waters JL, Clark AG and Ley RE 2016b Cross-species comparisons of host genetic associations with the microbiome. Science 352 532–535.

Gorvitovskaia A, Holmes SP and Huse SM 2016 Interpreting *Prevotella* and *Bacteroides* as biomarkers of diet and lifestyle. Microbiome 2016 15.

Govindaraj P, Nizamuddin S, Sharath A, Jyothi V, Rotti H, Raval R, Nayak J, Bhat BK, et al. 2015 Genome-wide analysis correlates Ayurveda Prakriti. Sci. Rep. 5 1–12.

Gupta SK, Thakar AB, Dudhamal TS and Nema A 2015 Management of Amavata (rheumatoid arthritis) with diet and Virechanakarma. Ayu 36 413–415.

Gupta VK, Kim M, Bakshi U, Cunningham KY, Davis JM, Lazaridis KN, Nelson H, Chia N and Sung J 2020 A predictive index for health status using species-level gut microbiome profiling. Nat. Commun. 11 1–16.

Haas KN and Blanchard JL 2017 *Kineothrix alysoides*, gen. nov., sp. nov., a saccharolytic butyrate-producer within the family Lachnospiraceae. Int. J. Syst. Evol. Microbiol. 67 402–410.

Hansen AK, Krych Ł, Nielsen DS and Hansen CHF 2015 A review of applied aspects of dealing with gut microbiota impact on rodent models. ILAR J. 56 250–264.

Heberle H, Meirelles GV, da Silva FR, Telles GP and Minghim R 2015 InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. BMC Bioinformatics 16 1–7.

Heintz-Buschart A and Wilmes P 2018 Human gut microbiome: function matters. Trends Microbiol. 26 563–574.

Herlemann DPR, Geissinger O, Ikeda-Ohtsubo W, Kunin V, Sun H, Lapidus A and Brune A 2009 Genomic analysis of '*Elusimicrobium minutum*', the first cultivated representative of the phylum '*Elusimicrobia*' (formerly termite group 1). Appl. Environ. Microbiol. 75 2841–2849.

Hiippala K, Kainulainen V, Kalliomaki M, Arkkila P and Satokari R 2016 Mucosal prevalence and interactions with the epithelium indicate commensalism of *Sutterella* spp. Front. Microbiol. 7 1706.

Hiippala K, Kainulainen V, Suutarinen M, Heini T, Bowers JR, Jasso-Selles D, Lemmer D, Valentine M, et al. 2020 Isolation of anti-inflammatory and epithelium reinforcing *Bacteroides* and *Parabacteroides* spp. from a healthy fecal donor. Nutrients 12 935.

Hsieh Y, Tung S, Pan H, Yen C, Xu H, Lin Y, Deng Y, Hsu W, et al. 2018 Increased abundance of *Clostridium* and *Fusobacterium* in gastric microbiota of patients with gastric cancer in Taiwan. Sci. Rep. 8 1–11.

Illumina 2013 16S Metagenomic Sequencing Library Preparation Retrieved from https://support.illumina.com/ documents/documentation/chemistry_documentation/16s/ 16s-metagenomic-library-prep-guide-15044223-b.pdf on 6th February 2021.

Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, Tyson JR, Beggs AD, et al. 2018 Nanopore sequencing and assembly of a human genome with ultra-long reads. Nat. Biotechnol. 36 338–345.

Jaquet M, Rochat I, Moulin J, Cavin C and Bibiloni R 2009 Impact of coffee consumption on the gut microbiota: A human volunteer study. Int. J. Food Microbiol. 130 117–121.

Jayasudha R, Chakravarthy SK, Prashanthi GS, Sharma S, Garg P, Murthy SI and Shivaji S 2018 Alterations in gut bacterial and fungal microbiomes are associated with bacterial Keratitis, an inflammatory disease of the human eye. J. Biosci. 43 835–856.

Joshipura KJ, Hung HC, Rimm EB, Willet WC and Ascherio A 2003 Periodontal disease, tooth loss, and incidence of ischemic stroke. Stroke 34 47–52.

Karatay M, Koktekir E, Celik H, Erdem Y, Sertbas I and Bayar MA 2015 Spinal intradural abscess caused by hematogenous spread of *Prevotella oralis* in a 3-year-old child with an asymptomatic congenital spinal abnormality. Spinal Cord 53 S13–S15.

Kashyap PC, Chia N, Nelson H, Segal E and Elinav E 2017 Microbiome at the frontier of personalized medicine. Mayo Clin. Proc. 92 1855–1864.

Kemis JH, Linke V, Barrett KL, Boehm FJ, Traeger LL, Keller MP, Rabaglia ME, Schueler KL, et al. 2019 Genetic determinants of gut microbiota composition and bile acid profiles in mice. PLoS Genet. 15 e1008073.

King CH, Desai H, Sylvetsky AC, Lo Tempio J, Ayanyan S, Carrie J, Crandall KA and Fochtman BC 2019 Baseline human gut microbiota profile in healthy people and standard reporting template. PLoS One 14 e0206484.

Kopylova E, Noé L and Touzet H 2012 SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics 28 3211–3217.

Klindworth A, Pruesse E, Schweer T, Peplies J, Quast C, Horn M, et al. 2013 Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Research 41 e1.

Lagkouvardos I, Lesker TR, Hitch TCA, Gálvez EJC, Smit N, Neuhaus K, Wang J, Baines JF, et al. 2019 Sequence and cultivation study of Muribaculaceae reveals novel species, host preference, and functional potential of this yet undescribed family. Microbiome 7 1–15.

Lakhotia SC 2014 Translating Ayurveda's Dosha-Prakriti into objective parameters. J. Ayurveda Integr. Med. 5 176.

Lan GQ, Abdullah N, Jalaludin S and Ho YW 2002 Culture conditions influencing phytase production of *Mitsuokella jalaludinii*, a new bacterial species from the rumen of cattle. J. Appl. Microbiol. 93 668–674.

Laviad S, Lapidus A, Han J, Haynes M, Reddy T, Huntemann M, Pati A, Ivanova NN, et al. 2015 High quality draft genome sequence of *Brachymonas chironomi* AIMA4 T (DSM 19884 T) isolated from a Chironomus sp. egg mass. Stand. Genomic Sci. 10 1–7.

Liu B, Faller LL, Klitgord N, Mazumdar V, Ghodsi M, Sommer DD, Gibbons TR, Treangen TJ, et al. 2012 Deep sequencing of the oral microbiome reveals signatures of periodontal disease. PLoS One 7 e37919.

Liu X, Zeng B, Zhang J, Li W, Mou F, Wang H, Zou Q, Zhong B, et al. 2016 Role of the gut microbiome in modulating arthritis progression in mice. Sci. Rep. 6 1–11.

Ma Q, Li Y, Wang J, Li P, Duan Y, Dai H, An Y, Cheng L, et al. 2020 Investigation of gut microbiome changes in type 1 diabetic mellitus rats based on high-throughput sequencing. Biomed. Pharmacother. 124 109873.

Maeda Y and Takeda K 2019 Host–microbiota interactions in rheumatoid arthritis. Exp. Mol. Med. 51 1–6.

Mailhe M, Ricaboni D, Benezech A, Cadoret F, Fournier PE and Raoult D 2017 '*Millionella massiliensis*' gen. nov., sp. nov., a new bacterial species isolated from human right colon. New Microbes New Infect. 17 11–12.

Maki JJ and Looft T 2018 *Megasphaera stantonii* sp. nov., a butyrate-producing bacterium isolated from the cecum of a healthy chicken. Int. J. Syst. Evol. Microbiol. 68 3409–3415.

Mangalam A, Shahi SK, Luckey D, Karau M, Marietta E, Luo N, Choung RS, Ju J, et al. 2017 Human gut-derived commensal bacteria suppress CNS inflammatory and demyelinating disease. Cell Rep. 20 1269–1277.

McNulty NP, Wu M, Erickson AR, Pan C, Erickson BK, Martens EC, Pudlo NA, Muegge BD, et al. 2013 Effects of diet on resource utilization by a model human gut microbiota containing *Bacteroides cellulosilyticus* WH2, a symbiont with an extensive glycobiome. PLoS Biol. 11 e1001637.

Mishra LC, Singh BB and Dagenais S 2001 Ayurveda: a historical perspective and principles of the traditional healthcare system in India. Altern. Ther. Health Med. 7 36–43.

Mobeen F, Sharma V and Prakash T 2019 Functional signature analysis of extreme *Prakriti* endophenotypes in gut microbiome of western Indian rural population. Bioinformation 15 490–505.

Mobeen F, Sharma V and Prakash T 2020 Comparative gut microbiome analysis of the *Prakriti* and Sasang systems reveals functional level similarities in constitutionally similar classes. Biotech 10 1–15.

Modi SR, Collins JJ and Relman DA 2014 Antibiotics and the gut microbiota. J. Clin. Invest. 124 4212–4218.

Murthy SKR 2009 Sushrutha Samhita commentary by Dalhana. Chaukhambha Orientalia, Varanasi, Vol II: Sutra Sthana 15/3.

Nicholson JK, Holmes E and Kinross J 2012 Host-gut microbiota metabolic interactions. Science 336 1262–1267.

Nishijima S, Suda W, Oshima K, Kim S, Hirose Y, Morita H and Hattori M 2016 The gut microbiome of healthy Japanese and its microbial and functional uniqueness. DNA Res. 23 125–133.

Ong IM, Gonzalez JG, McIlwain SJ, Sawain EA, Schoen AJ, Adluru N, Alexander AL and Yu JJ 2018 Gut microbiome populations are associated with structurespecific changes in white matter architecture. Transl. Psychiatry 8 1–11.

Ormerod KL, Wood DLA, Lachner N, Gellatly SL, Daly JN, Parsons JD, Dal'Molin CGO, Palfreyman RW, et al. 2016 Genomic characterization of the uncultured Bacteroidales family S24-7 inhabiting the guts of homeothermic animals. Microbiome 4 1–17.

Patras KA, Wescombe PA, Rösler B, Hale JD, Tagg JR and Doran KS 2015 *Streptococcus salivarius* K12 limits group B *Streptococcus* vaginal colonization. Infect. Immun. 83 3438–3444.

Pellock SJ, Walton WG, Biernat KA, Torres-Revera D, Creekmore BC, Xu Y, Liu J, Tripathy A, et al. 2018 Three structurally and functionally distinct b-glucuronidases from the human gut microbe *Bacteroides uniformis*. J. Biol. Chem. 293 18559–18573.

Podar M, Abulencia CB, Walcher M, Hutchison D, Zengler K, Garcia JA, Holland T, Cotton D, et al. 2007 Targeted access to the genomes of low-abundance organisms in complex microbial communities. Appl. Environ. Microbiol. 73 3205–3214.

Prasher B, Aggarwal S, Mandal AK, Sethi TP, Deshmukh SR, Purohit SG, Sengupta S, Khanna S, et al. 2008 Whole genome expression and biochemical correlates of extreme constitutional types defined in Ayurveda. J. Transl. Med. 6 1–12.

Prasher B, Gibson G and Mukerji M 2016 Genomic insights into ayurvedic and western approaches to personalized medicine. J. Genet. 95 209–228.

Qiagen 2016 QIAamp® DNA mini and blood mini handbook (Cat no. 51304). Retrieved from https://www.qiagen.com/us/resources/

Qiagen 2010 QIAamp® DNA Stool Handbook (Cat no. 51504). Retrieved from https://www.qiagen.com/us/resources/

Rheims H, Rainey FA and Stackebrandt E 1996 A molecular approach to search for diversity among bacteria in the environment. J. Ind. Microbiol. 17 159–169.

Ribeiro RM, de Souza-Basqueira M, de Oliveira LC, Salles FC, Pereira NB and Sabino EC 2018 An alternative storage method for characterization of the intestinal microbiota through next generation sequencing. Rev. Inst. Med. Trop. Sao Paulo 60 e77.

Rodríguez JM, Murphy K, Stanton C, Ross RP, Kober OI, Juge N, Avershina E, Rudi K, et al. 2015 The composition of the gut microbiota throughout life, with an emphasis on early life. Microb. Ecol. Health Dis. 26 26050.

Rogier R, Evans-Marin H, Manasson J, van der Kraan PM, Walgreen B, Helsen MM, van der Bersselaar LA, van de Loo FA, et al. 2017 Alteration of the intestinal microbiome characterizes preclinical inflammatory arthritis in mice and its modulation attenuates established arthritis. Sci. Rep. 7 1–12.

Rognes T, Flouri T, Nichols B, Quince C and Mahé F 2016 VSEARCH: a versatile open source tool for metagenomics. PeerJ 4 e2584–e2584.

Rotti H, Mallya S, Kabekkodu PS, Chakrabarty S, Bhale S, Bharadwaj R, Bhat BK, Dedge AP, et al. 2015 DNA methylation analysis of phenotype specific stratified Indian population. J. Transl. Med. 13 1–12.

Rotti H, Raval R, Anchan S, Bellampalli R, Bhale S, Bharadwaj R, Bhat BK, Dedge AP, et al. 2014 Determinants of *prakriti*, the human constitution types of Indian traditional medicine and its correlation with contemporary science. J. Ayurveda Integr. Med. 5 167–175.

Rowland I, Gibson G, Heinken A, Scott K, Swann J, Thiele I and Tuohy K 2018 Gut microbiota functions: metabolism of nutrients and other food components. Eur. J. Nutr. 57 1–24.

Sakamoto M, Ikeyama N, Kunihiro T, Lino T, Yuki M and Ohkuma M 2018 *Mesosutterella multiformis* gen. nov., sp. nov., a member of the family Sutterellaceae and *Sutterella megalosphaeroides* sp. nov., isolated from human faeces. Int. J. Syst. Evol. Microbiol. 68 3942–3950.

Salonen A, Lahti L, Salojärvi J, Holtrop G, Korpela K, Duncan SH, Date P, Farquharson F, et al. 2014 Impact of diet and individual variation on intestinal microbiota composition and fermentation products in obese men. ISME J. 8 2218–2230.

Schmidt TSB, Raes J and Bork P 2018 The human gut microbiome: from association to modulation. Cell 172 1198–1215.

Segata N, Izard J, Waldron L, Gevers D, Miropolsky L, Garrett WS and Huttenhower C 2011 Metagenomic biomarker discovery and explanation. Genome Biol. 12 1–18.

Shankar V, Gouda M, Moncivaiz J, Gordon A, Nv Reo, Hussein L and Paily O 2017 Differences in gut metabolites and microbial composition and functions between Egyptian and U.S. children are consistent with their diets. mSystems 2 e00169-16.

Sharma H 2016 Ayurveda: Science of life, genetics, and epigenetics. Ayu 37 87–91.

Sharma RK and Dash B 2009 Charaka Samhita of Agnivesha. Ayurveda Deepika Commentary, Chaukhambha Sanskrit Series Vol II, Shareera Sthana 3/6, 7 pp 370.

Soares-Castro P, Araújo-Rodrigues H, Godoy-Vitorino F, Ferreira M, Covelo P, López A, Vingada J, Eira C, et al. 2019 Microbiota fingerprints within the oral cavity of cetaceans as indicators for population biomonitoring. Sci. Rep. 9 1–15.

Szafrański SP, Deng ZL, Tomasch J, Jarek M, Bhuju S, Meisinger C, Kühnisch J, Sztajer H, et al. 2015 Functional biomarkers for chronic periodontitis and insights into the roles of *Prevotella nigrescens* and *Fusobacterium nucleatum*; a metatranscriptome analysis. NPJ Biofilms Microbiomes 1 1–13.

Ticinesi A, Lauretani F,Milani C,Nouvenne A, Tana C, Del Rio D, MaggioM,Ventura M, et al. 2017Aging gut microbiota at the cross-road between nutrition, physical frailty, and sarcopenia: is there a gut-muscle axis? Nutrients 9 1303.

Tito RY, Cypers H, Joossens M, Varkas G, Praet LV, Glorieus E, Van den Bosch F, De Vos M, et al. 2017 Brief report: *Dialister* as a microbial marker of disease activity in spondyloarthritis. Arthritis Rheumatol. 69 114–121.

Tschöp MH, Hugenholtz P and Karp CL 2009 Getting to the core of the gut microbiome. Nat. Biotechnol. 27 344–346.

Valles-Colomer M, Falony G, Darzi Y, Tigchelaar EF, Wang J, Tito RY, Schwiek C, Kurilshikov A, et al. 2019 The neuroactive potential of the human gut microbiota in quality of life and depression. Nat. Microbiol. 4 623–632.

Verma D, Garg PK and Dubey AK 2018 Insights into the human oral microbiome. Arch. Microbiol. 200 525–540.

Wang M, Wichienchot S, He X, Fu X, Huang Q and Zhang B 2019a *In vitro* colonic fermentation of dietary fibers: Fermentation rate, short-chain fatty acid production and changes in microbiota. Trends Food Sci. Technol. 88 1–9.

Wang YJ, Xu XJ, Zhou N, Sun Y, Liu C, Liu S and You X 2019b *Parabacteroides acidifaciens* sp. nov., isolated from human faeces. Int. J. Syst. Evol. Microbiol. 69 761–766.

Ward T, Larson J, Meulemans J, Hilmann B, Lynch J, Sidiropoulos D, Spear JR, Caporaso G, et al. 2017 BugBase predicts organism-level microbiome phenotypes. BioRxiv 133462.

Waters JL and Ley RE 2019 The human gut bacteria Christensenellaceae are widespread, heritable, and associated with health. BMC Biol. 17 1–11.

Wegmann U, Louis P, Goesmann A, Henrissat B, Duncan SH and Flint HJ 2014 Complete genome of a new Firmicutes species belonging to the dominant human colonic microbiota (‘*Ruminococcus bicirculans’*) reveals two chromosomes and a selective capacity to utilize plant glucans. Environ. Microbiol. 16 2879–2890.

Wei B, Dalwadi H, Gordon LK, Landers C, Bruckner D, Targan SR and Braun J 2001 Molecular cloning of a *Bacteroides caccae* TonB-linked outer membrane protein identified by an inflammatory bowel disease marker antibody. Infect. Immun. 69 6044–6054.

Welch JLM, Rossetti BJ, Rieken CW, Dewhirst FE and Borisy GG 2016 Biogeography of a human oral microbiome at the micron scale. Proc. Natl. Acad. Sci. U.S.A. 113 E791–E800.

Wen L and Duffy A 2017 Factors influencing the gut microbiota, inflammation, and type 2 diabetes. J. Nutr. 147 1468S-1475S.

Wu GD, Chen J, Hoffmann C, Bittinger K, Chen YY, et al. 2011 Linking long-term dietary patterns with gut microbial enterotypes. Science 334 105–108.

Yang C, Yeh Y, Yu H, Chin C, Hsu C, Liu H, Huang P, Hu S, et al. 2018 Oral microbiota community dynamics associated with oral squamous cell carcinoma staging. Front. Microbiol. 9 862.

Chapter

3

Cancer Insights: A Pathway and Network Analysis to Understand Oral Cancer

Shetty, Smitha Sammith, Mohit Sharma, Felipe Paiva Fonseca, Pradyumna Jayaram, **Ankit Singh Tanwar**, Shama Prasada Kabekkodu, Kapaettu Satyamoorthy, and Raghu Radhakrishnan. "Signaling pathways promoting epithelial mesenchymal transition in oral submucous fibrosis and oral squamous cell carcinoma." *Japanese Dental Science Review* 56, no. 1 (2020): 97-108.

DOI: 10.1016/j.jdsr.2020.07.002; **IF 6.4 (2021)**

Signaling Pathways Promoting Epithelial Mesenchymal Transition in Oral Submucous Fibrosis and Oral Squamous Cell Carcinoma

Summary

Epithelial-mesenchymal transition (EMT) is a critical process that occurs during the embryonic development, wound healing, organ fibrosis and the onset of malignancy. Emerging evidence suggests that the EMT is involved in the invasion and metastasis of cancers. The inflammatory reaction antecedent to fibrosis in the onset of oral submucous fibrosis (OSF) and the role of EMT in its malignant transformation indicates a hitherto unexplored involvement of EMT. This review focuses on the role of EMT markers which are regulators of the EMT mediated complex network of molecular mechanisms involved in the pathogenesis of OSF and OSCC. Further the gene enrichment analysis and pathway analysis support the association of the upregulated and downregulated genes in various EMT regulating pathways.

1. Introduction

Oral submucous fibrosis (OSF) is a chronic mucosal condition occurring predominantly among the Indians, possibly due to prolonged use of areca nut, leading to marked rigidity and inability to open the mouth (Ekanayaka and Tilakaratne, 2013). The pathogenesis of OSF is not clearly understood, but there is compelling evidence to suggest that OSF is a result of collagen deregulation (Rajalalitha and Vali, 2005). Therefore, an increase in collagen formation concomitant with reduced collagen degradation is one of the plausible explanations for the onset of this condition (Rajalalitha and Vali, 2005). An alarming complication associated with OSF is the higher risk of transforming to oral squamous cell carcinoma (OSCC). It has been described that the pathological changes in the connective tissue of OSF are likely to affect the overlying epithelium and induce EMT (Ekanayaka and Tilakaratne, 2013).

OSCC is the most common type of oral cancer, which accounts for 3.8% of all the cancers and 3.6% of cancer deaths (Shield et al, 2017). Despite its ease of access for diagnosis and treatment, the mortality rate remains high because of the increased risk of developing second primary malignancy, which is the leading cause of death in patients with head and neck cancer (Krisanaprakornkit and Iamaroon, 2012). It is well-established that EMT contributes to the acquisition of invasive behavior, essential for metastasis and invasion (Gonzalez and Medici, 2014). Although the experimental evidence points to the association of EMT in the mechanism of metastasis, its specific role in human cancer needs to be further explored (Zidar et al, 2018). Identification of signature genes influencing EMT may unravel novel pathways, which are critical to the progression of oral cancer. These markers of EMT may prove to be efficient targets to control the further spread and improve the prognosis of OSCC. With this background, this review focuses on the mechanisms and signaling pathways that direct the change in the gene expression signatures inducing EMT in OSF and OSCC.

2. Epithelial to mesenchymal transition (EMT)

Epithelial to mesenchymal transition (EMT) is a biological process involving the transition of a polarized epithelial cell into a cell that has the characteristics of a mesenchymal phenotype (Kalluri and Weinberg, 2009). EMT is crucial for developmental milestones such as gastrulation of the metazoans, neural crest formation, and heart morphogenesis (Larue and Bellacosa, 2005). EMT is shown to be elicited following chronic inflammation and during wound healing (Yanjia and Xinchun, 2007). The role of EMT has increasingly gained significance as an essential process in fibrosis and carcinogenesis. Considering the involvement of EMT in various physiological and pathological mechanisms, three types of EMTs have been described (Kalluri and Weinberg, 2009).

Type 1 EMT is associated with implantation, embryo formation and organ development. This type of EMT generates mesenchymal cells that have the ability to undergo a further transition to form secondary epithelia (Kalluri and Weinberg, 2009). Type II EMT is associated with inflammation related to wound healing, organ fibrosis and tissue regeneration. Type III EMT is associated with tumor formation, progression and metastasis (Kalluri and Weinberg, 2009). Although the molecular basis for all the three types of EMT remains the same, the type I EMT is a physiological process during organogenesis, which further undergoes MET (Mesenchymal to epithelial transition) to form the secondary epithelia, hence reversal in the expression of EMT inducing genes may be seen. Type II EMT presents with upregulation of ECM proteins and transcription factors bring about the phenotypic switch to induce fibrosis. Type III EMT exhibits change in the phenotype with the upregulation of EMT associated genes to promote tissue invasion and metastasis of cancer cells (Zeisberg and Neilson, 2009).

EMT is a process in which there is a reduced expression of epithelial genes (E-cadherin) and an increase in the expression of mesenchymal genes (N-cadherin) and EMT transcription fac-tors. Together with an altered localization of the β-catenin, the epithelial cells lose their phenotype and intercellular adhesions (Das et al, 2013). Besides, there is an increased expression of Vimentin (Mendez et al, 2016) signifying a mesenchymal change in the cytoskeleton. An increase in Tenascin (Tak et al, 2015) implies that the matrix deposition enables the migration of cells. Significantly, the matrix metalloproteinase 9 (MMP9) (Das et al, 2013) overexpression demonstrates the disruption of the basement membrane and the proneness of cells to infiltrate the underlying stroma (Lee and Nelson, 2012; Scanlon et al, 2013).

The inflammatory reaction antecedent to fibrosis and the role of EMT in fibrogenesis and malignant transformation in other organs (Kriz et al, 2011; Hyun et al, 2016; Horowitz and Thannickal, 2006), points to the involvement of EMT in the pathogenesis of OSF and its malignant transformation. The inflammatory cytokines produced in response to the inflammation may mediate the progression of OSF via various EMT pathways (Sharma et al, 2018). The membranous loss of E-cadherin (Pal et al, 2010), β-catenin (Das et al, 2013), Cytokeratin 5(CK5), and Cytokeratin 14(CK14) (Ranganathan et al, 2006) with an overwhelming expression of vimentin (Nayak et al, 2013), N-cadherin (Das et al, 2013), and α-Smooth muscle actin (α-SMA) (Chang et al, 2014; Angadi et al, 2011) seen in OSF further confirms the role of EMT in OSF (Table 1).

Table 1: Molecular events regulating EMT in the pathogenesis of OSF and OSCC.

	OSF	OSCC
Epithelial markers	• ↓ E –cadherin (membranous loss) (CDH1) • ↓ Cytokeratin 5, 14 (CK5, CK14) (KRT 5, 14)	• ↓ E –cadherin (membranous loss) (CDH1) • ↑ Cytokeratin 5, 14 (CK5, CK14) (KRT 5, 14)
Mesenchymal markers	• ↑Vimentin (VIM) • ↑N- cadherin (CDH2) • ↑Alpha Smooth muscle Actin (α-SMA)(ACTA2)	• ↑ Vimentin (VIM) • ↑ N- Cadherin (CDH2) • ↑Alpha Smooth muscle Actin (α-SMA)(ACTA2)
Transcription factors	• ↑ SNAI-1 • ↑ Twist • ↑ ZEB1	• ↑ SNAI-1 SNAI-2 • ↑ Twist • ↑ ZEB1 • ↑ LEF-1

3. Transcription factors inducing EMT

Several transcription factors have been shown to induce EMT by repressing the transcription of cell adhesion molecules and driving epithelial cell reprogramming. These transcription factors bind to the promoter region of the CDH1 gene encoding E-cadherin and thus initiate EMT (Gonzalez and Medici, 2014). An important attribute of EMT is the loss of expression of cell-cell adhesion molecule, E-cadherin. Among the transcription factors directly contributing to this process includes snail super family of zinc-finger transcription factors, Snail1 and Snail2 (also known as Slug) (Wang et al, 2013), zinc finger E-box–binding homeobox (ZEB) family with the ZF (zinc finger) class of homeodomain transcription factors ZEB1 (Liu et al, 2008), ZEB2 (Hanrahan et al, 2017) and TWIST1 gene, which encodes a basic helix-loop-helix (bHLH) transcription factor (Wang et al, 2016) and lymphoid enhancer binding factor-1 (LEF-1) (Sun et al, 2017) (Table 1).

The Snail1 and Snail2 (Slug) belong to the snail superfamily consisting of highly conserved C terminal domain with zinc fingers that bind to the E-box motif in the target gene promoters (Wang et al, 2013). The transcription factors of the Snail family plays an important role in EMT through the functional inhibition and suppression of E-cadherin expression by binding to the promoter region of the E-cadherin gene (Batlle et al, 2000). However, in the fibrotic buccal mucosa fibroblasts, snail binds to the E-box in the α-SMA promoter and brings about upregulation of myofibroblasts expression, thus perpetuating fibrosis (Yang et al, 2018). Stromal Snail positivity has been reported in OSCC due to the presence of fibro-/myofibroblasts generated from the dedifferentiated carcinoma cells (Franz et al, 2009). An inverse correlation between the E-cadherin and snail expression has been observed in OSCC cells in vitro (Yokoyama et al, 2001). The cells lines that exhibit upregulated Snail expression along with the downregulation of E-cadherin and desmoglein 2 show higher invasive potential implicating the role of Snail transcription factor in driving the epithelial cell reprogramming (Yokoyama et al, 2001; Kume et al, 2013). TGF-β was shown to upregulate Snail (SNAI1) and Slug (SNAI2) expression in OSCCs and thereby promote chemo- resistance to anti-cancer drugs (Nakamura et al, 2018). The upregulation of both Snail and Slug in OSCC cells showed decreased sensitivity to anti-cancer drugs. Hence knock down of Slug and Snail would suppress its chemo-resistance to anti-cancer drugs (Nakamura et al, 2018).

Twist, a basic helix-loop-helix domain-containing transcription factor functions as a transcription repressor to activate EMT (Yang et al, 2006; Lee et al, 2016). Ectopic expression of Twist has resulted in the loss of E-cadherin mediated cell-cell adhesion, activation of mesenchymal markers, and gain of cell motility (Yang et al, 2006). Twist was shown to be upregulated in fibroblasts of lung tissue in idiopathic pulmonary fibrosis patients (Tan et al, 2017). Further, the arecoline treated cells show enhanced expression of Twist and

myofibroblast transdifferentiation, with the silencing of Twist being able to reverse this phenomenon (Lee et al, 2016). The role of Twist playing a critical role in the progression and metastasis of head and neck carcinomas (Zhuo et al, 2015), including OSCC, has been demonstrated (Wushou et al, 2012; de Freitas et al, 2012). Twist overexpression has been associated with clinical outcomes such as advanced clinical stage, presence of lymph node metastasis, distant metastasis and local recurrence (Zhuo et al, 2015).

Zinc finger E-box binding homeobox 1 (ZEB1) is a well-known activator of the EMT programme (Chang et al, 2014). ZEB1 functions as a transcription repressor that negatively regulates the expression of polarity markers, such as E-cadherin, MucI and Pkp3 (Franz et al, 2009). ZEB1 plays a pathogenic role in the induction of the myofibroblast activity of buccal mucosal fibroblasts (BMFs) by binding to the promoter region of α-SMA and hence inducing myofibroblasts transdifferentiation and promoting fibrosis (Chang et al, 2016). In OSCC, a negative correlation exists between ZEB1 and E- cadherin expression (Yao et al, 2017). Overexpression of ZEB1 and loss of E-cadherin expression is shown to be associated with local recurrence, lymph node metastasis and advanced pathological grading (Yao et al, 2017). ZEB-1 promotes EMT by interacting with acetyltransferases p300/pCAF and SMADs to activate the target genes that contribute to mesenchymal differentiation (Schmalhofer et al, 2009). Upregulated of ZEB1 has been noted in recurrent OSCC cases compared to primary lesions, indicating its role as marker of tumor recurrence (Ho et al, 2015).

Lymphoid enhancer-binding factor 1 (LEF1), a member of the T-cell Factor (TCF)/LEF1 family of transcription factors, is a down- stream mediator of the Wnt/β-catenin signaling pathway that promotes the transcription of the Wnt target genes (Santiago et al, 2017). It has an essential role in EMT by activating the transcription of N-cadherin, Vimentin and Snail (Santiago et al, 2017). During embryogenesis, EMT is executed by the binding of SMAD2-P-SMAD4-LEF1 complex to three binding regions in the E- cadherin promoter leading to its transcriptional silencing (Nawshad et al, 2007). Activation of the Wnt-β catenin pathway promotes the transcription of downstream target genes such as *c-myc*, LBH, Oct4, Nanog and LEF1 in various carcinomas (Santiago et al, 2017). In OSCC, LEF1 is over expressed in the moderate and poorly differentiated carcinomas (Su et al, 2014). Overexpressed LEF1 maintains the cancer cells in undifferentiated embryonic stem cell morphology. This may be the mechanism by which LEF1 promotes tumor invasion and its overexpression is associated with poor prognosis (Su et al, 2014).

4. Signaling pathways in EMT

4.1. TGF-β activated Smad signaling in EMT

The canonical signaling pathway for TGF-β involves the Smad transcription activators. TGF- β pathway is the most common pathway that induces EMT. The signaling pathway is activated by TGF-β superfamily of ligands, which includes the 3 isoforms for TGF- β and 6 isoforms of Bone morphogenetic protein (BMP2−7). Binding of TGF-β to the cell membrane receptor TGF βR, activates type II TGF-βR to Trans phosphorylate and activate the type I TGF- βR. This recruits the receptor-activated Smads, the R- Smads (Smad2/3), which are phosphorylated to form a complex with Smad4. In the BMP signaling, Smad 1/Smad 5 form complex with

Smad4. This trimeric Smad complex translocates into the nucleus and bind to the promoter region of the target genes and thus activates or represses the transcription of regulatory genes (Lamouille et al, 2014).

The inhibitory Smads (I-Smads), Smad6 and Smad7 are negative regulators. They negatively regulate the Smad activation by competing with Smad2 and Smad3 or Smad1 and Smad5 for binding to the type I TGF-βR (Derynck et al, 2003). I- Smads also recruits the E3 ubiquitin- protein ligases Smurf1 and Smurf2 for proteasomal degradation of Smad proteins. It acts by forming complexes with smurfs in the nucleus, translocates to the plasma membrane and induces ubiquitination and proteasomal degradation of the TGFβ receptors hence terminating Smad-mediated signaling (Biernacka et al, 2011) (Fig. 1).

TGF-β induces EMT through transcription factors like Snail1 and ZEB1. The downstream mediator Smad3-Smad4 complex translocate into the nucleus and interacts with the transcriptional repressor Snail1 to form a complex which then targets the promoters of genes encoding E-cadherin and Occludin (Albanel et al, 2009). Several feedback loops between transcription factors and microRNAs also regulate the TGF-β induced EMT. A Double-negative feedback loop exists between Snail1/miR-34 and ZEB1/miR-200 and an autocrine feedback loop between TGF-β /miR-200, which brings about EMT changes (Yan et al, 2014).

TGF-β is upregulated in OSF tissues (Rai et al, 2020; Pant et al, 2016; Kamath et al, 2015) and its activation is shown by the nuclear localization of p-SMAD2 (Khan et al, 2011). The extracts of areca nut induce TGF-β signaling in epithelial cells with increased levels of p-SMAD2, indicating the induction of TGF-β ligand (TGF-2) and its activator Thrombospondin1 (THBS-1) leading to activation of TGF-β pathway (Khan et al, 2012). Thus, there is a pro fibrotic cascade involving TGF-β pathway triggered in epithelium that influences the underlying submucosa for a fibrotic response (Pant et al, 2015). Also there is down regulation of BMP7 in OSF as induced by TGF-β, suppressing the antifibrotic effect of BMP7 (Khan et al, 2011). In OSCCs, BMP 7, 2 (Titidej et al, 2018) expression is associated with the tumor differentiation and lymph node metastasis and hence indicative of poor prognosis (Titidej et al, 2018).

Defective TGF-β Smad signaling pathway may lead to loss of proliferation inhibitory effect of TGF-β. Loss or decreased expression of the TGF-β receptors effect the regulatory function of TGF-β, hence is associated with carcinogenesis and tumor progression (Peng et al, 2006). TGF-β treated OSCC cell lines showed EMT changes characterized by transformation to fibroblasts like cells with downregulation of E-cadherin and upregulation of Vimentin (Meng et al, 2011). TGF-β also induces THBS-1 in OSCC, which promotes the migration of cancer cells and upregulates the MMPs thereby favoring the OSCC invasion (Titidej et al, 2018) (Table 2).

Table 2: Signaling pathways regulating EMT in OSF and OSCC.

Pathways	OSF	OSCC
TGF-β signaling	↑ Transforming growth factor β1 & β2 ↑ Collagen type 1(COL1A1) ↑ Type I plasminogen activator inhibitor (PAI-1) ↑ Thrombospondin1 (THBS1) ↓ Bone Morphogenetic protein 7 (BMP7) ↑ SMAD2	↑ Transforming growth factor β1 & β2 (TGFB1 & TGFB2) ↑ Collagen type 1, 4 (COL1A1, COL4A1) ↑ Type I plasminogen activator inhibitor (PAI-1) ↑ Thrombospondin1 (THBS1) ↑ Bone Morphogenetic protein 7, 2 (BMP7, 2) ↓ SMAD7
RTK signaling	↑ Insulin-like growth factor-1 (IGF1) ↑ Vascular endothelial growth factor (VEGF) ↑ Basic fibroblast growth factor (bFGF) ↑ Fibroblast growth factors 2 (FGF2) ↑ Epidermal growth factor (EGF) ↑ Tumor necrosis factor α (TNF-α)(TNFa)	↑ Insulin-like growth factor-1 (IGF1) ↑ Vascular endothelial growth factor (VEGF) ↑ Basic fibroblast growth factor (bFGF) ↑ Fibroblast growth factors 2(FGF2) ↑ Epidermal growth factor (EGF) ↑ Tumor necrosis factor α (TNF-α) (TNFa)
Hypoxia signaling pathway	↑ Hypoxia-inducible factor-1α (HIF-1α)	↑ Hypoxia-inducible factor-1α (HIF-1α)
Wnt signaling	↓ Secreted frizzled related proteins 1, 5 (SFRP1,5) ↓ -β - catenin (membranous loss) (CTNNB1) ↓ Wnt Inhibitory factor (WIF1) ↓Dickkopf WNT signaling pathway inhibitor 3 (DKK3)	↓Secreted frizzled related proteins 1,4, 5 (SFRP1, SFRP4, SFRP5) ↓ -β - catenin (membranous loss) (CTNNB1) ↓ Wnt Inhibitory factor (WIF1) ↓ Dickkopf WNT signaling pathway inhibitor 3 (DKK3)
MAPk/ERK pathway	↑c-myc (MYC)	↑ c-myc (MYC)
Akt/mTOR pathway	↓ Phosphatase and tensin homolog (PTEN) ↑ PHLPP2	↓ Phosphatase and tensin homolog (PTEN)
Matrix signaling pathway	↑ αvβ6 Integrin (ITGB6) ↑ β1 integrin (ITGB1) ↓ Matrix Metallopeptidase 1 (MMP1) ↑ Matrix Metallopeptidase 2, 9 (MMP2, MMP9) ↑ Nuclear factor-kappa B (NF-κB) (NFKB1)	↑ αvβ6 Integrin (ITGB6) ↑ β1 integrin (ITGB1) ↑ Matrix Metallopeptidase 1, 2, 7, 9 (MMP1, MMP2, MMP7, MMP9) ↑ Nuclear factor-kappa B (NF-κB) (NFKB1)

4.2. Non-Smad signaling in TGF-β-induced EMT

TGF-β induces EMT alternatively by initiating non-Smad signaling, which leads to the activation of pathways that are more commonly considered as the effectors pathways of receptor tyrosine kinase (RTK) signaling, such as PI3K/Akt, Erk, and p38 (Mitogen-activated protein kinase) MAPK, and Rho-GTPases path- ways (Zhang et al, 2017). Activation of non-Smad pathways can occur as an indirect response to Smad-mediated gene expression induced by TGF-β. Direct activation of non-Smad signaling can occur through the interaction of signaling mediators directly with the TGF-β receptors or through other adopter proteins (Derynck et al, 2003).

4.3. PI3 kinase/Akt /mTOR signaling in EMT

Activation of PI3 kinase/Akt signaling by TGF-β plays a significant role in inducing EMT. TGF-β activates phosphoinositide 3-kinase (PI3K) through its receptors or trans-activation through epidermal growth factor (EGF) and platelet –derived growth factor (PDGF) receptors. PI3K on activation phosphorylates phosphatidylinositol 4,5-bisphosphate (PIP2) to phosphatidylinositol 3,4,5- trisphosphate (PIP3), a phospholipid membrane protein that binds Akt. Upon binding, Akt is phosphorylated and activated by phosphoinositide- dependent kinase 1 (PDK1). Phosphatase and tensin homolog (PTEN) facilitate the dephosphorylation of PIP3 (Gonzalez and Medici, 2015). Loss of function and mutation of PTEN is observed in various carcinomas (Carracedo and Pandolfi, 2008). Integrin-linked kinase (ILK) on activation may alternatively mediate the phosphorylation of Akt through integrins. Akt induces EMT in SCCs by promoting the transcription of snail through nuclear factor-kB (NF-kB) (Julien et al, 2007). TGF-β brings about the change in cell size and protein content during EMT by activation of mammalian target of rapamycin complex 1 (mTORC1) and mTORC2 through Akt to bring about cell migration and invasion (Lamouille et al, 2007; Lamouille et al, 2012) (Fig. 1).

Arecoline, the major active ingredient in the betel nut is involved in the pathogenesis of OSF. Downregulation of Akt/mTOR pathway in HacaT cells is seen on treatment with arecoline, executed through suppression in

phosphorylation of AKT, mTOR and eukaryotic initiation factor 4E–binding protein 1(4E-BP1) (Gu et al, 2019). The arecoline induced downregulation of the Akt/mTOR pathway is mediated through the upregulation of PH domain Leucine-rich repeat Protein Phosphatase 2 (PHLPP2), an upstream target of Akt. siRNA-mediated knockdown of PHLPP2 recovered the phosphorylation state of Akt, as well as attenuated the effect of arecoline on cell viability (Gu et al, 2019). Inverse correlation between p-Akt and E-cadherin expression is observed in OSCC. Akt activation represses E-cadherin gene transcription by upregulation of the transcription repressors SNAIL, TWIST (de Freitas et al, 2012) and Smad interacting protein 1 (SIP1) (Grille et al, 2003). Akt is associated with invasiveness, enhancement of proliferation, growth and anti-apoptosis, hence upregulation of Akt was associated with poor prognosis in patient with OSCC (Lim et al, 2004). Upregulation of Akt and PI3K with inactivation of PTEN is reported to be induced by tobacco components such as nicotine (West et al, 2003). Progressive decrease in expression of PTEN is observed in OSF, suggesting TGF-β mediated loss of PTEN that results in decreased apoptosis, increased survival of fibroblast leading to fibrosis (Angadi and Krishnapillai, 2012). The genes associated with the PI3K/AKT pathway, including PI3K, Akt, RAS and PTEN, are infrequently found to be mutated in Head and neck squamous cell carcinoma (HNSCC) and are rarely reported in OSCC cases (Cohen et al, 2011) (Table 2).

4.4. MAPK/ERK pathway in EMT

MAP kinases represent the cytoplasmic components of the signaling pathway that are activated by tyrosine kinases and the G protein-coupled receptors. The activation of the MAPK pathway by the family of TGF-β proteins are weaker than those induced by the receptor tyrosine kinase (RTK) ligands. Erk1/2 MAPK signaling is activated by TGF-β through the association of ShcA (Src homology and collagen A adaptor protein A) with TGF-βRI and subsequent phosphorylation at tyrosine and serine, which provides a docking site for the growth factor receptor-bound protein 2 (Grb2) and the son of sevenless (SOS) proteins hence initiating the MAPK pathway (Lamouille et al, 2014; Lee et al, 2007). The ShcA/Grb2/SOS complex converts G protein, such as Ras into its active GTP-bound form, which binds Raf kinase. The MAP kinase pathway is composed of three consecutive kinases (MAPKKK, MAPKK, and MAPK) leading to its phosphorylation to MEK (MAPKK), which on further phosphorylation forms MAPK (ERK). MAPK now functions as an enzyme and translocate into the nucleus to bring about phosphorylation and activation of various transcription factors to induce EMT (McCain, 2012; Zhang and Liu, 2006). The TGFβ induced pathway stabilizes SNAIL1 by inhibiting GSK3, thus increasing SNAIL1 activity, and repression of E- cadherin (Lamouille et al, 2014) (Fig. 1).

Areca nut and arecoline induce the activation of the ERK/JNK/p38 MAPK pathways, through phosphorylation of p-Akt, p-ERK and p-p38 and activation of these pathways play an important role in OSF and OSCC (Dai et al, 2014; Chang et al, 2004). These pathways regulate the expression of matrix metalloproteinases (MMPs) and tissue inhibitors of metalloproteinases (TIMPs) to promote wound healing and fibrosis (Dai et al, 2014). Upregulation of the downstream targets like *c-myc* has also been reported in OSF, where the expression of *c-myc* may be correlated with the progressive cellular transformation in these precancerous conditions (Eversole and Philip, 1955; Srinivasan and Jewell, 2001). Induction of Ras/ERK pathways by EGF reduces the

interferon-α mediated apoptosis of epidermoid carcinoma cells, indicating the survival of DNA damaged cells via this pathway (Caraglia et al, 2003).

Mutation in Ras or Raf oncogenes leads to the activation of ERK1/ERK2 pathways in many cancers (Schreck and Rapp, 2006). But there have been discrepancies pertaining to ERK1/ERK2 expression in oral cancer and HNSCC (Aguzzi et al, 2009), however, this may be due to the differentiation stage of the tumor, where poorly differentiated tumors present with decreased phosphorylation of ERK leading to increase in cell proliferation and cancer progression (Uzgare et al, 2003).

4.5. Receptor tyrosine kinase (RTK) signaling in EMT

RTK signaling pathway can be activated by various growth factors such as epidermal growth factor (EGF), vascular endothelial growth factor (VEGF), platelet-derived growth factor (PDGF), fibroblast growth factor (FGF) and insulin-like growth factor (IGF). These growth factors bind to the external domain of RTK, inducing the dimerization and subsequent auto phosphorylation of the tyrosine residue in the receptor. Hence activating the downstream signaling pathways such as PI3K/Akt/mTOR and ERK/MAPK pathway (Gonzalez and Medici, 2014; Lamouille et al, 2014) (Fig. 1).

Basic Fibroblast growth factor-2 (bFGF-2) induces EMT by decreasing the expression of cytokeratin and E-cadherin and inducing the expression of vimentin, FSP-1 and αSMA (Strutz et al, 2002). bFGF is upregulated in the early stages of OSF (Bishen et al, 2008), with increased expression in fibroblasts and endothelial cells. The expression in fibroblasts may be due to the heparan sulfate, which shows enrichment of bFGF-binding domains in fibrotic lesions, and these regions may play an important role in the fibrogenesis through their interaction with endogenous bFGF (Baird et al, 1988). The increased bFGF expressivity in endothelial cells along with fibroblasts may potentiate the leukocyte recruitment to inflammation by enhancing endothelial adhesion molecule expression (Zittermann and Issekutz, 2006). OSCCs showed the increased intensity of bFGF staining in the invasive fonts indicating the role of cancer cells in producing the bFGF. However, the expression was regardless of its clinical characteristics. bFGF promotes the production of proteinases and enhances the invasive capabilities of the cancer cells (Hase et al, 2006).

While increased expression of EGFR was noted in the stratum spinosum of the epithelium, TGF-α was restricted to stratum germinativum, indicating an upregulation of TGF-α initially and then exerting a paracrine effect of the non-proliferative cells to increase the expression of cell surface receptor (Srinivasan and Jewell, 2001). There was an upregulation of both the TGF-α and EGF in the precancerous lesions like OSF and oral leukoplakia, seen with the increase in the degree of dysplasia, implying the activation of RTK pathways and activation of oncogenes such as c-fos and *c-myc* subsequently (Srinivasan and Jewell, 2001). Areca nut extract (ANE) induces activation of RTK signaling by activating the upstream epidermal growth factor receptor (EGFR), Src and Ras signaling pathways (Chang et al, 2014). Increased expression of EGFR has been reported in OSCC (Bernardes et al, 2010) and is usually associated with poor prognosis and outcome (Temam et al, 2007).

IGF-1 is a profibrogenic growth factor which is overexpressed in fibroblasts derived from OSF (Tsai et al, 2005) similar to that seen in lung fibrosis (Uh et al, 1998) and systemic sclerosis (Harrison et al, 1994). This induces excessive production of collagen and ECM in OSF (Tsai et al, 2005). The OSCC cell lines express high levels of IGF-2 and IGF-1R, while the normal mucosa expresses IGF-1. Thus IGF-2 has a significant role in controlling the proliferation of oral carcinoma cells (Brady et al, 2007). Studies on head and neck cancer have shown no alteration in the level of expression of IGF2, E2F3 and IGF1 but there is an evidence of the upregulation of pro-apoptotic IGF1 binding protein 3 (IGFBP3) (Zhi et al, 2014). IGFBP3 regulates the IGF1 by blocking its anti-apoptotic function and increasing its half –life. IGFBP3 may also effect activation of IGF1 signaling in these carcinomas (Zhi et al, 2014) (Table 2).

Angiogenesis is an important phenomenon in precancerous conditions as it favors the nutrition and growth of the dysplastic cells, usually initiated through angiogenic stimulants such as Vascular endothelial growth factor (VEGF) (Hunasgi et al, 2018). Neovascularization induced by VEGF is important for tumor growth and metastasis (Hunasgi et al, 2018). OSF demonstrated increased expression of VEGF, indicative of hypoxia-induced angiogenesis in fibrous connective tissue stroma (Desai et al, 2012). VEGF expression by epithelial cells in OSF may promote growth via an autocrine proliferative effect on the atrophic epithelium supporting its survival and potential to undergo malignant change (Desai et al, 2012). In OSCCs, higher co-localisation of VEGF was seen in tumoral blood microvessels in the invasive fronts, while few studies showed an association with tumor differentiation (Pirici et al, 2010; Margaritescu et al, 2009).

4.6. Wnt signaling in EMT

Wnt signals are transduced through the binding of Wnt proteins to the extracellular domain of Fizzled (Fz) protein, in the presence of co-factors such as low-density-lipoprotein-related pro- tein5/6 (LRP5/6) which is required to mediate the canonical Wnt signal (Komiya and Habas, 2008). In the absence of signaling, β-catenin is degraded by the β-catenin destruction complex, which includes Axin, tumor suppressor adenomatosis polyposis coli (APC), glycogen synthase kinase 3β (GSK- 3β) and casein kinase 1 (CK1α) (Komiya and Habas, 2008). Phosphorylation of β-catenin by this complex drives it for ubiquitination and subsequent proteolytic degradation (Komiya and Habas, 2008). In case of Wnt signaling, the binding of Wnt protein to the receptor complex will result in the phosphorylation of LRP5/6 by glycogen synthase kinase 3β and recruitment of cytoplsmic phosphoprotein Dishevelled (Dsh/Dvl) and Axin, which prevents the formation destruction complex unable to phosphorylate β-catenin thereby leading to its accumulation in the cytoplasm and translocation into the nucleus. The nuclear β-catenin interacts with transcription factors T-cell factor/lymphocyte enhancer factor (TCF/LEF) and inhibits the transcription of E-cadherin to bring about EMT (Gonzalez and Medici, 2014) (Fig. 1).

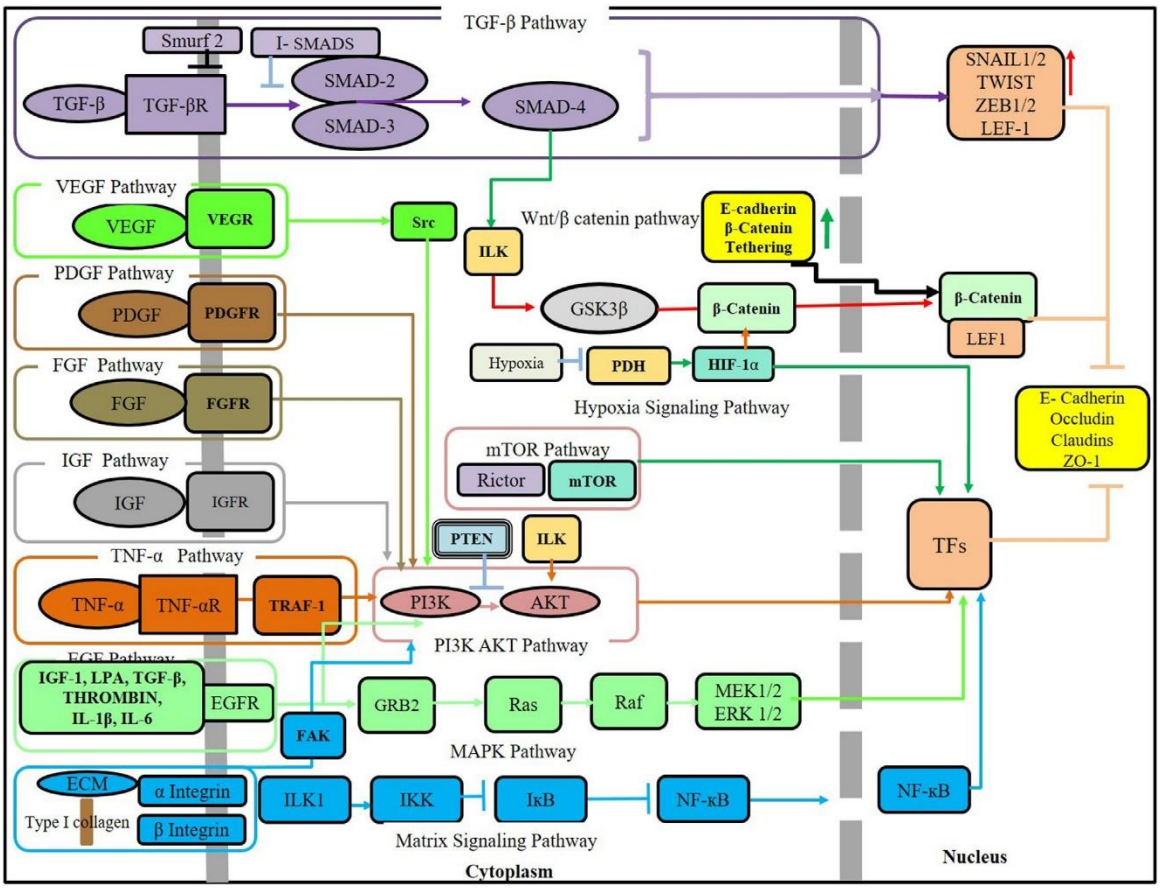


Figure 1: Signaling pathways promoting epithelial mesenchymal transition (EMT) in oral submucous fibrosis.

A group of secreted Wnt antagonists has been implicated in the regulation of the Wnt/β-catenin-signaling pathway, including Wnt inhibitory factor 1, secreted frizzled-related protein (SFRP), and the Dickkopf families (Zhou et al, 2015). The expression of SERP1 and SERP5 was seen to reduce in OSF undergoing malignant change and was associated with the loss of membranous β-catenin expression. The loss of SERP1 and SERP 5 expression was due to promoter methylation (Zhou et al, 2015). Dickkopf WNT signaling pathway inhibitor 3 (DKK3) showed upregulation in OSF progressing to OSCC and a rare mutation of DKK3 was observed in OSCC, along with increased copy numbers (Mashrah et al, 2016). However, in OSF there was decrease in DKK3 expression seen with further decline with progression in disease (Mashrah et al, 2016). The expression of Wnt Inhibitory factor (WIF1), an antagonist of the Wnt signaling is downregulated in OSF and OSCC due to methylation (Zhou et al, 2015). WIF1 methylation is associated with a poorer prognosis in OSCC patients (Sarbak et al, 2015). WIF1 is considered a potential epigenetic biomarker indicative of early changes in OSCC (Zhou et al, 2015). 4-Nitroquinoline carcinogen used to generate premalignant lesions and OSCC in Axin2-CreER; YFP mice. Tamoxifen was applied to induce Cre activity, which leads to the labeling of cancer-initiating cells (CICs). Immunohistochemical studies revealed co-expression of catenin and LEF1 in OSCC, suggesting activation of Wnt/β-catenin signaling. Increased Axin2 fluorescence was visualized in basal cells in OSCC, thus being able to confirm that Wnt-responsive CICs in OSCC contribute its malignant progression (Menon et al, 2017).

The downstream signaling molecules such as Wnt3a, β-catenin, secreted frizzled-related proteins sFRP-1, sFRP-2, sFRP-4, sFRP-5, Wnt inhibitory factor 1, dickkopf-1, *c-myc*, and cyclin-D1 studied in OSCC did not show significant expression except for *c-myc* (Marimuthu et al, 2018). Aberrant expression of β-catenin in OSCC was considered not to be due to mutation or epigenetic changes but due to focal or transient expression of β-catenin in OSCC due to various underlying mechanisms (Muzio et al, 2005) (Table 2).

4.7. Matrix signaling in EMT

The signaling pathways initiated in ECM is the result of its interaction with epithelium and it may facilitate the motility of the cells exhibiting migratory phenotype in the connective tissue through the remodeling of matrix (Gonzalez and Medici, 2014). EMT signaling pathways induces various MMPs such as MMP-2 and MMP-9 which cleaves the type IV collagen in the basal lamina and also has a role in weakening the adherens junctions of the epithelial cells and thereby promoting the phenotypic change and tumor invasion (Radisky and Radisky, 2010). In OSF, the upregulation of MMP-2 and MMP-9 has resulted in the decrease in type IV collagen on the progression of the disease, however the subsequent accumulation of type I collagen was evident (Katarkar et al, 2018). Further, the overexpression of MMP9 may result in complete destruction of the basement membrane (BM) due to degradation of collagen type-IV, which may stimulate the OSF towards malignancy (Katarkar et al, 2018). However, the expression of MMP1 was decreased in OSF, favoring the condition of fibrosis (Mishra and Ranganathan, 2010). In OSCC, upregulation of MMP1, MMP2, MMP7 and MMP9 is seen, owing to their role in the degradation of BM and ECM and association with the clinical outcome such as regional lymph node and/or distant metastasis (de Vicente et al, 2007; Katayama et al, 2004) (Table 2).

Integrin binding to ECM proteins will activate intercellular cascades that induce EMT. The importance of type I collagen in matrix signaling inducing metastasis has been reported in carcinomas of lungs, breast and esophagus (Heino, 2014; Fang et al, 2019). Binding of type I collagen to integrin activates intercellular cascades causing phosphorylation of IkB (inhibitor of kB) in an ILK- dependent manner. This, in turn increases the nuclear translocation of active NF- κB, which promotes the expression of SNAI1 and LEF1 to induce EMT (Medicia and Nawshad, 2011). Similarly, the binding of type I collagen to Discoidin domain receptor 1/2 (DDR1/2) increases NF- κB and activation of transcription factor LEF-1 to initiate EMT (Walsha et al, 2011). Abundance of collagen activates JNK pathways (Shintani et al, 2006) and also canonical and noncanonical TGF-β signaling pathways (Garamszegi et al, 2010) (Fig. 1).

ECM remodeling is noted in OSF and significant alteration in the expression of ECM molecules is demonstrable as the disease progresses. With the progression of OSF from early to advanced stages, Type III collagen and type IV (Reichart et al, 1994) are replaced with type I collagen leading to the accumulation of type I collagen in the connective tissue (Utsunomiya et al, 2005). Upregulation of v 6 in oral keratinocytes induced by arecoline promotes the invasive and migratory characteristics. 80% of the OSF transformed to OSCC shows a higher αvβ6 expression suggestive of an underlying EMT change promoting malignant transformation in OSF (Moutasim et al, 2011). Arecoline induces the upregulation of αvβ6, promoting the

trans-differentiation of oral fibroblasts into myofibroblasts (Moutasim et al, 2011). β1 integrin is has also shown to be upregulated in OSF (Veeravarmal et al, 2018). Activation of integrin by type I collagen to induce EMT changes in OSF needs to be studied further. The binding of integrin αvβ8 with type I collagen activates the downstream MEK/ERK signaling pathway, thereby facilitating the proliferation and invasion of OSCC cells in vivo (Hayashido et al, 2014). The expression of αvβ6 essentially has a role in cell proliferation, adhesion and migration in the progression of OSCC (Hayashido et al, 2014). The downstream NF- κB is upregulated in epithelial cells, fibroblasts and inflammatory cells in OSF. NF- κB overexpression is associated with persistent chronic inflammation indicating the role of inflammation in inducing fibrosis (Ni et al, 2007). The invasive and metastatic potential of OSCC cells is enhanced by Tumor necrosis factor (TNFα) via the NF-κB signaling pathway. NF-κB regulates the expression of MMPs especially MMP9 which degrades the ECM to enhance the tumor invasion (Tang et al, 2017).

4.8. Hypoxia signaling in EMT

Fibrosis and cancerous tissues are triggered by hypoxia resulting in a phenotypic change promoting EMT (Gonzalez and Medici, 2014). Hypoxia-inducible factor-1α (HIF-1α) is one such transcription factor that regulates oxygen homeostasis. Normally, it undergoes ubiquitination and subsequent proteasomal degradation with a short half-life of 5 min (Masoud and Li, 2015). Under hypoxic condition it stabilizes and interacts with coactivators such as p300/CBP to modulate its transcriptional activity (Masoud and Li, 2015). HIF-1α induces EMT by binding to the promoter region of ZEB1 (Zhang et al, 2015), SNAIL1 (Zhu et al, 2016) and hence increasing its trans activity and expression (Fig. 1).

OSF exhibits the upregulation of HIF-1α, which further progresses with dysplastic changes. Functional HIF-1α helps in cell survival and proliferation during the early stages of carcinogenesis under hypoxic conditions (Tilakaratna et al, 2008; Chaudhary et al, 2015). There is an increase in HIF-1α in OSCC which is attributed to the genetic changes in the tumor cells as well the tumor hypoxia which results in the stabilization of HIF-1α (Chaudhary et al, 2015). HIF-1α regulates various angiogenic-stimulating cytokines, growth factors and genes regulating angiogenesis (Liang et al, 2014). (Table 2).

5. Gene enrichment analysis

Gene enrichment analysis to analyze the molecular functions, biological processes and cellular components of the upregulated and downregulated genes was performed with the g:Profiler (Kull et al, 2007). Association of gene with disease and protein-protein interaction was identified by Metascape tool (Zhou et al, 2019). Pathway enrichment on wikiPathway cancer was performed by WebGestalt (Zhou et al, 2019).

Gene enrichment and pathway analysis identified the upregulated genes to be involved in 47 molecular functions, 885 biological process, 24 cellular components significantly (FDR < 0.05 (Fig. 2A). The downregulated genes were involved in 2 Molecular functions, 56 Biological process, 24 cellular components significantly (FDR < 0.05) (Fig. 2B). Top 20 gene ontology features based on FDR is represented in Fig. 3.

Gene ontology exposed that the genes were commonly involved in growth factor activity, signaling receptor binding, regulation of cell differentiation, cadherin binding and epithelial cell differentiation.

Figure 2: Gene Ontology covering three domains showing top 20 biological processes, molecular functions and cellular components among (A) Upregulated gene sets (B) Downregulated gene sets.

Evaluation of the association of differentially expressed genes with disease showed the involvement of upregulated genes with oral submucous fibrosis (OSF) and squamous cell carcinoma (SCC) (Fig. 3A). The genes are also seen to be associated with neoplasm metastasis, neoplasm invasiveness etc. Protein-protein interaction (PPI) network was constructed with Molecular Complex Detection (MCODE) algorithm. For upregulated genes, the interaction of ITGB1, MMP2, COL1A1, THBS1, and TGFB1 was significant. (Indicated in red, Fig. 3C). In downregulated genes, CDH1, CTNNB1, KRT14, KRT18, PTEN were seen to be interacting significantly (Indicated in red, Fig. 3D). The pathway analysis revealed the participation of overexpressed genes in TGF-β signaling pathway and Epithelial-Mesenchymal Transition pathways. Downregulated genes were mainly involved in Wnt Signaling Pathway, DNA Damage Response and CDK-β catenin activity (Table 3).

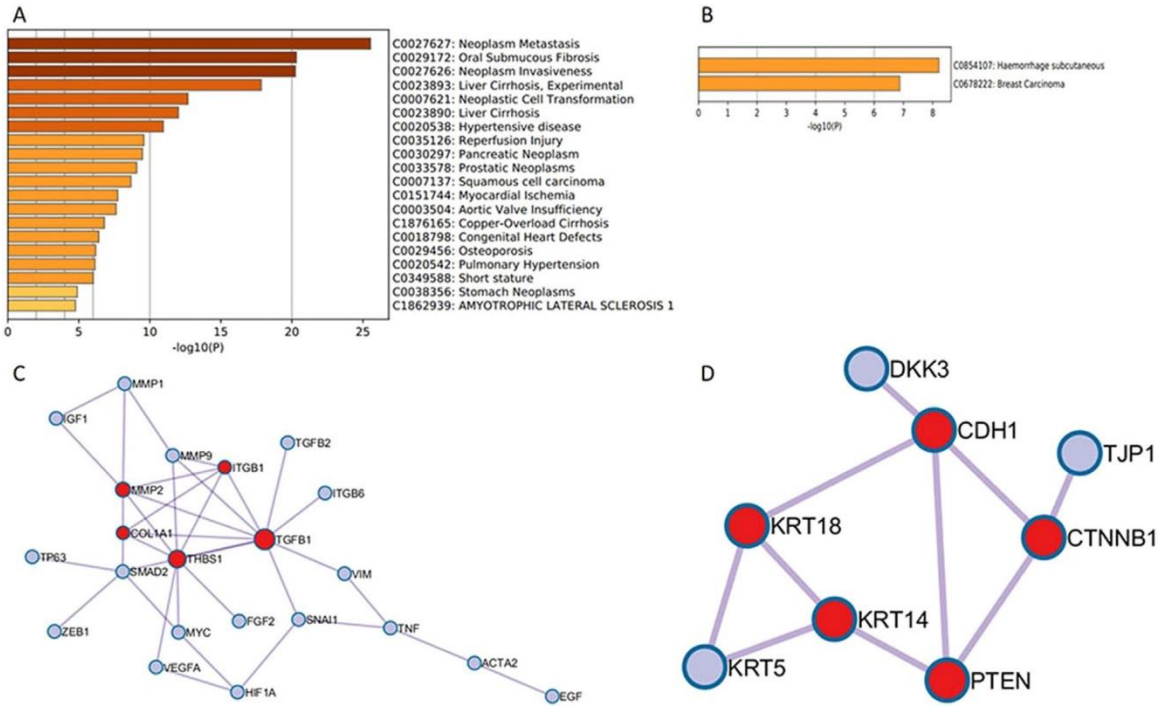


Figure 3: Gene Ontology and Protein–Protein Interactions showing (A) Bar graph of enriched Gene ontology for overexpressed list of genes (B) Bar graph of enriched Gene ontology for downregulated list of genes. (C) Protein–Protein Interaction of overexpressed genes, (D) Protein- Protein Interaction of downregulated genes.

Table 3: The pathways involved in upregulated genes (A) downregulated genes (B).

Description	P Value	FDR
(A) Pathways involved in upregulated genes		
TGF-β Signaling in Thyroid Cells for Epithelial-Mesenchymal Transition	0.0000010501	0.000078755
Pathways in Bladder Cancer	0.0000042588	0.00012569
Epithelial to mesenchymal transition in colorectal cancer	0.0000050276	0.00012569
TGF-β Receptor Signaling	0.00041368	0.0077565
TGF- β Signaling Pathway	0.00052635	0.0078953
Angiogenesis	0.0024025	0.030031
Photodynamic therapy-induced NF-kB survival signaling	0.0071459	0.072530
Pathways in Type 2 papillary renal cell carcinoma	0.0077366	0.072530
Chromosomal and microsatellite instability in colorectal cancer	0.0094417	0.078681
Pathways in clear cell renal cell carcinoma	0.016513	0.11709
(B) Pathways involved in downregulated gene		
Wnt Signaling Pathway	0.049612	0.41343
DNA Damage Response (only ATM dependent)	0.047317	0.41343
Epithelial to mesenchymal transition in colorectal cancer	0.010704	0.16056
LncRNA involvement in canonical Wnt signaling and colorectal cancer	0.039286	0.41343
ncRNAs involved in Wnt signaling in hepatocellular carcinoma	0.028737	0.35921
Pathways in Endometrial cancer	0.00066794	0.031752
TGF-β Signaling in Thyroid Cells for Epithelial-Mesenchymal Transition	0.056162	0.42122
Wnt/ β-catenin Signaling Pathway in Leukemia	0.0026673	0.050012
Regulation of Wnt/B-catenin Signaling by Small Molecule Compounds	0.0012701	0.031752
H19 action Rb-E2F1 signaling and CDK- β-catenin activity	0.00099900	0.031752

Pathway enrichment performed with WebGestalt.

6. Conclusions

Transcription factors act synergistically to bring about the epithelial cell reprogramming. Regulation of these factors control the expression of critical genes and identification of the downstream targets. Evidence suggests

a cross talk between various signaling pathways and some studies suggest the inhibition of single transcription factor is enough to block EMT (Xie et al, 2004). EMT has detrimental role in the progression of fibrosis and cancer metastasis. Poor prognosis clinical outcomes of oral cancer combined with the development of drug resistance makes it critical to identify suitable targets to prevent the induction of EMT. A thorough understanding of signaling pathways involved in EMT and the tumor microenvironment in OSF and OSCC paves for newer therapeutic strategies. Whilst the systematic analysis of the association of genes with disease showed its involvement in OSF and SCC, pathway analysis showed the participation of upregulated and downregulated genes with various EMT regulating pathways.

Conflict of interest

The authors declare that they have no conflicts of interest to disclose.

Funding

Science and Engineering Research Board (SERB) - EMR/2017/002792.

References:

Aguzzi A, Maggioni D, Nicolini G, Tredici G, Gaini RM, Garavello W. MAP kinase modulation in squamous cell carcinoma of the oral cavity. Anticancer Res 2009;29:303–8.

Albanell J, Pietras K, Virtanen I, Philipson L, Philip L, Crystal RG, et al. A SNAIL1– SMAD3/4 transcriptional repressor complex promotes TGF-β mediated epithelial–mesenchymal transition. Nat Cell Biol 2009;11:943–50.

Andressa F, Ribeiro P, Noguti J, Tizuko C, Oshima F, Ribeiro DA. Effective target-ing of the epidermal growth factor receptor (EGFR) for treating oral Cancer:a promising approach. Anticancer Res 2014;34:1547–52.

Angadi PV, Kale AD, Hallikerimath S. Evaluation of myofibroblasts in oral submucous fibrosis: Correlation with disease severity. J Oral Pathol Med 2011;40:208–13.

Angadi PV, Krishnapillai R. Evaluation of PTEN immunoexpression in oral submucous fibrosis: role in pathogenesis and malignant transformation. HeadNeck Pathol 2012;6:314–21.

Angadi PV, Patil PV, Angadi V, Mane D, Shekar S, Hallikerimath S, et al. Immunoexpression of epithelial mesenchymal transition proteins in oral squamous cell carcinoma. Int J Surg Pathol 2016;24.

Baird A, Schubert D, Ling N, Guillemin R. Receptor- and heparin-binding domains of basic fibroblast growth factor. Proct Natl Acad Sci USA 1988;85:2324–8.

Baker EA, Leaper DJ, Hayter JP, Dickenson AJ. Plasminogen activator systemin oral squamous cell carcinoma. Br J Oral Maxillofac Surg 2007;45:623–7.

Bano N, Yadav M, Mohania D, Das Id BC. The role of NF-κB and miRNA in oral cancer and cancer stem cells with or without HPV16 infection. PLoS One2018;13:e025518.

Batlle E, Sancho E, Francí C, Domínguez D, Monfar M, Baulida J, et al. The transcription factor Snail is a repressor of E-cadherin gene expression in epithelial tumour cells. Nat Cell Biol 2000;2:84–9.

Bernardes VF, Gleber-Netto FO, Sousa SF, Silva TA, Aguiar MCF. Clinical significance of EGFR, Her-2 and EGF in oral squamous cell carcinoma: a case control study. J Exp Clin Cancer Res 2010;29:1–7.

Biernacka A, Dobaczewski M, Frangogiannis NG. TGF-β signaling in fibrosis. Growth Factors 2011;29:196–202.

Bishen KA, Radhakrishnan R, Satyamoorthy K. The role of basic fibroblast growth factor in oral submucous fibrosis pathogenesis. J Oral Pathol Med 2008;37:402–11.

Brady G, Crean S, Naik P, Kapas S. Upregulation of IGF-2 and IGF-1 receptor expression in oral cancer cell lines. Int J Oncol 2007;31:875–81.

Caraglia M, Tagliaferri P, Marra M, Giuberti G, Budillon A, Di Gennaro E, et al. EGF activates an inducible survival response via the RAS-& Erk-1/2 pathway to counteract interferon-α-mediated apoptosis in epidermoid cancer cells. Cell Death Differ 2003;10:218–29.

Carracedo A, Pandolfi PP. The PTEN-PI3K pathway: of feedbacks and cross-talks. Oncogene 2008;27:5527–41.

Chang MC, Chen YJ, Chang HH, Chan CP, Yeh CY, Wang YL, et al. Areca nut components affect COX-2, cyclin B1/cdc25C and keratin expression, PGE2 production in keratinocyte is related to reactive oxygen species, CYP1A1, Src,EGFR and Ras signaling. PLoS One 2014;9:1–12.

Chang MC, Wu HL, Lee JJ, Lee PH, Chang HH, Hahn LJ, et al. The induction of prostaglandin E 2 production, interleukin-6 production, cell cycle arrest, and cytotoxicity in primary oral keratinocytes and KB cancer cells by areca nut ingredients is differentially regulated by MEK/ERK activation. J Biol Chem2004;279:50676–83.

Chang Y-C, Lin C-W, Yu C-C, Wang B-Y, Huang Y, Hsieh Y, et al. Resveratrol sup-presses myofibroblast activity of human buccal mucosal fibroblasts through the epigenetic inhibition of ZEB1 expression. Oncotarget 2016;7:12137–49.

Chang Y-C, Tsai C-H, Lai Y-L, Yu C-C, Chi W-Y, Li JJ, et al. Arecoline-induced myofibroblast transdifferentiation from human buccal mucosal fibroblastsis mediated by ZEB1. J Cell Mol Med 2014;18:698–708.

Chaudhary AK, Pandya S, Mehrotra R, Bharti AC, Jain S, Singh M. Functional polymorphism of the MMP-1 promoter (-1607 1G/2G) in potentially malignant and malignant head and neck lesions in an Indian population. Biomarkers2010;15:684–92.

Chaudhary M, Bajaj S, Bohra S, Swastika N, Hande A. The domino effect: role of hypoxia in malignant transformation of oral submucous fibrosis. JOral Maxillofac Pathol 2015;19:122–7.

Chen C, Méndez E, Houck J, Fan W, Doody D, Yueh B, et al. Gene expression profiling identifies genes predictive of oral squamous cell carcinoma. Cancer Epidemiol Biomarkers Prev 2008;17:2152–62.

Chiu C-J, Chiang C-P, Chang M-L, Chen H-M, Hahn L-J, Hsieh L-L, et al. Association between generic polymorphism of tumor necrosis Factor-α and risk of oral submucous fibrosis, a pre-cancerous condition of oral Cancer. J Dent Res 2001;80:2055–9.

Cohen Y, Goldenberg-Cohen N, Shalmon B, Shani T, Oren S, Amariglio N, et al. Mutational analysis of PTEN/PIK3CA/AKT pathway in oral squamous cell carcinoma. Oral Oncol 2011;47:946–50.

Dai JP, Chen XX, Zhu DX, Wan QY, Chen C, Wang GF, et al. Panax notoginseng saponins inhibit areca nut extract-induced oral submucous fibrosis in vitro. J Oral Pathol Med 2014;43:464–70.

Das RK, Anura A, Pal M, Bag S, Majumdar S, Barui A, et al. Epithelio-mesenchymal transitional attributes in oral sub-mucous fibrosis. Exp MolPathol 2013;95:259–69.

de Freitas Silva BS, Yamamoto FP, Pontes FSC, Cury SEV, Fonseca FP, Pontes HAR, et al. TWIST and p-Akt immune expression in normal oral epithelium, oral dysplasia and in oral squamous cell carcinoma. Med Oral Patol Oral CirBucal 2012;17:29–34.

de Vicente JC, Lequerica-Fernández P, Santamaría JFM. Expression of MMP-7and MT1-MMP in oral squamous cell carcinoma as predictive indicator for tumor invasion and prognosis. J Oral Pathol Med 2007;36:4115–24.

Derynck R, Zhang YE. Smad-dependent and Smad-independent pathways in TGF-β family signalling. Nature 2003;425:577–84.

Desai RS, Mamatha GS, Khatri MJ, Shetty SJ. Immunohistochemical expression of vascular endothelial growth factor (VEGF) and its possible role in tumour progression during malignant transformation of atrophic epithelium in oral submucous fibrosis. Curr Angiogenes 2012;1:347–53.

Eckert AW, Schütze A, Lautner MHW, Taubert H, Schubert J. HIF-1α isa prognostic marker in oral squamous cell carcinomas. Int J Biol Markers 2010;25:87–92.

Ekanayaka RP, Tilakaratne WM. Oral submucous fibrosis: review on mechanisms of pathogenesis and malignant transformation oral submucous fibrosis. Oral Surg Oral Med Oral Pathol Oral Radiol 2013;122:192–9.

Eversole LR, Philip Sapp J. c-myc Oncoprotein expression in oral precancerous and early cancerous lesions. Eur J Cancer Part B Oral Oncol 1993;29:131–5.

Fang S, Dai Y, Mei Y, Yang M, Hu L, Yang H, et al. Clinical significance and biological role of cancer-derived Type I collagen in lung and esophageal cancers. Thorac Cancer 2019;10:277–88.

Franz M, Spiegel K, Umbreit C, Richter P, Berndt A, Sven AA, et al. Expression of Snail is associated with myo W broblast phenotype development in oral squamous cell carcinoma. Histochem Cell Biol 2009;131:651–60.

Frohwitter G, Buerger H, Diest PJVAN, Korsching E, Kleinheinz J, Fillies T. Cytokeratin and protein expression patterns in squamous cell carcinoma of the oral cavity provide evidence for two distinct pathogenetic pathways.Oncol Lett 2016;12:107–13.

Gao Q, Tong W, Luria JS, Wang Z, Nussenbaum B, Effects PHK. Effects of bone morphogenetic protein-2 on proliferation and angiogenesis in oral squamous cell carcinoma. Int J Oral Maxillofac Surg 2010;39:266–71.

Garamszegi N, Garamszegi SP, Walford E, Schneiderbauer MM. Extracellularmatrix-induced transforming growth factor- b receptor signaling dynamics. Oncogene 2010;29:2368–80.

Ge WL, Xu JF, Hu J. Regulation of oral squamous cell carcinoma proliferation through crosstalk between SMAD7 and CYLD. Cell Physiol Biochem2016;38:1209–17.

Gonzalez DM, Medici D. Signaling mechanisms of the epithelial-mesenchymal transition. Sci Signal 2014;7.

Grille SJ, Bellacosa A, Upson J, Klein-szanto AJ, Van Roy F, Lee-kwon W, et al. The protein kinase Akt induces epithelial mesenchymal transition and promotes enhanced motility and invasiveness of squamous cell carcinoma lines. Cancer Res 2003;63:2172–8.

Gu L, Xie C, Peng Q, Zhang J, Li J, Tang Z. Arecoline suppresses epithelial cell viability through the Akt/mTOR signaling pathway via upregulation of PHLPP2.Toxicology 2019;419:32–9.

Hanrahan K, O'Neill A, Prencipe M, Bugler J, Murphy L, Fabre A, et al. The role of epithelial-mesenchymal transition drivers ZEB1 and ZEB2 in mediating docetaxel-resistant prostate cancer. Mol Oncol 2017;11:251–65.

Harrison NK, Cambrey AD, Myers AR, Southcolts AM, Black CM, Du Bois RM, et al. Insulin-like growth factor-I is partially responsible for fibroblast proliferation induced by bronchoalveolar lavage fluid from patients with systemic sclerosis. Clin Sci 1994;86:141–8.

Hase T, Kawashiri S, Tanaka A, Nozaki S, Noguchi N, Kato K, et al. Correlation of basic fibroblast growth factor expression with the invasion and the prognosis of oral squamous cell carcinoma. J Oral Pathol Med 2006;35:136–9.

Hayashido Y, Kitano H, Sakaue T, Fujii T. Overexpression of integrin αv facilitates proliferation and invasion of oral squamous cell carcinoma cells via MEK/ERK signaling pathway that is activated by interaction of integrin αvβ8 with type I collagen. Int J Oncol 2014;45:1875–82.

Heino J. Cellular signaling by collagen-binding integrins. Adv Exp Med Biol2014;819:143–55.

Ho CM, Hu FW, Lee SS, Shieh TM, Yu CH, Lin SS, et al. ZEB1 as an indicator of tumor recurrence for areca quid chewing-associated oral squamous cell carcinomas. J Oral Pathol Med 2015;44:693–8.

Horowitz JC, Thannickal VJ. Epithelial-mesenchymal interactions in pulmonary fibrosis. Semin Respir Crit Care Med 2006;27:600–12, http://dx.doi.org/10.1055/s-2006-957332.

Hu Y, Jian X, Peng J, Jiang X, Li N, Zhou S. Gene expression profiling of oral submucous fibrosis using oligonucleotide microarray. Oncol Rep 2008;20:287–94.

Hunasgi S, Koneru A, Vanishree M, Manvikar V. Coalition of E-cadherin and vascular endothelial growth factor expression in predicting malignant trans-formation in common oral potentially malignant disorders. J Oral MaxillofacPathol 2018;22:40–7.

Hyun K, Koo G, Han H, Sohn J, Choi W. Epithelial-to-mesenchymal transition leads to loss of EpCAM and different physical properties in circulating tumor cells from metastatic breast cancer. Oncotarget 2016;7:24677–87.

Iamaroon A, Pattamapun K, Piboonniyom S. Aberrant expression of Smad4, a TGF-β signaling molecule, in oral squamous cell carcinoma. J Oral Sci 2006;48:105–9.

Julien S, Puig I, Caretti E, Bonaventure J, Nelles L, Van Roy F, et al. Activation of NF-κB by Akt upregulates Snail expression and induces epithelium mesenchyme transition. Oncogene 2007;26:7445–56.

Kalluri R, Weinberg RA. The basics of epithelial-mesenchymal transition. JClin Invest 2009;119:1420–8.

Kamath VV, Krishnamurthy S, Satelur KP, Rajkumar K. Transforming growthfactor-β1 and TGF-β2 act synergistically in the fibrotic pathway in oral sub-mucous fibrosis: an immunohistochemical observation. Indian J Med Paediatr Oncol 2015;36:111–6.

Kamath VV, Satelur KP, Rajkumar K, Krishnamurthy S. Transforming growth factor beta 1 in oral submucous fibrosis: An immunohistochemical study –understanding the pathogenesis. J Dent Res Rev 2014;1:75–80.

Katarkar A, Prodhan C, Mukherjee S, Ray JG, Chaudhuri K. Role of matrixmetalloproteinase-9 polymorphisms in basement membrane degradation and pathogenesis of oral submucous fibrosis. Meta Gene 2018;16:255–63.

Katayama A, Bandoh N, Kishibe K, Takahara M, Ogino T, Nonaka S, et al. Expressions of matrix metalloproteinases in early-stage oral squamous cell carcinoma as predictive indicators for tumor metastases and prognosis. Clin Cancer Res 2004;10:634–40.

Khan I, Agarwal P, Thangjam GS, Radhesh R, Rao SG, Kondaiah P. Role of TGF-β and BMP7 in the pathogenesis of oral submucous fibrosis. Growth Factors2011;29:119–27.

Khan I, Kumar N, Pant I, Narra S, Kondaiah P. Activation of TGF pathway by Areca nut constituents: a possible cause of oral submucous fibrosis. PLoS One2012;7:1–12.

Komiya Y, Habas R. Wnt signal transduction pathways. Organogenesis2008;4:68–75.

Krisanaprakornkit S, Iamaroon A. Epithelial-mesenchymal transition in oral squamous cell carcinoma. ISRN Oncol 2012;2012:10.

Kriz W, Kaissling B, Le Hir M. Epithelial-mesenchymal transition (EMT) in kidney fibrosis: fact or fantasy? J Clin Invest 2011;121:468–74.

Kull M, Peterson H, Hansen J, Vilo JG. Profiler — a web-based toolset for functional profiling of gene lists from large-scale experiments. Nucleic Acids Res 2007;35:193–200.

Kume K, Haraguchi M, Hijioka H, Ishida T, Miyawaki A, Nakamura N, et al. The transcription factor Snail enhanced the degradation of E-cadherin and desmoglein 2 in oral squamous cell carcinoma cells. Biochem Biophys Res Commun 2013;430:889–94.

Kurasawa Y, Shiiba M, Nakamura M, Fushimi K, Ishigami T, Bukawa H, et al. PTEN expression and methylation status in oral squamous cell carcinoma. Oncol Rep 2008;19:1429–34.

Lamouille S, Connolly E, Smyth JW, Akhurst RJ, Derynck R. TGF—induced activation of mTOR complex 2 drives epithelial-mesenchymal transition and cell invasion. J Cell Sci 2012;125:1259–73.

Lamouille S, Derynck R. Cell size and invasion in TGF-β-induced epithelial to mesenchymal transition is regulated by activation of the mTOR pathway. JCell Biol 2007;178:437–51.

Lamouille S, Xu J, Derynck R. Molecular mechanisms of epithelial–mesenchymal transition. Nat Rev Mol Cell Biol 2014;15:178–96.

Larue L, Bellacosa A. Epithelial – mesenchymal transition in development and cancer: role of phosphatidylinositol 3' kinase / AKT pathways. Oncogene 2005;24:7443–54.

Lee K, Nelson CM. New insights into the regulation of epithelial – mesenchymal transition and tissue fibrosis, vol. 294, 1st ed. Elsevier Inc.; 2012.

Lee MK, Pardoux C, Hall MC, Lee PS, Warburton D, Qing J, et al. TGF-β activates Erk MAP kinase signalling through direct phosphorylation of ShcA. EMBO J2007;26:3957–67.

Lee SS, Tsai CH, Yu CC, Chang YC. Elevated snail expression mediates tumor progression in Areca quid chewing-associated oral squamous cell carcinoma via reactive oxygen species. PLoS One 2013;8.

Lee Y, Yang L-C, Hu F-W, Peng C-Y, Yu C-H, Yu C-C. Elevation of Twist expression by arecoline contributes to the pathogenesis of oral submucous fibrosis. J Formos Med Assoc 2016;115:311–7.

Liang J, Zhang Z, Liang L, Shen Y, Ouyang K. HIF-1α regulated tongue squamous cell carcinoma cell growth via regulating VEGF expression in a xenograft model. Ann Transl Med 2014;2:1–7.

Lim J, Kim JH, Paeng JY, Kim MJ, Hong SD, Lee JI, et al. Prognostic value of activated Akt expression in oral squamous cell carcinoma. J Clin Pathol 2005;58:1199–205.

Liu Y, El-Naggar S, Darling DS, Higashi Y, Dean DC. ZEB1 links epithelial-mesenchymal transition and cellular senescence. Development 2008;7135:579–88.

Lo Muzio L, Goteri G, Vinella A, Mastrangelo F, Carcinoma O. Beta-catenin gene analysis in oral squamous cell carcinoma. Int J Immunopathol Pharmacol 2005;18:33–8.

Margaritescu C, Pirici D, Simionescu C, Mogoanta L, Raica M, Stinga A, et al. VEGF and VEGFRs expression in oral squamous cell carcinoma. Rom J Morphol Embryol 2009;50:527–48.

Marimuthu M, Andiappan M, Wahab A, MR M, Balakrishnan A, Shanmugam S. Canonical wnt pathway gene expression and their clinical correlation in oral squamous cell carcinoma. Indian J Dent Res 2018;29:291–7.

Mashrah M, Yao Z, Zhang X, Zhang C, Zhou S, Guo F, et al. Expression pattern of DKK3, dickkopf WNT signaling pathway inhibitor 3, in the malignant progression of oral submucous fibrosis. Oncol Rep 2016;37:979–85.

Mashrah M, Yao Z. Aberrant DKK3 expression in the oral leukoplakia and oral submucous fibrosis : a comparative immunohistochemical study. Eur J Histochem 2016;60:2629.

Masoud GN, Li W. HIF-1α pathway: role, regulation and intervention for cancer therapy. Acta Pharm Sin B 2015;5:378–89.

McCain J. The MAPK (ERK) pathway. Pharm Ther 2012;38:346–53.

Medicia D, Nawshad A. Type I collagen promotes epithelial-mesenchymal transition through ILK-dependent activation of NF-κB and LEF-1. Matrix Biol2011;29:161–5.

Mendez MG, Kojima S, Goldman RD. Vimentin induces changes in cell shape, motility, and adhesion during the epithelial to mesenchymal transition. FASEB J 2016;24:1838–51.

Meng W, Xia Q, Wu L, Chen S, He X, Zhang L, et al. Downregulation of TGF-beta receptor types II and III in oral squamous cell carcinoma and oral carcinoma-associated fibroblasts. BMC Cancer 2011;11:88.

Menon R, Li CC, Li MZ. Wnt signaling in the oral cancer initialing cells. Oral Maxillofac Pathol 2017;124:e202.

Mishra G, Ranganathan K. Matrix metalloproteinase-1 expression in oral submucous fibrosis: an immunohistochemical study. Indian J Dent Res2010;21:320–5.

Moutasim KA, Jenei V, Sapienza K, Marsh D, Weinreb PH, Violette SM,et al. Betel-derived alkaloid up-regulates keratinocyte alphavbeta6 integrin expression and promotes oral submucous fibrosis. J Pathol 2011;223:366–77.

Nakamura R, Ishii H, Endo K, Hotta A, Fujii E, Miyazawa K, et al. Reciprocal expression of slug and snail in human oral cancer cells. PLoS One 2018;13:1–14.

Nakano Y, Kobayashi W, Sugai S, Kimura H. Expression of tumor necrosis factor-α and Interleukin-6 in oral squamous cell carcinoma. Jpn J Cancer Res 1999;90:858–66.

Nawshad A, Medici D, Liu C-C, Hay ED. TGFbeta3 inhibits E-cadherin gene expression in palate medial-edge epithelial cells through a Smad2-Smad4-LEF1 transcription complex. J Cell Sci 2007;120:1646–53.

Nayak MT, Singh A, Desai RS, Vanaki SS. Immunohistochemical analysis of vimentin in oral submucous fibrosis. J Cancer Epidemiol 2013.

Nayak S, Goel MM, Makker A, Bhatia V, Chandra S, Kumar S, et al. Fibroblast growth factor (FGF-2) and its receptors FGFR-2 and FGFR-3 may Be putative biomarkers of malignant transformation of potentially malignant oral lesions into oral squamous cell carcinoma. PLoS One 2015;10:e0138801.

Ni W-F, Tsai C-H, Yang S-F, Chang Y-C. Elevated expression of NF- kB in oral submucous fibrosis – evidence for NF- kB induction by safrole in human buccal mucosal fibroblasts. Oral Oncol 2007;43:557–62.

PA R, CW VW, J B, D S. Distribution of procollagen type III, collagen type VI and tenascin in oral submucous fibrosis (OSF). J Oral Pathol Med 1994;23:394–8.

Pal M, Ray AK, Sengupta S, Paul RR, Barui A, Chatterjee J, et al. Assessment of malignant potential of oral submucous fibrosis through evaluation of p63,E-cadherin and CD105 expression. J Clin Pathol 2010;63:894–9.

Pal SK, Thi C, Nguyen K, Morita K, Miki Y, Kayamori K, et al. THBS1 is induced by TGFB1 in the cancer stroma and promotes invasion of oral squamous cell carcinoma. J Oral Pathol Med 2016;45:730–9.

Pant I, Kumar N, Khan I, Rao SG, Kondaiah P. Role of areca nut induced TGF-β and epithelial-mesenchymal interaction in the pathogenesis of oral sub-mucous fibrosis. PLoS One 2015;10:1–19.

Pant I, Rao SG, Kondaiah P. Role of areca nut induced JNK / ATF2 / Jun axis in the activation of TGF-β pathway in precancerous Oral Submucous Fibrosis. Sci Rep 2016;6:34314.

Peng H, Shintani S, Kim Y, Wong DT. Loss of p12 CDK2-AP1expression in human oral squamous cell carcinoma with disrupted transforming growth. Neoplasia 2006;8:1028–36.

Pirici D, Simionescu C, Raica M, Stepan A, Ribatti D, Mogoanta L. VEGF expression and angiogenesis in oral squamous cell carcinoma: an immunohistochemical and morphometric study. Clin Exp Med 2010;10:209–14.

Radisky ES, Radisky DC. Matrix metalloproteinase-induced epithelial-mesenchymal transition in breast cancer. J Mammary Gland Biol Neoplasia2010;15:201–12.

Rahmani A, Alzohairy M, Babiker AY, Rizvi MA, Elkarimah- HG. Clinico-pathological significance of PTEN and bcl2 expressions in oral squamous cell carcinoma. Int J Clin Exp Pathol 2012;5:965–71.

Rai A, Ahmad T, Parveen S, Parveen S, Faizan MI, Ali S. Expression of trans-forming growth factor beta in oral submucous fibrosis. J Oral Biol Craniofacial Res 2020;10:166–70.

Rajalalitha P, Vali S. Molecular pathogenesis of oral submucous fibrosis—a collagen metabolic disorder. J Oral Pathol Med 2005;34:321–8.

Ranganathan K, Kavitha R, Sawant SS, Vaidya MM. Cytokeratin expression in oral submucous fibrosis–an immunohistochemical study. J Oral Pathol Med 2006;35:25–32.

Santiago L, Daniels G, Wang D, Deng F-M, Lee P. Wnt signaling pathway protein LEF1 in cancer, as a biomarker for prognosis and a target for treatment. Am J Cancer Res 2017;7:1389–406.

Sarbak J, Kostrzewska-poczekaj M, Mielcarek-kuchta D, Baer-dubowska W. The negative regulators of Wnt pathway — DACH1, DKK1, and WIF1 are methylated in oral and oropharyngeal cancer and WIF1 methylation pre-dicts shorter survival. Tumor Biol 2015;36:2855–61.

Scanlon CS, Van Tubergen EA, Inglehart RC, D'Silva NJ. Biomarkers of epithelial- mesenchymal transition in squamous cell carcinoma. J Dent Res 2013;92:114–21.

Schmalhofer O, Brabletz S, Brabletz T. E-cadherin, β-catenin, and ZEB1 in malignant progression of cancer. Cancer Metastasis Rev 2009;28:151–66.

Schreck R, Rapp UR. Raf kinases: oncogenesis and drug discovery. Int J Cancer 2006;119:2261–71.

Sharada P, Swaminathan U, Nagamalini BR, Kumar KV, Ashwini BK, Lavanya V. Coalition of E-cadherin and vascular endothelial growth factor expression in predicting malignant transformation in common oral potentially malignant disorders. J Oral Maxillofac Pathol 2018;22:40–7.

Sharma M, Shetty SS, Radhakrishnan R. Oral submucous fibrosis as an overhealing wound: implications in malignant transformation. Recent Pat Anticancer Drug Discov 2018;13.

Shield KD, Ferlay J, Jemal A, Sankaranarayanan R, Chaturvedi AK, Bray F, et al. The global incidence of lip, oral cavity, and pharyngeal cancers by subsite in2012. CA Cancer J Clin 2017;67:51–64.

Shintani Y, Hollingsworth MA, Johnson KR, Wheelock MJ. Collagen I promotes metastasis in pancreatic Cancer by activating c-Jun NH 2 -Terminal kinase1 and up-regulating N-Cadherin expression. Cancer Res 2006;66:11745–54.

Smitha A, Rao K, Umadevi HS, Smitha T, Sheethal HS, Vidya MA. Immunohistochemical study of α-smooth muscle actin expression in oral leukoplakia and oral squamous cell carcinoma. J Oral Maxillofac Pathol 2019;23:59–64.

Srinivasan M, Jewell SD. Evaluation of TGF-α and EGFR expression in oral leukoplakia and oral submucous fibrosis by quantitative immunohistochemistry. Oncology 2001;61:284–92.

Srinivasan M, Jewell SD. Quantitative estimation of PCNA, c-myc, EGFR and TGF-α in oral submucous fibrosis - an immunohistochemical study. OralOncol 2001;37:461–7.

Strutz F, Zeisberg M, Ziyadeh FN, Yang C-Q, Kalluri R, Muller GA, et al. Role of basic fibroblast growth factor-2 in epithelial-mesenchymal transformation. Kidney Int 2002;61:1714–28.

Su MC, Chen CT, Huang FI, Chen YL, Jeng YM, Lin CY. Expression of LEF1 is an independent prognostic factor for patients with oral squamous cell carcinoma. J Formos Med Assoc 2014;113:934–9.

Su TR, Liao YW, Hsieh PL, Tsai LL, Fang CY, Lin T, et al. Butylidenephthalide abrogates the myofibroblasts activation and mesenchymal transdifferentiation in oral submucous fibrosis. Environ Toxicol 2018;33:686–94.

Sun L, Liu T, Zhang S, Guo K, Liu Y. Oct4 induces EMT through LEF1/β-catenin dependent WNT signaling pathway in hepatocellular carcinoma. Oncol Lett2017;13:2599–606.

Tak J, Rao NN, Chandra A, Gupta N. Immunohistochemical analysis of tenascin expression in different grades of oral submucous fibrosis. J Oral MaxillofacPathol 2015;19:291–6, http://dx.doi.org/10.4103/0973-029X.174645.

Tan J, Tedrow JR, Nouraie M, Dutta JA, Miller DT, Li X, et al. Loss of Twist1 in the mesenchymal compartment promotes increased fibrosis in experimental lung injury by enhanced expression of CXCL12. J Immunol 2017;198:2269–85.

Tang D, Tao D, Fang Y, Deng C, Xu Q, Zhou J. TNF-alpha promotes invasion and metastasis via NF-Kappa B pathway in oral squamous cell carcinoma. Med Sci Monit Basic Res 2017;23:141–9.

Temam S, Kawaguchi H, El-Naggar AK, Jelinek J, Tang H, Liu DD, et al. Epi-dermal growth factor receptor copy number alterations correlate with poor clinical outcome in patients with head and neck squamous cancer. J Clin Oncol2007;25:2164–70.

Tilakaratne WM, Iqbal Z, Teh MT, Ariyawardana A, Pitiyage G, Cruchley A,et al. Upregulation of HIF-1α in malignant transformation of oral submucous fibrosis. J Oral Pathol Med 2008;37:372–7.

Titidej A, Eshghyar N, Jolehar M, Jolehar M. Prognostic new marker (bone morphogenetic protein 7) in squamous cell carcinoma. J Contemp Dent Pract 2018;19:675–9.

Tsai C, Yang S, Chen Y-J, Chou M-Y, Chang Y-C. The upregulation of insulin-like growth factor-1 in oral submucous fibrosis. Oral Oncol 2005;41:940–6.

Uh S, Inoue Y, King TE, Chan ED, Newman LS, Riches DWH. Morphometric analysis of insulin-like growth Factor-I localization in lung tissues of patients with idiopathic pulmonary fibrosis. Am J Respiir Crit Care Med 1998;158:1626–35.

Utsunomiya H, Tilakaratne WM, Oshiro K, Maruyama S, SuzukiM, Ida-Yonemochi H, et al. Extracellular matrix remodeling in oral submucous fibrosis: its stage-specific modes revealed by immunohistochemistry and in situ hybridization. J Oral Pathol Med 2005;34:498–507.

Uzgare AR, Kaplan PJ, Greenberg NM. Differential expression and/or activation of p38MAPK, erk1/2, and jnk during the initiation and progression of prostate cancer. Prostate 2003;55:128–39.

Veeravarmal V, Austin RD, Nagini S, Nassar MHM. Expression of 1integrinin normal epithelium, oral submucous fibrosis and oral squamous cell carcinoma. Pathol Res Pract 2018;214:273–80.

Walsha LA, Nawshadb A, Medicia D. Discoidin domain receptor 2 is a critical regulator of epithelial-mesenchymal transition. Matrix Biol 2011;30:243–7.

Wang Y, Liu J, Ying X, Lin PC, Zhou BP. Twist-mediated epithelial-mesenchymal transition promotes breast tumor cell invasion via inhibition of Hippo Pathway. Sci Rep 2016;6:24606.

Wang Y, Shi J, Chai K, Ying X, Zhou B. The role of snail in EMT and tumorigenesis. Curr Cancer Drug Targets 2013;13:963–72.

West KA, Brognard J, Clark AS, Linnoila IR, Yang X, Swain SM, et al. Rapid Akt activation by nicotine and a tobacco carcinogen modulates the phenotype of normal human airway epithelial cells. J Clin Invest 2003;111:81–90.

Wushou A, Pan HY, Liu W, Tian Z, Wang LZ, Shali S, et al. Correlation of increased twist with lymph node metastasis in patients with oral squamous cell carcinoma. J Oral Maxillofac Surg 2012;70:1473–9.

Xie L, Law BK, Chytil AM, Brown KA, Aakre ME, Moses HL. Activation of the erk pathway is required for TGF-1-Induced EMT in vitro. Neoplasia 2004;6:603–10.

Yan Z, Kui Z, Ping Z. Reviews and prospectives of signaling pathway analysis in idiopathic pulmonary fibrosis. Autoimmun Rev 2014;13:1020–5.

Yang H-W, Lu M-Y, Chiu Y-W, Liao Y-W, Huang Y-F, Ju Chueh P, et al. Hinokitiolablates myofibroblast activation in precancerous oral submucous fibrosis by targeting Snail. Environ Toxicol 2018;33:454–62.

Yang J, Mani SA, Weinberg RA. Exploring a new twist on tumor metastasis. Cancer Res 2006;66:4549–52.

Yang S, Hsieh Y, Tsai C. The upregulation of type I plasminogen activator inhibitor in oral submucous fibrosis. Oral Oncol 2003;39:367–72.

Yanjia H, Xinchun J. The role of epithelial – mesenchymal transition in oral squamous cell carcinoma and oral submucous fibrosis. Clin Chim Acta 2007;383:51–6.

Yao X, Sun S. Clinicopathological significance of ZEB-1 and E-cadherin proteins in patients with oral cavity squamous cell carcinoma. Onco Targets Ther2017;10:781–90.

Yokoyama K, Kamata N, Hayashi E, Hoteiya T, Ueda N, Fujimoto R, et al. Reverse correlation of E-cadherin and snail expression in oral squamous cell carcinoma cells in vitro. Oral Oncol 2001;37:65–71.

Zagabathina S, Ramadoss R, Ah HP, Krishnan R. Comparative evaluation of SMAD-2 expression in oral submucous fibrosis and reactive oral lesions. Asian Pac J Cancer Prev 2020;21:399–403.

Zeisberg M, Neilson EG. Biomarkers for epithelial-mesenchymal transitions. J Clin Invest 2009;119:1429–37.

Zhang W, Liu HT. MAPK signal pathways in the regulation of cell proliferation in mammalian cells. Cell Res 2006;12:9–18.

Zhang W, Shi X, Peng Y, Wu M, Zhang P, Xie R, et al. HIF-1α promotes epithelial-mesenchymal transition and metastasis through direct regulation of ZEB1 in colorectal Cancer. PLoS One 2015;10:e0129603.

Zhang YE. Non-Smad signaling pathways of the TGF-β family. Cold SpringHarb Perspect Biol 2017;9:1–18.

Zheng M, Jiang Y, Chen W, Li K, Liu X, Gao S, et al. Snail and Slug collabo-rate on EMT and tumor metastasis through miR-101-mediated EZH2 axis in oral tongue squamous cell carcinoma. Oncotarget 2015;6:6797–810.

Zhi X, Lamperska K, Golusinski P, Schork NJ, Luczewski L, Golusinski W, et al. Expression levels of insulin-like growth factor 1 and 2 in head and neck squamous cell carcinoma. Growth Horm IGF Res 2014;24:137–41.

Zhou S, Chen L, Mashrah M, Zhu Y, He Z, Hu Y, et al. Expression and promoter methylation of Wnt inhibitory factor-1 in the development of oral submucous fibrosis. Oncol Rep 2015;34:2636–42.

Zhou S, Chen L, Mashrah M, Zhu Y, Liu J, Yang X, et al. Deregulation of secreted frizzled-related proteins is associated with aberrant beta-catenin activation in the carcinogenesis of oral submucous fibrosis. Onco Targets Ther 2015;8:2923–31.

Zhou S, Mashrah M, Zhu Y, Liu J, He Z, Xiang T, et al. Deregulation of secreted frizzled-related proteins is associated with aberrant β-catenin activation in the carcinogenesis of oral submucous fibrosis. Onco Targets Ther2015;8:2923–31.

Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, et al. Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat Commun 2019;10:1523.

Zhu Y, Tan J, Xie H, Wang J, Meng X, Wang R. HIF-1α regulates EMT via the Snail and b -catenin pathways in paraquat poisoning-induced early pulmonary fibrosis. J Cell Mol Med 2016;20:688–97.

Zhuo X, Luo H, Chang A, Li D, Zhao H, Zhou Q. Is overexpression of TWIST, a transcriptional factor, a prognostic biomarker of head and neck carcinoma? Evidence from fifteen studies. Sci Rep 2015;5:18073.

Zidar N, Boštjančič E, Malgaj M, Gale N, Dovšak T, Didanovič V. The role of epithelial-mesenchymal transition in squamous cell carcinoma of the oral cavity. Virchows Arch 2018;472:237–45.

Zittermann SI, Issekutz AC. Basic fibroblast growth factor (bFGF, FGF-2) potentiates leukocyte recruitment to inflammation by enhancing endothelial adhesion molecule expression. Am J Pathol 2006;168:835–46.

Chapter

4

Emerging Pathogens: Comparative Genome Analysis of *Clostridia* Species

Tanwar, Ankit Singh*, Padival Shruptha*, Apoorva Jnana, Angela Brand, Mamatha Ballal, Kapaettu Satyamoorthy, and Thokur Sreepathy Murali.
"Emerging Pathogens in Planetary Health and Lessons from Comparative Genome Analyses of Three *Clostridia* Species." *OMICS: A Journal of Integrative Biology* 27, no. 6 (2023): 247–259.

DOI: 10.1089/omi.2023.0034; **IF 3.3 (2023)**

*Equal first authors

Emerging Pathogens in Planetary Health and Lessons from Comparative Genome Analyses of Three *Clostridia* Species

Abstract:

Clostridioides difficile (CD) is a major planetary health burden. A gram-positive opportunistic pathogen, CD colonizes the large intestine and is implicated in sepsis, pseudomembranous colitis and colorectal cancer. *C. difficile* infections typically following antibiotic exposure, result in dysbiosis of the gut microbiome, and is one of the leading causes of diarrhoea in the elderly population. While several studies have focused on the toxigenic strains of CD, gut commensals such as *Clostridium butyricum* (CB) and *Clostridium tertium* (CT) could harbour toxin/virulence genes and thus, pose a threat to human health. In this study, we sequenced and characterized three isolates, namely, *C. tertium* (MALS001), *C. butyricum* (MALS002) and *C. difficile* (MALS003) for their antimicrobial, cytotoxic, antiproliferative, genomic and proteomic profiles. Although *in vitro* cytotoxic and antiproliferative potential was observed predominantly in CD MALS003, genome analysis revealed pathogenic potential of CB MALS002 and CT MALS001. Pangenome analysis revealed the presence of several accessory genes typically involved in fitness, virulence and resistance characteristics in the core genomes of sequenced strains. The presence of an array of virulence and antimicrobial resistance genes in CB MALS002 and CT MALS001 suggests their potential role as emerging pathogens with significant impact on planetary health.

Keywords: planetary health, whole genome sequencing, omics, antimicrobial resistance, virulence, pangenome analysis

Introduction:

Clostridioides difficile is a Gram-positive opportunistic pathogen that resides in the human gut. It causes severe gastrointestinal diseases such as diarrhoea, pseudomembranous colitis, as well as sepsis and multiple organ dysfunction with major outbreaks worldwide and is the most significant cause of diarrhoea in the elderly population (Leffler and Lamont, 2015; Jin et al, 2017). Treatment with broad spectrum antibiotics and longer periods of hospitalization result in dysbiosis of normal gastrointestinal flora and creates a niche suitable for overgrowth of *C. difficile* (Rineh et al, 2014). Continued use of broad-spectrum antibiotics and old age are major risk factors for *C. difficile* infections (CDI).

Reports suggest at least 1-3% of hospitalized patients acquire CDI while 25% of these patients experience recurrent infections due to high antibiotic resistance among these strains. Approximately, half-a-million CDI cases are reported every year from USA alone, with health care costs for treatment approaching US$1.5 billion (Chen et al, 2017), while all-cause mortality within a month of infection is reported to be 20% (Smits et al, 2016). However, most individuals with *C. difficile* remain asymptomatic and only few exhibits severe disease symptoms. *C. difficile* spores are highly resistant and upon exposure to bile salts in the gastrointestinal tract, germinate to form vegetative cells which colonize the colon (Edwards et al, 2016).

In clinical settings, CDI is generally diagnosed by culture-based methods that involve faecal material, testing for non-specific antigens such as glutamate dehydrogenase (GDH) and/or specific toxins (toxins A and B) by cytotoxicity or immunoassays, which are time consuming, technically demanding and show low sensitivity and specificity (Bartlett and Gerding, 2008). The gold standard assays take longer time to diagnose the infection while severe cases might require immediate surgical intervention. Molecular techniques such as polymerase chain reaction (to detect toxin genes such as *tcdA*, *tcdB*) (Chen et al, 2017), 16S rRNA ribotyping (Bidet et al, 1999), pulsed field gel electrophoresis and microarray analysis (Al-Khaldi et al, 2012) have been used for the diagnosis and delineating different species of *Clostridioides*. Molecular assays provide a distinctive advantage over culture-based methods with their rapid turn-around time in addition to identifying asymptomatic carriers often missed by culture assays.

However, exclusive reliance on molecular tests without detecting the presence of toxins can lead to false positives (Smits et al, 2016). Asymptomatic carriage has also been reported which further complicates detection of pathogenic *Clostridia* (Cassir et al, 2016a). The presence of asymptomatic carriers in the hospital settings is one of the major reasons for epidemic outbreaks and the nosocomial nature of CDI (Riggs et al, 2007). In addition, the severity of CDI cannot be explained merely by the genetic factors of *C. difficile* alone. CDI is often associated with dysbiosis of gut microbiome which exponentially worsens the infection (Abbas and Zackular, 2020). Hence, it is critical to study the interactions of *C. difficile* with co-occurring pathogens such as *C. butyricum* and *C. tertium* (Ferraris et al, 2012). In addition, study of the genomes of co-occurring pathogens may offer insights into pathogen evolution and improve our understanding of the genetic basis of various process critical for pathogenesis such as virulence and resistance. It will be interesting to explore genomes of *C. butyricum*, a probiotic producing human gut commensal with reports of being both beneficial and harmful (Cassir et al, 2016b, Kanai et al, 2015) and *C. tertium*, a largely non-toxigenic strain occasionally associated with cases of bacteremia, meningitis and several others (Moore and Lacey, 2019, Shah et al, 2016, Kourtis et al, 1997) for pathogenic elements. This can provide insights into emergence of pathogens with improved virulence, antimicrobial resistance and overall pathogenic potential. The goal of the current work was to study the genomes of *C. tertium* (MALS001), *C. butyricum* (MALS002) and compare it with that of *C. difficile* (MALS003) for genomic surveillance to uncover virulence-associated factors and their potential role as emerging pathogens. Emerging pathogens have the capacity to spread, infect large populations, devastate health systems, raise morbidity and mortality rates and add to the economic burden, making them a serious planetary health concern.

Materials and Methods

This study was approved by the Kasturba Medical College and Kasturba Hospital Institutional Ethics committee in Manipal, India. Informed consent was obtained from all individuals included in the study.

Culturing of *Clostridia* isolates

For this study, a clinical isolate of *C. tertium* (CT MALS001) and *C. butyricum* (CB MALS002) were isolated from stool samples of patients visiting Kasturba Hospital, Manipal (Enteric Diseases Division, Department of Microbiology, Kasturba Medical College, Manipal) and *C. difficile* CD MALS003 (from *C. difficile* ATCC

9689) was included. The strains were revived in Brain Heart Infusion agar (BHIA) (HiMedia, India) and incubated anaerobically at 37°C in a GasPak EZ Anaerobe system (Becton Dickinson, USA) for 48 hours.

Starch hydrolysis assay

Cultures of isolates CB MALS002, CD MALS003, and CT MALS001 (OD_{600} - 0.1) were streaked onto BHIA plates supplemented with 2% starch (HiMedia, India) and incubated at 37°C for 16 hours. Culture plates were flooded with Gram's iodine solution and examined for the presence of clear zones (Mahon and Manuselis, 1995).

Motility test

The single colonies of CB MALS002, CD MALS003, and CT MALS001 grown on BHIA, were picked using straight wire and stab inoculated into 0.175% BHI agar in a glass vial. Following incubation at 37°C for 16 hours, vials were removed from the anaerobic chamber and photographed to record the swimming motility (Dingle et al., 2011, Chen et al., 2019).

Minimal inhibitory concentration assay

Minimal inhibitory concentration (MIC) testing was performed using the gradient method wherein plastic strips with predefined concentrations of antibiotic are employed to determine the exact MIC for chosen antibiotics (Kowalska-Krochmal and Dudeck-Wicher, 2021). In this study, MIC assay was done for ciprofloxacin and vancomycin based on the corresponding presence of resistance genes in the genomes of isolates sequenced in the current study. Towards this, following revival of CB MALS002, CD MALS003, and CT MALS001 as described before, isolates were streaked (OD_{600} - 0.1) onto Mueller Hinton Agar plates. HiComb gradient MIC strips (HiMedia, Mumbai, India) of vancomycin (0.016 µg to 256 µg gradient) and ciprofloxacin (0.001 µg to 240 µg gradient) were placed at the centre of the inoculated plates which were subsequently incubated at 37°C in a GasPak EZ Anaerobe system (Becton Dickinson, USA) for 48 hours. Average MIC values obtained from triplicates for each strain was assessed for antimicrobial resistance. Interpretation of antimicrobial resistance was performed as per CLSI and EUCAST guidelines where applicable. For ciprofloxacin, interpretation was done with breakpoints assigned for moxifloxacin defined by CLSI (30^{th} edition, 2020) since neither CLSI nor EUCAST had interpretation criteria for ciprofloxacin (Büchler et al, 2014). MIC values of ≤ 2 µg, 4 µg and ≥8 µg of ciprofloxacin indicates sensitivity, intermediate resistance and resistance accordingly (CLSI guidelines for moxifloxacin were considered). For vancomycin, MIC values of ≤ 2 µg and > 2 µg indicate sensitivity and resistance accordingly (EUCAST, 2023; Büchler et al., 2014).

Biofilm formation analysis

Biofilm formation is a key ability that enhances the pathogenesis of *C. difficile* (Vuotto et al., 2018). CB MALS002, CD MALS003 and CT MALS001 strains were assessed for biofilm formations in anaerobic static condition at 24 and 48 hours post incubation. Tissue culture plate method was used to determine the static biofilm by staining with 0.1% crystal violet. Interspecies interactions have been reported to enhance *C. difficile* virulence and resistance (Abbas and Zackular, 2020). Hence, polymicrobial biofilms (pairwise combinations of the three isolates (CB+CD, CT+CD, CB+CT) and multiple combinations (CB+CD+CT)) were also assessed by tissue culture method. Briefly, following culturing as before in BHI broth, 100 µl of each isolate was

combined with the other at an OD_{600} of 0.05 for pairwise and multiple interactions. 200 µl of each combination was seeded into a 96 well plate. Following seeding, plates were incubated for 24h and 48h followed by staining with 0.1% crystal violet (Prasad et al, 2020).

Cytotoxicity assay

Cultures were grown to a concentration of 1×10^7 cells/mL and the culture supernatant was syringe filtered with the (0.2µm pore size) (Sartorius Stedim). SiHa cells were seeded in 96-well tissue culture plates containing Dulbecco's Modified Eagle Medium (DMEM) supplemented with 10% Fetal Bovine Serum (FBS) at a density of 1×10^4 cells/mL. Following adherence, cells were treated with 10 µl of bacterial supernatant in 90 µl of DMEM medium with 10% FBS and cell survival was measured at 24, 48 and 72 h. After treatment, $1/10^{th}$ volume of Cell Counting Kit-8 (CCK-8) was added to the cells, incubated for 4 hours and absorbance was measured at 450 nm using a plate reader (Varioskan™ Thermo Scientific, USA) (Prasad et al, 2020). Mitomycin (5 µg/µl) was used as positive control while peptone water was used as vehicle control.

Transfection and TCF reporter assay

We tested for the antiproliferative activity of toxins A and B via Polyethylenimine (PEI) transfection and TCF reporter assay. The reporter plasmids (TOPflash and FOPflash) were transfected to SiHa cells (6×10^5 cells) using PEI mediated transfection. After 12 hours of transfection, cells were treated with 200 µL of bacterial crude extracts corresponding to 10^6 CFU/mL and luciferase activities were measured after 36 hours. SiHa cells co-transfected with pRL-TK (thymidine kinase basal promoter) were used as control (Roehl et al, 1990). Dual luciferase reporter system (Promega) was used to quantify the Luciferase and Renilla by luminescence signal measurements (Lima et al, 2014).

Whole genome sequencing

Genomic DNA of CB MALS002, CD MALS003, and CT MALS001 were extracted using the standard phenol-chloroform protocol (Green and Sambrook, 2017) with some slight modifications. Cells were grown in anaerobic conditions as mentioned earlier for a period of 48 hours and the cell pellets were lysed with 1.5 ml of pre-lytic buffer (20mM Tris, 2mM EDTA, 1% Triton X-100, 60µg/ml of lysozyme) and incubated overnight at 37°C before genomic DNA was extracted. Quality and quantity of DNA was assessed with a Qubit 2.0 fluorometer (Thermo Fisher Scientific, USA). Whole genome sequencing was performed on ION torrent PGM (Thermo Fisher Scientific, USA). 100ng of genomic DNA was fragmented to 400bp using Ion Shear Plus Reagents followed by clean up with Agencourt™ AMPure™ XP (Beckman Coulter, USA). Further, the prepared libraries (100pM) were amplified with Ion One Touch 2, enriched with Ion One Touch ES and sequenced on an Ion Torrent Personal Genome Machine using a 318v2 chip following standard protocol (Utturkar et al, 2015).

Genome assembly, annotation and comparison

After read quality check and trimming using FASTQC (Andrews, 2010) and FASTX toolkit (Gordon and Hannon, 2010), a reference-based assembly using SPAdes v3.13.0 (Prjibelski et al, 2020) was performed. References utilized for genome assembly are listed in **Supplementary Table S1**. Genome annotation was performed using NCBI Prokaryotic Genome Annotation Pipeline (PGAP, v6.1) (Tatusova et al, 2016).

Assembled contig sequences were used as input for functional annotation using eggNOG-mapper (Huerta-Cepas et al, 2017) and Rapid Annotation using Subsystem Technology (RAST) (Aziz et al, 2008).

A maximum-likelihood based phylogenetic tree was constructed with bootstrap support of 1000 replicates using MEGA (Kumar et al, 2018). Nucleotide sequence level comparison between the clinical isolates and their respective reference genomes (**Supplementary Table S1**) was performed using Mauve (v20150226) (Darling et al, 2004). Annotation of raw sequences was performed using prokka v1.14.6 (Seemann, 2014) followed by pangenome analysis using roary v3.13.0 (Page et al, 2015) to obtain core and accessory genes. To predict single copy gene content, BUSCO v4.1.2 (Seppey et al, 2019) analysis was done using lineage-specific (*Clostridia* class level) dataset. Reference genomes used for BUSCO and pangenome analysis are listed in **Supplementary Table S1**.

Functional profiling

Genes coding for antimicrobial resistance was identified using AMRFinderPlus v3.10.5 (Feldgarden et al, 2019) and ABRicate v1.0.1 (Seemann, 2021). Unique hits across two approaches were identified and filtered (≥75% query cover and ≥70% identity). Virulence genes were predicted with ABRicate tool compared to core and full data set of Virulence Factor Database (VFDB) (Chen et al, 2016) and visualized with CGView (Stothard and Wishart 2005). Genes corresponding to sporulation process were extracted from functional annotation (eggNOG-mapper) files manually. A blastp alignment was performed for common sporulation proteins between three strains to find sequence similarity.

Results

Biochemical and antimicrobial characterization

CB MALS002, CD MALS003 and CT MALS001 were revived and tested for their starch degradation and motility characteristics. *Clostridia* species have been shown to have differential starch degradation abilities. We observed saccharolytic activity (zones of clearance around bacterial growth) among CB MALS002 and CT MALS001 in decreasing order respectively, while CD MALS003 did not display saccharolytic activity **(Figure 1a).** CB MALS002, CD MALS003 and CT MALS001 displayed differential rates of swimming motility **(Figure 1b).** In stab culture, the isolates showed varying degrees of motility with CD MALS003 showing the highest swimming motility with vastly diffuse growth followed by CB MALS002 and CT MALS001. In terms of antimicrobial profiles, CD MALS003 was resistant to vancomycin while CB MALS002 and CT MALS001 were sensitive. For ciprofloxacin, CB MALS002 was found to be resistant, while CD MALS003 and CT MALS001 was sensitive to the antibiotic **(Supplementary Table S2).**

Biofilm formation analysis

Biofilm formation in the three strains was measured with crystal violet staining. CD MALS003 showed enhanced biofilm production compared to CB MALS002 and CT MALS001. While dual cultures of CD MALS003 and CT MALS001 showed significantly higher levels of biofilm production compared to monocultures, CB MALS002 and CT MALS001 dual cultures showed similar levels of biofilm production as their respective individual cultures. Biofilm levels were higher in combined cultures having the presence of all the three strains at 24 and 48 hours post incubation **(Figure 2a)**.

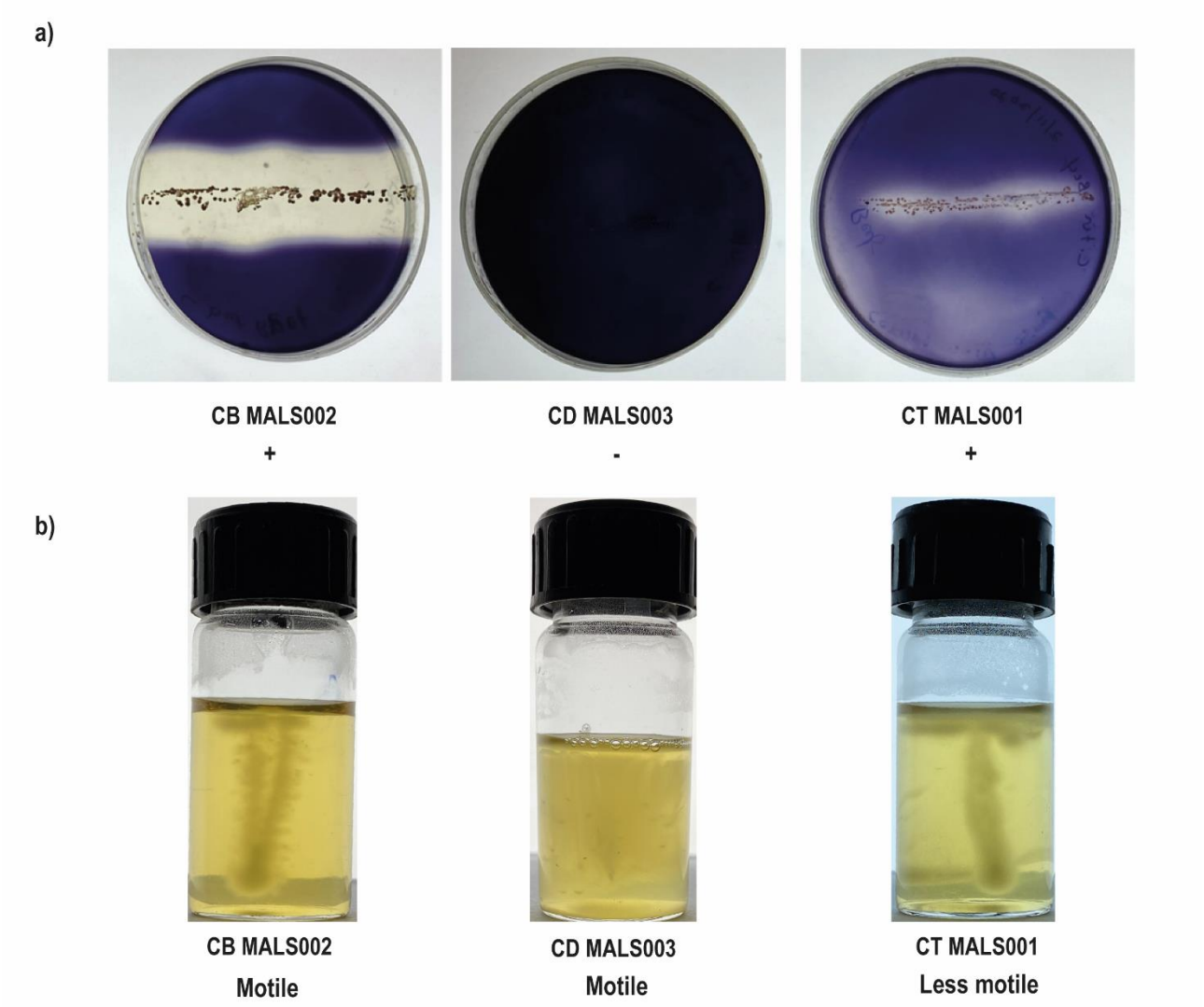


FIG. 1. Biochemical characterization. (a) Starch hydrolysis assay for CB MALS002, CD MALS003 and CT MALS001. Clear zones indicate ability of bacteria to hydrolyse starch. "+/-" indicate presence and absence of clear zones respectively. (b) Motility test for CB MALS002, CD MALS003 and CT MALS001. Spreading phenotype was visually assessed to determine degree of motility. Diffuse spreading phenotype from the line of stab inoculum was considered as visual confirmation of swimming motility. Experiments were performed in triplicates.

Cytotoxic effect of *Clostridia* secretome on SiHa cells

Cytotoxic effect of the culture supernatants of CB MALS002, CD MALS003 and CT MALS001 on SiHa cells were tested by CCK-8 assay. Different concentrations of the culture supernatant (crude secretome, 1:5, 1:10 dilutions) were tested. The crude secretome of all three bacterial strains showed significant cytotoxic effect at 24h, 48h and 72h **(Figure 2b)** while the diluted extracts did not show significant difference in cytotoxicity compared to control **(Supplementary Figure S1).**

Effects on Wnt pathway reporter gene expression by *Clostridia* toxins

The antiproliferative effect of CB MALS002, CD MALS003 and CT MALS001 secretome was investigated by β-catenin reporter assay using SiHa cells. The comparison of the ratio of TOPFlash to FOPFlash luciferase

activity in SiHa cells treated with CD MALS003 secretome showed a significant downregulation in the β-catenin transcriptional regulatory response compared to control. However, there was no significant difference in expression in cells treated with CB MALS002 and CT MALS001 when compared to control **(Figure 2c).**

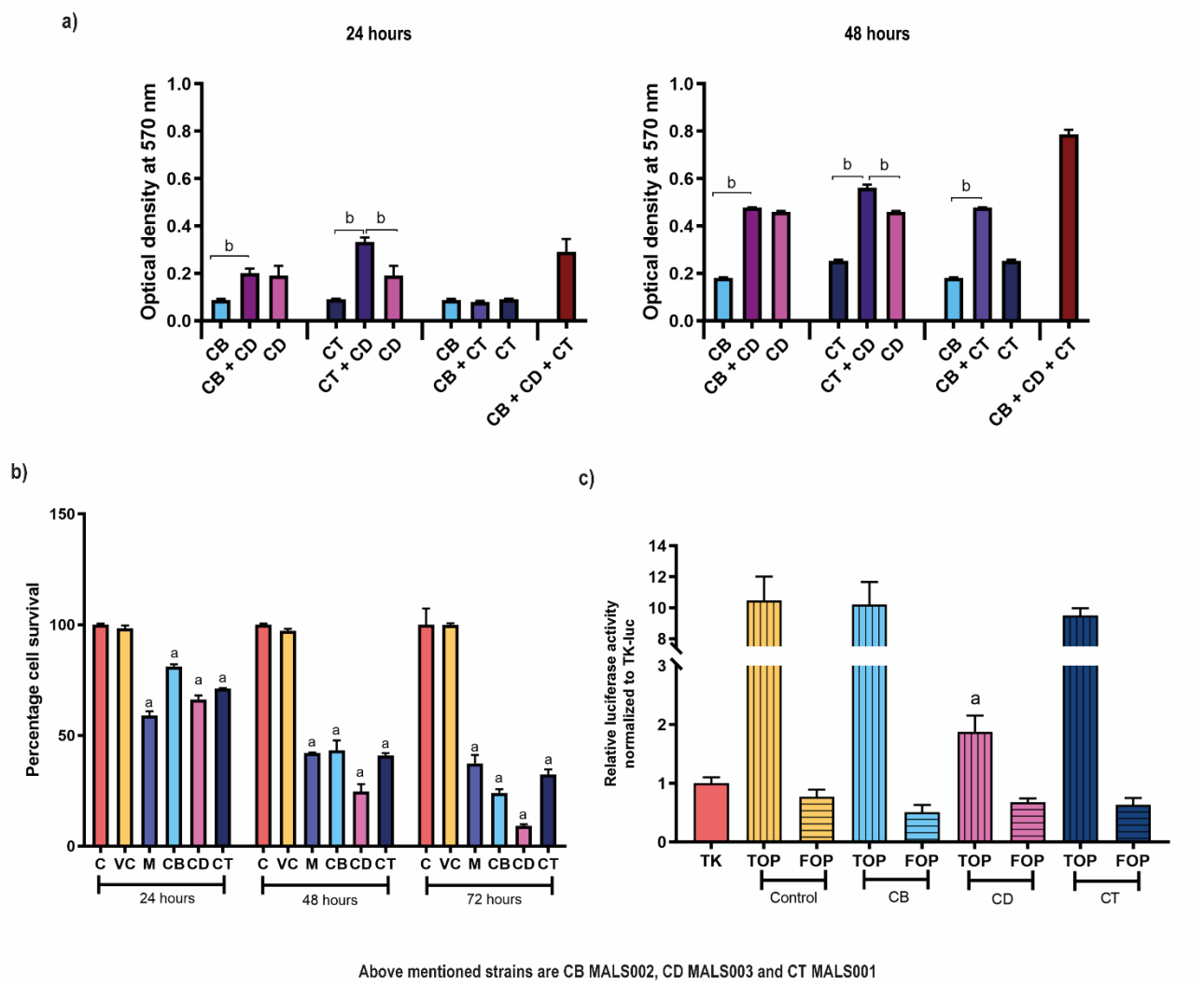


FIG. 2. Biofilm, Cytotoxicity and antiproliferative assay of CB MALS002, CD MALS003, and CT MALS001. (a) Detection of biofilm produced by three *Clostridia* in static condition using crystal violet assay at 24 and 48 hours. CD MALS003 showed enhanced biofilm production compared to other strains. Combination of CB MALS002, CD MALS003, and CT MALS001 promoted biofilm formation compared to biofilm of mono and co-cultured species. Data presented as mean ± SE of n=3 independent experiment. Statistical significance was analysed by student t test, b - P = < 0.01 (b) Cytotoxicity of *Clostridia* crude supernatant on cell survival of SiHa cells was tested by CCK-8 assay at 24, 48 and 72 hours. a - P = < 0.001 compared to control. (c) Antiproliferative effect of *Clostridia* crude extracts on SiHa cells was tested by TCF reporter assay SiHa cells were co-transfected with either TOPFlash or FOPFlash luciferase reporter constructs to assess decrease/increase in proliferation of SiHa cells via changes in luciferase expression. Data presented as mean ± SE of n=3 independent experiment. Statistical significance was analysed by student t test, a - P = < 0.001

Genome assembly, annotation and whole genome comparison

Reference based genome assembly and refining resulted in a genome size of 4.5 Mb for CB MALS002 with 125 contigs, 4.4 Mb for CD MALS003 with 132 contigs and 4.5 Mb for CT MALS001 with 169 contigs. Phylogenetic analysis of our sequences with related sequences confirmed that the strains belonged to CB, CD and CT genomes **(Supplementary Figure S2).** CT MALS001 genome harboured 4332 genes followed by CB MALS002 (4330) and CD MALS003 (4138). Sequence annotation identified 4186 proteins in CT MALS001, followed by CB MALS002 (4131) and CD MALS003 (3973). Strain CB MALS002 had the highest number of tRNA (83) sequences followed by CD MALS003 (77) and CT MALS001 (53). CB MALS002 also had the highest number of rRNA (23) sequences followed by CT MALS001 (7) and CD MALS003 (5) **(Table 1)**.

Table 1. Genome assembly and annotation features of sequenced strains.

Strain	*Clostridium butyricum* MALS002	*Clostridioides difficile* MALS003	*Clostridium tertium* MALS001
Total length (bp)	4551027	4403852	4532340
No. of contigs	125	132	169
N50 (bp)	107911	187823	70613
GC content (%)	28.7	28.7	27.8
Genes	4330	4138	4332
Proteins	4131	3973	4186
rRNA	23	5	7
tRNA	83	77	53

The most abundant RAST SEED subsystem (functionally related protein families) across all *Clostridia* species belonged to ‘carbohydrates’, ‘amino acids and its derivatives’ and ‘protein metabolism’ categories with an average of 16.85%, 15.57% and 11.25% respectively. Relative contribution of protein families involved in ‘dormancy and sporulation’ was higher in CB MALS002 (4.1%) compared to CD MALS003 (2.1%, 2-fold decrease) and CT MALS001 (1.3%, 3-fold decrease). A 3-fold increase of protein families involved in ‘phosphorus metabolism’ and ‘nitrogen metabolism’ was observed in CB MALS002 and CT MALS001 when compared with CD MALS003 **(Supplementary Figure 3a).** Comparison of annotated genes using PGAP, RAST and eggNOG-mapper showed that eggNOG could annotate higher number of genes **(Supplementary Figure 3b).** A total of 757 protein encoding genes were identified as "core proteome" while a total of 167, 234 and 97 unique protein encoding genes were found in CB MALS002, CD MALS003, and CT MALS001 genomes respectively **(Supplementary Figure 3c)**. CD MALS003 shared 72.6% and 71.9% of genes with CB MALS002 and CT MALS001 respectively while CB MALS002 and CT MALS001 shared higher number of protein-encoding genes.

Multiple genome alignment using Mauve highlighted several genomic rearrangements including inversions, shifts, deletions and gaps in assembled genomes relative to reference genomes. CD MALS003 had an average of 99.67% identity with its reference genomes followed by CB MALS002 (99.35% identity) while the CT MALS001 genomes showed less identity (99.18%) **(Supplementary Figure S4).**

Pangenome analysis

Pangenome analysis of three isolates with respective references showed 3228 core genes that were shared among five different CD genomes while CB (2836 among 5 taxa) and CT (2741 among 4 taxa) genomes shared fewer core genes. Presence of accessory genes was highest among CT MALS001 genome with 1311 unique genes out of which 884 genes were hypothetical while 39 encoded putative proteins. CB MALS002 genome had 552 (331 hypothetical + 18 putative) unique genes. CD reference genomes had fewer unique genes with strain FDAARGOS 267 having no unique gene while ATCC 9689 strain had only one unique gene **(Figure 3a).** Among our strains, CD MALS003 had fewer unique genes in comparison with CB MALS002 and CT MALS001 (110 of which 92 were hypothetical). After removing hypothetical and putative genes from core genome, CD MALS003 assembly contained 1840 genes followed by CT MALS001 (1697genes) and CB MALS002 (1594 genes). The core genome in individual strains ranged from 67-82% in CB genomes, 80-90% in CD genomes and 65-77% in CT genomes **(Figure 3b).**

Based on BUSCO analysis of 247 single copy genes from 385 *Clostridia* genomes, strains MALS001, MALS002 and MALS003 were found to have 244, 244 and 245 single copy genes respectively **(Figure 3c).** This included DHHA1 domain protein, which was absent in CD MALS003 and ATCC 9689 compared to the R20291 strain; pyridoxal phosphate homeostasis protein, which was absent in CB MALS002 compared to NBRC 13949 strain, and cytidine deaminase enzyme, which was absent in CT MALS001 when compared to Src5 strain. A total of 5 fragmented genes coding for helicase-exonuclease AddAB subunit *addA* in CD MALS003, 2,3-bisphosphoglycerate-independent phosphoglycerate mutase and Dephospho-CoA kinase in CB, *recN* and Phosphohydrolase in CT MALS001 were identified.

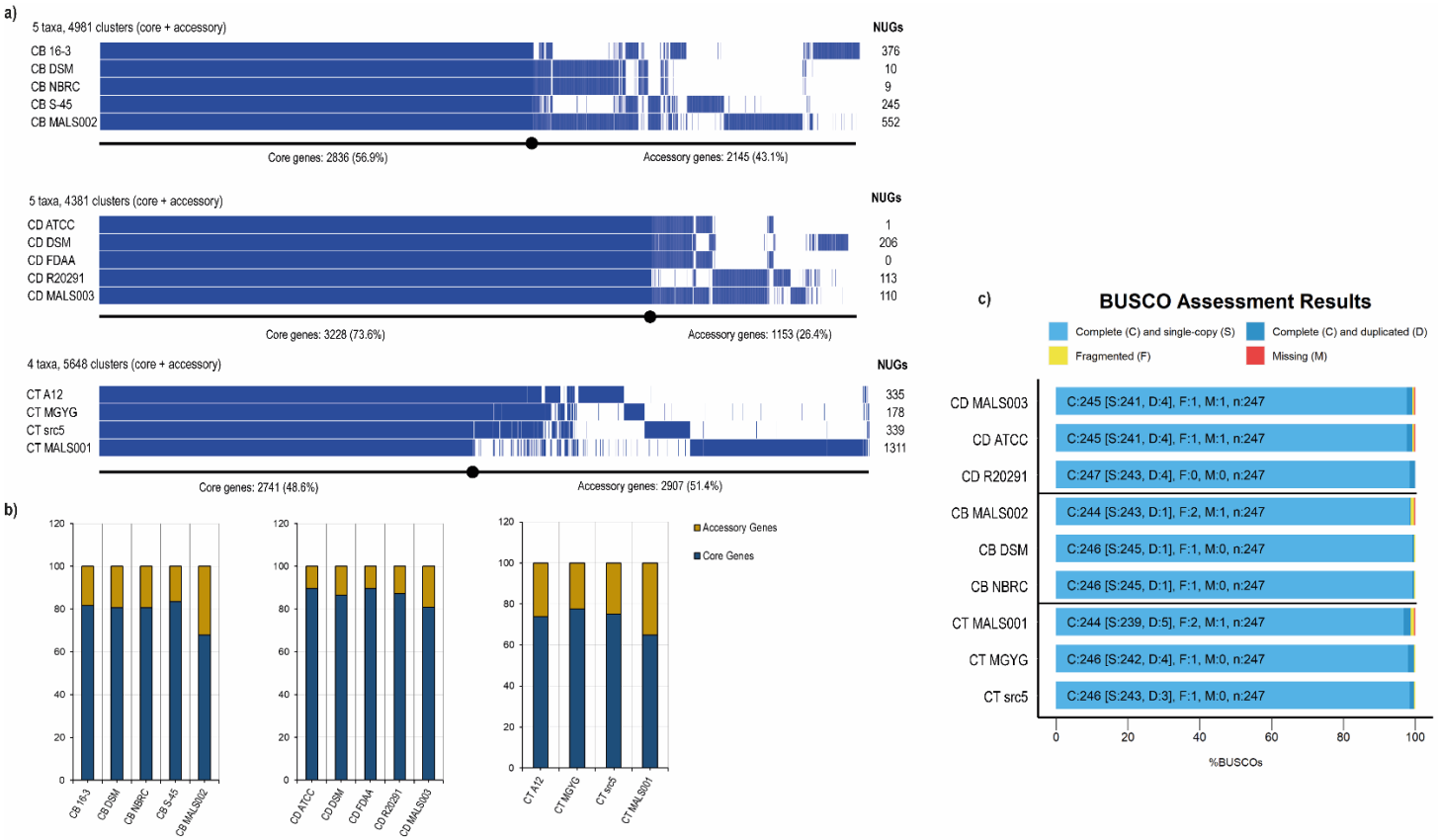


FIG. 3. Pangenome analysis of CB MALS002, CD MALS003, and CT MALS001. (a,b) Pangenome analysis for *Clostridia* genomes as performed with Roary. Parameters for selection: Core genes: 99% ≤ strains ≤ 100% -accessory genes (15% ≤ strains < 95%). NUG: Number of unique genes (c) Genomic data completeness was analysed with prediction of *Clostridia* single copy genes with BUSCO. Split of complete genes (single copy and duplicate), fragmented and missing genes are shown in the bar graphs

Antimicrobial resistance profile

A total of 9 unique genes coding for antimicrobial resistance (AMR) were predicted from all three strains. A variety of antibiotic resistance genes conferring resistance to glycopeptide, fluoroquinolone, streptogramin, beta-lactam and phenicol were recorded **(Supplementary Table S3).** Highest number of resistance genes were found in CD MALS003 (7), followed by CT MALS001 and CB MALS002 with one AMR gene each. No common AMR genes were found across all three strains.

Antimicrobial resistance predicted for two antibiotic classes, fluoroquinolone and glycopeptide, were selected for phenotypic confirmation by minimum inhibitory concentration assay. Strain CB MALS002 showed phenotypic resistance to ciprofloxacin despite no resistance genes being predicted for these drug classes **(Supplementary Table S2 and S3)**. Strain CD MALS003 showed phenotypic resistance to vancomycin. Concurrent with *in-silico* analysis 5 genes (*vanG*, *vanR*, *vanS*, *vanT*, and *vanZ1*) encoding resistance to vancomycin (Stogios and Savchenko et al., 2020) were predicted in CD MALS003 genome **(Supplementary Table S2 and S3)**.

Toxin/virulence profile

Sixty-six unique toxin or virulence genes were predicted across all three strains with *groEL*, *plr* and *fbpA* being common among all three strains. Toxin genes were highest in CD MALS003 (54) followed by CT MALS001 (14) and CB MALS002 (13). In CD MALS003, a large cluster consisting of 31 toxin genes were identified that coded for flagellar proteins and few cell adhesion proteins **(Supplementary Table S4).** Two exotoxin coding sequences located in pathogenicity locus (PaLoc) coding for toxin A (*tcdA*) and toxin B (*tcdB*), which are implicated in pathogenicity of CD, as well as three other accessory genes *tcdC*, *tcdE* and *tcdR* were present in our CD MALS003 strain. Toxin genes unique to CB MALS002 were *Cbei_1707*, *Cbei_0023*, *pfoA*, *ureB*, *rmlA* and *ZP_02950902*, while toxin genes unique to CT MALS001 were *nagK*, *pilT*, *CLL_A2400*, and *SSP0068*. A hyaluronidase exoenzyme (*nagK*) and hemolysin A (*CLL_A2400*) were major predicted toxins along with few flagellar proteins (*cheY, flhA, flip, fliS2, flil, pilT*) in CT MALS001 genome. In CB MALS002, perfringolysin O (*pfoA*) an exotoxin, two hemolysins (*Cbei_1707* and *Cbei_0023*) were predicted along with very few flagellar proteins (*fliS1, flil, cheY*). Genomic locations of core and full toxins, AMR genes and genes coding for flagellar assembly of all three strains were plotted **(Figure 4).**

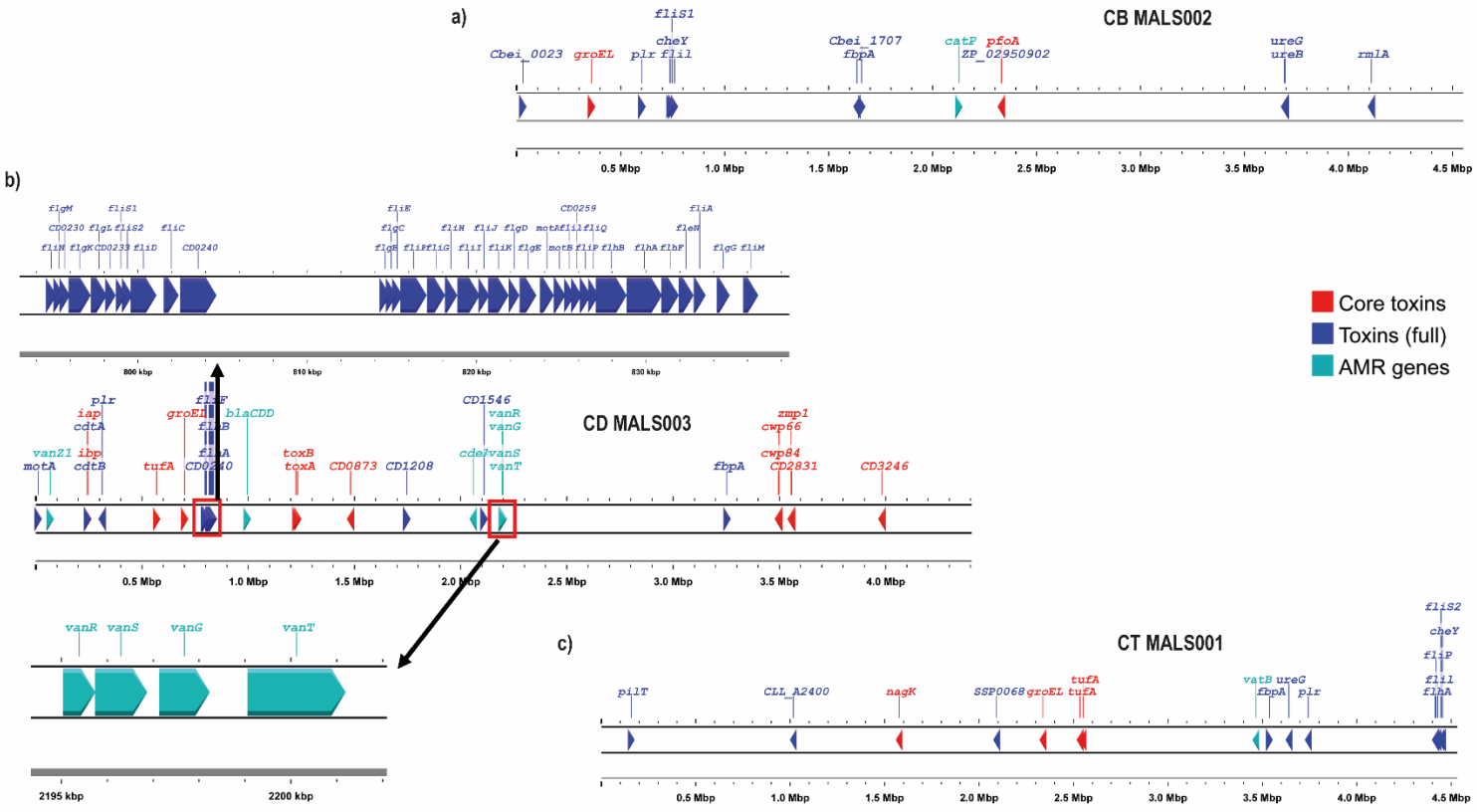


FIG. 4. Genomic locations of genes encoding antimicrobial resistance and virulence/toxins. Details of antimicrobial genes and virulence genes obtained using AMRFinderPlus, ABRicate and Virulence Factor Database (VFDB) (query coverage ≥75%, percentage identity ≥70%) respectively were visualized as different tracks onto the assembled genome sequences of (a) CB MALS002, (b) CD MALS003 and (c) CT MALS001

Sporulation genes

From functional annotation files obtained using eggNOG-mapper, a total of 105 genes corresponding to sporulation process were extracted **(Supplementary Table S5).** Following removal of duplicates based on gene copies in each isolate, 53 genes were retained, of which, 26 genes were shared among all the three isolates **(Supplementary Figure 5a).** Highest number of sporulation genes were present in CB MALS002 (48) followed by CT MALS001 (47) and CD MALS003 (32). A comparison of twenty-six common sporulation proteins using protein blast (blastp) identified *spoIIE* to have the maximum sequence dissimilarity across the three strains while RNA polymerase sporulation specific sigma factors (*sigG*, *sigE*, *sigA* and *sigK*) shared the highest sequence similarity **(Supplementary Figure 5b).**

Discussion

The class *Clostridia* comprises of heterogeneous species which are phylogenetically distantly related. The major human pathogens include *C. difficile*, *C. botulinum*, *C. tetani*, and *C. perfringens* while several other species such as *C. tertium* and *C. butyricum* are emerging as new pathogens and can cause human infections. In the current study, *C. butyricum* MALS002, *C. difficile* MALS003 and *C. tertium* MALS001 strains were characterized by culture-based and genomic methods. CB is a predominantly saccharolytic bacterium that

plays a crucial role in regulation of intestinal homeostasis and conversion of dietary fibre to beneficial short chain fatty acids (SCFAs) (Zhang et al, 2018). Contrary to current literature, CD MALS003 from our study did not display any saccharolytic activity. However, CD has a diverse metabolic capacity and has been known to exhibit varying patterns of carbohydrate fermentation (Neumann-Schaal et al, 2019). Biofilm production which significantly contributes to virulence potential was found to be higher in CD MALS003 strain compared to CB MALS002 and CT MALS001 (Figure 2a). In addition, combined cultures of CT MALS001 and CD MALS003 produced significantly higher levels of biofilm when compared to their monocultures. However, biofilm levels of CD MALS003 on co-culture with CB MALS002 did not show any significant variation. While all three strains displayed varied motility, CT MALS001 displayed limited motility in comparison to the diffuse phenotype observed in CB MALS002 and CD MALS003. In addition to biofilm production and flagellar motility that are critical for host colonization (Chen et al, 2019), antimicrobial resistance is a significant factor in determining pathogen epidemiology (Spigaglia et al, 2018). Our genomic data showed that CD MALS003 harboured 5 genes encoding vancomycin resistance (*vanG, vanR, vanS, vanT,* and *vanZ1*) (Stogios and Savchenko et al, 2020) while no such genes were detected in CT MALS001 and CB MALS002. This was concurrent with an observation of phenotypic resistance **(Table S2).** However, despite possessing *cdeA* gene which is known to encode resistance to ciprofloxacin and other fluoroquinolone drugs (Saxton et al, 2009), CD MALS003 displayed phenotypic susceptibility. This phenotypic susceptibility despite the presence of corresponding resistance genes may be explained by presence of compensatory mutations (Melnyk et al, 2015) or by gene silencing via transcriptional control (Enne et al, 2015). However, these genes can get activated by metabolites produced by co-occurring bacteria such as *C. butyricum* and *C. tertium* (Schmidt et al, 2015; Ferraris et al, 2012). Of note, genome analysis did not predict any resistance genes encoding fluoroquinolone resistance in CB MALS002 (**Table S3**). However, it demonstrated a high degree of phenotypic resistance towards the same. This phenotypic resistance in the absence of genotypic evidence in the form of presence of genes encoding antimicrobial resistance could hint towards the presence on "non-inherited antibiotic resistance" (Levin and Rozen, 2006). This finding also emphasizes the need to explore emergence of antimicrobial resistance in relation with microbial communities (Bottery et al, 2020). CT MALS001 exhibited genomic and phenotypic susceptibility to both ciprofloxacin and vancomycin. This is concordant with the pathogenic profile of *C. tertium* which is largely considered as a non-pathogenic strain that is occasionally found in clinical cases (Moore and Lacey, 2019; Steensma et al, 2011).

We found a total of 66 different toxin genes in the three isolates, with CD MALS003 harbouring maximum number of toxin-coding genes (54) while CB MALS002 (13) and CT MALS001 (14) showed fewer toxin-coding genes. Based on genomic analysis, CD MALS003 was characterized as a toxigenic strain with the presence of *tcdA*, *tcdB* (encoding toxins A and B), *cdtA* and *cdtB* (encoding binary toxin) while these toxin genes were absent in CB MALS002 and CT MALS001. Accordingly, CD MALS003 exhibited significant cytotoxicity *in vitro* compared to other tested strains. Studies indicate that *C. butyricum* strains can be used to prevent CDI as well as in improving gastric ulcers and other bacterial infections due to their ability to produce bacteriocins and SCFAs. However, some toxigenic strains of *C. butyricum* have been implicated in botulism and necrotizing enterocolitis (Cassir *et al.* 2016a). In the current study, even though CB MALS002 and CT

MALS001 did not have any of the toxin genes commonly implied in the pathogenicity of CDI, we detected other toxin coding genes coding for hemolysins such as *pfoA* (cholesterol-dependent cytolysin) which have been previously reported in *C. perfringens* (Verherstraeten et al, 2015). GroEL has been reported to play an important role in stimulating inflammatory response and is known to contribute to bacterial infections (Ranford and Henderson, 2002). The presence of various genes coding for toxins and virulence factors along with reports on their ability to produce bacteriocins and butyric acid highlight the confusion in categorising *C. butyricum* as a pathogen or a beneficial gut microbe (Cassir et al., 2016b).

C. tertium which was long considered a non-toxin producer; occasionally found in clinical cases such as bacteremia and necrotizing fasciitis associated septicemia in patients with neutropenia and hematological malignancies (Shah et al, 2016), brain abscess (Lew et al, 1990) and meningitis (Kourtis et al, 1997) among others. However, recent reports indicate that *C. tertium* isolated from stool samples of patients with diarrhoea harboured sequences homologous to *tcdA*, *tcdB* and other toxin coding genes (Muñoz et al, 2019) suggesting their importance in disease outbreaks and pathogen emergence. The presence of *nagK*, a bacterial hyaluronidase reported earlier from *C. perfringens* (Geier et al, 2021) in CT MALS001, demonstrates the extent of dynamic mobility among the Clostridial genomes. Though bacterial hyaluronidase such as *nagH* have been reported in *C. tertium* (Muñoz et al, 2019), to the best of our knowledge, *nagK* has only been reported from *C. perfringens* and CD but not from CT strains (Low et al, 2018).

The importance of spore differentiation in pathogenesis of enteric pathogens cannot be underestimated. For instance, *spo0A* gene codes for a master regulator that turns on several downstream RNA polymerase sigma factors including *sigG* suggesting their role in persistence and transmission of highly resistant endospores thereby providing the bacteria with high resistance and resilience to survive in the gut environment (Pereira et al, 2013). Interestingly the highest number of sporulation genes were present in CB MALS002 followed by CT MALS001 and CD MALS003.

In summary, surveillance of infectious diseases, even those with a predominant single pathogen such as CDI may be severely limited if performed without considering microbial interaction with co-occurring pathogens such as *C. butyricum* and *C. tertium* (Ferraris et al., 2012, Bottery et al., 2020). Several elements of microbial interactions such as extracellular vesicles and microbial volatiles have been known to influence gene expression of microbial partners leading to enhanced antimicrobial resistance, biofilm production, stress defence, immunomodulation and pathogenic colonization among several other roles that favour pathogen emergence and increase in disease severity (Tarashi et al., 2022, Schmidt et al., 2015). Hence, it is critical to study the genomes of microbes known to co-occur with *C. difficile* such as *C. butyricum* and *C. tertium* to assess their genomic potential so as to offer insights into preventing epidemic outbreaks and study the evolution of pathogens. Towards this, our study based on whole-genome sequence analysis of bacteria frequently associated with CDI (*C. difficile* – known pathogen (Dawson et al., 2009), *C. butyricum* – known to be both beneficial and pathogenic (Cassir et al., 2016b), *C. tertium* – generally considered as non-toxigenic (Moore and Lacey, 2019) was performed to identify potential virulence factors, antimicrobial resistance genes, mobile genetic elements, and sporulation factors, which can provide insights into emergence of pathogens with improved virulence, antimicrobial resistance and overall pathogenic potential. Such studies will further

improve our understanding of development of antimicrobial resistance, provide new avenues in genomic monitoring of emerging pathogens and offer better treatment strategies for crippling infectious diseases worsened by multidrug resistance such as CDI that threaten planetary health.

Data availability:

Genomes of *C. butyricum* strain MALS002 (CB), *C. difficile* strain MALS003 (CD) and *C. tertium* strain MALS001 (CT) are submitted to NCBI genome database under BioProject ID PRJNA820142 and with following Genome accession numbers, JALGRX000000000 (CB), JALGRY000000000 (CD) and JALGRW000000000 (CT).

Acknowledgement:

Authors thank Manipal Academy of Higher Education, Technology Information Forecasting Assessment Council - Centre of Relevance and Excellence in Pharmacogenomics (TIFAC-CORE) and DBT BUILDER – Interdisciplinary Life Science Programme for Advance Research and Education (DB-ILSPARE) for the support. AST and AJ thank MAHE for Dr. TMA Pai PhD Scholarship and PS thanks Lady TATA Memorial Trust for Junior Research Fellowship.

Author Contributions:

Conceptualization: KS, MB and TSM; investigation: PS; software: AST; validation and visualization: PS and AST; writing – original draft: TSM and AJ; writing – review and editing: TSM, AB, MB and KS; supervision: TSM, AB, MB and KS.

Funding: This research was financially supported by Department of Science and Technology - Science for Equity, Empowerment and Development division (DST-SEED) (SEED/WS/2019/57S(G)), Indo-German Science and Technology Centre (IGSTC) and Fund for Improvement of S & T Infrastructure in Universities and Higher Educational Institutions (DST-FIST) (SR/FST/LSI-515/2011(C)20/03/2013).

Disclosure Statement: No potential conflict of interest was reported by the authors.

Declarations

Ethics approval: This study was approved by the Kasturba Medical College and Kasturba Hospital Institutional Ethics committee in Manipal, India.

Consent to participate: Informed consent was obtained from all individuals included in the study.

Consent for publication: All authors have provided their consent to publish.

References:

Abbas A and Zackular JP. Microbe–microbe interactions during Clostridioides difficile infection. Curr Opin Microbiol 2020;53:19–25.

Al-Khaldi SF, Mossoba MM, Allard MM, et al. Bacterial identification and subtyping using DNA microarray and DNA sequencing. Methods Mol Biol 2012;881:73–95.

Andrews S. FastQC: a quality control tool for high throughput sequence data. 2010. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ [Last accessed: 5/20/2021].

Aziz RK, Bartels D, Best A, et al. The RAST Server: Rapid annotations using subsystems technology. BMC Genomics 2008;9:1–15.

Bartlett JG, Gerding DN. Clinical Recognition and Diagnosis of *Clostridium difficile* Infection. Clin Infect Dis 2008;46:12–18.

Bidet P, Barbut F, Lalande V, et al. Development of a new PCR-ribotyping method for *Clostridium difficile* based on ribosomal RNA gene sequencing. FEMS Microbiol Lett 1999;175(2):261–266.

Bottery MJ, Pitchford JW and Friman VP. Ecology and evolution of antimicrobial resistance in bacterial communities. ISME J 2020 154 2020;15(4):939–948.

Büchler AC, Rampini SK, Stelling S, et al. Antibiotic susceptibility of *Clostridium difficile* is similar worldwide over two decades despite widespread use of broad-spectrum antibiotics: An analysis done at the University Hospital of Zurich. BMC Infect Dis 2014;14(1):1–9.

Cassir N, Benamar S, Croce O, et al. *Clostridium* species identification by 16S rRNA pyrosequencing metagenomics. Clin Infect Dis 2016;62(12):1616–1618.

Cassir N, Benamar S and La Scola B. *Clostridium butyricum*: from beneficial to a new emerging pathogen. Clin Microbiol Infect 2016b;22(1):37–45.

Chen KY, Rathod J, Chiu YC, et al. The transcriptional regulator Lrp contributes to toxin expression, sporulation, and swimming motility in *Clostridium difficile*. Front Cell Infect Microbiol 2019;9:356.

Chen L, Zheng D, Liu B, et al. VFDB 2016: Hierarchical and refined dataset for big data analysis - 10 years on. Nucleic Acids Res 2016;44:694–697.

Chen S, Gu H, Sun C, et al. Rapid detection of *Clostridium difficile* toxins and laboratory diagnosis of Clostridium difficile infections. Infection 2017;45(3):255–262.

CLSI. Clinical and Laboratory Standards Institute. 2020. Available from: https://www.nih.org.pk/wp-content/uploads/2021/02/CLSI-2020.pdf [Last accessed: 4/5/2023].

Darling ACE, Mau B, Blattner FR, et al. Mauve: Multiple alignment of conserved genomic sequence with rearrangements. Genome Res 2004;14(7):1394–1403.

Dawson LF, Valiente E and Wren BW. *Clostridium difficile*—A continually evolving and problematic pathogen. Infect Genet Evol 2009;9(6):1410–1417.

Dingle TC, Mulvey GL and Armstrong GD. Mutagenic analysis of the *Clostridium difficile* flagellar proteins, fliC and fliD, and their contribution to virulence in hamsters. Infect Immun 2011;79(10):4061.

Dilnessa T, Getaneh A, Hailu W, et al. Prevalence and antimicrobial resistance pattern of *Clostridium difficile* among hospitalized diarrheal patients: A systematic review and meta-analysis. PLoS One 2022;17(1):e0262597.

Edwards AN, Karim ST, Pascual RA, et al. Chemical and stress resistances of *Clostridium difficile* spores and vegetative cells. Front Microbiol 2016;7:1698.

Enne VI, Delsol AA, Roe JM, et al. Evidence of antibiotic resistance gene silencing in *Escherichia coli*. Antimicrob Agents Chemother 2006;50(9):3003–3010.

EUCAST. The European Committee on Antimicrobial Susceptibility Testing. 2023. Available from: https://www.eucast.org/fileadmin/src/media/PDFs/EUCAST_files/Breakpoint_tables/v_13.0_Breakpoint_Tables.pdf [Last accessed: 4/5/2023].

Feldgarden M, Brover V, Haft DH, et al. Validating the AMRFINder tool and resistance gene database by using antimicrobial resistance genotype-phenotype correlations in a collection of isolates. Antimicrob Agents Chemother 2019;63(11):1–19.

Ferraris L, Butel MJ, Campeotto F, et al. *Clostridia* in premature neonates' gut: incidence, antibiotic susceptibility, and perinatal determinants influencing colonization. PLoS One 2012;7(1):e30594.

Geier RR, Rehberger TG, Smith AH. Comparative genomics of *Clostridium perfringens* reveals patterns of host-associated phylogenetic clades and virulence factors. Front Microbiol 2021;12:1315.

Gordon A and Hannon G. FASTX-Toolkit: Fastq/a short-reads pre-processing tools. 2010. Available from: http://hannonlab.cshl.edu/fastx_toolkit/index.html [Last accessed: 5/20/2021].

Green MR, Sambrook J. Isolation of high-molecular-weight DNA using organic solvents. Cold Spring Harb Protoc 2017;pdb.prot093450.

Huerta-Cepas J, Forslund K, Coelho LP, et al. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Mol Biol Evol 2017;34(8):2115–2122.

Jin D, Luo Y, Huang C, et al. Molecular epidemiology of *Clostridium difficile* infection in hospitalized patients in eastern China. J Clin Microbiol 2017;55(3):801–810.

Kanai T, Mikami Y and Hayashi A. A breakthrough in probiotics: *Clostridium butyricum* regulates gut homeostasis and anti-inflammatory response in inflammatory bowel disease. J Gastroenterol 2015 509 2015;50(9):928–939.

Johnson JR and Tenover FC. *Clostridium tertium* bacteremia in a patient with aspiration pneumonia: an elusive diagnosis. J Infect Dis 1988;157(4):854–855.

Kourtis AP, Weiner R, Belson K, et al. *Clostridium tertium* meningitis as the presenting sign of a meningocele in a twelve-year-old child. Pediatr Infect Dis J 1997;16(5):527–529.

Kowalska-Krochmal B and Dudek-Wicher R. The minimum inhibitory concentration of antibiotics: methods, interpretation, clinical relevance. Pathog 2021, Vol 10, Page 165 2021;10(2):165.

Kumar S, Stecher G, Li M, et al. MEGA X: Molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol 2018;35(6):1547–1549.

Leffler DA, Lamont JT. *Clostridium difficile* Infection. N Engl J Med 2015;372(16):1539–1548.

Levin BR and Rozen DE. Non-inherited antibiotic resistance. Nat Rev Microbiol 2006 47 2006;4(7):556–562.

Lew JF, Wiedermann BL, Sneed J, et al. Aerotolerant *Clostridium tertium* brain abscess following a lawn dart injury. J Clin Microbiol 1990;28(9):2127–2129.

Lima BB, Fonseca BF, Amado N da G, et al. *Clostridium difficile* toxin A attenuates Wnt/β-catenin signaling in intestinal epithelial cells. Infect Immun 2014;82(7):2680–2687.

Low LY, Harrison PF, Gould J, et al. Concurrent host-pathogen transcriptional responses in *Clostridium perfringens* murine myonecrosis infection. MBio 2018;9(2).

Mahon C., Manuselis G. Textbook of diagnostic microbiology. W.B. Saunders Company, Tokyo; 1995.

Melnyk AH, Wong A, Kassen R. The fitness costs of antibiotic resistance mutations. Evol Appl 2015;8(3):273–283.

Moore RJ and Lacey JA. Genomics of the pathogenic *Clostridia*. Microbiol Spectr 2019;7(3).

Muñoz M, Restrepo-Montoya D, Kumar N, et al. Comparative genomics identifies potential virulence factors in *Clostridium tertium* and *C. paraputrificum*. Virulence 2019;10(1):657–676.

Neumann-Schaal M, Jahn D, Schmidt-Hohagen K. Metabolism the difficile way: The key to the success of the pathogen *Clostridioides difficile*. Front Microbiol 2019;10:219.

Page AJ, Cummins CA, Hunt M, et al. Roary: Rapid large-scale prokaryote pan genome analysis. Bioinformatics 2015;31(22):3691–3693.

Pereira FC, Saujet L, Tomé AR, et al. The spore differentiation pathway in the enteric pathogen *Clostridium difficile*. PLOS Genet 2013;9(10):e1003782.

Prasad ASB, Shruptha P, Prabhu V, et al. *Pseudomonas aeruginosa* virulence proteins pseudolysin and protease IV impede cutaneous wound healing. Lab Investig 2020;100(12):1532–1550.

Prjibelski A, Antipov D, Meleshko D, et al. Using SPAdes De Novo Assembler. Curr Protoc Bioinforma 2020;70(1):1–29.

Ranford JC, Henderson B. Chaperonins in disease: mechanisms, models, and treatments. Mol Pathol 2002;55(4):209.

Riggs MM, Sethi AK, Zabarsky TF, et al. Asymptomatic carriers are a potential source for transmission of epidemic and nonepidemic *Clostridium difficile* strains among long-term care facility residents. Clin Infect Dis 2007;45(8):992–998.

Rineh A, Kelso MJ, Vatansever F, et al. *Clostridium difficile* infection: Molecular pathogenesis and novel therapeutics. Expert Rev. Anti. Infect. Ther. 2014;12(1):131–150.

Roehl HH and Conrad SE. Identification of a G1-S-phase-regulated region in the human thymidine kinase gene promoter. Mol Cell Biol 1990;10(7):3834–3837.

Saxton K, Baines SD, Freeman J, et al. Effects of exposure of *Clostridium difficile* PCR ribotypes 027 and 001 to fluoroquinolones in a human gut model. Antimicrob Agents Chemother 2009;53(2):412–420.

Schmidt R, Cordovez V, De Boer W, et al. Volatile affairs in microbial interactions. ISME J 2015 911 2015;9(11):2329–2335.

Seemann T. Prokka: Rapid prokaryotic genome annotation. Bioinformatics 2014;30(14):2068–2069.

Seemann T. Abricate. 2021. Available from: https://github.com/tseemann/abricate [Last accessed: 5/20/2021].

Seppey M, Manni M and Zdobnov EM. BUSCO: Assessing genome assembly and annotation completeness. In: Methods in Molecular Biology. Humana Press Inc., 2019; pp. 227–245.

Shah DN, Aitken SL, Barragan LF, et al. Economic burden of primary compared with recurrent *Clostridium difficile* infection in hospitalized patients: a prospective cohort study. J Hosp Infect 2016;93(3):286–289.

Smits WK, Lyras D, Lacy DB, et al. *Clostridium difficile* infection. Nat Rev Dis Prim 2016;2(1):1–20.

Spigaglia P, Mastrantonio P and Barbanti F. Antibiotic resistances of *Clostridium difficile*. Adv Exp Med Biol 2018;1050:137–159.

Steensma EA, Ertl CW and Burke LH. *Clostridium tertium* isolated from a necrotizing soft tissue infection in a diabetic but otherwise nonimmunocompromised patient. J Am Col Certif Wound Spec 2011;3(2):42.

Stogios PJ and Savchenko A. Molecular mechanisms of vancomycin resistance. Protein Sci 2020;29(3):654–669.

Stabler RA, He M, Dawson L, et al. Comparative genome and phenotypic analysis of *Clostridium difficile* 027 strains provides insight into the evolution of a hypervirulent bacterium. Genome Biol 2009;10(9):1–15.

Stothard P and Wishart DS. Circular genome visualization and exploration using CGView. Bioinformatics 2005;21(4):537–539.

Tarashi S, Zamani MS, Omrani MD, et al. Commensal and pathogenic bacterial-derived extracellular vesicles in host-bacterial and interbacterial dialogues: two sides of the same coin. J Immunol Res 2022.

Tatusova T, Dicuccio M, Badretdin A, et al. NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res 2016;44(14):6614–6624.

Utturkar SM, Klingeman DM, Bruno-Barcena JM, et al. Sequence data for *Clostridium autoethanogenum* using three generations of sequencing technologies. Sci Data 2015;2(1):1–9.

Verherstraeten S, Goossens E, Valgaeren B, et al. Perfringolysin O: The underrated *Clostridium perfringens* toxin? Toxins (Basel) 2015;7(5):1702–1721.

Vuotto C, Donelli G, Buckley A, et al. *Clostridium difficile* biofilm. Adv Exp Med Biol 2018;1050:97–115.

Wu X, Santos RR and Fink-Gremmels J. Analyzing the antibacterial effects of food ingredients: model experiments with allicin and garlic extracts on biofilm formation and viability of Staphylococcus epidermidis. Food Sci Nutr 2015;3(2):158–168.

Zhang J, Sun J, Chen X, et al. Combination of *Clostridium butyricum* and corn bran optimized intestinal microbial fermentation using a weaned pig model. Front Microbiol 2018;9:3091.

Supplementary Data

Table S1: Reference strains utilized for different analysis for all three assembled genomes.

Strain	Reference Genome	Analysis
C. butyricum MALS002	*C. butyricum* strain NBRC 13949 (AP019716)	Genome Assembly
C. difficile MALS003	*C. difficile* strain R20291 (CP029423)	
C. tertium MALS001	*C. tertium* isolate src5 (OAOE01)	
C. butyricum MALS002	1. *C. butyricum* strain NBRC 13949 (AP019716) 2. *C. butyricum* strain DSM 10702 (CP040626)	Whole Genome Alignment using Mauve
C. difficile MALS003	1. *C. difficile* strain R20291 (CP029423) 2. *C. difficile* strain ATCC 9689 (CP011968)	
C. tertium MALS001	1. *C. tertium* isolate src5 (OAOE01) 2. *C. tertium* strain MGYG-HGUT-01328 (CABKOG01)	
C. butyricum MALS002	1. *C. butyricum* strain 16-3 (CP053292) 2. *C. butyricum* strain DSM 10702 (CP040626) 3. *C. butyricum* strain NBRC 13949 (AP019716) 4. *C. butyricum* strain S-45-5 (CP030775)	Pangenome
C. difficile MALS003	1. *C. difficile* strain ATCC 9689 (CP011968) 2. *C. difficile* strain DSM 29745 (CP019857) 3. *C. difficile* strain FDAARGOS 267 (CP020424) 4. *C. difficile* strain R20291 (CP029423)	
C. tertium MALS001	1. *C. tertium* strain src5 (OAOE01) 2. *C. tertium* strain MGYG-HGUT-01328 (CABKOG01) 3. *C. tertium* strain BSD2780120875b 170604 A12 (JADPEJ01)	
C. butyricum MALS002	1. *C. butyricum* strain NBRC 13949 (AP019716) 2. *C. butyricum* strain DSM 10702 (CP040626)	BUSCO
C. difficile MALS003	1. *C. difficile* strain R20291 (CP029423) 2. *C. difficile* strain ATCC 9689 (CP011968)	
C. tertium MALS001	1. *C. tertium* strain src5 (OAOE01) 2. *C. tertium* strain MGYG-HGUT-01328 (CABKOG01)	

Table S2: MIC values of all three species of *Clostridia* against select antibiotics.

Drug Class	Antibiotics	CB MALS002 (µg)	CD MALS003 (µg)	CT MALS001 (µg)
Glycopeptide	Vancomycin	S	R	S
Fluoroquinolone	Ciprofloxacin	R	S	S

R – Resistant; S – Sensitive; I – Intermediate

Table S3: Antimicrobial resistance genes predicted in all three strains are listed with respective drug classes they belong to, and number of gene copies predicted.

Drug Class	AMR gene	CB MALS002 % Identity	CD MALS003 % Identity	CT MALS001 % Identity
Acridine dye	cdeA		99.77	
Beta-lactam	blaCDD		94.95	
Fluoroquinolone	cdeA		99.77	
Glycopeptide	vanG		99.73	
	vanR		100	
	vanS		99.65	
	vanT		99.58	
	vanZ1		99.22	
Phenicol	catP	70.05		
Streptogramin	vatB			76.16

Table S4: Toxin genes predicted using VFDB core and full dataset resources for all three strains.

Strains	Category	Core toxins	Full dataset toxins
Clostridium butyricum MALS002	Adherence	groEL	fbpA
	Exotoxin	pfoA	
	Hemolysin		Cbei_0023, Cbei_1707, ZP_02950902
	Flagellar assembly / Chemotaxis		fliS1, flil, cheY
	Putative		rmlA, plr
	Other		ureB, ureG
Clostridioides difficile MALS003	Adherence	groEL, CD0873, CD2831, CD3246, cwp66, cwp84, tufA	fbpA
	Exotoxin	iap, ibp, toxA, toxB	
	Binary toxin	cdtA, cdtB	
	Exoenzyme	Zmp1	
	Flagellar assembly / Chemotaxis		motA, motB, fleN, flgB, flgC, flgD, flgE, flgG, flgK, flgL, flgM, flhB, flhF, fliA, fliC, fliD, fliE, fliF, fliG, fliH, fliI, fliJ, fliK, fliM, fliQ, flhA, fliP, fliN, fliS1, fliS2, flil

	Putative		plr, CD0233, CD0230, CD0255A, CD0259, CD0240, CD1546, CD1208
Clostridium tertium MALS001	Adherence	groEL, tufA	fbpA
	Exoenzyme	nagK	
	Hemolysin		CLL_A2400
	Flagellar assembly / Chemotaxis		cheY, flhA, flip, fliS2, flil, pilT
	Putative		plr, SSP0068
	Other		ureG

Table S5: Sporulation genes present in three strains with their functions and stages of sporulation process in which they play a role.

No.	Gene	Stage	Function	*Clostridium butyricum MALS002*	*Clostridioides difficile MALS003*	*Clostridium tertium MALS001*
1	disA		The rise in c-di-AMP level generated by DisA while scanning the chromosome, operates as a positive signal that advances sporulation	+	+	+
2	ftsX		Part of the ABC transporter FtsEX involved in asymmetric cellular division facilitating the initiation of sporulation	+	-	+
3	soj		sporulation initiation inhibitor protein Soj	+	-	+
4	spo0A	0	Spo0A may act in concert with spo0H (a sigma factor) to control the expression of some genes that are critical to the sporulation process	+	+	+
5	spoIID	2	stage II sporulation protein D	+	+	+
6	spoIIE	2	stage II sporulation protein E	+	+	+
7	spoIIGA	2	aspartic protease, responsible for the proteolytic cleavage of the RNA polymerase sigma E factor (SigE spoIIGB) to yield the active peptide in the mother cell during sporulation.	+	-	+
8	spoIIM	2	Stage II sporulation protein M	+	-	+
9	spoIIP	2	Stage II sporulation protein P (SpoIIP)	+	+	+
10	spoIIR	2	stage II sporulation protein R	+	+	+
11	spoIIIAA	3	stage III sporulation protein AA	+	-	+
12	spoIIIAB	3	Stage III sporulation protein AB	+	+	+
13	spoIIIAC	3	Stage III sporulation protein AC/AD protein family	+	+	-
14	spoIIIAD	3	Stage III sporulation protein AD	+	+	+
15	spoIIIAE	3	stage III sporulation protein AE	+	+	+
16	spoIIIAF	3	Stage III sporulation protein af	+	-	+
17	spoIIIAG	3	stage III sporulation protein AG	+	+	+
18	spoIIIAH	3	Stage III sporulation protein	+	-	+
19	spoIIID	3	Stage III sporulation protein D	+	+	+
20	spoiIIIAF	3	Stage III sporulation protein AF (Spore_III_AF)	-	+	-
21	spoIVB	4	Stage IV sporulation protein B	+	-	+
22	spoVAC	5	stage V sporulation protein AC	+	+	+
23	spoVAD	5	Stage V sporulation protein AD	+	+	+

24	spoVAE	5	stage V sporulation protein	+	+	+
25	spoVAEB	5	stage V sporulation protein	-	-	+
26	spoVB	5	Stage V sporulation protein B	+	+	+
27	spoVD	5	stage V sporulation protein D	+	-	+
28	spoVK	5	stage V sporulation protein K	+	-	+
29	spoVR	5	stage V sporulation protein R	+	-	+
30	spoVS	5	Stage V sporulation protein S	+	+	+
31	spoVT	5	stage V sporulation protein T	+	-	+
32	whiA		May be required for sporulation	+	+	+
33	yabG		sporulation peptidase YabG	+	-	+
34	yabP		Sporulation protein YabP	+	-	+
35	ylbJ		Sporulation integral membrane protein YlbJ	+	+	+
36	ylmC		Sporulation protein, YlmC YmxH	+	-	+
37	yqfC		sporulation protein YqfC	+	-	+
38	yqfD	4	Putative stage IV sporulation protein YqfD	+	+	-
39	yteA		TIGRFAM Sporulation protein YteA	-	-	+
40	ytfJ		Sporulation protein YtfJ	+	+	+
41	ytvI		sporulation integral membrane protein YtvI	-	+	+
42	ytxC		sporulation protein YtxC	+	-	+
43	yunB		sporulation protein YunB	+	+	+
44	yyaC		Sporulation protein YyaC	+	-	+
45	sigK	3	RNA polymerase sporulation specific sigma factor SigK	+	+	+
46	sigE	2	RNA polymerase sporulation specific sigma factor SigE	+	+	+
47	sigG	3	RNA polymerase sporulation specific sigma factor SigG	+	+	+
48	sigF	2	RNA polymerase sporulation specific sigma factor SigF	+	+	+
49	sigH	0	RNA polymerase sporulation specific sigma factor SigH	+	+	+
50	sigA		RNA polymerase sporulation specific sigma factor SigA	+	+	+
51	sigB		RNA polymerase sporulation specific sigma factor SigB	-	+	-
52	sigI		RNA polymerase sporulation specific sigma factor SigI	+	-	-
53	sigV		RNA polymerase sporulation specific sigma factor SigV	+	+	-

(+) Present

(-) Absent

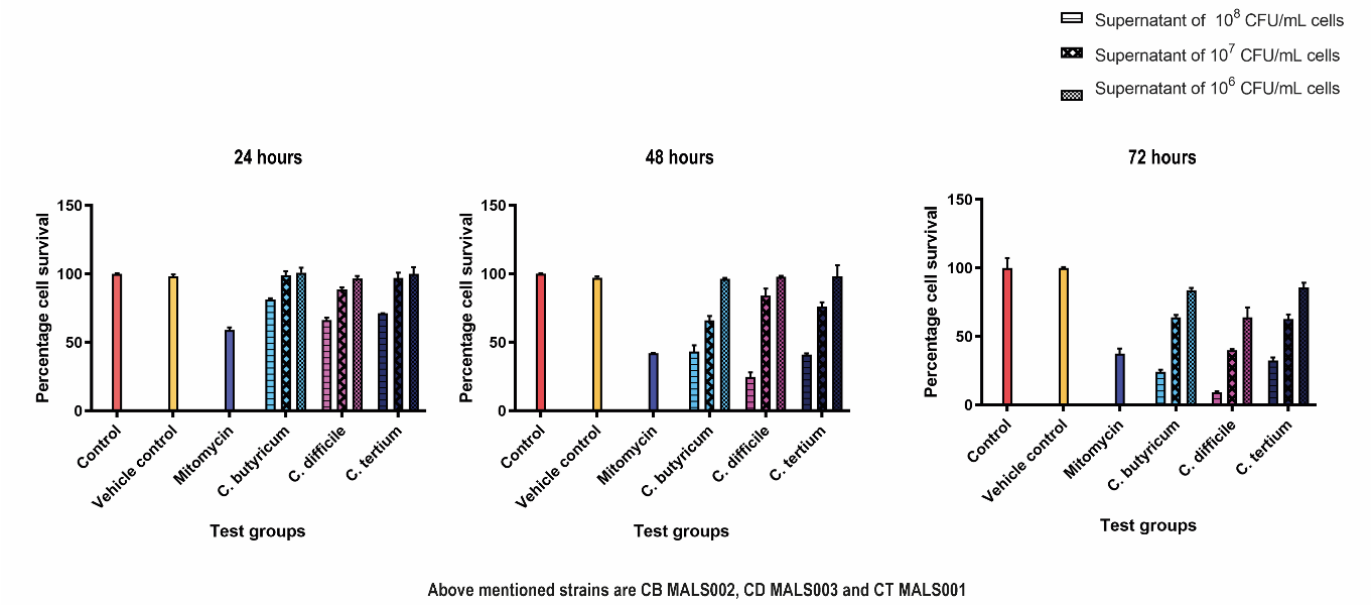


Figure S1: Cytotoxicity of CB MALS002, CD MALS003, and CT MALS001. Cytotoxicity of *Clostridia* supernatant (at different dilutions) on cell survival of SiHa cells was tested by CCK-8 assay at 24, 48 and 72 hours

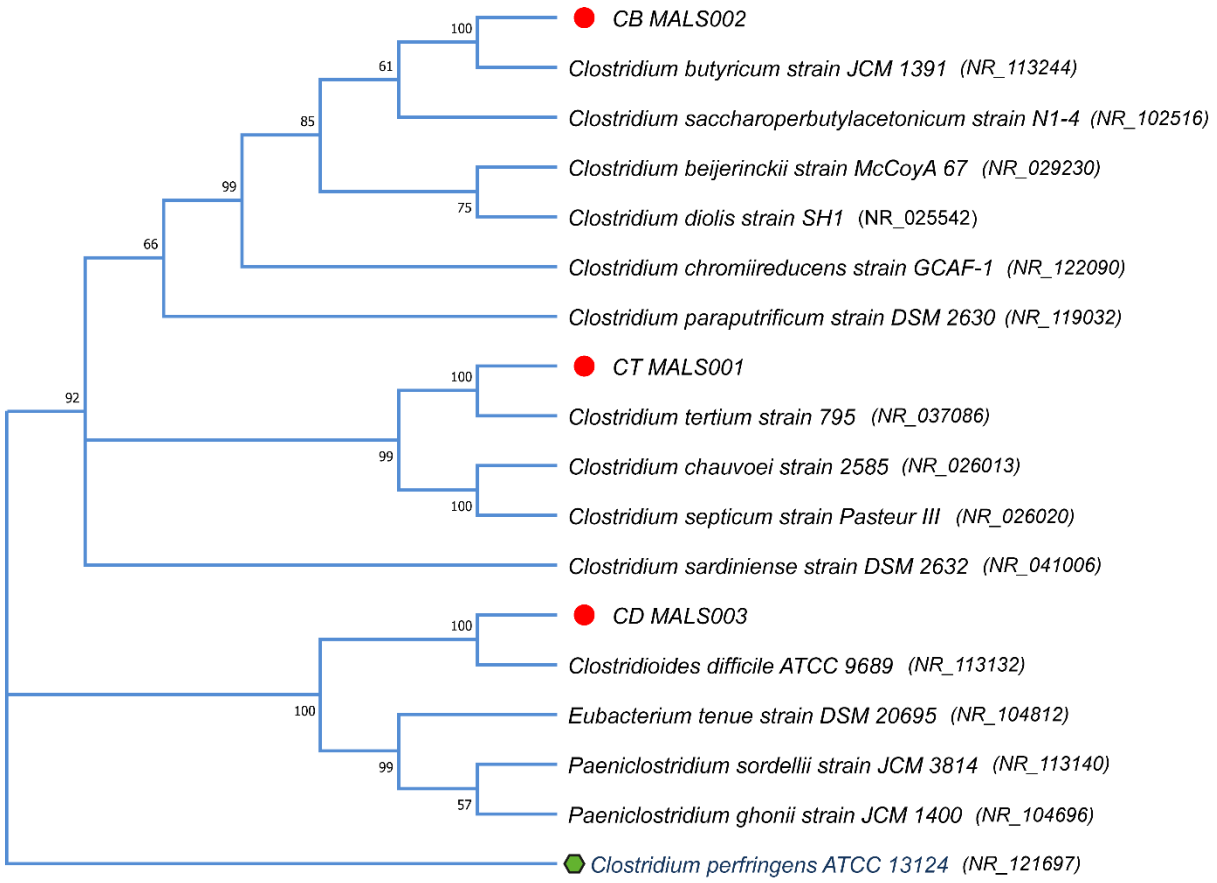
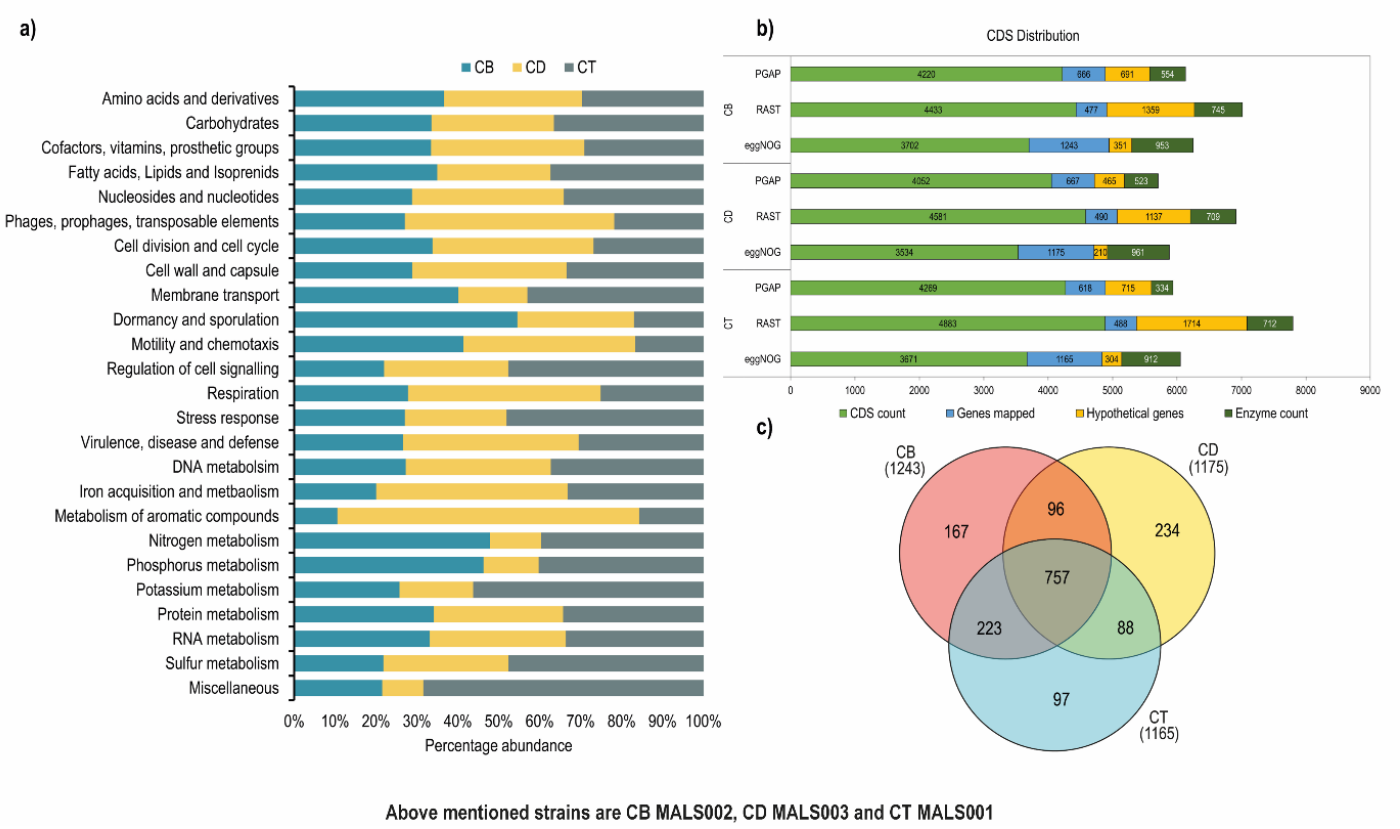


Figure S2: 16S rRNA phylogeny**.** Extracted 16S rRNA sequences of CB MALS002, CD MALS003 and CT MALS001 were blasted against rRNA/ITS databases to retrieve homologs. BlastN results with default parameters were filtered based on query coverage ≥ 97% and percentage identity ≥ 95%. Sequences that matched these criteria were used to construct a phylogenetic tree using MEGA X (Kumar et al., 2018) with maximum-likelihood method and 1000 bootstrap replicates.



Above mentioned strains are CB MALS002, CD MALS003 and CT MALS001

Figure S3: Functional annotation of CB MALS002, CD MALS003, and CT MALS001. (a) Subsytem category distribution as obtained by RAST. (b) Annotated CDS, mapped genes, hypothetical genes and enzyme count distribution across three strains for PGAP, RAST and eggNOG (c) EggNOG annotations indicating “core” and strain specific protein coding genes

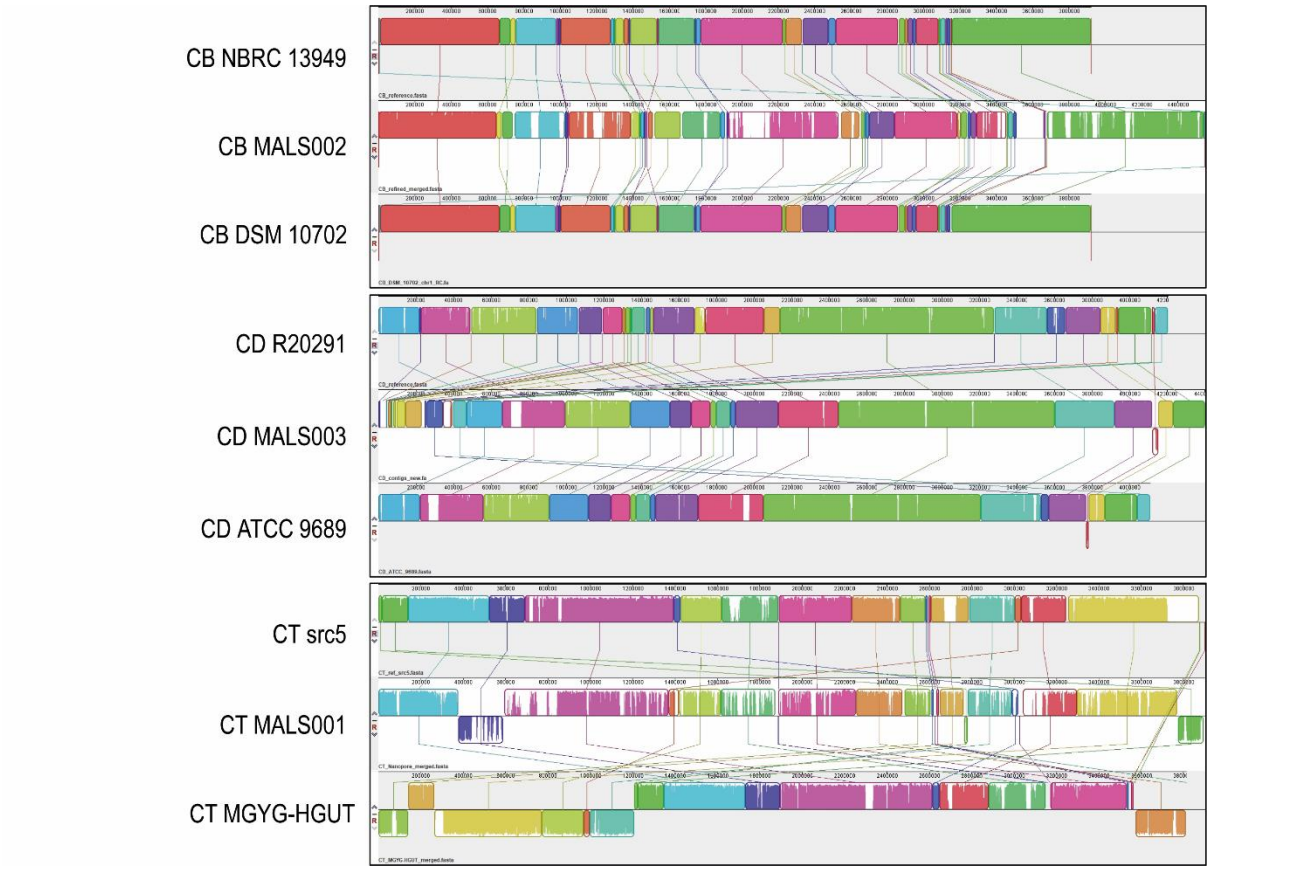


Figure S4: Whole genome alignment of CB MALS002, CD MALS003 and CT MALS001 with their respective reference genomes using Mauve. Coloured blocks and the respective connecting lines indicate regions of alignment. Alignments in reverse orientation are indicated by placement of the blocks below the central line. Similarity profiles within the blocks indicate level of conservation within the alignment.

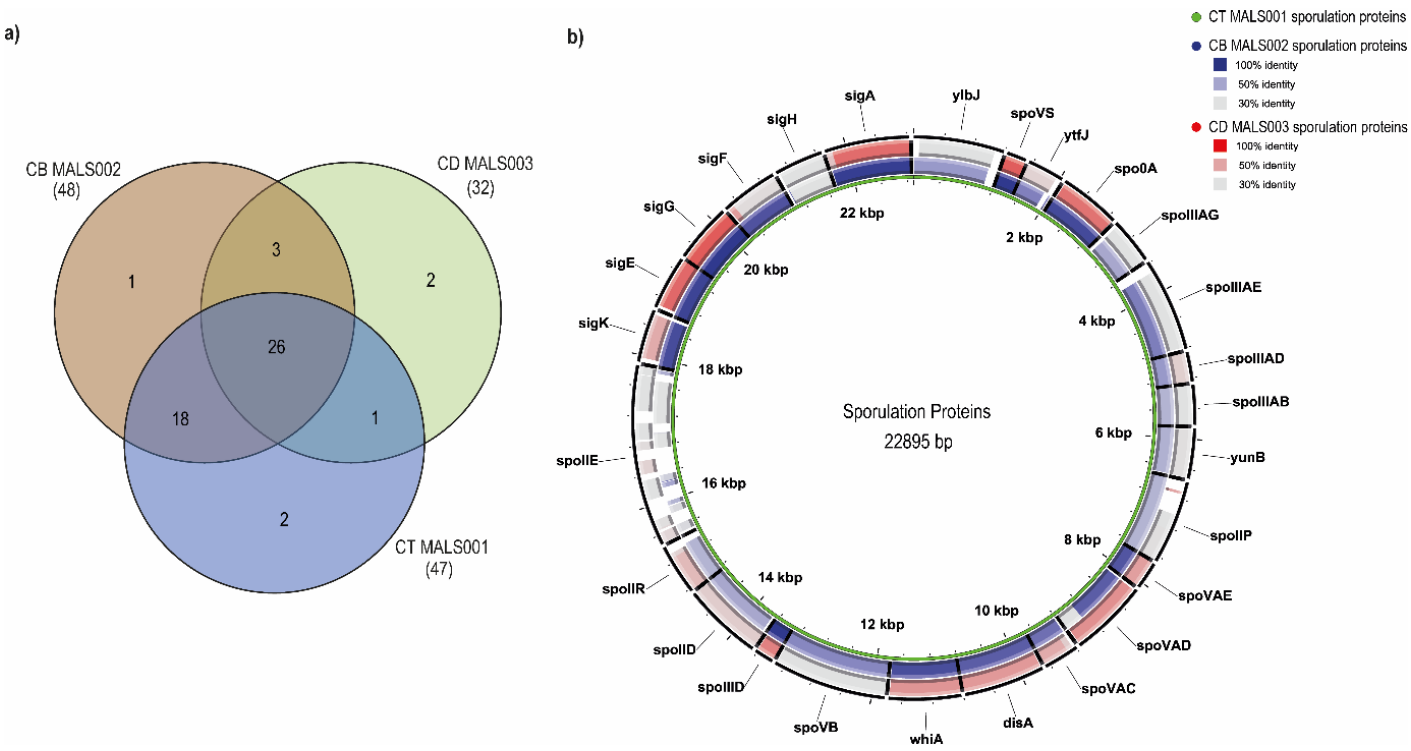


Figure S5: Sporulation proteins in CB MALS002, CD MALS003, and CT MALS001. (a) Distribution of core and unique genes encoding sporulation in the three *Clostridia* (b) Blastp comparison of sporulation proteins across the three *Clostridia* visualized by BRIG tool. Regions of high dissimilarity are indicated by gaps

Chapter

5

Antimicrobial Resistance and Virulence: Genome Comparison of *Staphylococcus aureus* Strains

Tanwar, Ankit Singh, Padival Shruptha, Bobby Paul, Thokur Sreepathy Murali, Angela Brand, and Kapaettu Satyamoorthy. "How Can Omics Inform Diabetic Foot Ulcer Clinical Management? A Whole Genome Comparison of Four Clinical Strains of *Staphylococcus aureus*." *OMICS: A Journal of Integrative Biology* 27, no. 2 (2023): 51-61.

DOI: 10.1089/omi.2022.0184; **IF 3.3 (2023)**

How Can Omics Inform Diabetic Foot Ulcer Clinical Management? A Whole Genome Comparison of Four Clinical Strains of *Staphylococcus aureus*

Abstract

Foot ulcers and associated infections significantly contribute to morbidity and mortality in diabetes. While diverse pathogens are found in the diabetes-related infected ulcers, *Staphylococcus aureus* remains one of the most virulent and widely prevalent pathogens. The high prevalence of *S. aureus* in chronic wound infections, especially in clinical settings, is attributed to its ability to evolve and acquire resistance against common antibiotics and to elicit an array of virulence factors. In this study, whole genome comparison of four strains of *Staphylococcus aureus* (MUF168, MUF256, MUM270 and MUM475) isolated from diabetic foot ulcer infections showing varying resistance patterns was carried out to study the genomic similarity, antibiotic resistance profiling, associated virulence factors and sequence variations in drug targets. The comparative genome analysis showed strains MUM475 and MUM270 to be highly resistant, MUF256 with moderate levels of resistance, and MUF168 to be the least resistant. Strain MUF256 and MUM475 harboured more virulence factors compared to other two strains. Deleterious sequence variants were observed suggesting potential role in altering drug targets and drug efficacy. This comparative whole genome study offers new molecular insights that may potentially inform evidence-based diagnosis and treatment of diabetic foot ulcers in the clinic.

Keywords: diabetes, antimicrobial resistance, biofilm, virulence factors, pangenome, drug targets

Introduction

Diabetic foot ulcer (DFU) is a serious comorbid condition associated with Type 2 diabetes and high mortality rates among the older population (Brennan et al, 2017). About 4-10% of diabetic patients are affected by foot ulceration and the elderly population are more susceptible to the DFU. The incidence of lower limb amputation is 155 times higher in diabetic individuals with foot ulcers compared to those without infection (Lavery et al, 2006). Management of chronic, non-healing foot ulcers is a major challenge due to the presence of microbiome communities at the site of infection (Percival et al, 2018).

Although Gram-negative bacilli (such as *Pseudomonas aeruginosa*) and obligate anaerobic bacteria (such as *Veillonella* spp.) are common, *Staphylococcus aureus*, an aerobic, Gram-positive coccus has been found to be one of the major pathogens reported from infected diabetic foot ulcers (Murali et al, 2014a; Shettigar et al, 2016). It has been shown that biofilm-forming *S. aureus* are capable of specifically inhibiting wound healing and worsening the wound infection (Bowling et al, 2009). The effective colonization and invasion of wound tissues by *S. aureus* is facilitated by the presence of *ica-AB* gene which confers the ability to produce biofilm and the ability to produce several toxins including pore-forming toxins that lyse the host cells, exfoliative toxin which facilitate bacterial skin invasion, enterotoxins that contribute to emetic and pyrogenic effects and epidermal cell differentiation inhibitors that promote their distribution in tissues (Dunyach-Remy et al, 2016;

Shettigar et al, 2016). In addition to its high virulence capabilities, high levels of antibiotic resistance to common antibiotics makes this pathogen highly successful in pathobiome wound environments.

The standard practices in treatment of diabetic foot ulcer involves the removal of the necrotic tissues, management of the infection and offloading of the ulcer. However, the recent widespread occurrence of multidrug-resistant bacteria can severely restrict devising proper wound management strategies. The key to combating the current antibiotic resistance crisis lies in the development of specific and sensitive methods to detect antibiotic resistance in clinical strains. Comparative genomics is a widely used computational approach to explore the mechanisms of evolution of virulence and drug resistance from clinically important bacterial species (Coll et al, 2015; Holden et al, 2004). The reduced costs associated with whole genome sequencing (WGS) has made it a viable alternative in understanding strain level differences at genome level (Price et al, 2013). Analysis of Whole Genome Sequence (WGS) data can be used as a primary tool for the detection of multidrug resistance, virulence capabilities, disruptive targets, candidate drug compounds, mechanisms and evolution of pathogenicity (Punina et al, 2015).

The current study focuses on whole genome comparative analysis of four *S. aureus* strains (MUF168, MUF256, MUM270, and MUM475) isolated from diabetic foot ulcer patients to understand their microbial resistance and virulence profiles. We also studied the presence of potential drug targets and their degree of sequence variation (nucleotide and corresponding amino acid changes) between four strains to explore altered drug binding mechanisms. The whole genome comparative studies of extremely close strains can highlight key genomic characteristics which can differentiate pathogens from nonpathogens.

Materials and Methods

Bacterial strains and isolation of genomic DNA

The four *S. aureus* strains were obtained in an earlier study from diabetic patients with infected foot ulcer visiting Kasturba Hospital, Manipal, a tertiary care hospital in southern India, over a period of three years between 2010 and 2012 (Murali et al, 2014b). Wound swabs were collected following debridement of superficial exudates from infected ulcers and cultured for bacteria in Blood agar and MacConkey agar. The present study was conducted under the full research ethics oversight of the authors institutions. The bacterial strains were revived in peptone water and kept in shaker incubator overnight at 37° C. The overnight grown cultures were streaked onto MacConkey Agar plates. The revived bacterial cultures were grown in 2X YT medium to get a concentration of 10^8 cells/ml. The cells were centrifuged to obtain a pellet and DNA was extracted by phenol-chloroform method. The quality of the DNA was checked in a 0.8% agarose gel (Green and Sambrook, 2017).

Antibiotic resistance profiling

The minimum inhibitory concentration of antibiotics against the four strains were determined by using E-strips (HiMedia, India). 10^7 CFU/ml of bacterial culture were grown in Muller Hinton Agar (MHA) and the antibiotic strips were placed on the lawn culture and incubated overnight at 37° C and checked for the zone of inhibition

and the results were interpreted as per the Clinical and Laboratory Standard Institute guidelines (CLSI, 2011). The following nine antibiotics were studied for their MIC values: amikacin, amoxicillin/clavulanate, ampicillin, linezolid, chloramphenicol, ciprofloxacin, erythromycin, teicoplanin and vancomycin.

Genome data extraction and annotation

The present study used genome assemblies and annotation data of four *S. aureus* strains MUF168, MUF256, MUM270 and MUM475 isolated from four different individuals affected with diabetic foot ulcer (Murali et al, 2014b). The genome assemblies of these four *S. aureus* strains are available at the NCBI Genome database with accession numbers AZQR00000000 (MUF168), AZSE00000000 (MUF256), AZSF00000000 (MUM270), and AZSG00000000 (MUM475). The quality of these four genome assemblies were assessed using Quast v5.0.2 (Gurevich et al, 2013). Whole genome of all four strains were submitted to RAST (Aziz et al, 2008) to obtain subsystem category distribution and NCBI annotated protein sequences were submitted to eggNOG-mapper (Huerta-Cepas et al, 2019) for taxa-level (family level: Staphylococcaceae) specific gene annotations.

Genome comparison and pangenome analysis

Whole-genome sequence comparison of four *S. aureus* strains with respect to the reference genome (*S. aureus* NCTC 8325 [accession: CP000253]) was performed using megablast (Altschul et al, 1990). The DNA–DNA hybridization (DDH) values of these genomes were calculated by using the GGDC (Genome-to-Genome Distance Calculator) v2.0 (Meier-Kolthoff et al, 2013) for genome-to-genome comparison. Raw fasta sequences from genome assembly data were annotated using Prokka v1.14.6 (Seemann, 2014) and were used for pangenome analysis with Roary v3.13.0 (Page et al, 2015) against reference *S. aureus* strains NCTC 83225, ATCC-12600 and WBG8287. To predict single-copy gene content and conservation along with the orthology status of predicted genes across four *S. aureus* strains, BUSCO v4.1.2 (Manni et al, 2021) tool was utilized with a lineage-specific (bacillales: order-level) dataset.

Antimicrobial resistance gene and virulence factor identification

Antibiotic resistance gene profiling was done using Abricate v1.0.1 (Seemann, 2021) and AMRFinder v3.10.5 (Feldgarden et al, 2019) tools. For Abricate, CARD (Alcock et al, 2020) and MEGARes (Doster et al, 2020) database resources were utilized while AMRfinder uses BARRGD (Bacterial Antimicrobial Resistance Reference Gene Database) for screening antimicrobial resistance genes. Antimicrobial resistance genes with a query coverage of ≥80% and sequence identity of ≥70% were taken into consideration. Validation of antibiotic resistance genes presence was done by single target PCR amplification for genes *mecA*, *ant(4')-Ib* and *ermC* and PCR products were visualized on 1.2% agarose gel.

To check the presence of virulence factors in four *S. aureus* strains, VFDB (Chen et al, 2016) resources using Abricate tool was applied. The nucleotide core dataset of VFDB was used to obtain experimentally verified virulence factors with the following filter criteria: minimum coverage of 85% and minimum identity of 70%. Also, a total of 95 virulence-associated protein sequences reported for *S. aureus* in VFDB protein core dataset

were extracted and aligned using protein-blast (blastp) against all four strains to obtain sequence identity at protein level.

Biofilm formation analysis

MUF168, MUF256, MUM270, and MUM475 strains were assessed for biofilm formations in dynamic and static condition. Dynamic biofilm formation ability was checked using BioFlux TM microfluidic system (Fluxion Bioscience Inc., Alameda, CA, USA) (Pouget et al, 2021). Tissue culture plate method was used to determine the static biofilm formation using crystal violet (0.1%) staining (Prasad et al, 2020).

Drug target study

The nucleotide and protein sequences of approved drug targets reported for *S. aureus* were downloaded from the DrugBank database (Wishart et al, 2018). A BLAST sequence similarity search between the approved drug targets and *S. aureus* strains was performed to confirm drug target presence in genomes of four strains. Further, bowtie2 aligner was used to align raw fastq reads against approved drug target sequences to verify blast results. The variation in drug targets at the nucleotide level were identified using UnifiedGenotyper of Genome Analysis Toolkit (GATK, version=3.8-1-0-gf15c1c3ef) pipeline (McKenna et al, 2010).

To identify variations at the protein level, the variants in VCF format were converted to fasta format using FastaAlternateReferenceMaker (GATK) and translated into proteins sequences using Transeq (Madeira et al, 2019). Multiple sequence alignment of translated protein sequences was performed using ClustalOmega (Sievers et al, 2011) and PROVEAN (Choi and Chan, 2015) tool was utilized to predict protein variant impact on the biological function of protein. Different online tools were utilized to summarize and visualize results obtained from above analysis (Heberle et al, 2015; Stothard and Wishart, 2005; Petkau et al, 2010).

Results

Genome annotation

The draft genome of MUF168, MUF256, MUM270, and MUM475 strains consisted of 59 (2.75 mb), 58 (2.78 mb), 125 (2.83 mb) and 74 (2.85 mb) contigs, respectively. The protein-coding density of MUM270 was found to be the highest with 2728 total proteins followed by MUM475 (2620 proteins) and the two MUF strains (2433 proteins each). The average number of ribosomal RNA genes predicted across four strains was sixteen. GenBank annotation details and other genomic characteristics are provided in **Supplementary Table S1**.

Functional annotations and subsystem category distribution of genes using RAST identified 97 (MUF168), 114 (MUF256), 100 (MUM270) and 97 (MUM475) genes in the virulence category. A total of 72 genes for MUF168 and MUM475 strains and 77 genes for MUF256 and MUM270 were categorized as stress response genes. A maximum number of genes involved in membrane transport was predicted for MUF strains, 101 (MUF256) and 100 (MUF168) followed by 88 genes in MUM475 and 58 genes in MUM270. All subsystem categories are summarized in **Figure 1a** along with subsystem coverage provided in **Figure 1b**. Highest

number of genes mapped to any subsystem were 1618 genes for strain MUF256 followed by MUF168 (1554 genes), MUM270 (1542 genes) and MUM475 (1524 genes). Taxa-level specific gene annotations by eggNOG-mapper mapped 979, 975, 1041 and 1044 genes for MUF168, MUF256, MUM270 and MUM475. A total of 888 genes from eggNOG plus GenBank annotations were shared between the four strains **(Figure 1c)**.

Both Cluster of Orthologues (COG) database and RAST functional annotation yielded similar results with majority of genes mapped to the amino acid transport and metabolism category while least number of genes were mapped to cell motility category **(Supplementary Figure S1)**. Higher abundance of genes in 'function unknown' category based on COG database and presence of many hypothetical genes in RAST indicated that a considerable amount of *S. aureus* genome function is unknown and remains to be explored.

Sequence comparison and pangenome analysis

Whole-genome sequence comparison using GGDC and megablast **(Supplementary Table S2)** identified high similarity between MUF256 and MUM475 strains with a query coverage of 99% and sequence identity of 99.91%. Sequence similarity between MUF256 and MUM270 strains was predicted to be the lowest with a query coverage of 93% and identity of 99.09%. A blastn comparison of nucleotide sequences revealed regions of high similarity and unique regions **(Supplementary Figure S2)** present in four strains compared to the reference sequence. Coding sequence (CDS) blast identified 185 unique CDS present in the reference genome (strain NCTC-8325). Out of these 185 unique CDS, 165 CDS were found to be hypothetical proteins and the description of the remaining 20 coding sequences is listed in **Supplementary Table S3**.

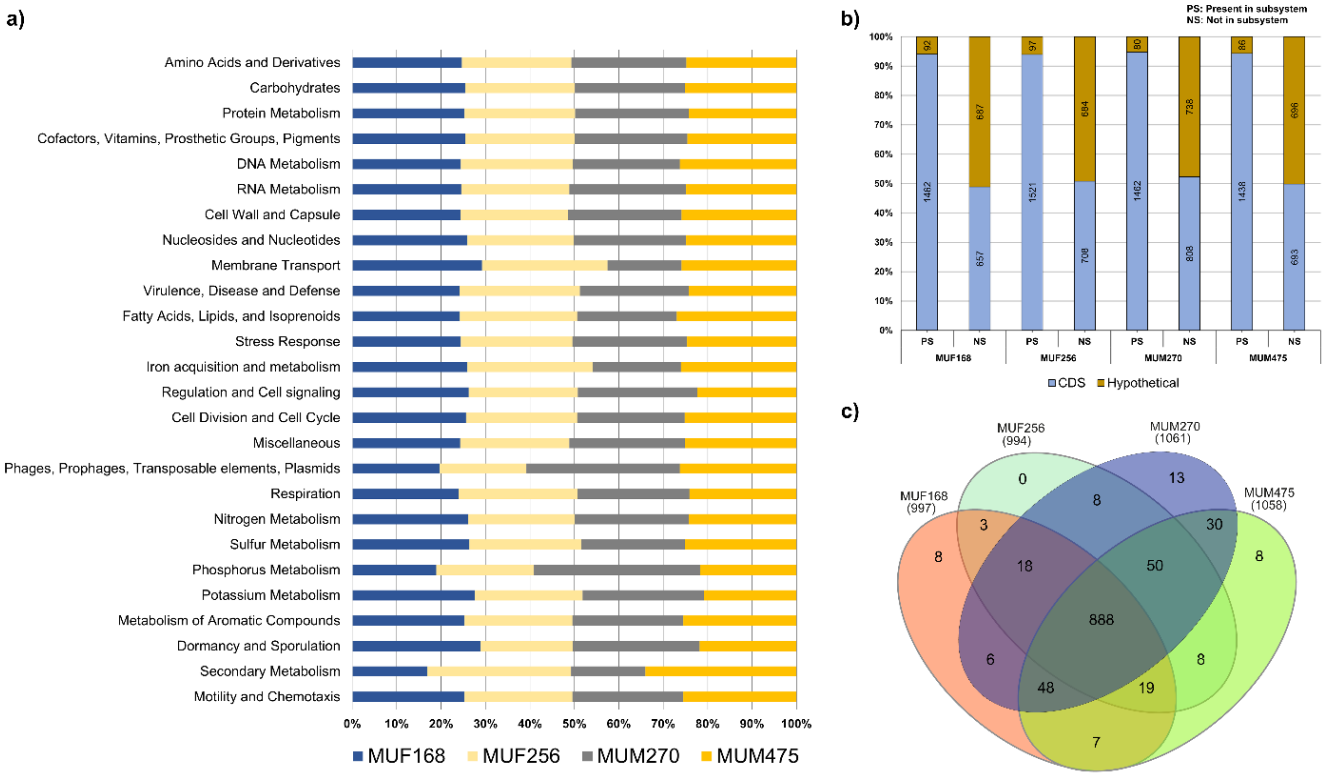


Figure 1: Gene annotation summary. (a) RAST subsystem category distribution of annotated genes in all four *S. aureus* strains. (b) Subsystem coverage summary with hypothetical gene count. (c) Core and unique genes shared across four strains.

Pangenome analysis of all four *S. aureus* strains along with three reference genomes (strain: NCTC-8325, ATCC-12600 and WBG8287) identified a core genome (genes that shared more than 99% similarity across genomes) of 1951 (45.5%) genes and accessory genome (genes with less than 99% similarity across genomes) of 2332 (54.5%) genes **(Supplementary Figure 3a)**. The highest number of unique genes were present in strain MUM270 (410) and only 12 unique genes were found in strain WBG8287. Highest number of accessory genes were also identified for strain MUM270 (966) followed by MUF256 (942), MUF168 (855) and MUM475 (850). Phylogenetic distance based on pangenome analysis clustered strains MUF256 and MUM475 in one clade while strain MUF168 and MUM270 clustered together in one clade.
BUSCO analysis using bacillales (order-level) lineage dataset which contains 450 single-copy orthologs from 412 species showed the presence of all 450 complete single-copy genes in strain ATCC-12600. Strain MUF256 had 14 single-copy genes missing followed by MUF168 (12 genes), MUM270 (10 genes) and MUM475 (8 genes). Both MUF strains had seven common genes (*adk*, *infC*, *dnaK*, *rplU*, *bfmBAA_2*, *uppS* and *pdhD*) that were absent from their genomes while MUM strains had three common genes (*cmk*, *rpll* and *whiA*) that were absent. Also, 31 genes were fragmented for strain MUF256 followed by MUF168 (26 genes), MUM270 (26 genes) and MUM475 (18 genes). Single copy ortholog status of 450 genes in all four strains along with reference strains is represented in **Supplementary Figure 3b**.

Antimicrobial resistance gene and virulence factor identification

All four strains displayed varying degrees of resistance to the nine antibiotics tested **(Supplementary Table S4)** and showed sensitivity against most of the tested drugs. MUF168 and MUM475 tested resistant for ciprofloxacin and MUF256 showed resistance against erythromycin while strain MUM270 showed sensitivity to all nine antibiotics.

A total of 34 different antimicrobial resistance (AMR) genes conferring resistance to 23 antibiotics were predicted to be present in the four *S. aureus* strains **(Supplementary Table S5)**. Out of the 34 antibiotic resistance genes, MUF168 harbored 17 genes, MUF256 had 20 genes, MUM270 had 27 genes, MUM475 had 26 genes and 14 genes were common to all four strains **(Figure 2a)**. AMR genes unique to different strains included *fosB* (MUF168), *dfrG* (MUF256), *23S_C2220T*, *aadD1*, *ant(4')-Ib*, *bleO*, and *ermC* (MUM270) and *mphC*, *msrA*, and *sat4* (MUM475). AMR genes *blaI*, *blaR* and *blaZ* that confer resistance to beta-lactams were found in MUF256, MUM270 and MUM475 strains **(Figure 2b)**. Presence of antibiotic resistance genes in four strains were validated by single-plex PCR for a panel of three antibiotic resistance genes, *mecA*, *ant(4')-Ib* and *ermC* and the results were identical to the bioinformatic analysis **(Figure 2c)**.

A total of 76 virulence factors, divided into eleven broad categories were identified across four strains, with the maximum number of virulence genes identified for MUF256 (71) followed by MUM475 (69), MUF168 (67) and MUM270 (60) **(Supplementary Figure S4)**. A total of 13 toxin-coding genes were identified out of which 12 were present in MUF256, 10 in MUM475, 9 in MUM270 and 8 in MUF168 **(Supplementary Figure S4)**. One of the major *S. aureus* virulence factors, α-toxin, which is encoded by *hla* was present in all four strains along with three γ-hemolysins (*hlgA*, *hlgB* and *hlgC*) and δ-hemolysin (*hld*). Virulence genes from enterotoxin superfamily were present across four strains – for instance, *sea* and *seh* in MUF256 and MUM475, *seb* in MUF256, while staphylococcal enterotoxin-like (SELs) proteins *selk* and *selq* were present in all. Bicomponent pore-forming toxin lukF-PV, also known as leukocidin (luk) was identified in all four strains while lukS-PV was found in MUM270 only. Fibronectin-binding proteins *fnbA* and *fnbB* that play an important role in host cell attachment were found in all four strains while *clfA* and *clfB* involved in clumping were present only in strains MUF168 and MUM475.

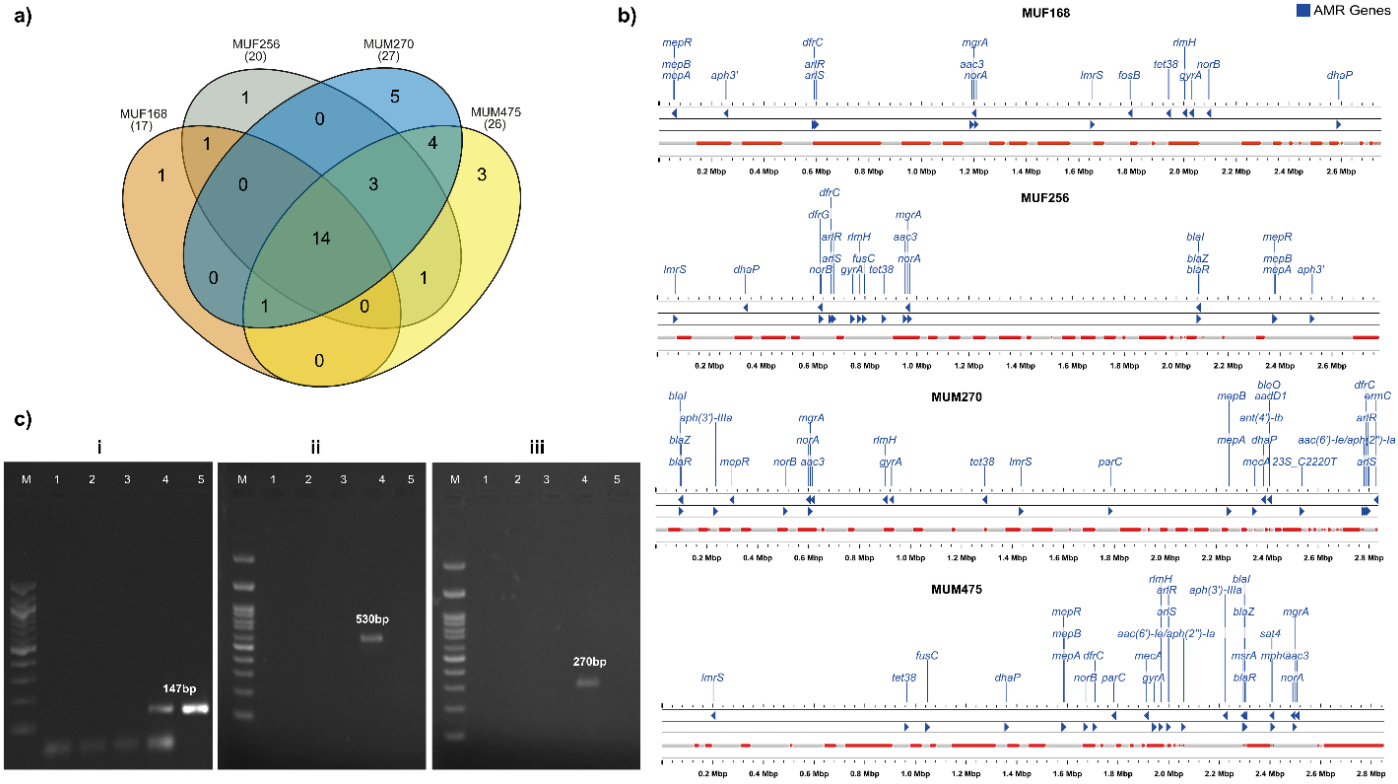


Figure 2: Antimicrobial resistance profiling. (a) Common and unique AMR genes shared across four S. aureus strains. (b) Genomic coordinates of identified AMR genes in all four strains. (c) Antibiotic resistance gene profiling for (i) *mecA*, (ii) *ant(4')-Ib* and (iii) *ermC*. Lane M: 100bp ladder, Lane 1: Negative control, Lane 2: MUF168, Lane 3: MUF256, Lane 4: MUM270 and Lane 5: MUF475.

The *S. aureus* gene regulatory network highlighted major virulence-associated regulatory systems (**Figure 3a)**. The genes in this network play an important role in *S. aureus* virulence by regulating major *S. aureus* toxins thereby overcoming host defense systems and increasing survival time. Among these, genes involved in *agr* quorum-sensing system (*agrB*, *agrC*, *agrA* and *hld*) were present in all the four strains but *agrD* (ribosomal peptide precursor of autoinducing peptide) was not found in any strains. Two-component system genes (*arlR*,

arlS, *srrA*, *srrB*, *saeR* and *saeS*) were present in the four strains while *saeP* and *saeQ* were absent from all four. All three major cytoplasmic *sarA*-family regulators (*sarA*, *mgrA* and *rot*) were found across four strains along with alternative sigma factors *sigB* and *sigH*. Blastp identity between *S. aureus* specific virulence factors (VFDB core protein dataset) and four strains is represented in **Figure 3b.** A total of 56 toxin genes were present in all the four strains **(Figure 3c)**, and a nucleotide blast of these 56 toxin genes among VFDB reference and four *S. aureus* strains showed a high similarity score **(Supplementary Figure S5)**.

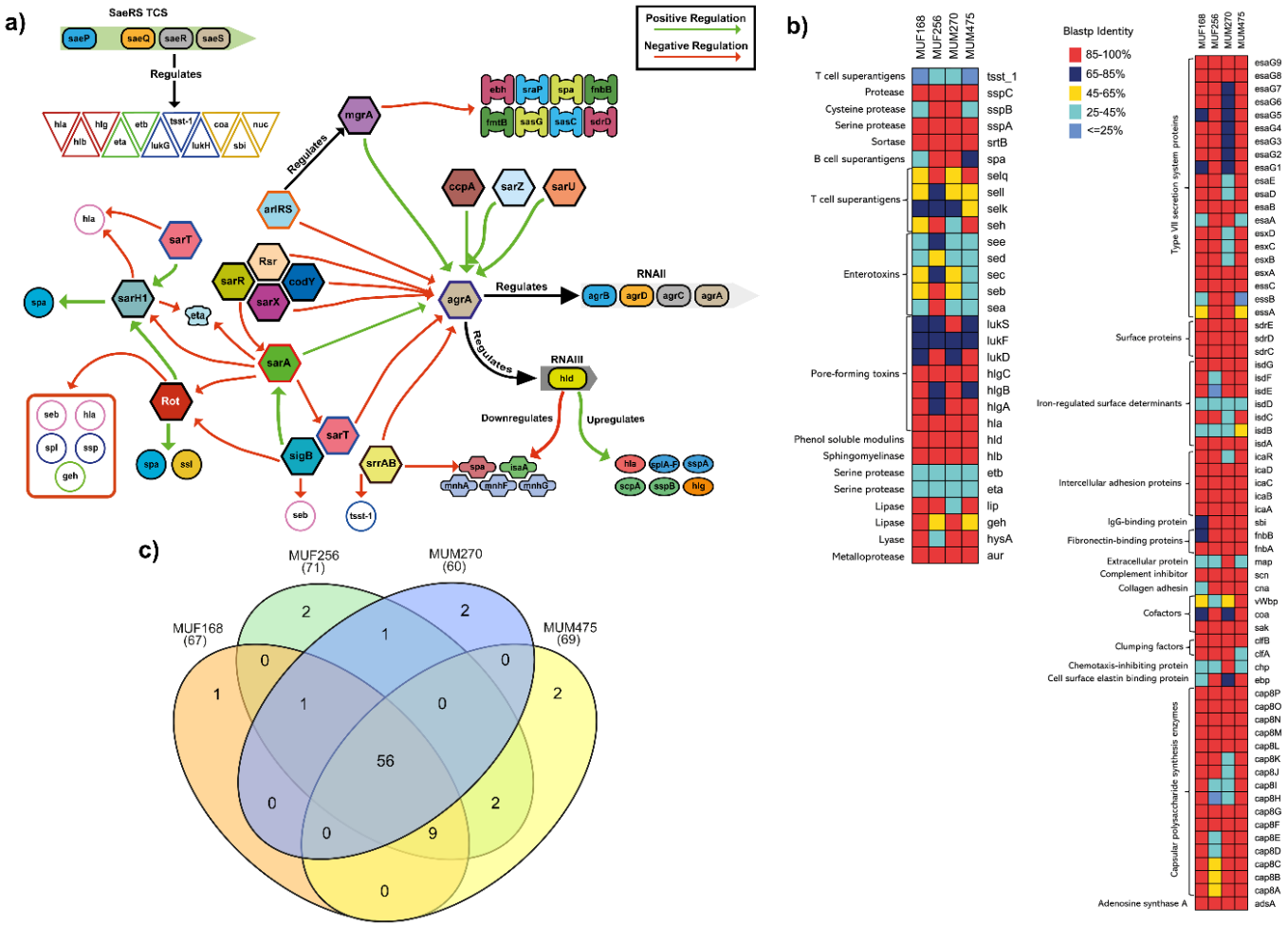


Figure 3: Virulence genes and regulation. (a) Represents virulence gene regulatory network of S. aureus. (b) Toxin protein identity (blastp) comparison against VFDB dataset. (c) Common and unique toxin genes shared across four strains.

Biofilm formation analysis

Biofilm production in bacterial strains were measured following crystal violet staining and monitoring the attachment of bacteria under sheared force. The bacterial strains were divided into three groups – high (>0.6), moderate (0.3-0.6) or low (<0.3) biofilm producers based on their absorbance values. Among the four strains of *S. aureus*, MUM475 was found to be a high biofilm producer, MUF168 and MUF256 as moderate biofilm producers and MUM270 as low biofilm forming organism **(Figure 4a)**. Thick and dense layer of biofilm was observed in MUM475 compared to other stains of *S. aureus* **(Figure 4b)**.

Drug target and variant analysis

A total of 16 drug targets were found in the DrugBank database for *S. aureus* out of which 15 targets showed significant sequence similarity to sequence contigs of four strains. Ten drug targets were found to be common among the four strains while gene coding for kanamycin nucleotidyltransferase (*knt*) was present in strain MUM270 only **(Supplementary Table S6)**.

All 15 drug targets across four strains displayed more variation at the nucleotide level compared to the protein level **(Supplementary Figure S6)**. Strain MUM270 contained 39, the highest number of nucleotide variations for beta-lactamase. Predicted deleterious amino acid changes were high for *gyrA* while no amino acid changes were predicted as deleterious for penicillin-binding proteins (*pbp2a*, *pbp3* and *pbp4*), *murB*, *mecA*, *ileS* and *topA*. No variations were predicted at the protein level for drug targets, *fabI*, *knt*, *pbp2* and *trxB*. Predicted single amino acid variation in *gyrB* for strains MUF168, MUM270 and MUM475 was identified to have a deleterious effect **(Table 1)**.

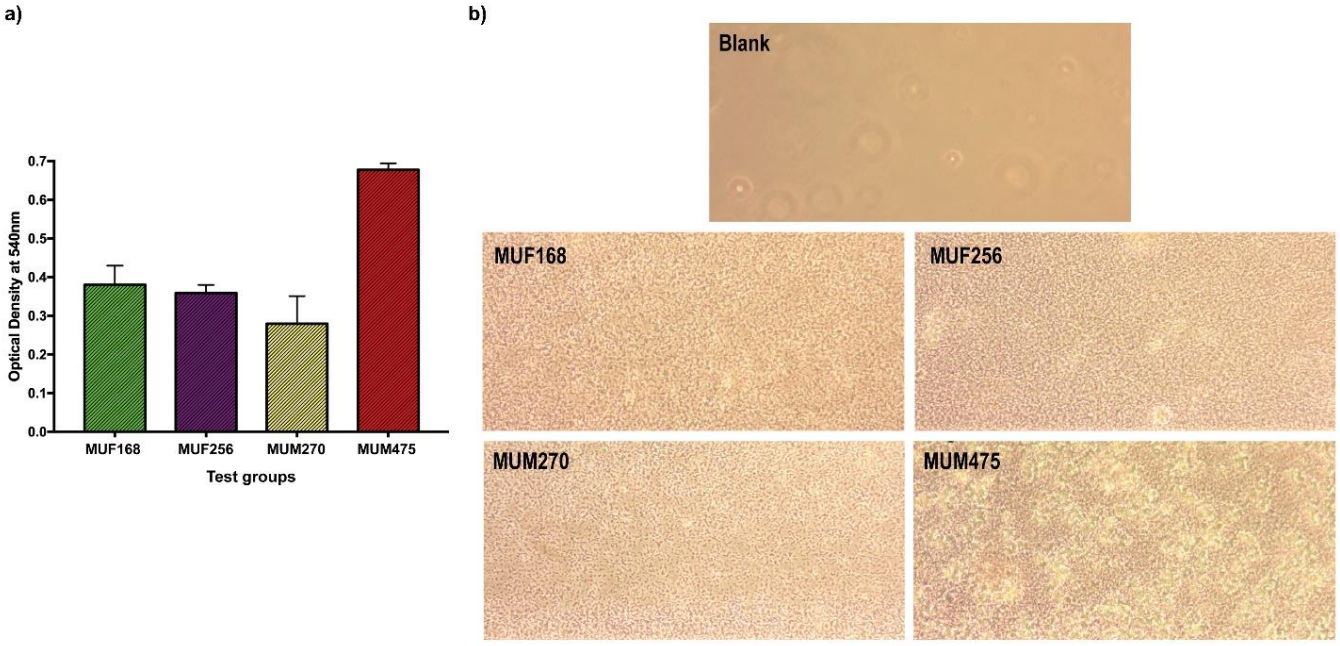


Figure 4: Biofilm formation assay. (a) Detection of biofilm produced by MUF168, MUF256, MUM270 and MUM475 strains in static condition using crystal violet assay. (b) Detection of biofilm in dynamic condition using bioflux.

Data availability

The whole genome sequences of four *Staphylococcus aureus* strains utilized in this study are available at NCBI Genome database with accession numbers AZQR00000000 (MUF168), AZSE00000000 (MUF256), AZSF00000000 (MUM270), and AZSG00000000 (MUM475).

Table 1: Variants in approved drug targets at nucleotide and protein levels.

		Number of variants							
		Genomic level				*Protein level, number of variants (neutral/deleterious)*			
Sl. no	*Target gene*	*MUF168*	*MUF256*	*MUM270*	*MUM475*	*MUF168*	*MUF256*	*MUM270*	*MUM475*
1	*blaZ*	—	13	39	15	—	3 (3/0)	11 (10/1)	5 (5/0)
2	*gyrA*	15	10	16	8	6 (4/2)	2 (1/1)	3 (1/2)	3 (1/2)
3	*gyrB*	13	11	11	11	1 (0/1)	NF	1 (0/1)	1 (0/1)
4	*topA*	11	11	12	11	1 (1/0)	1 (1/0)	1 (1/0)	1 (1/0)
5	*parC*	13	13	12	12	4 (3/1)	4 (3/1)	4 (3/1)	3 (2/1)
6	*fabI*	12	6	13	6	NF	NF	NF	NF
7	*ileS*	22	19	26	23	1 (1/0)	NF	1 (1/0)	1 (1/0)
8	*knt*	—	—	1	—	—	—	NF	—
9	*mecA*	—	—	NF	1	—	—	NF	1 (1/0)
10	*pbp2a*	—	—	3	4	—	—	2 (2/0)	3 (3/0)
11	*pbp2*	—	—	NF	NF	—	—	NF	NF
12	*pbp3*	11	6	27	6	3 (3/0)	3 (3/0)	5 (5/0)	3 (3/0)
13	*pbp4*	10	10	14	12	3 (3/0)	4 (4/0)	5 (5/0)	4 (4/0)
14	*trxB*	4	2	7	2	NF	NF	NF	NF
15	*murB*	4	NF	8	NF	1 (1/0)	NF	NF	NF

NF, not found.

Discussion

Reports suggest that the lifetime incidence of foot ulcers in diabetic patients to be as high as 25% while diabetic individuals are at higher risk of developing foot ulcers and lower extremity amputation compared to non-diabetic individuals (Lavery et al, 2006; Singh et al, 2005). The current infection management strategies adopted for diabetic foot ulcers are often insufficient, resulting in delayed healing and ultimately lower limb amputation (Leung, 2007). Infections in diabetic foot ulcers are among the major comorbid conditions leading to higher mortality in diabetic individuals. The pathogenesis of foot ulceration is complex with varied clinical presentation, and hence management requires early expert assessment (Jeffcoate and Harding, 2003). Among all the pathogens causing foot ulcer infections, *Staphylococcus aureus* is considered a major pathogen due to its high virulence capabilities (Murali et al, 2014a).

Antimicrobial resistance gene profiling of the four strains included in the present study revealed that none of the strains carried any resistance genes against vancomycin, though several resistance genes (*parC, norA, norB, gyrA, arlR, arlS* and *mgrA*) for fluoroquinolones (Redgrave et al, 2014) were present across all four strains. Fosfomycin resistance protein, encoded by gene *fosB* (Cao et al, 2001) was found specifically in strain MUF168 while its prevalence in other *S. aureus* genomes (available in NCBI database) was found to be approximately 70%. Gene *dfrG* (coding for dihydrofolate reductase) that confers resistance to trimethoprime (Nurjadi et al, 2014) was found in strain MUF256 only, while the prevalence of this gene in other strains (NCBI genomes) was found to be only 9%. Macrolide resistance genes (*msrA, mphC* and *sat4*) (Otarigho and Falade, 2018) were present in MUM475 and gene *ermC* in MUM270, indicating high resistance to macrolide, lincosamide and streptogramin class antibiotics. In all three strains, MUF256, MUM270 and MUM475, genes (*mecA, blaI, blaR* and *blaZ*) coding for resistance against beta-lactams (Hackbarth and Chambers, 1993) were present while in strain MUF168 none of these genes were present. Gene *bleO* that confers resistance to

glycopeptide bleomycin (Scholar, 2007) was found in strain MUM270 only. All four strains encoded *mgrA* gene, a major global regulator in *S. aureus* which can act as a repressor or activator of multiple genes involved in antibiotic resistance and autolysis. MgrA is an important member of the *marR/sarA* protein family and acts as a transcriptional regulator by directly binding to the promoter sites of certain target genes (Luong et al, 2006) and plays a significant role in biofilm production and in the regulation of resistance and virulence genes.

Staphylococcus aureus is an opportunistic pathogen with a complex regulatory network to control its virulence potency. Using its virulence network, *S. aureus* can manipulate the host's innate and adaptive immune response by various mechanisms including inducing cytolysis, triggering a major inflammatory response, hijacking host's coagulation system, disrupting cell membrane and producing a host of exoenzymes that are involved in proteolytic activity (Jenul and Horswill, 2019; Tam and Torres, 2019; Zhang et al, 2017). Tam and Torres (2019) classified *S. aureus* exotoxins that cause immense damage to host cells further into three broad categories namely, cytotoxins, superantigens and cytotoxic enzymes. In the current study, major exotoxin genes (*hla, hlb, hld, hlgA, hlgB, hlgC, lukS, lukF, sea, seb, seh* and *spa*) were present across all the four strains investigated. Pore-forming cytotoxins (*hla, hlgA, hlgB, hlgC, lukS,* and *lukF*) and cytotoxic enzymes (*hlb*) act on host cell membrane, resulting in cell lysis and inflammation. Superantigens (*sea, seb, seh, sek* and *seq*) trigger B and T cell proliferation by mediating cytokine production and are commonly resistant to proteolysis, heat and desiccation. Interestingly, cytotoxic enzyme *hlb* and superantigen *seb* were only found in the strain MUF256. None of the exfoliative toxins were found in any of the strains studied but exoenzymes such as proteases (*aur, sspA, sspB* and *sspC*), lipase (*geh and lip*), lyase (*hysA*) and other associated cofactors (*coa, vWbp* and *sak*) were present across the four strains (Jenul and Horswill, 2019; Tam and Torres, 2019).

Based on *in-silico* analysis, we identified nucleotide and protein variants present in approved drug targets in the four strains studied and predicted whether those variants have a neutral or deleterious effect on their fitness. The maximum number of nucleotide variants were obtained for beta-lactamase (*blaZ*) and Isoleucine-tRNA ligase (*ileS*). Surprisingly protein variants for *ileS* were low while *blaZ* in strain MUM270 had the highest (11) number of variants at the protein level. Recent studies in *Mycobacterium tuberculosis* have identified molecular targets that contribute to drug resistance mechanisms (Dookie et al, 2018; Hameed et al, 2018). Similarly, rare genetic variants in human drug-related genes are known to contribute to complex diseases (Schärfe et al, 2017; Verma et al, 2018). Sequence variations in drug targets may lead to sub-optimal binding of drugs to their targets and therefore might contribute to the development of antibiotic resistance in bacterial communities.

In this comparative genome analysis study, we compared four *S. aureus* strains for carriage of genes involved in conferring multidrug resistance and virulence potential, biofilm forming potential and variants of potential drug targets that can contribute to antibiotic resistance development. We found that strains MUM270 and MUM475 coded for a greater number of antibiotic resistance genes (ARGs) compared to MUF strains based on our *in-silico* analysis. Strain MUM270 was sensitive to all nine antibiotics tested suggesting that the presence of ARGs alone does not necessarily contribute to increased resistance to antibiotics. Strain MUF256 coded for highest number of virulence genes followed by strains MUM475, MUF168 and MUM270, while strain MUM475 was found to be a high biofilm producer. We also found that both the MUM strains had a

higher number of deleterious variants than MUF strains. Further studies using omics approach can provide critical clues on key regulators of microbial virulence and factors that contribute to antimicrobial resistance in clinically relevant pathogenic isolates.

Acknowledgements

Authors thank Manipal Academy of Higher Education, Technology Information Forecasting Assessment Council - Centre of Relevance and Excellence in Pharmacogenomics (TIFAC-CORE), DST-FIST and DBT BUILDER – Interdisciplinary Life Science Programme for Advance Research and Education (DB-ILSPARE) for the support and Lady TATA Memorial Trust for the fellowship.

Funding

This research was financially supported by Department of Science and Technology - Science for Equity, Empowerment and Development division (DST-SEED) (SEED/WS/2019/57S(G)) and Indo-German Science and Technology Centre (IGSTC).

Author's Contribution

Conceptualization, KS, BP and TSM; methodology, all the authors; validation and experiments, PS; data analysis, AST; writing - original draft, AST, PS, and BP; review and editing, KS, BP, TSM and AB; supervision, KS and AB.

Disclosure Statement

No potential conflict of interest was reported by the authors.

References:

Alcock BP, Raphenya AR, Lau TTY, et al. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic Acids Res 2020;48(D1):D517–D525.

Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. J Mol Biol 1990;215(3):403–410.

Aziz RK, Bartels D, Best A, et al. The RAST server: rapid annotations using subsystems technology. BMC Genomics 2008;9:1–15.

Bowling FL, Jude EB and Boulton AJM. MRSA and diabetic foot wounds: contaminating or infecting organisms? Curr Diab Rep 2009;9(6):440–444.

Brennan MB, Hess TM, Bartle B, et al. Diabetic foot ulcer severity predicts mortality among veterans with type 2 diabetes. J Diabetes Complications 2017;31(3):556–561.

Cao M, Bernat BA, Wang Z, et al. FosB, a cysteine-dependent fosfomycin resistance protein under the control of σw, an extracytoplasmic-function σ factor in *Bacillus subtilis*. J Bacteriol 2001;183(7):2380–2383.

Chen L, Zheng D, Liu B, et al. VFDB 2016: hierarchical and refined dataset for big data analysis - 10 years on. Nucleic Acids Res 2016;44(D1):D694–D697.

Choi Y and Chan AP. PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics 2015;31(16):2745–2747.

Coll F, McNerney R, Preston MD, et al. Rapid determination of anti-tuberculosis drug resistance from whole-genome sequences. Genome Med 2015;7(1):1–10.

Dookie N, Rambaran S, Padayatchi N, et al. Evolution of drug resistance in *Mycobacterium tuberculosis*: a review on the molecular determinants of resistance and implications for personalized care. J Antimicrob Chemother 2018;73(5):1138–1151.

Doster E, Lakin SM, Dean CJ, et al. MEGARes 2.0: a database for classification of antimicrobial drug, biocide and metal resistance determinants in metagenomic sequence data. Nucleic Acids Res 2020;48(D1):D561–D569.

Dunyach-Remy C, Essebe CN, Sotto A, et al. *Staphylococcus aureus* toxins and diabetic foot ulcers: role in pathogenesis and interest in diagnosis. Toxins (Basel) 2016;8(7):1–20.

Feldgarden M, Brover V, Haft DH, et al. Validating the AMRFINder tool and resistance gene database by using antimicrobial resistance genotype-phenotype correlations in a collection of isolates. Antimicrob Agents Chemother 2019;63(11):1–19; doi: 10.1128/AAC.00483-19.

Green MR and Sambrook J. Isolation of high-molecular-weight dna using organic solvents. Cold Spring Harb Protoc 2017;2017(4):pdb.prot093450.

Gurevich A, Saveliev V, Vyahhi N, et al. QUAST: quality assessment tool for genome assemblies. Bioinformatics 2013;29(8):1072–1075.

Hackbarth CJ and Chambers HF. BlaI and blar1 regulate β-lactamase and pbp 2a production in methicillin-resistant *Staphylococcus aureus*. Antimicrob Agents Chemother 1993;37(5):1144–1149.

Hameed HMA, Islam MM, Chhotaray C, et al. Molecular targets related drug resistance mechanisms in mdr-, xdr-, and tdr- *Mycobacterium tuberculosis* strains. Front Cell Infect Microbiol 2018;8:114.

Jeffcoate WJ and Harding KG. Diabetic foot ulcers. 2003;361(9368):1545–1551.

Heberle H, Meirelles G V, Silva FR, et al. InteractiVenn: A web-based tool for the analysis of sets through venn diagrams. BMC Bioinformatics 2015;16:1–7.

Holden MTG, Feil EJ, Lindsay JA, et al. Complete genomes of two clinical *Staphylococcus aureus* strains: evidence for the rapid evolution of virulence and drug resistance. Proc Natl Acad Sci 2004;101(26):9786–9791.

Huerta-Cepas J, Szklarczyk D, Heller D, et al. EggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res 2019;47(D1):D309–D314.

CLSI. Clinical and Laboratory Standards Institute. Performance standards for antimicrobial susceptibility testing: twenty-first informational supplement M100-S21. CLSI, Wayne, PA, USA, 2011.

Jenul C and Horswill AR. Regulation of *Staphylococcus aureus* virulence. Microbiol Spectr 2019;669–686.

Lavery LA, Armstrong DG, Wunderlich RP, et al. Risk factors for foot infections in individuals with diabetes. Diabetes Care 2006;29(6):1288–1293.

Leung P. Diabetic foot ulcers - a comprehensive review. Surgeon 2007;5(4):219–231.

Luong TT, Dunman PM, Murphy E, et al. Transcription profiling of the mgra regulon in *Staphylococcus aureus*. J Bacteriol 2006;188(5):1899–1910.

Madeira F, Park YM, Lee J, et al. The EMBL-EBI search and sequence analysis tools apis in 2019. Nucleic Acids Res 2019;47(W1):W636–W641.

Manni M, Berkeley MR, Seppey M, et al. BUSCO update: novel and streamlined workflows along with broader and deeper phylogenetic coverage for scoring of eukaryotic, prokaryotic, and viral genomes. Mol Biol Evol 2021;38(10):4647–4654.

McKenna A, Hanna M, Banks E, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 2010;20(9):1297–1303.

Meier-Kolthoff JP, Auch AF, Klenk HP, et al. Genome Sequence-Based Species Delimitation with Confidence Intervals and Improved Distance Functions. BMC Bioinformatics 2013;14.

Murali TS, Kavitha S, Spoorthi J, et al. Characteristics of microbial drug resistance and its correlates in chronic diabetic foot ulcer infections. J Med Microbiol 2014a;63:1377–1385.

Murali TS, Paul B, Parikh H, et al. Genome sequences of four clinical *Staphylococcus aureus* strains with diverse drug resistance profiles isolated from diabetic foot ulcers. Genome Announc 2014b;2(2):1–2.

Nurjadi D, Olalekan AO, Layer F, et al. Emergence of trimethoprim resistance gene dfrG in *Staphylococcus aureus* causing human infection and colonization in sub-saharan africa and its import to Europe. J Antimicrob Chemother 2014;69(9):2361–2368.

Otarigho B and Falade MO. Analysis of antibiotics resistant genes in different strains of *Staphylococcus aureus*. Bioinformation 2018;14(03):113–122.

Page AJ, Cummins CA, Hunt M, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 2015;31(22):3691–3693.

Percival SL, Malone M, Mayer D, et al. Role of anaerobes in polymicrobial communities and biofilms complicating diabetic foot ulcers. Int Wound J 2018;15(5):776–782.

Petkau A, Stuart-Edwards M, Stothard P, et al. Interactive microbial genome visualization with GView. Bioinformatics 2010;26(24):3125–3126.

Pouget C, Dunyach-Remy C, Pantel A, et al. New adapted in vitro technology to evaluate biofilm formation and antibiotic activity using live imaging under flow conditions. Diagnostics 2021;11(10).

Prasad ASB, Shruptha P, Prabhu V, et al. *Pseudomonas aeruginosa* virulence proteins pseudolysin and protease iv impede cutaneous wound healing. Lab Investig 2020;100(12):1532–1550.

Price J, Gordon NC, Crook D, et al. The usefulness of whole genome sequencing in the management of *Staphylococcus aureus* infections. Clin Microbiol Infect 2013;19(9):784–789.

Punina N V., Makridakis NM, Remnev MA, et al. Whole-genome sequencing targets drug-resistant bacterial infections. Hum Genomics 2015;9:19.

Redgrave LS, Sutton SB, Webber MA, et al. Fluoroquinolone resistance: mechanisms, impact on bacteria, and role in evolutionary success. Trends Microbiol 2014;22(8):438–445.

Schärfe CPI, Tremmel R, Schwab M, et al. Genetic variation in human drug-related genes. Genome Med 2017;9(1):1–15.

Scholar E. Bleomycin. In: XPharm: The Comprehensive Pharmacology Reference. Elsevier, 2007; pp. 1–6.

Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics 2014;30(14):2068–2069.

Seemann T. Abricate: mass screening of contigs for antimicrobial and virulence genes; 2021. Available from: https://github.com/tseemann/abricate [Last accessed: 10/26/2021].

Shettigar K, Jain S, Bhat D V., et al. Virulence determinants in clinical *Staphylococcus aureus* from monomicrobial and polymicrobial infections of diabetic foot ulcers. J Med Microbiol 2016;65(12):1392–1404.

Sievers F, Wilm A, Dineen D, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 2011;7(539).

Singh N, Armstrong DG and Lipsky BA. Preventing foot ulcers in patients with diabetes. Jama 2005;293(2):217–228.

Stothard P and Wishart DS. Circular genome visualization and exploration using CGView. Bioinformatics 2005;21(4):537–539.

Tam K and Torres VJ. *Staphylococcus aureus* secreted toxins and extracellular enzymes. Microbiol Spectr 2019;640–668.

Verma SS, Josyula N, Verma A, et al. Rare variants in drug target genes contributing to complex diseases, Phenome-Wide. Sci Rep 2018;8(1):1–16.

Wishart DS, Feunang YD, Guo AC, et al. DrugBank 5.0: a major update to the drugbank database for 2018. Nucleic Acids Res 2018;46(D1):D1074–D1082.

Zhang X, Hu X and Rao X. Apoptosis induced by *Staphylococcus aureus* toxins. Microbiol Res 2017;205:19–24.

Supplementary Data

Table S1: Genome assembly and annotation features of all four *S. aureus* strains.

Strain	MUF168	MUF256	MUM270	MUM475
Total length (bp)	2,752,244	2,783,966	2,834,436	2,852,107
No. of contigs	59	58	125	74
N50 (bp)	96843	87410	52369	124901
GC content (%)	32.8	32.7	32.8	32.7
Genes	2770	2825	2991	2852
Proteins	2433	2433	2728	2620
rRNA	13	15	19	17
tRNA	60	59	59	59

Table S2: Nucleotide sequence identity comparison using GGDC and Blast.

		Formula 1 (D0)	Formula 2 (D4)	Formula 3 (D6)	Blast (megablast)	
Reference	**Query**	**Prob. DDH >=70%**	**Prob. DDH >=70%**	**Prob. DDH >=70%**	**Query Coverage (%)**	**Percentage Identity (%)**
MUF168	MUF256	98.65	95.76	99.88	95	99.42
MUF168	MUM270	98.63	92.17	99.84	95	98.97
MUF168	MUM475	98.38	95.61	99.85	95	99.40
MUF256	MUM270	98.06	92.38	99.74	93	99.09
MUF256	MUM475	99.4	98.14	99.97	99	99.91
MUM270	MUM475	98.22	91.87	99.77	94	99.09

Table S3: Coding sequences in unique regions of CDS blast.

Locus Tag	Gene	Length	Protein Name
SAOUHSC_00001	*dnaA*	453	chromosomal replication initiation protein
SAOUHSC_00009		428	seryl-tRNA synthetase
SAOUHSC_00352		103	integrase-like protein
SAOUHSC_01523		527	SLT orf 527-like protein
SAOUHSC_01524		274	holin-like protein
SAOUHSC_01528		151	bacteriophage L54aIg-like domain-containing protein
SAOUHSC_01529		213	major tail protein
SAOUHSC_01532		110	SLT orf 110-like protein
SAOUHSC_01542		455	SNF2 family protein
SAOUHSC_01543		96	phi-like protein

SAOUHSC_01563		650	phage encoded DNA polymerase I
SAOUHSC_01566		388	phi APSE P51-like protein
SAOUHSC_01615	*recN*	559	DNA repair protein RecN
SAOUHSC_01649		336	rhomboid family protein
SAOUHSC_01759	*mreC*	280	rod shape-determining protein MreC
SAOUHSC_01952		997	lantibiotic epidermin biosynthesis protein EpiB
SAOUHSC_01972		320	protein export protein PrsA
SAOUHSC_01993		400	transposase, putative
SAOUHSC_02392		67	truncated resolvase
SAOUHSC_02410		400	transposase, putative

Table S4: Antibiotic resistance profile of the four S. aureus strains against 9 different antibiotics based on Minimum Inhibitory Concentration results.

Antibiotics tested	MUF168	MUF256	MUM270	MUM475
Amikacin (AK)	S	S	S	S
Amoxicillin Clavulanate (AMC)	S	S	S	I
Ampicillin (AMP)	S	S	S	S
Chloramphenicol (C)	S	S	S	S
Ciprofloxacin (CIP)	R	S	S	R
Erythromycin (E)	S	R	S	S
Linezolid (LZ)	S	S	S	S
Teicoplanin (TEI)	S	S	S	S
Vancomycin (VA)	S	S	S	S

S = Sensitive, I = Intermediate, R = Resistant.

Table S5: Antimicrobial resistance gene presence/absence summary for all four strains.

Sl. No	AMR gene	Length (bp)	Antibiotic	Present in strains
1	*aac(6')-Ie/aph(2'')-Ia*	1440	Aminoglycoside	MUM270, MUM475
2	*aac3*	444		MUF168, MUF256, MUM270, MUM475
3	*aadD1*	962		MUM270
4	*ant(4')-Ib*	762		MUM270
5	*aph(3')-IIIa*	795		MUM475, MUM270
6	*aph3'*	801		MUF168, MUF256, MUM270, MUM475
7	*lmrS*	1443	Aminoglycoside; Diaminopyrimidine; Macrolide; Oxazolidinone; Phenicol	MUF168, MUF256, MUM270, MUM475
8	*blaI*	381	Beta-lactam	MUF256, MUM270, MUM475
9	*blaR*	1758		MUF256, MUM270, MUM475

10	*blaZ*	888		MUF256, MUM270, MUM475
11	*mecA*	2007		MUM270, MUM475
12	*bleO*	405	Bleomycin	MUM270
13	*gyrA*	2661	Fluoroquinolone	MUF168, MUF256, MUM270, MUM475
14	*norA*	1167		MUF168, MUF256, MUM270, MUM475
15	*norB*	1392		MUF168, MUF256, MUM270, MUM475
16	*parC*	2400		MUM270, MUM475
17	*arlR*	660	Fluoroquinolone; Acridine Dye	MUF168, MUF256, MUM270, MUM475
18	*arlS*	1356		MUF168, MUF256, MUM270, MUM475
19	*mgrA*	444	Fluoroquinolone; Penam; Peptide; Tetracycline; Acridine Dye; Cephalosporin	MUF168, MUF256, MUM270, MUM475
20	*fosB*	420	Fosfomycin	MUF168
21	*fusC*	639	Fusidic Acid	MUF256, MUM475
22	*mepA*	1356	Glycycline; Tetracycline	MUF168, MUF256, MUM270, MUM475
23	*mepR*	420		MUF168, MUF256, MUM270, MUM475
24	*msrA*	1467	Lincosamide; Macrolide; Oxazolidinone; Phenicol; Pleuromutilin; Streptogramin; Tetracycline	MUM475
25	*ermC*	735	Lincosamide; Macrolide; Streptogramin	MUM270
26	*rlmH*	480		MUF168, MUF256, MUM270, MUM475
27	*mphC*	900	Macrolide	MUM475
28	*mepB*	441	Multi-Drug MATE Efflux Pump	MUF168, MUF256, MUM270, MUM475
29	*23S_C2220T*	2926	Oxazolidinone	MUM270
30	*dhaP*	1188	Phenicol	MUF168, MUF256, MUM270, MUM475
31	*sat4*	543	Streptothricin; Nucleoside	MUM475
32	*tet38*	1353	Tetracycline	MUF168, MUF256, MUM270, MUM475
33	*dfrC*	486	Trimethoprim; Diaminopyrimidine	MUM270, MUF168, MUF256, MUM475
34	*dfrG*	498		MUF256

Table S6: Sequence similarity and number of raw reads aligned between approved drug targets and *S. aureus* strains.

Drug Targets Length (bp/aa)	Gene Symbol	Drugs	Percentage Identity / Number of Reads Aligned			
			MUF168	MUF256	MUM270	MUM475
Beta-lactamase (846/281)	*blaZ*	Sulbactam, Clavulanate, Imipenem	-	**98 / 292**	**95 / 497**	**98 / 814**
DNA gyrase subunit A (2667/889)	*gyrA*	Ciprofloxacin	**99 / 2254**	**99 / 2501**	**99 / 2495**	**99 / 1623**
DNA gyrase subunit B (1935/644)	*gyrB*	Novobiocin	**99 / 1634**	**99 / 1704**	**99 / 1787**	**99 / 1191**
DNA topoisomerase 1 (2067/689)	*topA*	Pefloxacin, Novobiocin	**99 / 1226**	**99 / 1451**	**99 / 1485**	**99 / 903**
DNA topoisomerase 4 subunit A (2400/800)	*parC*	Ciprofloxacin	**99 / 1432**	**99 / 1534**	**99 / 1641**	**99 / 1007**
Enoyl-[acyl-carrier-protein] reductase [nadph] fabi (771/256)	*fabI*	Triclosan, Triclocarban	**98 / 444**	**99 / 549**	**98 / 529**	**99 / 340**
Hth-type transcriptional regulator qacr (567/188)	*qacR*	Proflavine	-	-	-	-
Isoleucine-tRNA ligase (2754/917)	*ileS*	Mupirocin	**99 / 1713**	**99 / 2094**	**99 / 2065**	**99 / 1353**
Kanamycin nucleotidyltransferase (708/253)	*knt*	Kanamycin	-	-	**99 / 587**	-
mecA (2007/668)	*mecA*	Phenoxymethylpenicillin, Meticillin, Ceftizoxime, Ceftobiprole	-	-	**100 / 1463**	**99 / 1030**
Penicillin binding protein 2a (1458/486)	*pbp2a*	Cefpiramide, Cyclacillin, Cefmetazole, Ticarcillin	-	-	**99 / 1064**	**99 / 764**
Penicillin-binding protein 2 (285/95)	*pbp2*	Oxacillin	-	-	**99 / 252**	**100 / 167**
Penicillin-binding protein 3 (2076/691)	*pbp3*	Benzylpenicillin	**99 / 1315**	**99 / 1457**	**99 / 1503**	**99 / 912**
Penicillin-binding protein 4 (1293/431)	*pbp4*	Doripenem	**99 / 885**	**99 / 871**	**99 / 973**	**99 / 629**
Thioredoxin reductase (936/311)	*trxB*	Azelaic Acid	**99 / 609**	**99 / 650**	**99 / 762**	**99 / 447**
Udp-n-acetylenolpyruvoylglucosamine reductase (924/307)	*murB*	Flavin adenine dinucleotide	**99 / 629**	**99 / 641**	**100 / 749**	**100 / 471**

Supplementary Figures

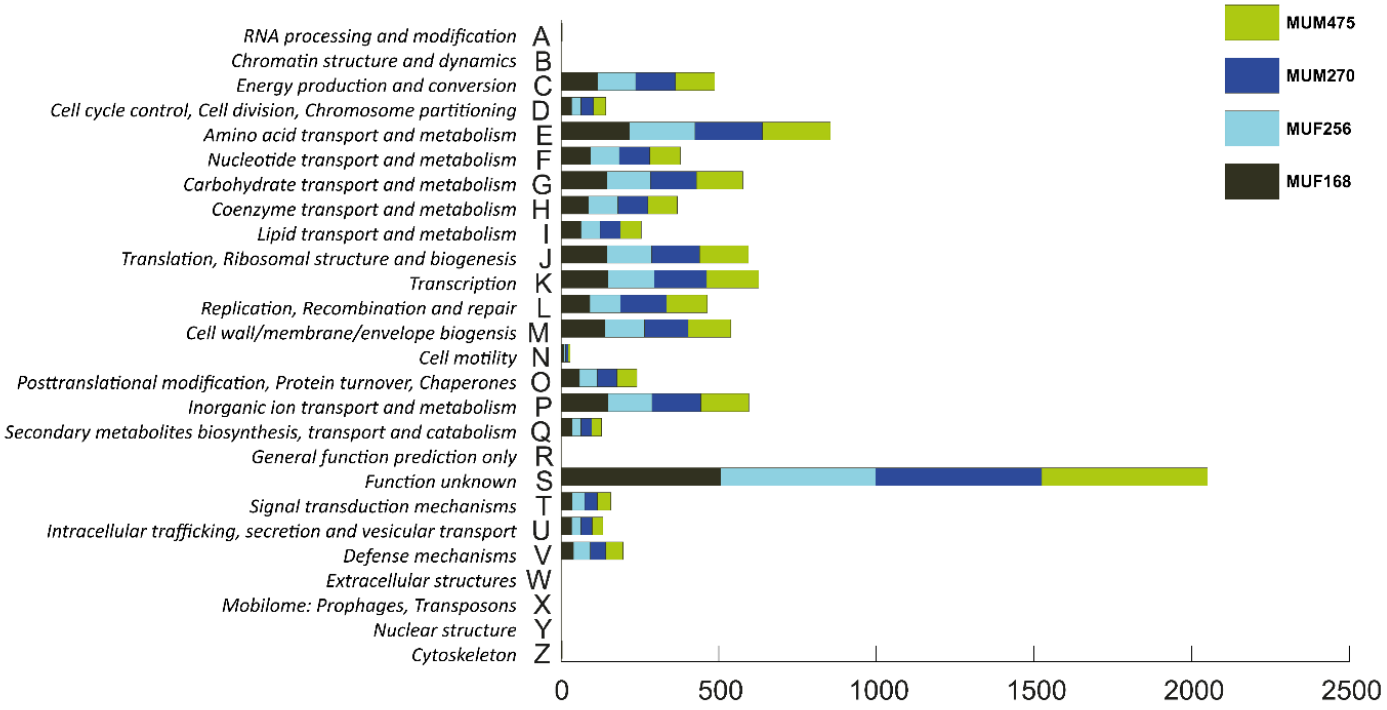


Figure S1: Clusters of Orthologous Genes (COGs) database categories distribution across four strains.

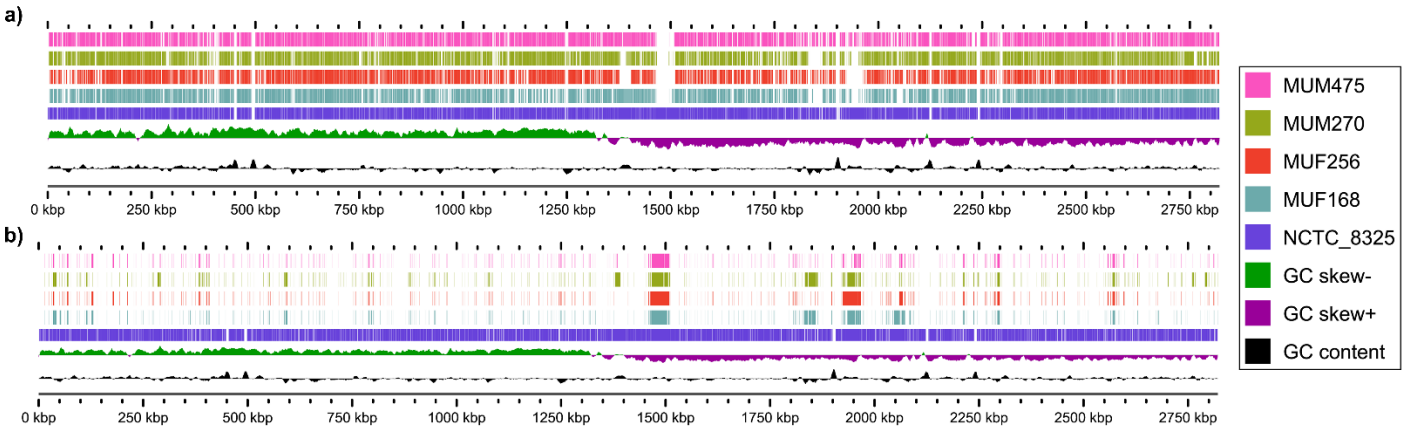


Figure S2: Blast sequence comparison. (a) Highly conserved regions across four strains compared to reference genome (strain NCTC-8325). (b) Unique regions present across four strains compared to strain NCTC-8325.

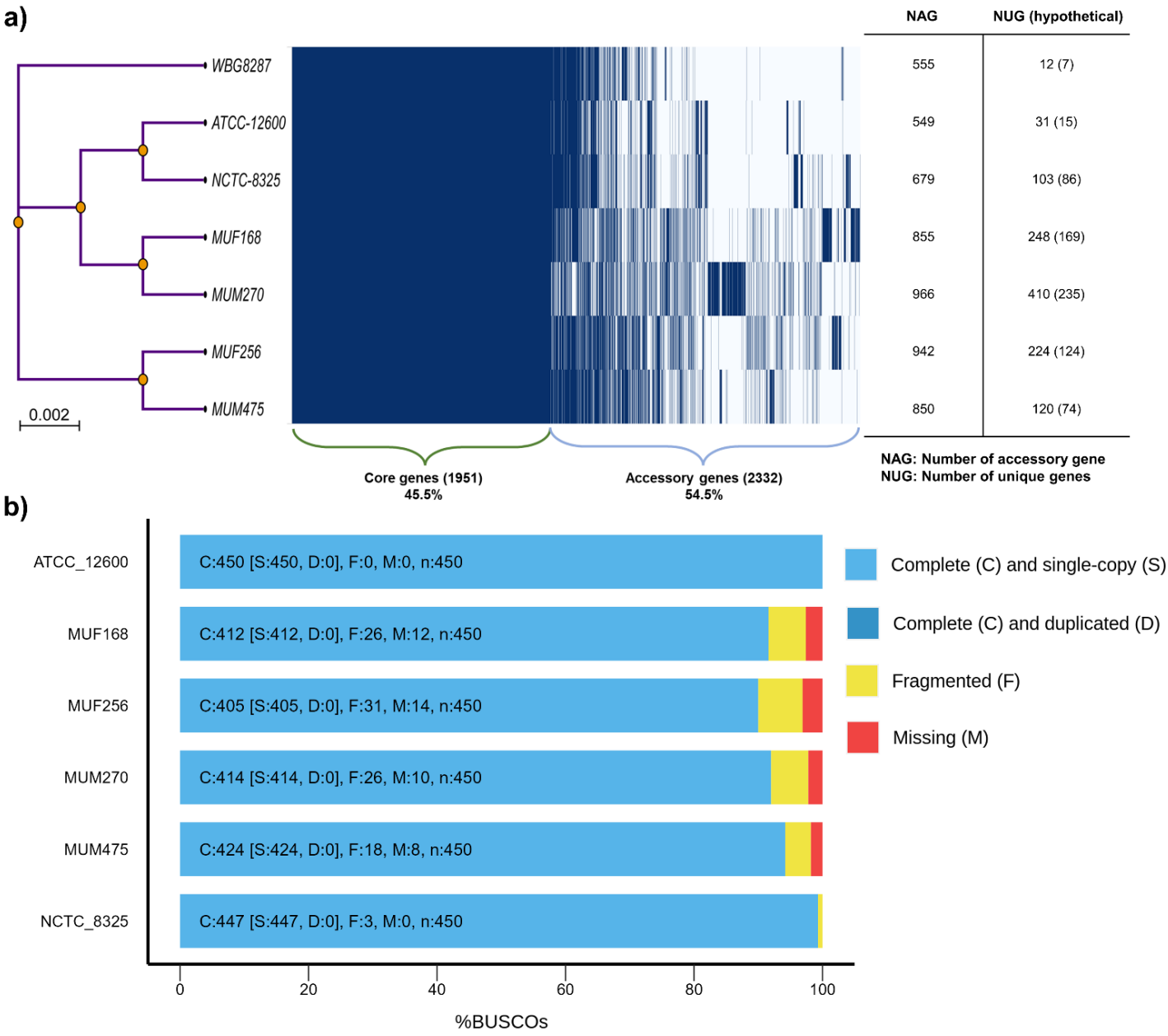


Figure S3: Pangenome analysis. (a) Roary pangenome analysis highlighting core, accessory and unique gene abundance across four strains compared to three reference strains (NCTC-8325, ATCC-12600 and WBG8287). (b) Single copy orthology status of 450 genes across tested and reference strains to visualize gene conservation.

Category	Gene	Product	Function
Adherence	*clfA, clfB*	Clumping factor	fibrinogen binding
	Can	collagen adhesin precursor	collagen binding
	ebp	cell surface elastin binding protein	elastin binding
	fnbA, fnbB	fibronectin-binding protein	host cell attachment
	icaA, icaB, icaC, icaD, icaR	intercellular adhesion protein	intercellular binding
	map	extracellular proteins Map	extracellular binding
	sdrD, sdrE	Ser-Asp rich fibrinogen-binding bone sialoprotein-binding protein	attachment
ATP-binding	*clpC*	endopeptidase Clp ATP-binding chain C	biofilm formation
Carbohydrate binding	*hysA*	hyaluronate lyase precursor	
Coagulase	*coa*	staphylocoagulase precursor	ability to coagulate mammalian plasma
Exoenzyme	*adsA*	Adenosine synthase A	converts adenosine monophosphate to adenosine
	aur	zinc metalloproteinase aureolysin	surface proteins modifier
	sak	Staphylokinase precursor	activates plasminogen
	sspA	serine protease; V8 protease; glutamyl endopeptidase	autolytic, degrades FnBP
	sspB	staphopain cysteine proteinase SspB	pathogenesis
	sspC	Staphostatin B	pathogenesis
	vWbp	secreted von Willebrand factor-binding protein precursor	anti-opsonophagocytic
Immune evasion	*cap8A, cap8B, cap8C, cap8D, cap8E, cap8F, cap8G, cap8H, cap8I, cap8J, cap8K, cap8L, cap8M, cap8N, cap8O, cap8P*	capsular polysaccharide synthesis enzyme	Antiphagocytosis: protects cells from engulfment, avoids desiccation, maintains carbohydrates reserve and adhesion
	sbi	IgG-binding protein SBI	
	spa	Immunoglobulin G binding protein A precursor	
Iron uptake	*isdA, isdB, isdC, isdD, isdE, isdF, isdG,*	iron-regulated surface determinant	reservoir for heme-irons
	srtB	NPQTN specific sortase B	
Pathogenesis	*chp*	chemotaxis-inhibiting protein CHIPS	counter host defence mechanisms
	scn	complement inhibitor SCIN	
Proteolysis	*clpP*	ATP-dependent Clp protease proteolytic subunit	degradation of misfolded proteins
Secretion system	*esaA, esaB, esaC, essA, essB, essC, esxA, esxB*	type VII secretion system protein	secretion
Toxin	*hlb*	beta-hemolysin	pathogenesis
	hld	delta-hemolysin	
	hlgA, hlgB, hlgC	gamma-hemolysin	
	hly/hla	Alpha-Hemolysin precursor	
	lukF-PV, lukS-PV	Panton-Valentine leukocidin chain F & chain S precursor	
	sea, seb, seh, selk, selq	staphylococcal enterotoxin	
	geh	glycerol ester hydrolase	
	lip	triacylglycerol lipase precursor	triglyceride lipase activity

Figure S4: Total number of toxin genes predicted for all strains are categorized in figure. Red square represents gene present with 100% identity. Blue square represents gene present with <100% identity. Aqua square represents gene absent for that strain.

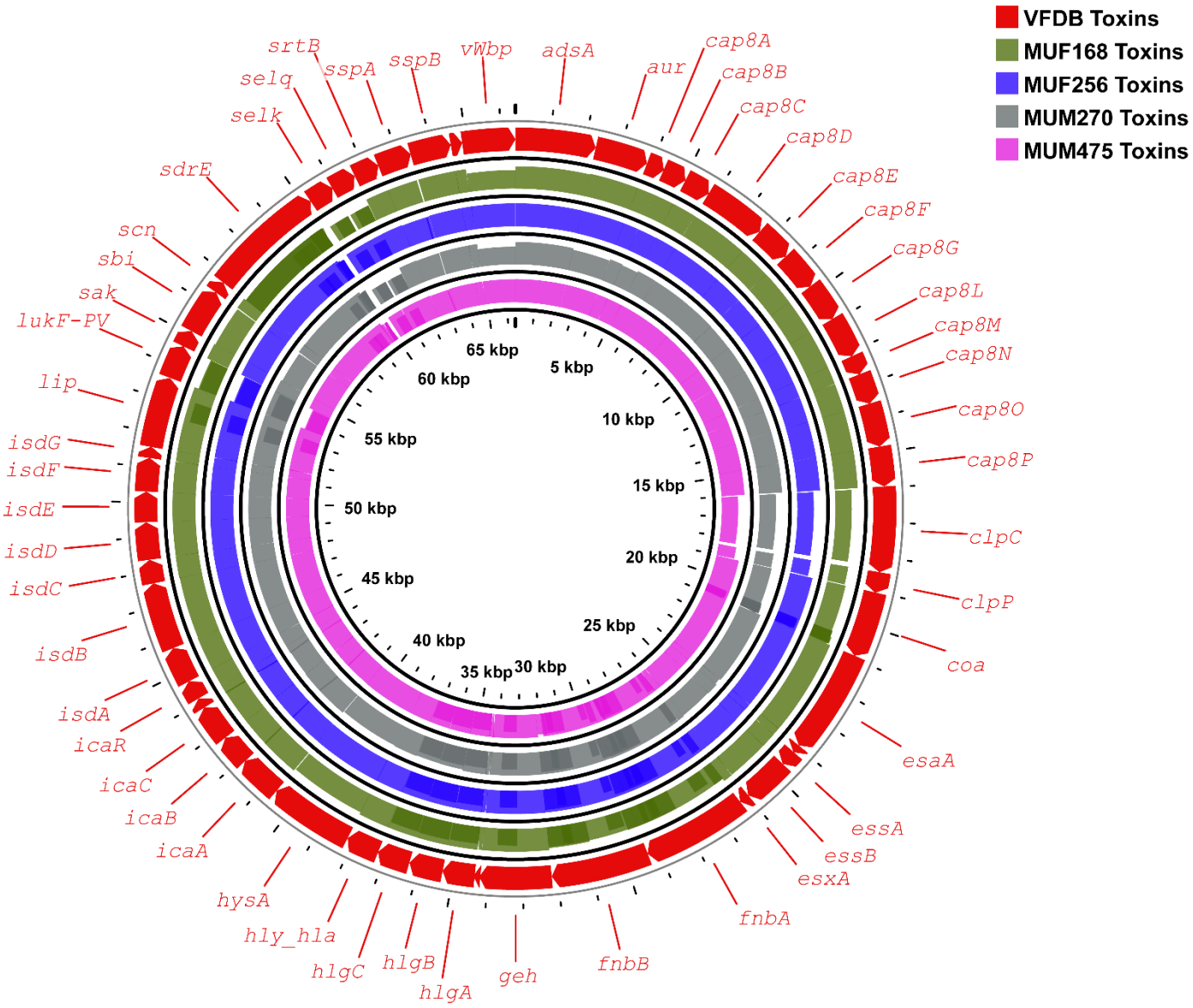


Figure S5: Blast similarity scores are represented for 56 common toxin genes against their reference toxin gene in VFDB. Height and colour of the blocks represent regions of high similarity.

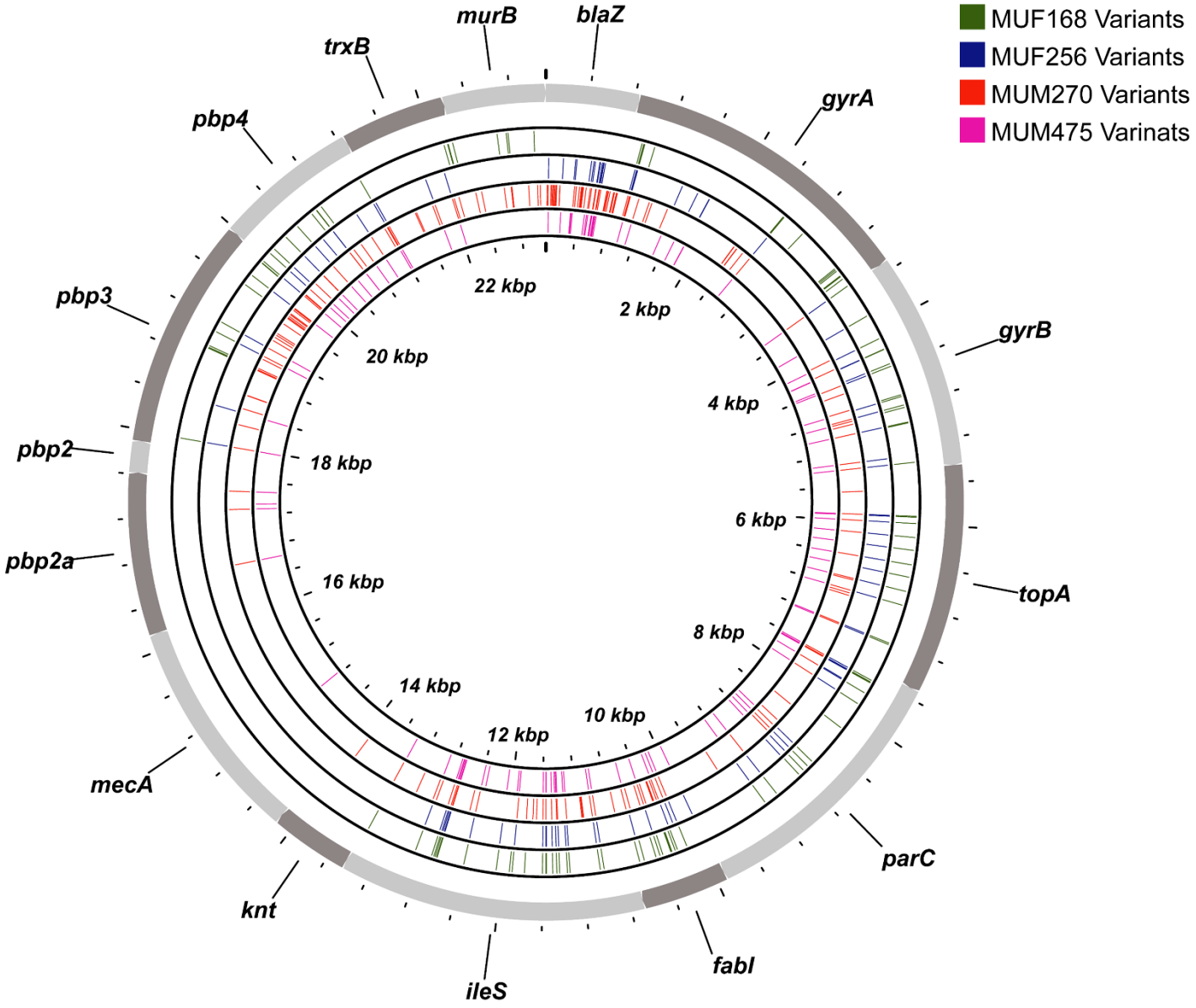


Figure S6: Variant analysis. Nucleotide variants present across selected drug targets in all four strains.

Chapter

6

Microbial Genomics: A Comprehensive Guide for Microbial Genome Research

Samantray, Debyani*, **Ankit Singh Tanwar***, Thokur Sreepathy Murali, Angela Brand, Kapaettu Satyamoorthy, and Bobby Paul. "A Comprehensive Bioinformatics Resource Guide for Genome-Based Antimicrobial Resistance Studies." *OMICS: A Journal of Integrative Biology* 27, no. 10 (2023): 445-460.

DOI: 10.1089/omi.2023.0140; **IF 3.3 (2023)**

*Equal first authors

A Comprehensive Bioinformatics Resource Guide for Genome-Based Antimicrobial Resistance Studies

Abstract

High-throughput sequencing technologies and bioinformatic tools have revolutionized the microbial genome research and, with several sophisticated computational tools, has facilitated whole genome assembly, advanced genome-based species identification, comparative genomics, and the identification or prediction of genes that code for proteins, antimicrobial resistance (AMR), and toxins. These bioinformatics resources are likely to continuously improve in quality, become more user-friendly to analyze the multiple genomic data, efficient in generating information and translating it into meaningful knowledge, and enhance our understanding of the genetic mechanism of AMR. In this manuscript, we provide an essential guide for selecting the popular resources for microbial research, such as genome assembly and annotation, antibiotic resistance gene profiling, identification of virulence factors, and drug interaction studies. Additionally, we discuss the best practises in computer-oriented microbial genome research, emerging trends in microbial genomic data analysis, integration of multi-omics data, the appropriate use of machine learning algorithms, and open-source bioinformatics resources for genome data analytics.

Keywords: Antibiotic resistance, microbial genomics, computational biology, next generation sequencing.

Introduction

Advancements in genome sequencing have revolutionized microbial genome research and generated a large number of whole genome sequences, and this information is publicly available in the Genome database (www.ncbi.nlm.nih.gov/genome/). Currently, the Genome database has more than 510,000 bacterial genome assemblies available for scientific data analysis. However, only 6.9% of these genome assemblies are in the 'complete' stage, while most of the genome assemblies are in the contig stage (62.08%), scaffold stage (29.86%) or at chromosome level (1.15%). The genomes in the 'complete' category and listed as reference sequences are more appropriate for analysis for research studies. The genome sequence data can be used for comparative genomics studies such as classifying an organism, profiling antimicrobial resistance, identifying potential drug targets, and determining genetic relatedness. While this is extremely useful information for researchers, the inclusion of genome assemblies with cross-species contamination or those belonging to taxonomically misclassified isolates remains a major problem. Therefore, it is important to select the proper genome assemblies for downstream analysis. Considering the high bacterial diversity, only a fraction of the global bacterial population is sequenced and assembled with whole genome sequences of 1882 different genera belonging to 3732 distinct species.

Antimicrobial resistance (AMR) is a major global problem, particularly in low-and middle-income countries (Wangai et al., 2019). Microbial species can develop antimicrobial resistance mechanisms and play a critical role in spread of resistance genes across bacterial populations (Woolhouse et al., 2015). Antibiotic

resistance alone contributed to more than 35,000 deaths in the United States of America (Solomon and Oliver, 2014), and AMR might cause for yearly death of 10 million people in the United Kingdom (O'Neill, 2016). Recent global research on antimicrobial resistance data from 204 countries reported that 4.95 million deaths were due to AMR in 2019 and 1.27 million deaths attributable to AMR, and also stated that AMR-related deaths are much higher than HIV-AIDS or malaria-associated deaths (Murray et al., 2022). Therefore, it is crucial to examine the bacterial genomes from a wider perspective, covering all aspects including genome sequencing and assembly, genome annotation and gene prediction, identification of bacterial communities in an environment, and genetic variations on drug targets. For this, reliable and effective bioinformatics tools for whole genome sequence data analysis and proper interpretation of results is essential. Hence, in this review, we intend to provide an overview of sophisticated bioinformatics resources for microbial genome research, exclusively for predicting AMR. All major steps are outlined in **figure 1** along with popular tools involved at each of these processes.

Review methodology:

Search strategy

The two literature databases Scopus and Google Scholar were searched in February 2023. During the search, five search keywords were used: AMR, databases, tools, NGS, analysis, and annotation in the fields of title, keywords, and abstract. Book chapters, letters, reports, conference proceedings, and articles in other languages were excluded using the "Advanced search" settings. For databases and tools, the year was not limited, however the search was only allowed to include articles published between 2014 and 2022, with certain exceptions.

Inclusion criteria

Original research publications that discuss AMR resistance genes, databases, tools, and servers needed for bioinformatics analysis, annotation, and designing of bacterial genomic data were taken into consideration.

Exclusion criteria

Using the "Advanced search" criteria in databases, Scopus, Google Scholar, and PubMed, book chapters, letters, reports, conference proceedings, and articles written in other languages were removed. To exclude out papers that did not fit the inclusion criteria, the title, keywords, abstracts, and complete texts were once again examined. Additionally, papers that did not provide a thorough explanation of the bioinformatics techniques that was utilised, especially for their analysis, annotation, or design, were also excluded.

Study selection

The authors reviewed the title and abstract of the studies followed by a full-text screening to identify the studies based on the inclusion criteria. Initially, 390 studies (223 studies from Scopus and 167 studies from PubMed and Google Scholar) were identified by searching the above-mentioned databases based on the search keywords. A total of 89 duplicates were removed following which 210 studies were shortlisted. Further, from the 210 studies, 12 studies were excluded based on the exclusion criteria, and 198 studies were shortlisted by screening and analysing the title, keyword, and abstracts of the articles. Finally, 151 studies that met the inclusion criteria with the required data were selected and systematically reviewed.

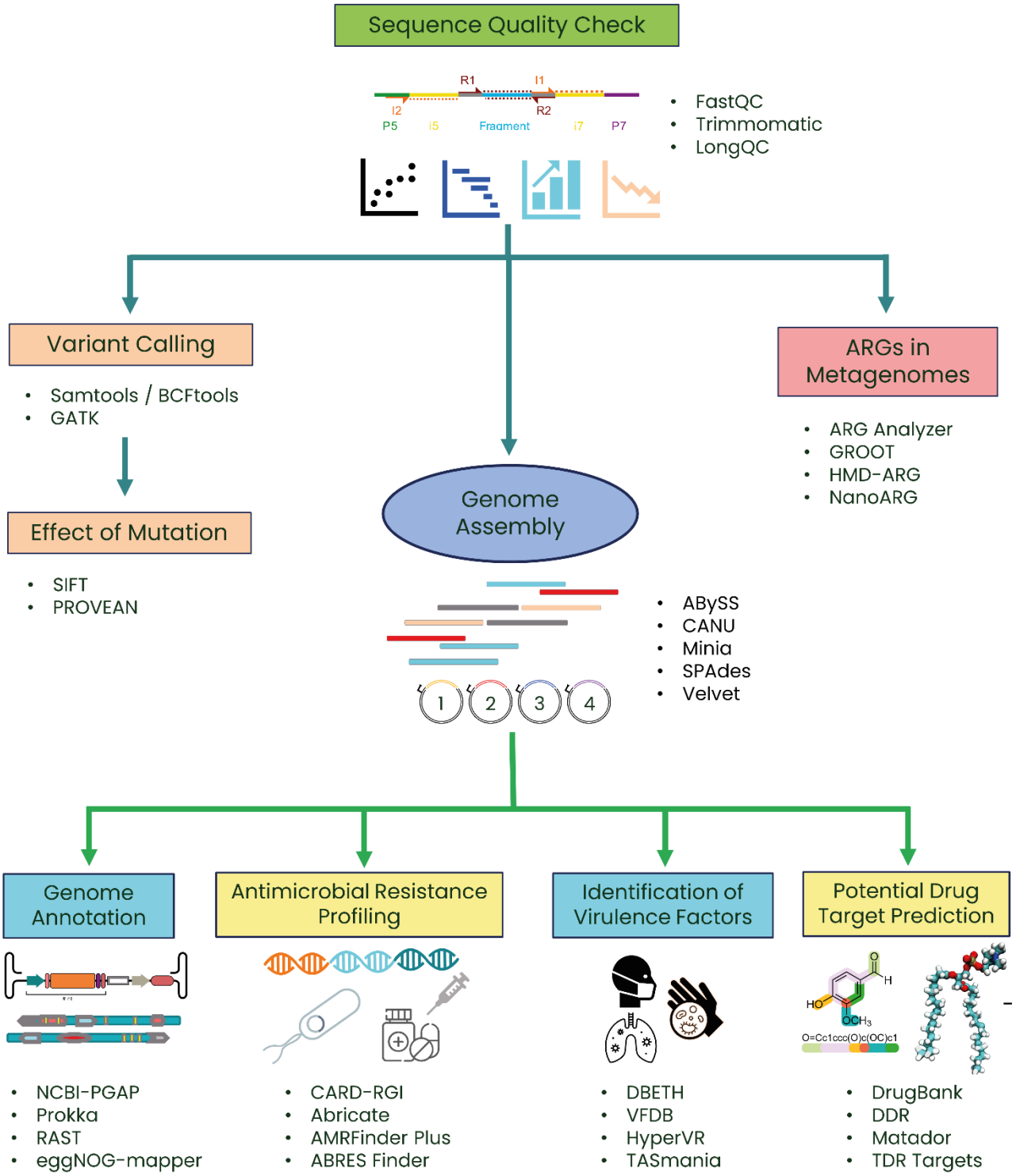


FIG. 1. A basic workflow highlighting major steps and tools involved in microbial genome research. Functional URLs of tools and databases mentioned above are provided in Table 1-7. ARGs, Antibiotic Resistance Genes.

High-throughput genome sequencing technologies

The evolution of DNA sequencing began in 1970, when Sanger and co-workers introduced nucleotide identification using chain termination. Over the years, innovations in sequencing chemistry and automation in technologies have drastically reduced the cost of sequencing, and at the same time improved the sequencing output in order to generate a few hundred to kilobases of long DNA molecules with gigabases of data per run. Advancements in sequencing technologies have revolutionized microbial genome research by enabling scientists to obtain high-quality sequencing data from a large number of samples in a relatively short time and, reduced cost. This breakthrough has led to the discovery of novel genomic features and molecular mechanisms underlying complex biological processes (Krehenwinkel et al., 2019).

Applications of high-throughput sequencing in microbial genomics include understanding the genetic mechanism of an organism, identification and classification of microbial populations from different environments, prediction of virulence factors and antibiotic resistance genes (ARGs), development of diagnostic assays and vaccines (Deurenberg et al., 2017; Besser et al., 2018). The data generated through next generation sequencing (NGS) can be used for more accurate pathogen identification and characterization, as well as identifying genetic mutations, and predicting pathogenicity (Varshney et al., 2009). Further, high-throughput sequencing has transformed many areas of genomics research, such as genotyping, epigenomics, transcriptomics, and metagenomics. Integration of high-throughput sequencing with other omics technologies, such as proteomics and metabolomics, has the potential to provide a more comprehensive understanding of complex biological systems (Ariey et al., 2013; Schmidt et al., 2016; Berry et al., 2019).

High-throughput sequencing has rapidly become an essential tool in microbial research due to its unprecedented speed, accuracy, cost-effectiveness, and capability to deliver tens of thousands of genomes in a short span of time. However, high-throughput sequencing faces certain challenges in widespread usage, data analysis, and interpretation of results. Although many research institutions can purchase high-throughput sequencing equipment, they need more high-performance computational resources and staffing to analyze and clinically interpret the analyzed data. Choosing the best sequencing platform is essential, as there are a couple of high-throughput technologies with different sequencing chemistry. A combination of short and long read sequencing is more appropriate to cover the whole genome of an organism.

Computational tools for genome assembly

Whole genome assembly refers to the computational process of placing the raw sequence data generated through high-throughput sequencing technologies in correct order. The quality of the genome assembly determines the success rate of the genome assembly and downstream genome data analysis. High quality genome assembly is required for accurate genome annotation, predicting the genes accurately and analysing their copy numbers encoded in the genome. The success of genome assembly process depends on the quality of the sequencing data, the choice of assembly algorithms, and the availability of appropriate bioinformatics tools and pipelines.

The whole genome assembly process consists of mainly three steps: i) quality check of the raw sequence reads, ii) genome assembly either by *de novo* or mapping approach, iii) post-genome assembly assessment. Each step requires a specific bioinformatic tool, and the choice of the tool may affect the accuracy

and completeness of the assembly. The abundance of bioinformatics tools available today for various stages of data analysis is another challenge. Pre-assembly quality check of raw sequence data and trimming of low-quality reads are essential. Tools like FastQC (Andrews, 2010) and Trimmomatic (Bolger et al., 2014) can be used for quality control and trimming the raw data sequence, respectively.

The genome assembly process consists of error correction, read alignment, scaffolding or mapping. Over the years, several *de novo* and reference-based assemblers have been derived **(Table 1)**. *De novo* genome assemblers are very specific on the raw data type and their backend algorithm; few assemblers prefer short-read data while others excel in assembly of long read sequencing data. Further, these tools are very specific for sequencing technologies, and parameters are very sensitive. *De novo* genome assembly approach is time-consuming and requires high performance computing facilities; however, it is more appropriate for every strain. Reference-based mapping requires high quality reference genomes and is easier to assemble. The approach can be utilised for the assembly of the same strains, post *de novo* assembly improvements, and variant calling. Tools like St. Petersburg genome assembler SPAdes (Bankevich et al., 2012) and Canu (Koren et al., 2017) can be used for *de novo* genome assembly and scaffolding. Reference-based mapping or identifying the mutations on drug targets can be easily employed with the Bowtie2 aligner (Langmead, 2012) followed by variant calling pipelines such as Samtools/BCFtools (Li, 2011), and GATK pipeline (DePristo et al., 2011). Finally, potential single nucleotide variations can be reviewed by visualization of the sample reads mapped against the reference genome with Integrative Genome Viewer (Thorvaldsdóttir et al., 2013).

Genome assembly at the draft level or inaccurate assembly may result in partial genes or pseudogenes, which will compromise the genome quality (Didelot et al., 2012; Wattam et al., 2017). Hence, post genome assembly assessment is crucial and needs to follow the recent guidelines for genome submission to the public domain (https://www.ncbi.nlm.nih.gov/genbank/genomesubmit). Many scientific journals require newly assembled genome sequence data to be deposited in the International Nucleotide Sequence Database Collaboration. However, submitting genomes to public databases in a prescribed format remains a burden for many data science and comparative genome researchers. The QUAST tool can be utilized for post assembly genome assessment (Gurevich et al., 2013). The Pilon tool (Walker et al., 2014) can be used for post assembly genome improvement while DDBJ Fast Annotation and Submission Tool (DFAST) (Tanizawa et al., 2018) supports genome submission to public sequence databases.

Table 1: Computational resources available for microbial genome assembly.

Sl. No.	Software	Supporting Platform	Description	URL & Reference
1	A5	Illumina	An integrated automated pipeline for cleaning reads and assembling.	https://bio.tools/a5 (Tritt et al., 2012)
2	ABySS	Illumina Ion Torrent	A tool for *de novo* genome assembly using short read data. It implements a distributed representation of de Bruijn graphs, which enables parallel computation of the assembly algorithm.	https://github.com/bcgsc/abyss (Jackman et al., 2017)
3	ALLPATHS-LG	Illumina	A tool for assembling both small and large genomes.	http://www.broadinstitute.org/science/programs/genome-biology/crd. (Gnerre et al., 2011)

4	Bambus 2	Illumina PacBio Oxford Nanopore	The program orders and orients contigs into scaffolds based on various types of linking information.	http://www.cbcb.umd.edu/software/bambus (Koren et al., 2011)
5	CANU	PacBio Oxford Nanopore	Derived from Celera Assembler, mainly for long read assembly.	https://github.com/marbl/canu (Koren et al., 2017)
6	CAP3	PacBio Oxford Nanopore Sanger	To assemble a set of contiguous sequences (contigs) or long reads sequence data.	https://sourceforge.net/projects/staden/ (Huang and Madan et al., 1999)
7	Edena	Illumina	A *de novo* assembly framework that can be used for extremely short reads.	https://mybiosoftware.com/edena-de-novo-short-reads-assembler.html (Hernandez et al., 2008)
8	FALCON	PacBio	*De novo* assembler for long read sequencing data.	https://github.com/PacificBiosciences/FALCON/ (Chin et al., 2016)
9	HIFIASM	PacBio	A *de novo* assembler that uses long high-fidelity sequence reads to represent the haplotype information in an assembly.	https://github.com/chhylp123/hifiasm (Cheng et al., 2021)
10	MaSuRCA	Illumina Ion Torrent PacBio Oxford Nanopore	Assemble data sets containing only short reads or a mixture of short reads and long reads.	https://github.com/alekseyzimin/masurca (Zimin et al., 2013)
11	MECAT	PacBio Oxford Nanopore	A tool enables reference mapping and *de novo* assembly of large genomes.	https://github.com/xiaochuanle/MECAT (Xiao et al., 2017)
12	MEGAHIT	Illumina	*De novo* assembler for large and complex metagenomics data.	https://github.com/voutcn/megahit (Li et al., 2015)
13	metAMOS	Illumina PacBio Oxford Nanopore	An integrated assembly and analysis pipeline for metagenomic data.	https://github.com/marbl/metAMOS (Treangen et al., 2013)
14	Minia	Illumina	A short-read assembler based on a de Bruijn graph method.	http://minia.genouest.org/ (Chikhi and Rizk, 2013)
15	Miniasm	PacBio Oxford Nanopore	An ultrafast *de novo* assembler for noisy long reads.	https://github.com/lh3/miniasm (Li, 2016)
16	MIRA	Illumina Ion Torrent	*De novo* genome and transcriptome assembler and can be used for mapping and genome polishing.	https://sourceforge.net/projects/mira-assembler/ (Chevreux et al., 2004)
17	PBcR pipeline	PacBio	A pipeline for read correction and assembly.	http://www.cbcb.umd.edu/software/PBcR/ (Koren et al., 2012)
18	Racon	PacBio Oxford Nanopore	An intensive error-correction tool for long reads to obtain high-quality assemblies.	https://github.com/lbcb-sci/racon (Vaser et al., 2017)
19	Ray Meta	Illumina	A tool for *de novo* assembly of metagenome sequence data.	https://denovoassembler.sourceforge.net/ (Boisvert et al., 2012)
20	Redundans	Illumina Roche 454	A pipeline that improves the genome assembly of heterozygous genomes.	https://github.com/Gabaldonlab/redundans (Pryszcz and Gabaldón, 2016)
21	SGA	Illumina	Performs contig assembly using SGA and builds scaffolds using BESST.	https://github.com/jts/sga (Simpson and Durbin, 2012)
22	SHORTY	SOLiD	A tool for *de novo* genome assembly of short reads.	https://www3.cs.stonybrook.edu/~skiena/shorty/

				(Hossain et al., 2009)
23	Smart denovo	PacBio Oxford Nanopore	A rapid assembler for long reds from single-molecule sequencing platforms.	https://github.com/ruanjue/smartdenovo (Liu et al., 2021)
24	SOAPdenovo2	Illumina	A short-read assembly method that can build a *de novo* draft assembly for large plant and animal genomes, also works well on bacteria and fungi genomes.	https://sourceforge.net/projects/soapdenovo2/ (Luo et al., 2012)
25	SPAdes	Illumina Ion Torrent PacBio Oxford Nanopore Sanger	A novel assembly toolkit containing various assembly pipelines and it support hybrid genome assembly.	https://cab.spbu.ru/software/spades/ (Prjibelski et al., 2020)
26	SSAKE	Illumina	A *de novo* assembler for short DNA sequence reads.	https://www.bcgsc.ca/resources/software/ssake/ (Warren et al., 2007)
27	SSPACE	Illumina	A tool for scaffolding contigs using paired-end reads.	https://github.com/nsoranzo/sspace_basic (Boetzer and Pirovano, 2014)
28	SSPACE-LongRead	PacBio Oxford Nanopore	A tool to upgrade incomplete draft genomes using long reads.	http://www.baseclear.com/bioinformatics-tools/ (Boetzer and Pirovano, 2014)
29	Velvet	Illumina	A De Bruijn graph assembler works fairly rapidly on short (microbial) genomes.	https://github.com/dzerbino/velvet (Zerbino and Birney, 2008)
30	wtdbg2	PacBio Oxford Nanopore	A long-read *de novo* assembler essentially uses an all-versus-all read alignment procedure to progress the overlap-layout-consensus method.	https://github.com/ruanjue/wtdbg2 (Ruan and Li, 2019)

Bioinformatics resources for genome annotation

Genome annotation consists of several computational processes to deliver the structural and functional information from genome assembly using differential analysis, comparison, estimation, precision, prediction, and other data mining techniques. The process involves identifying and classifying functional components such as genes that code for proteins, non-coding RNAs, genetic signatures that regulate gene expression, and conserved and non-conserved genomic regions. Genome annotation can shed light on pathogenicity, metabolism, antibiotic resistance, evolutionary relationships, microbial adaptations in host and other environments. Bacterial genome annotation also aids in carriage of genes coding for toxins which can distinguish pathogenic strains from non-pathogens, and more importantly in identifying strains with desired genetic characteristics for an industrial application.

Genome annotation is time-consuming and a multi-step process and requires skills in high performance computing. Genome annotation pipelines are well equipped to analyse large amounts of genomic data by automating crucial steps that are much faster, more efficient, and more accurate. Automated pipelines for various genome annotation processes are listed in **Table 2**. Web-based and automated pipelines are intended to reduce manual errors, which helps in file format conversion, makes the data presentable, and helps prepare the data to be submit into the public domain.

Table 2: Computational resources available for genome annotation.

Sl.No.	Tool	Description	URL
1	DFAST	A pipeline for annotating bacterial genomes that incorporates techniques for taxonomy and quality evaluation.	https://dfast.ddbj.nig.ac.jp/ (Tanizawa et al., 2018)
2	eggNOG-Mapper	A tool for fast functional annotation of novel genomes.	http://eggnog-mapper.embl.de/ (Huerta-Cepas et al., 2019)
3	Eugene	An integrative gene finder applicable to both prokaryotic and eukaryotic genomes	http://eugene.toulouse.inra.fr/Configuration (Sallet et al., 2019)
4	FlaGs	A flexible tool for sensitive detection of flanking gene conservation at any evolutionary distance, and displays results in an intuitive, publication-quality vector graphics format.	https://github.com/GCA-VH-lab/FlaGs (Saha et al., 2021)
5	FunGeCo	A web-based platform for gene-context-based functional inference in microbial genomes and metagenomes.	https://web.rniapps.net/fungeco (Anand et al., 2020)
6	GeneMark	A fully automatic integrated tool use Bayesian formalism for gene prediction.	http://opal.biology.gatech.edu/GeneMark/ (Besemer et al., 2005)
7	GhostKOALA	Web server designed with KEGG Orthology for functional characterization of genome and metagenome sequences.	http://www.kegg.jp/blastkoala/ (Kanehisa et al., 2016)
8	Glimmer	A system for finding genes in microbial genome.	http://www.cbcb.umd.edu/software/glimmer/ (Delcher et al., 1999)
9	InterProScan	A server that integrates various protein signature recognition methods from multiple member databases of the InterPro for protein analysis and annotation.	http://www.ebi.ac.uk/InterProScan/ (Quevillon et al., 2005)
10	MAKER2	An automated pipeline for comprehensive functional annotation and data management.	http://www.yandell-lab.org/software/maker.html (Holt and Yandell, 2011)
11	MG-RAST API	Web application server that suggests automatic phylogenetic and functional analysis of metagenomes.	http://api.metagenomics.anl.gov/api.html (Keegan et al., 2016)
12	MicrobeAnnotator	An easy-to-use pipeline coupled with several reference protein databases for the annotation of microbial genomes.	https://github.com/cruizperez/MicrobeAnnotator (Ruiz-Perez et al., 2021)
13	NCBI - PGAP	NCBI Prokaryotic Genome Annotation Pipeline is designed to annotate bacterial and archaeal genomes.	https://github.com/ncbi/pgap (Tatusova et al., 2016)
14	PANNZER2	A web server designed for fast Gene Ontology (GO) annotations and predictions	http://ekhidna2.biocenter.helsinki.fi/sanspanz/ (Törönen et al., 2018)
15	Prokka	A command line tool to fully annotate bacterial genome and generate standards-compliant output files for further analysis or viewing in genome browsers.	https://github.com/tseemann/prokka (Seemann, 2014)
16	RAST	A webserver with fully automated annotation service for archaeal and bacterial genomes.	https://rast.nmpdr.org/ (Aziz et al., 2008)
17	REPET/ PASTEC	An automated pipeline for the identification and classification of transposable elements.	https://urgi.versailles.inrae.fr/Tools/REPET (Quesneville et al., 2005; Hoede et al., 2014)
18	tRNAscan	A widely used tool for predicting transfer RNA (tRNA) genes in genomic sequences.	http://lowelab.ucsc.edu/tRNAscan-SE/ (Chan et al., 2021)

Resources for antimicrobial resistance gene profiling

The genomes of bacterial species are comparatively small in size, and with simple structural organization, though, they have a remarkable genetic mechanism that allows them to respond to a wide array of

environmental threats, including antibiotics. Antimicrobial resistance (AMR) is a major public health threat worldwide that can lead to longer hospital stays, higher healthcare costs, and increased mortality and morbidity rates. Most pathogenic bacterial species have genetic mechanisms to develop resistance to at least a few antibiotics. The main mechanisms of antimicrobial resistance are: i) restricting the uptake of drug molecules, ii) structural modification of the drug and target, iii) inactivation of drug, and iv) active efflux system. Bacteria acquire resistance to antibiotics via the transfer of AMR genes from other organisms through genetic exchange mechanisms such as conjugation, transduction, or transformation. Examples of these mechanisms include the acquisition of the *mecA* gene encoding methicillin resistance in *Staphylococcus aureus* and the various *van* genes in *Enterococcus* encoding resistance to glycopeptides (Tunstall et al., 2020).

The quest for rapid and effective AMR diagnosis and testing is of paramount relevance in the current global context. Traditional antimicrobial susceptibility testing (AST or WGS-AST) is labour-intensive, low throughput, and limited to bacteria that can be cultured in a lab. The present emphasis in AMR prediction is on enhancing technologies and methodologies that enable prompt identification of resistant strains. With the aid of bioinformatics tools, researchers can analyze the large amount of whole genome sequence data generated, in order to implement certain strategies, improve prevention, control and treatment measures for infections caused by antibiotic resistant pathogens.

Artificial intelligence and machine learning methods are being used more frequently to predict AMR genes. By predicting the AMR resistance profile of a pathogen, clinicians can select the most appropriate treatment and reduce the risk of treatment failure or the spread of AMR. For example, some systems are now able to predict phenotypic resistance based on genomic sequencing data, enabling faster and more accurate diagnosis and treatment. In addition, accurate AMR prediction can help healthcare systems optimize their use of antibiotics and reduce the incidence of healthcare-associated infections (Eschlböck et al., 2017; Walker et al., 2019). Nevertheless, there are still obstacles to overcome, such as ensuring data sharing, standardizing methods, and regulating oversight to guarantee the precision and dependability of predictive models (Fanelli et al., 2020; Shankarnarayan et al., 2022; Wang et al., 2022).

Several computational tools and web servers are available to predict AMR genes, ranging from simple rule-based algorithms to complex machine learning models. These tools use various types of data such as whole genome sequence information, phenotypic characteristics, and epidemiological information to predict AMR patterns **(Table 3)**. Some of the commonly used AMR prediction tools and databases include ResFinder, a web-based tool that uses genome sequence data to predict the presence of antimicrobial resistance genes (Zankari et al., 2012), KmerResistance, that uses k-mer frequencies to predict antimicrobial resistance from genome sequences (Rowe and Winn, 2018), Antibiotic Resistance Gene-ANNOTation (ARG-ANNOT), a database that contains information on antimicrobial resistance genes and mutations responsible for AMR from bacterial genomes (Gupta et al., 2014), The Comprehensive Antibiotic Resistance Database (CARD), or comprehensive database of antibiotic resistance genes, mutations, and associated metadata (Alcock et al., 2020), and AMRmap, a platform that provides information on the prevalence and patterns of antimicrobial resistance worldwide. These tools provide interactive data analysis, and visualization tools are continuously updated with information from multiple sources (Kuzmenkov et al., 2021).

Table 3: Computational resources available for antimicrobial resistance profiling

Sl. No.	Resource	Description	URL
1	ABRES Finder	A webserver to predict 36467 antibiotic resistance genes belonging to 37 antibiotics.	http://scbt.sastra.edu/ABRES/ (Xavier et al., 2016)
2	ABRICATE	Pipeline coupled with multiple databases to screen contigs for antimicrobial resistance or virulence genes.	https://github.com/tseemann/abricate (Seemann, 2018)
3	AMRFinderPlus	A tool that identifies AMR genes, resistance-associated point mutations from assembled genome sequence.	https://github.com/ncbi/amr (Feldgarden et al., 2019)
4	AMRmap	A web platform for analysis of antimicrobial resistance surveillance data.	https://amrmap.net/ (Kuzmenkov et al., 2021)
5	ARG-ANNOT	A tool to detect existing and putative new ARGs in bacterial genomes.	https://www.mediterranee-infection.com/acces-ressources/base-de-donnees/arg-annot-2/ (Gupta et al., 2014)
6	ARGDIT	A toolkit to minimize the efforts in validating, curating and merging multiple ARG protein or coding sequence databases.	https://github.com/phglab/ARGDIT (Chiu and Ong, 2019)
7	ARIBA	A standalone tool to identify AMR-associated genes and single nucleotide polymorphisms directly from short reads with respect to the reference genome.	https://github.com/sanger-pathogens/ariba (Hunt et al., 2017)
8	BacMet	A tool for identification of biocide and metal-resistance genes from genomes.	http://bacmet.biomedicine.gu.se/ (Pal et al., 2014)
9	β-lactamases Database	A specialized manually curated public resource providing up-to-date structural and functional information focused on β-lactamase with a great impact on antibiotic resistance.	http://bldb.eu (Naas et al., 2017)
10	CARD	A database of ARGs that are peer-reviewed. It includes software to predict resistome from protein, genome, or metagenomics datasets.	https://card.mcmaster.ca/ (Alcock et al., 2020)
11	DRAGdb	A manually curated, repository of mutational data of drug resistance associated genes.	http://bicresources.jcbose.ac.in/ssaha4/drag/ (Ghosh et al., 2020)
12	FARMEDB	A database of publically available DNA sequences and predicted protein sequences conferring antibiotic resistance as well as regulatory elements, mobile genetic elements.	http://staff.washington.edu/jwallace/farme (Wallace et al., 2017)
13	IRIDA	A pipeline makes use of the RGI/CARD and staramr tools for detection of antimicrobial resistance genes.	https://github.com/phac-nml/irida-plugin-amr-detection (Matthews et al., 2018)
14	LREfinder	A web tool to detect mutations in 23S rRNA, *optrA*, *cfr*, *cfr(B)*, and *poxtA* genes encoding linezolid resistance from enterococci.	https://bitbucket.org/genomicepidemiology/lre-finder/src/master/ (Hasman et al., 2019)
15	MARA	A database includes antibiotic resistance genes and selected mobile elements from Gram-negative bacteria, distinguishing important variants.	https://galileoamr.arcbio.com/mara/ (Partridge and Tsafnat, 2018)
16	MEGARes	An antimicrobial resistance database for high-throughput sequencing: A database that includes ARGs as well as a biocide and heavy metal resistance genes.	https://github.com/cdeanj/resistome analyzer (Lakin et al., 2017)
17	Mustard	A tool to predict resistome by a three-dimensional structure-based approach.	http://mgps.eu/Mustard/ (Ruppé et al., 2018)
18	Mykrobe	An ultrafast search tool for predicting resistance genes from bacterial and viral genomic data.	https://www.mykrobe.com/ (Bradley et al., 2015)
19	PARGT	A software tool for predicting antimicrobial resistance in Gram-negative bacteria.	https://github.com/abu034004/PARGT

			(Chowdhury et al., 2020)
20	Patric	A comprehensive bioinformatics resource to analyse genomes for AMR prediction.	https://www.patricbrc.org/ (Wattam et al., 2017)
21	PointFinder	A web tool to predict antimicrobial resistance associated point mutations from bacterial pathogens.	https://bitbucket.org/genomicepidemiology/pointfinder_db/src/master/ (Zankari et al., 2017)
22	ResFams	A curated database of protein families for profiling antibiotic resistance function and ontology.	http://www.dantaslab.org/resfams (Gibson et al., 2015)
23	ResFinder	A web based antimicrobial susceptibility testing identifies AMR phenotypes.	http://www.genomicepidemiology.org/ (Zankari et al., 2012)
24	SRST2	A read mapping-based tool for fast and accurate detection of genes, alleles and multi-locus sequence types from whole genome sequence data.	https://github.com/katholt/srst2 (Inouye et al., 2014)
25	SSTAR	A pipeline to identify known antimicrobial resistance genes and detect putative new variants from whole genome sequencing data.	https://github.com/tomdeman-bio/Sequence-Search-Tool-for-Antimicrobial-Resistance-SSTAR- (de Man and Limbago, 2016)
26	VAMPr	A tool for mapping and variant prediction using machine learning approaches.	https://github.com/jiwoongbio/VAMPr (Kim et al., 2020)

Tools for virulence profiling

Toxins are potent virulence factors produced by organisms such as bacteria, fungi, plants, and animals that can cause harm or death to other organisms (Greenfield et al., 2002). Surprisingly, bacterial toxins were the first substances to be attributed to acute bacterial infections in both humans and animals. These toxins can act on a variety of targets within the host, including the nervous system, immune system, and various organs. While some toxins damage cellular membranes, others prevent the creation of proteins or interfere with cellular signalling pathways (Blanco, 2018; Kim, 2010; Zhang et al., 2019). Certain bacterial toxins are created locally and predominantly affect cells close to the site of infection, while others are released by the bacterium and have an effect far from the site of infection.

There are several bacterial species that produce toxins, like *Clostridium botulinum*, which produces the potent neurotoxin botulinum, causing the disease botulism. Botulinum toxin acts by inhibiting the release of acetylcholine, a neurotransmitter that is necessary for muscle contraction (Guzmán-Gómez et al., 2013). *Vibrio cholerae* produces cholera toxin, which causes the disease cholera. The toxin acts by causing the secretion of large amounts of water and electrolytes into the intestinal lumen, leading to severe diarrhea and dehydration (Chowdhury et al., 2011). *Staphylococcus aureus* produces several toxins, including staphylococcal enterotoxins and toxic shock syndrome toxin-1 (TSST-1), which can cause food poisoning, toxic shock syndrome, and other diseases (Chakraborty et al., 2022).

Understanding the mechanisms of these toxins and the organisms that produce them is crucial for developing effective treatments and preventing the spread of these diseases. So, the bioinformatic approaches are essential for analysing toxins and predicting their properties. These approaches include sequence analysis, structure-based modelling, and functional analysis. Sequence analysis involves identifying genes and proteins involved in toxin production and studying their evolution and diversity. Structure-based modelling involves

predicting the three-dimensional structure of toxins and their interactions with target molecules. Functional analysis involves studying the biological activity and toxicity of toxins and their mechanisms of action.

We have listed the various toxin prediction tools and database resources available **(Table 4)**. Some of the popular toxin databases include the Toxin and Toxin-Target Database (T3DB) (Lim et al., 2010) and the Comparative Toxicogenomics Database (CTD) (Davis et al., 2023). Toxin databases are important resources for researchers studying and gaining knowledge on toxins. These databases provide valuable information on toxins, including their mechanisms of action, targets, potential health effects, the structure of the toxin molecule, biological activity, and level of toxicity. Various algorithms are used to predict the presence of toxins based on the organism's genome or proteome. Some of the popular toxin prediction tools include virulence factor database VFDB (Chen et al., 2005), TASmania (Akarsu et al., 2019) which are useful for identifying potential toxins in new organisms and understanding the diversity and evolution of toxin-producing organisms.

Table 4: Computational resources available for toxins or virulence factor prediction.

Sl. No.	Resource	Description	URL
1	DBETH	A database of bacterial exotoxins sequences, structures, interaction networks, and analytical results for 229 exotoxins, representing 26 distinct bacterial genera that are pathogenic to human.	http://www.hpppi.iicb.res.in/btox/ (Chakraborty et al., 2012)
2	HyperVR	A pipeline for accurately predict virulence factors, antibiotic resistant and negative genes.	https://github.com/jiboyalab/HyperVR (Ji et al., 2023)
3	RASTA-Bacteria	A web-based tool for identifying toxin and antitoxin genes prediction in prokaryotic genomes.	http://genoweb1.irisa.fr/duals/RASTA-Bacteria/ (Sevin and Barloy-Hubler, 2007)
4	T1TAdb	A web database of type I toxin-antitoxin which contains1894 toxin-antitoxin loci from 493 bacterial strains.	https://d-lab.arna.cnrs.fr/t1tadb (Tourasse and Darfeuille, 2021)
5	TADB	A database to predict type II toxin-antitoxin from bacterial genomes.	http://bioinfo-mml.sjtu.edu.cn/TADB2/ (Xie et al., 2018)
6	TASmania	A database system for predicting bacterial toxin-antitoxins.	https://shiny.bioinformatics.unibe.ch/apps/tasmania/ (Akarsu et al., 2019)
7	VFDB	An integrated and comprehensive online resource for curating information about virulence factors of bacterial pathogens.	http://www.mgc.ac.cn/VFs/ (Chen et al., 2005)

Drug target interaction studies

Genetic mutation in the target sites of drugs is a common mechanism of AMR. Drug-target site alterations often result from spontaneous mutation on the target gene in the presence of antibiotics. Examples include structural alteration in RNA polymerase and DNA gyrase, resulting in resistance to rifamycin and quinolones, respectively (Tunstall et al., 2020). Advancements in bioinformatic have provided different tools to describe the effect of mutation on drug targets induced antimicrobial resistance. While a few predict functional effects based on the amino acid substitution, other algorithms consider the variation within the protein structure

exclusively on the drug binding site. Therefore, when analysing specific mutations on drug targets, it is better to use multiple tools with different methodologies, which may give complementary information. Several tools are available for analysing the effects of mutation in AMR genes **(Table 5)** and are very useful in identifying the known AMR phenotypes and explaining the cause of resistance.

Table 5: Computational tools available for prediction of effect of mutation.

Sl. No.	Database/Tools	Description	URL
1	DeepDDG	A neural network-based tool for predicting the stability change of protein point mutations.	http://protein.org.cn/ddg.html (Cao et al., 2019)
2	DUET	A web server for an integrated computational approach to study missense mutations in proteins.	http://biosig.unimelb.edu.au/duet/ (Pires et al., 2014)
3	DynaMut	A web server for predicting the impact of mutations on protein conformation, flexibility and stability.	http://biosig.unimelb.edu.au/dynamut/ (Rodrigues et al., 2018)
4	ELASPIC	A webserver to predict stability effects of mutations on protein folding and interactions	http://elaspic.kimlab.org/ (Witvliet et al., 2016)
5	I-Mutant	An automatic prediction tool for predicting stability changes upon mutation from the protein sequence or structure.	https://bio.tools/i-mutant_suite (Capriotti et al., 2005)
6	INPS-MD	A web server to predict stability of protein variants from sequence and structure.	https://inpsmd.biocomp.unibo.it/inpsSuite (Savojardo et al., 2016)
7	MAESTRO	A versatile tool in the field of stability change prediction upon point mutations.	https://pbwww.che.sbg.ac.at/maestro/web (Laimer et al., 2016)
8	PROVEAN	A web server tool to predict the functional effect of single or multiple amino acid substitutions, insertions and deletions.	http://provean.jcvi.org/seq_submit.php (Choi et al., 2012)
9	SIFT	An online tool to predict the effects of non-synonymous variants on protein function.	https://sift.bii.a-star.edu.sg/ (Kumar et al., 2009)
10	SNAP2	A neural network-based classifier to forecast how changes in single amino acid would affect the protein's functionality.	https://www.rostlab.org/services/SNAP/ (Hecht et al., 2015)
11	STRUM	A web tool for structure-based prediction of protein stability changes upon single-point mutation.	https://zhanglab.ccmb.med.umich.edu/STRUM/ (Quan et al., 2016)

Predicting novel drug targets is a crucial task in the drug discovery process. There are several phases involved in predicting drug targets, including identifying potential targets, estimating the binding affinity between a drug and those targets, and determining how well the drug inhibits the activity of the target. While there are many different approaches to predict drug targets, from experimental methods to computational techniques, the latter relies on the use of algorithms and machine learning models to predict drug targets. It involves the analysis of data from various sources, including genomics, proteomics, and metabolomics, to identify potential drug targets. Further, computational techniques have the potential to greatly accelerate the drug discovery process and provide important insights into the mechanism of action of drugs. One of the major advantages of using computational techniques for predicting drug targets is that they can help identify potential targets much more quickly and in a cost-effective manner than experimental methods.

In recent years, a large number of tools and databases have been developed to aid in the prediction of drug targets. There are also many databases that are designed to aid in the prediction of drug targets. These databases include resources such as DrugBank (Wishart et al., 2008), which provides information on drug-target interactions and drug metabolism, and STITCH (Kuhn et al., 2014), which integrates information on protein interactions, chemical structures, and genomic data to predict drug-target interactions. Other databases, such as ChEMBL (Gaulton et al., 2012) provide information on chemical compounds and their biological activities and can be used to identify potential drug targets based on chemical structure and activity data. A comprehensive list of all resources available for drug-target interaction studies are provided **(Table 6)**. Overall, the availability of these tools and databases has greatly facilitated the prediction of drug targets and has accelerated the drug discovery process. While these resources have limitations, such as reliance on the quality and availability of data, they represent important tools for researchers in the field of drug discovery.

Table 6: Computational resources available for Drug Designing and other medical databases.

Sl. No.	Resource	Description	URL
1	DDR	A computational method to predict drug–target interactions using graph mining and machine learning approaches.	https://bitbucket.org/RSO24/ddr/src/master/ (Olayan et al., 2018)
2	DrugBank	A richly annotated resource that combines detailed drug data with comprehensive drug target and drug action information.	http://www.drugbank.ca (Wishart et al., 2008)
3	IUPHAR/BPS Guide	An open knowledgebase that provides the information of approved targets and experimental drugs.	http://www.guidetopharmacology.org/ (Pawson et al., 2014)
4	Matador	A database for drug–target interactions and identifying potential drug targets.	https://www.hsls.pitt.edu/obrc/index.php?page=URL1209757429 (Günther et al., 2008)
5	SuperPred	A machine learning approach for identifying drug targets.	http://prediction.charite.de/ (Nickel et al., 2014)
6	TDR Targets	A web tool facilitates the identification and prioritization of candidate drug targets for pathogens.	http://tdrtargets.org (Agüero et al., 2008)
7	TTD	A web tool to facilitate information on known and undiscovered therapeutic proteins.	https://db.idrblab.net/ttd/ (Qin et al., 2014)

Computational tools for metagenome based global resistome profiling

As a culture independent method, the metagenome approach has several advantages in the global profiling of AMRs. Many reports on AMR profiling using the metagenome approach have been published. A study identified ARGs against 53 antibiotics from 275 individuals belonging to eight countries (Ghosh et al., 2013). Metagenome based resistome profile of 207 fecal samples from three different countries identified 50 AMR genes covering 68 classes of antibiotics (Forslund et al., 2013). Another study identified 1625 different ARGs belonging to 408 groups from 1546 genera (Hendriksen et al., 2019). Despite numerous advantages, metagenome-based approach still poses quite a few challenges such as sequencing errors, lack of advanced computing facilities for data analysis, improper interpretation of results, and taxonomic assignments. Moreover, analysis of metagenome sequence data requires sophisticated statistical and computational tools, as

well as domain-specific knowledge of microbial ecology and physiology. The computational tools for metagenome based antimicrobial gene profiling is listed in **Table 7**.

Table 7: Databases for identifying potential ARGs in metagenomic data.

Sl. No.	Database/Tools	Description	URL
1	AgroSeek	A computational tool for comparing and analysing metagenomic data from agriculture sector in monitoring and reducing the spread of antibiotic resistance genes.	https://agroseek.cs.vt.edu/ (Liang et al., 2021)
2	ARG Analyzer	A pipeline designed to identify, and annotate antibiotic resistance genes from environmental metagenomes.	http://mem.rcees.ac.cn:8083/ (Wei et al., 2019)
3	ARGpore2	A pipeline for predicting antibiotic resistance genes from metagenome data.	https://github.com/sustc-xylab/ARGpore2 (Wu et al., 2018)
4	ARGs-OAP	An online pipeline that was developed to efficiently annotate and categorise antibiotic resistance genes from metagenomic data.	http://smile.hku.hk/SARGs (Yin et al., 2018)s
5	GROOT	A pipeline for antibiotic resistance gene profiling from metagenomic datasets.	https://github.com/will-rowe/groot (Rowe and Winn, 2018)
6	HMD-ARG	A web tool to find possible antimicrobial resistance gene from metagenomic contigs.	http://www.cbrc.kaust.edu.sa/HMDARG/ (Li et al., 2021)
7	NanoARG	A pipeline to detect the antibiotic resistance genes from the Oxford Nanopore sequenced metagenome data.	https://github.com/gaarangoa/nanoARG (Arango-Argoty et al., 2019)
8	PathoFact	A simple-to-use, modular, and repeatable programme that can accurately predict virulence factors, bacterial toxins, and antibiotic resistance genes from metagenomic data.	https://git-r3lab.uni.lu/laura.denies/PathoFact/ (de Nies et al., 2021)
9	ResCap	A pipeline for in-depth resistome profiling from targeted metagenomics data.	https://github.com/valflanza/ResCap (Lanza et al., 2018)
10	ShortBRED	A system for profiling protein families of interest at very high specificity in shotgun metagenomics sequencing data.	https://huttenhower.sph.harvard.edu/shortbred/ (Kaminski et al., 2015)

Conclusion

Advancements in whole genome sequencing and bioinformatics has revolutionized the way of genome assembly, annotations, and drug discovery process. Numerous bioinformatics tools with its strengths and weaknesses are available for genome bacterial antimicrobial resistance profiling. This review provides a comprehensive list of widely used bioinformatics tools and databases, covering a wide range of techniques starting from genome assembly to drug designing. An overview of computational approaches to predict antimicrobial resistance and pros and cons will help the researchers to plan their research, analyse the genomic data, interpret their findings, and form valid conclusions.

Acknowledgments

The authors would like to thank Manipal Academy of Higher Education (MAHE), and the Department of Biotechnology (DBT-BUILDER scheme), Government of India for the infrastructure and support. Ms. Debyani gratefully acknowledges MAHE, Manipal for the Dr. TMA Pai Ph.D. fellowship.

Author Disclosure Statement

The authors declare they have no conflicting financial interests.

Funding Information

No funding was received by the authors of this article.

References

Agüero F, Al-Lazikani B, Aslett M, et al. Genomic-scale prioritization of drug targets: The TDR Targets database. Nat Rev Drug Discov 2008;7(11):900–907.

Akarsu H, Bordes P, Mansour M, et al. TASmania: A bacterial toxin-antitoxin systems database. PLoS Comput Biol 2019;15(4):e1006946.

Alcock BP, Raphenya AR, Lau TTY, et al. CARD 2020: Antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic Acids Res 2020;48(D1):D517–D525.

Anand S, Kuntal BK, Mohapatra A, et al. FunGeCo: a web-based tool for estimation of functional potential of bacterial genomes and microbiomes using gene context information. Bioinformatics 2020;36(8):2575–2577.

Andrews S. FastQC A Quality Control tool for High Throughput Sequence Data. 2010. Available from: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Arango-Argoty GA, Dai D, Pruden A, et al. NanoARG: a web service for detecting and contextualizing antimicrobial resistance genes from nanopore-derived metagenomes. Microbiome 2019;7(1):88.

Ariey F, Witkowski B, Amaratunga C, et al. A molecular marker of artemisinin-resistant Plasmodium falciparum malaria. Nature 2014;505(7481):50–55.

Aziz RK, Bartels D, Best A, et al. The RAST Server: Rapid annotations using subsystems technology. BMC Genomics 2008;9:1–15.

Bankevich A, Nurk S, Antipov D, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. J Comput Biol 2012;19(5):455–477.

Berry IM, Eyase F, Pollett S, et al. Global Outbreaks and Origins of a Chikungunya Virus Variant Carrying Mutations Which May Increase Fitness for *Aedes aegypti*: Revelations from the 2016 Mandera, Kenya Outbreak. Am J Trop Med Hyg 2019;100(5):1249–1257.

Besemer J and Borodovsky M. GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Res 2005;33:W451–W454.

Besser J, Carleton HA, Gerner-Smidt P, et al. Next-generation sequencing technologies and their application to the study and control of bacterial infections. Clin Microbiol Infect 2018;24(4):335–341.

Blanco J. Accumulation of Dinophysis Toxins in Bivalve Molluscs. Toxins (Basel) 2018;10(11):453.

Boetzer M and Pirovano W. SSPACE-LongRead: Scaffolding bacterial draft genomes using long read sequence information. BMC Bioinformatics 2014;15(1).

Boisvert S, Raymond F, Godzaridis É, et al. Ray Meta: Scalable de novo metagenome assembly and profiling. Genome Biol 2012;13(12):1–13; doi: 10.1186/gb-2012-13-12-r122.

Bolger AM, Lohse M and Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;30(15):2114–2120; doi: 10.1093/bioinformatics/btu170.

Bradley P, Gordon NC, Walker TM, et al. Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis*. Nat Commun 2015;6; doi: 10.1038/ncomms10063.

Cao H, Wang J, He L, et al. DeepDDG: Predicting the Stability Change of Protein Point Mutations Using Neural Networks. J Chem Inf Model 2019;59(4):1508–1514; doi: 10.1021/acs.jcim.8b00697.

Capriotti E, Fariselli P and Casadio R. I-Mutant2.0: Predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res 2005;33:W306–W310; doi: 10.1093/nar/gki375.

Chakraborty A, Ghosh S, Chowdhary G, et al. DBETH: A database of bacterial exotoxins for human. Nucleic Acids Res 2012;40(D1):D615–D620; doi: 10.1093/nar/gkr942.

Chakraborty N, Srinivasan S, Yang R, et al. Comparison of Transcriptional Signatures of Three Staphylococcal Superantigenic Toxins in Human Melanocytes. Biomedicines 2022;10(6):1402; doi: 10.3390/biomedicines10061402.

Chan PP, Lin BY, Mak AJ, et al. tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. Nucleic Acids Res 2021;49(16):9077–9096; doi: 10.1093/nar/gkab688.

Chen L, Yang J, Yu J, et al. VFDB: A reference database for bacterial virulence factors. Nucleic Acids Res 2005;33:D325–D328; doi: 10.1093/nar/gki008.

Cheng H, Concepcion GT, Feng X, et al. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods 2021;18(2):170–175; doi: 10.1038/s41592-020-01056-5.

Chevreux B, Pfisterer T, Drescher B, et al. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. Genome Res 2004;14(6):1147–1159; doi: 10.1101/gr.1917404.

Chikhi R and Rizk G. Space-efficient and exact de Bruijn graph representation based on a Bloom filter. Algorithms Mol Biol 2013;8(1):1–9; doi: 10.1186/1748-7188-8-22.

Chin CS, Peluso P, Sedlazeck FJ, et al. Phased diploid genome assembly with single-molecule real-time sequencing. Nat Methods 2016;13(12):1050–1054; doi: 10.1038/nmeth.4035.

Chiu JKH and Ong RTH. ARGDIT: A validation and integration toolkit for Antimicrobial Resistance Gene Databases. Bioinformatics 2019;35(14):2466–2474; doi: 10.1093/bioinformatics/bty987.

Choi Y, Sims GE, Murphy S, et al. Predicting the Functional Effect of Amino Acid Substitutions and Indels. PLoS One 2012;7(10):e46688.

Chowdhury AS, Call DR and Broschat SL. PARGT: a software tool for predicting antimicrobial resistance in bacteria. Sci Rep 2020;10(1):1–7.

Chowdhury F, Rahman MA, Begum YA, et al. Impact of Rapid Urbanization on the Rates of Infection by *Vibrio cholerae* O1 and Enterotoxigenic *Escherichia coli* in Dhaka, Bangladesh. PLoS Negl Trop Dis 2011;5(4):e999.

Davis AP, Wiegers TC, Johnson RJ, et al. Comparative Toxicogenomics Database (CTD): update 2023. Nucleic Acids Res 2023;51(D1):D1257–D1262.

de Man TJB and Limbago BM. SSTAR, a Stand-Alone Easy-To-Use Antimicrobial Resistance Gene Predictor. mSphere 2016;1(1).

de Nies L, Lopes S, Busi SB, et al. PathoFact: a pipeline for the prediction of virulence factors and antimicrobial resistance genes in metagenomic data. Microbiome 2021;9(1):49.

Delcher AL, Harmon D, Kasif S, et al. Improved microbial gene identification with GLIMMER. Nucleic Acids Res 1999;27(23):4636–4641.

DePristo, M. A., Banks, E., Poplin, R., et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 2011;43(5):491–498.

Deurenberg RH, Bathoorn E, Chlebowicz MA, et al. Application of next generation sequencing in clinical microbiology and infection prevention. J Biotechnol 2017;243:16–24.

Didelot X, Bowden R, Wilson DJ, et al. Transforming clinical microbiology with bacterial genome sequencing. Nat Rev Genet 2012;13(9):601–612.

Eschlböck S, Wenning G and Fanciulli A. Evidence-based treatment of neurogenic orthostatic hypotension and related symptoms. J Neural Transm 2017;124(12):1567–1605.

Fanelli U, Pappalardo M, Chinè V, et al. Role of Artificial Intelligence in Fighting Antimicrobial Resistance in Pediatrics. Antibiotics 2020;9(11):767.

Feldgarden M, Brover V, Haft DH, et al. Using the NCBI AMRFinder Tool to Determine Antimicrobial Resistance Genotype-Phenotype Correlations Within a Collection of NARMS Isolates. bioRxiv 2019;550707.

Forslund K, Sunagawa S, Kultima JR, et al. Country-specific antibiotic use practices impact the human gut resistome. Genome Res 2013;23(7):1163–1169.

Gaulton A, Bellis LJ, Bento AP, et al. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 2012;40(D1):D1100–D1107.

Ghosh A, Saran N and Saha S. Survey of drug resistance associated gene mutations in *Mycobacterium tuberculosis*, ESKAPE and other bacterial species. Sci Rep 2020;10(1):1–11.

Ghosh TS, Gupta S Sen, Nair GB, et al. In Silico Analysis of Antibiotic Resistance Genes in the Gut Microflora of Individuals from Diverse Geographies and Age-Groups. PLoS One 2013;8(12):e83823.

Gibson MK, Forsberg KJ and Dantas G. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. ISME J 2015;9(1):207–216.

Gnerre S, MacCallum I, Przybylski D, et al. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci U S A 2011;108(4):1513–1518.

Greenfield RA, Slater LN, Bronze MS, et al. Microbiological, Biological, and Chemical Weapons of Warfare and Terrorism. Am J Med Sci 2002;323(6):326–340.

Günther S, Kuhn M, Dunkel M, et al. SuperTarget and Matador: Resources for exploring drug-target relationships. Nucleic Acids Res 2008;36:D919–D922.

Gupta SK, Padmanabhan BR, Diene SM, et al. ARG-annot, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. Antimicrob Agents Chemother 2014;58(1):212–220.

Gurevich A, Saveliev V, Vyahhi N, et al. QUAST: Quality assessment tool for genome assemblies. Bioinformatics 2013;29(8):1072–1075.

Guzmán-Gómez G, Ayala Valdovinos MA, Cabrera-Díaz E, et al. Frequency of *Salmonella* and *Listeria monocytogenes* in Five Commercial Brands of Chicken Eggs Using a Combined Method of Enrichment and Nested-PCR. J Food Prot 2013;76(3):429–434.

Hasman H, Clausen PTLC, Kaya H, et al. LRE-Finder, a Web tool for detection of the 23S rRNA mutations and the optrA, cfr, cfr(B) and poxtA genes encoding linezolid resistance in enterococci from whole-genome sequences. J Antimicrob Chemother 2019;74(6):1473–1476.

Hecht M, Bromberg Y and Rost B. Better prediction of functional effects for sequence variants. BMC Genomics 2015;16(8):1–12.

Hendriksen RS, Munk P, Njage P, et al. Global monitoring of antimicrobial resistance based on metagenomics analyses of urban sewage. Nat Commun 2019;10(1):1124.

Hernandez D, François P, Farinelli L, et al. De novo bacterial genome sequencing: Millions of very short reads assembled on a desktop computer. Genome Res 2008;18(5):802–809.

Hoede C, Arnoux S, Moisset M, et al. PASTEC: An Automatic Transposable Element Classification Tool. PLoS One 2014;9(5):e91929.

Holt C and Yandell M. MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 2011;12(1):1–14.

Hossain M, Azimi N and Skiena S. Crystallizing short-read assemblies around seeds. BMC Bioinformatics 2009;10:1–12.

Huang X and Madan A. CAP3: A DNA sequence assembly program. Genome Res 1999;9(9):868–877.

Huerta-Cepas J, Szklarczyk D, Heller D, et al. EggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Res 2019;47(D1):D309–D314.

Hunt M, Mather AE, Sánchez-Busó L, et al. ARIBA: Rapid antimicrobial resistance genotyping directly from sequencing reads. Microb Genomics 2017;3(10).

Inouye M, Dashnow H, Raven LA, et al. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. Genome Med 2014;6(11):1–16.

Jackman SD, Vandervalk BP, Mohamadi H, et al. ABySS 2.0: Resource-efficient assembly of large genomes using a Bloom filter. Genome Res 2017;27(5):768–777.

Ji B, Pi W, Liu W, et al. HyperVR: a hybrid deep ensemble learning approach for simultaneously predicting virulence factors and antibiotic resistance genes. NAR Genomics Bioinforma 2023;5(1):lqad012.

Kaminski J, Gibson MK, Franzosa EA, et al. High-Specificity Targeted Functional Profiling in Microbial Communities with ShortBRED. PLoS Comput Biol 2015;11(12):e1004557.

Kanehisa M, Sato Y and Morishima K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. J Mol Biol 2016;428(4):726–731.

Keegan KP, Glass EM and Meyer F. MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function. In: Microbial Environmental Genomics (MEG). Methods in Molecular Biology Springer: New York, NY; 2016; pp. 207–233.

Kim J, Greenberg DE, Pifer R, et al. VaMPR: VAriant Mapping and Prediction of antibiotic resistance via explainable features and machine learning. PLoS Comput Biol 2020;16(1):e1007511.

Kim KS. Acute bacterial meningitis in infants and children. Lancet Infect Dis 2010;10(1):32–42.

Koren S, Schatz MC, Walenz BP, et al. Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nat Biotechnol 2012;30(7):693–700.

Koren S, Treangen TJ and Pop M. Bambus 2: scaffolding metagenomes. Bioinformatics 2011;27(21):2964–2971.

Koren S, Walenz BP, Berlin K, et al. Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. Genome Res 2017;27(5):722–736.

Krehenwinkel H, Pomerantz A and Prost S. Genetic Biomonitoring and Biodiversity Assessment Using Portable Sequencing Technologies: Current Uses and Future Directions. Genes (Basel) 2019;10(11):858.

Kuhn M, Szklarczyk D, Pletscher-Frankild S, et al. STITCH 4: integration of protein–chemical interactions with user data. Nucleic Acids Res 2014;42(D1):D401–D407.

Kumar P, Henikoff S and Ng PC. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 2009;4(7):1073–1082.

Kuzmenkov AY, Trushin I V, Vinogradova AG, et al. AMRmap: An Interactive Web Platform for Analysis of Antimicrobial Resistance Surveillance Data in Russia. Front Microbiol 2021;12:620002.

Laimer J, Hiebl-Flach J, Lengauer D, et al. MAESTROweb: A web server for structure-based protein stability prediction. Bioinformatics 2016;32(9):1414–1416.

Lakin SM, Dean C, Noyes NR, et al. MEGARes: An antimicrobial resistance database for high throughput sequencing. Nucleic Acids Res 2017;45(D1):D574–D580.

Langmead B and Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;9(4):357–359.

Lanza VF, Baquero F, Martínez JL, et al. In-depth resistome analysis by targeted metagenomics. Microbiome 2018;6(1):1–14.

Li D, Liu CM, Luo R, et al. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 2015;31(10):1674–1676.

Li H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 2011;27(21):2987–2993.

Li H. Minimap and miniasm: Fast mapping and de novo assembly for noisy long sequences. Bioinformatics 2016;32(14):2103–2110.

Li Y, Xu Z, Han W, et al. HMD-ARG: hierarchical multi-task deep learning for annotating antibiotic resistance genes. Microbiome 2021;9(1):40.

Liang X, Akers K, Keenum I, et al. AgroSeek: a system for computational analysis of environmental metagenomic data and associated metadata. BMC Bioinformatics 2021;22(1):117.

Lim E, Pon A, Djoumbou Y, et al. T3DB: a comprehensively annotated database of common toxins and their targets. Nucleic Acids Res 2010;38:D781–D786.

Liu H, Wu S, Li A, et al. SMARTdenovo: a de novo assembler using long noisy reads. Gigabyte 2021;2021:1–9.

Luo R, Liu B, Xie Y, et al. SOAPdenovo2: An empirically improved memory-efficient short-read de novo assembler. Gigascience 2012;1(1).

Matthews TC, Bristow FR, Griffiths EJ, et al. The Integrated Rapid Infectious Disease Analysis (IRIDA) Platform. bioRxiv 2018;381830.

Murray CJL, Ikuta KS, Sharara F, et al. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. Lancet 2022;399(10325):629–655.

Naas T, Oueslati S, Bonnin RA, et al. Beta-lactamase database (BLDB)–structure and function. J Enzyme Inhib Med Chem 2017;32(1):917–919.

Nickel J, Gohlke BO, Erehman J, et al. SuperPred: Update on drug classification and target prediction. Nucleic Acids Res 2014;42(W1):W26–W31.

O'Neill J. Tackling drug-resistant infections globally: final report and recommendations. 2016.

Olayan RS, Ashoor H and Bajic VB. DDR: Efficient computational method to predict drug-Target interactions using graph mining and machine learning approaches. Bioinformatics 2018;34(7):1164–1173.

Pal C, Bengtsson-Palme J, Rensing C, et al. BacMet: Antibacterial biocide and metal resistance genes database. Nucleic Acids Res 2014;42(D1):D737–D743.

Partridge SR and Tsafnat G. Automated annotation of mobile antibiotic resistance in Gram-negative bacteria: The Multiple Antibiotic Resistance Annotator (MARA) and database. J Antimicrob Chemother 2018;73(4):883–890.

Pawson AJ, Sharman JL, Benson HE, et al. The IUPHAR/BPS Guide to PHARMACOLOGY: An expert-driven knowledgebase of drug targets and their ligands. Nucleic Acids Res 2014;42(D1):D1098–D1106.

Pires DE V, Ascher DB and Blundell TL. DUET: A server for predicting effects of mutations on protein stability using an integrated computational approach. Nucleic Acids Res 2014;42(W1):W314–W319.

Prjibelski A, Antipov D, Meleshko D, et al. Using SPAdes De Novo Assembler. Curr Protoc Bioinforma 2020;70(1):1–29.

Pryszcz LP and Gabaldón T. Redundans: An assembly pipeline for highly heterozygous genomes. Nucleic Acids Res 2016;44(12):e113.

Qin C, Zhang C, Zhu F, et al. Therapeutic target database update 2014: A resource for targeted therapeutics. Nucleic Acids Res 2014;42(D1):D1118–D1123.

Quan L, Lv Q and Zhang Y. STRUM: Structure-based prediction of protein stability changes upon single-point mutation. Bioinformatics 2016;32(19):2936–2946.

Quesneville H, Bergman CM, Andrieu O, et al. Combined Evidence Annotation of Transposable Elements in Genome Sequences. PLoS Comput Biol 2005;1(2):e22.

Quevillon E, Silventoinen V, Pillai S, et al. InterProScan: protein domains identifier. Nucleic Acids Res 2005;33:W116–W120.

Rodrigues CHM, Pires DE V and Ascher DB. DynaMut: Predicting the impact of mutations on protein conformation, flexibility and stability. Nucleic Acids Res 2018;46(W1):W350–W355.

Rowe WPM and Winn MD. Indexed variation graphs for efficient and accurate resistome profiling. Bioinformatics 2018;34(21):3601–3608.

Ruan J and Li H. Fast and accurate long-read assembly with wtdbg2. Nat Methods 2020;17(2):155–158.

Ruiz-Perez CA, Conrad RE and Konstantinidis KT. MicrobeAnnotator: a user-friendly, comprehensive functional annotation pipeline for microbial genomes. BMC Bioinformatics 2021;22(1):11.

Ruppé E, Ghozlane A, Tap J, et al. Prediction of the intestinal resistome by a three-dimensional structure-based method. Nat Microbiol 2019;4(1):112–123.

Saha CK, Sanches Pires R, Brolin H, et al. FlaGs and webFlaGs: discovering novel biology through the analysis of gene neighbourhood conservation. Bioinformatics 2021;37(9):1312–1314.

Sallet E, Gouzy J and Schiex T. EuGene: An automated integrative gene finder for eukaryotes and prokaryotes. In: Methods in Molecular Biology Humana Press Inc.; 2019; pp. 97–120.

Savojardo C, Fariselli P, Martelli PL, et al. INPS-MD: A web server to predict stability of protein variants from sequence and structure. Bioinformatics 2016;32(16):2542–2544.

Schmidt K, Mwaigwisya S, Crossman LC, et al. Identification of bacterial pathogens and antimicrobial resistance directly from clinical urines by nanopore-based metagenomic sequencing. J Antimicrob Chemother 2017;72(1):104–114.

Seemann T. Abricate: mass screening of contigs for antimicrobial and virulence genes; 2021. Available from: https://github.com/tseemann/abricate.

Seemann T. Prokka: Rapid prokaryotic genome annotation. Bioinformatics 2014;30(14):2068–2069.

Sevin EW and Barloy-Hubler F. RASTA-Bacteria: A web-based tool for identifying toxin-antitoxin loci in prokaryotes. Genome Biol 2007;8(8):1–14.

Shankarnarayan SA, Guthrie JD, Charlebois DA, et al. Machine Learning for Antimicrobial Resistance Research and Drug Development. In: The Global Antimicrobial Resistance Epidemic - Innovative Approaches and Cutting-Edge Solutions IntechOpen; 2022.

Simpson JT and Durbin R. Efficient de novo assembly of large genomes using compressed data structures. Genome Res 2012;22(3):549–556.

Solomon SL and Oliver KB. Antibiotic Resistance Threats in the United States: Stepping Back from the Brink. Am Fam Physician 2014;89(12):938–941.

Tanizawa Y, Fujisawa T and Nakamura Y. DFAST: A flexible prokaryotic genome annotation pipeline for faster genome publication. Bioinformatics 2018;34(6):1037–1039.

Tatusova T, Dicuccio M, Badretdin A, et al. NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res 2016;44(14):6614–6624.

Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 2013;14(2):178–192.

Törönen P, Medlar A and Holm L. PANNZER2: a rapid functional annotation web server. Nucleic Acids Res 2018;46(W1):W84–W88.

Tourasse NJ and Darfeuille F. T1TAdb: The database of type I toxin-antitoxin systems. RNA 2021;27(12):1471–1481.

Treangen TJ, Koren S, Sommer DD, et al. MetAMOS: A modular and open source metagenomic assembly and analysis pipeline. Genome Biol 2013;14(1).

Tritt A, Eisen JA, Facciotti MT, et al. An Integrated Pipeline for de Novo Assembly of Microbial Genomes. PLoS One 2012;7(9):e42304.

Tunstall T, Portelli S, Phelan J, et al. Combining structure and genomics to understand antimicrobial resistance. Comput Struct Biotechnol J 2020;18:3377–3394.

Varshney RK, Nayak SN, May GD, et al. Next-generation sequencing technologies and their implications for crop genetics and breeding. Trends Biotechnol 2009;27(9):522–530.

Vaser R, Sović I, Nagarajan N, et al. Fast and accurate de novo genome assembly from long uncorrected reads. Genome Res 2017;27(5):737–746.

Walker BJ, Abeel T, Shea T, et al. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. PLoS One 2014;9(11):e112963.

Walker GT, Quan J, Higgins SG, et al. Predicting Antibiotic Resistance in Gram-Negative Bacilli from Resistance Genes. Antimicrob Agents Chemother 2019;63(4):10.1128/aac.02462-18.

Wallace JC, Port JA, Smith MN, et al. FARME DB: A functional antibiotic resistance element database. Database 2017;2017(1).

Wang S, Zhao C, Yin Y, et al. A Practical Approach for Predicting Antimicrobial Phenotype Resistance in *Staphylococcus aureus* Through Machine Learning Analysis of Genome Data. Front Microbiol 2022;13.

Wangai FK, Masika MM, Lule GN, et al. Bridging antimicrobial resistance knowledge gaps: The East African perspective on a global problem. PLoS One 2019;14(2):e0212131.

Warren RL, Sutton GG, Jones SJM, et al. Assembling millions of short DNA sequences using SSAKE. Bioinformatics 2007;23(4):500–501.

Wattam AR, Davis JJ, Assaf R, et al. Improvements to PATRIC, the all-bacterial bioinformatics database and analysis resource center. Nucleic Acids Res 2017;45(D1):D535–D542.

Wei Z, Wu Y, Feng K, et al. ARGA, a pipeline for primer evaluation on antibiotic resistance genes. Environ Int 2019;128:137–145.

Wishart DS, Knox C, Guo AC, et al. DrugBank: A knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res 2008;36:D901–D906.

Witvliet DK, Strokach A, Giraldo-Forero AF, et al. ELASPIC web-server: Proteome-wide structure-based prediction of mutation effects on protein stability and binding affinity. Bioinformatics 2016;32(10):1589–1591.

Woolhouse M, Ward M, van Bunnik B, et al. Antimicrobial resistance in humans, livestock and the wider environment. Philos Trans R Soc B Biol Sci 2015;370(1670):20140083.

Wu Z, Che Y, Dang C, et al. Nanopore-based long-read metagenomics uncover the resistome intrusion by antibiotic resistant bacteria from treated wastewater in receiving water body. Water Res 2022;226:119282.

Xavier BB, Das AJ, Cochrane G, et al. Consolidating and exploring antibiotic resistance gene data resources. J Clin Microbiol 2016;54(4):851–859.

Xiao C Le, Chen Y, Xie SQ, et al. MECAT: Fast mapping, error correction, and de novo assembly for single-molecule sequencing reads. Nat Methods 2017;14(11):1072–1074.

Xie Y, Wei Y, Shen Y, et al. TADB 2.0: An updated database of bacterial type II toxin-antitoxin loci. Nucleic Acids Res 2018;46(D1):D749–D753.

Yin X, Jiang X-T, Chai B, et al. ARGs-OAP v2.0 with an expanded SARG database and Hidden Markov Models for enhancement characterization and quantification of antibiotic resistance genes in environmental metagenomes. Bioinformatics 2018;34(13):2263–2270.

Zankari E, Allesøe R, Joensen KG, et al. PointFinder: A novel web tool for WGS-based detection of antimicrobial resistance associated with chromosomal point mutations in bacterial pathogens. J Antimicrob Chemother 2017;72(10):2764–2768.

Zankari E, Hasman H, Cosentino S, et al. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother 2012;67(11):2640–2644.

Zerbino DR and Birney E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 2008;18(5):821–829.

Zhang F, Lammi MJ, Shao W, et al. Cytotoxic Properties of HT-2 Toxin in Human Chondrocytes: Could T3 Inhibit Toxicity of HT-2? Toxins (Basel) 2019;11(11):667.

Zimin A V, Marçais G, Puiu D, et al. The MaSuRCA genome assembler. Bioinformatics 2013;29(21):2669–2677.

Chapter

7

Genomic Surveillance: Linking Omics Data for Pandemic Preparedness

Geanta, Marius, **Ankit Singh Tanwar**, Hans Lehrach, Kapaettu Satyamoorthy, and Angela Brand. "Horizon scanning: rise of planetary health genomics and digital twins for pandemic preparedness." *OMICS: A Journal of Integrative Biology* 26, no. 2 (2022): 93-100.

DOI: 10.1089/omi.2021.0062; **IF 3.9 (2021)**

Horizon Scanning: Rise of Planetary Health Genomics and Digital Twins for Pandemic Preparedness

Abstract

The Covid-19 pandemic accelerated research and development not only in infectious diseases but also in digital technologies to improve monitoring, forecasting, and intervening on planetary and ecological risks. In the European Commission, the Destination Earth (DestinE) is a current major initiative to develop a digital model of the Earth (a ‘‘digital twin’’) with high precision. Moreover, omics systems science is undergoing digital transformation impacting nearly all dimensions of the field, including real-time phenotype capture to data analytics using machine learning and artificial intelligence, to name but a few emerging frontiers. We discuss the ways in which the current ongoing digital transformation in omics offers synergies with digital twins/-DestinE. Importantly, we note here the rise of a new field of scholarship, planetary health genomics. We conclude that digital transformation in public and private sectors, digital twins/DestinE, and their convergence with omics systems science are poised to build robust capacities for pandemic preparedness and resilient societies in the 21st century.

Keywords:

Digital transformation, digital twins, public health genomics, SARS-CoV-2 sequencing, pandemic preparedness, genomic surveillance.

Introduction

Over the past decade, digital transformation in public health have evolved into a broader, planetary scale, scope, and relevance. The COVID-19 pandemic has further catapulted digital transformation around the world that is now taking off in a variety of large-scale planetary health initiatives. In the European Commission, the Destination Earth (DestinE) is a current and major initiative to develop a digital model of the Earth (a ‘‘digital twin’’) with a very high precision (https://digital-strategy.ec.europa.eu/en/library/destination-earth) (Fig. 1).

DestinE is part of a broader vision and growing awareness that the COVID-19 pandemic is a kind reminder for future planetary health crises looming on the horizon. Omics systems science offers veritable prospects for genomic surveillance of emerging pathogens, including new zoonosis threats. Omics is itself undergoing digital transformation impacting nearly all dimensions of the field, including realtime phenotype capture to data analytics using machine learning and artificial intelligence, to name but a few emerging frontiers.

We discuss the digital transformation in omics and the ways in which this offers synergies with digital twins. In addition, the present article introduces and highlights the rise of a new field of scholarship, planetary health genomics.

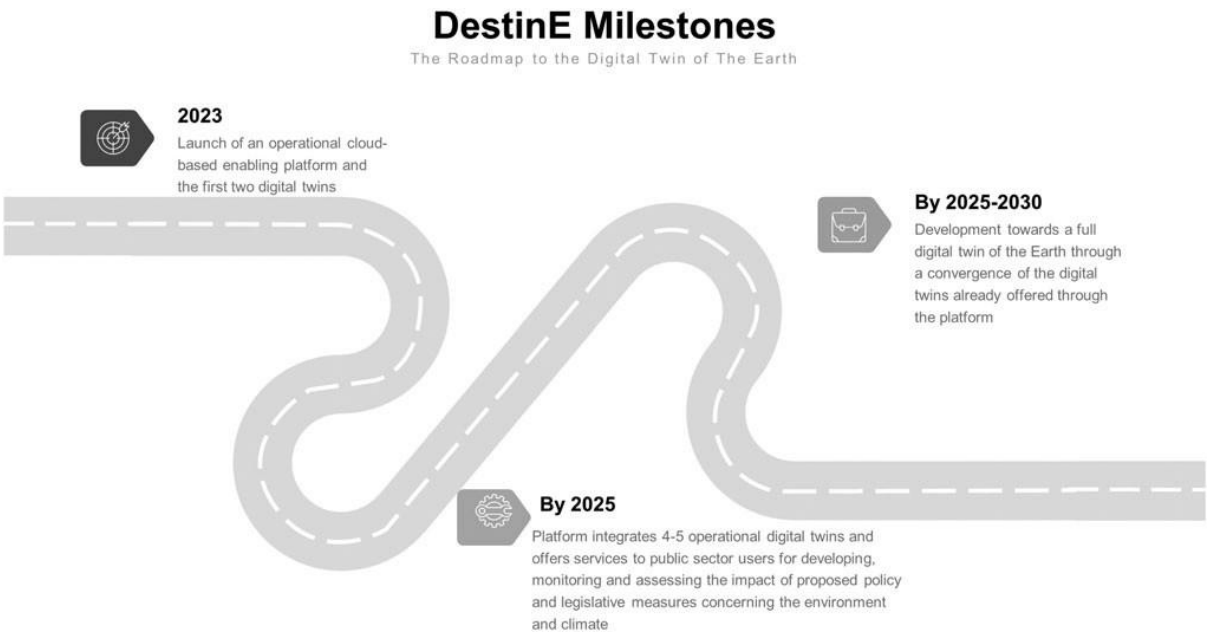


Figure 1: DestinE milestones—the road to digital twins of the Earth. Source of information: European Commission, 2021. DestinE, Destination Earth.

Digital Transformation in Omics from Phenomics to Data-to-Knowledge Trajectory

The Covid-19 pandemic is the most important public health crisis in the last century and one of the three major infectious outbreaks in the first decades of this century (World Health Organization, 2021a). An important characteristic of the Covid-19 pandemic was the early and widespread use of digital health technologies such as telemedicine platforms, contact tracing apps, wearables, genome sequencing, artificial intelligence and machine learning, genomic data sharing platforms, data dashboards, real-time and real-world data from mobile devices (including global positioning systems), electronic health records, disease and vaccination registries, e-prescriptions, and Internet of Things. These developments, in part, built on earlier efforts to incorporate digital technologies in public health (Evangelatos et al., 2020a).

Digital health technologies and extensive use of data have been incorporated into several layers of the health systems: implementation of public health measures for containment and mitigation, planning and tracking of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) variants, clinical management, and political and administrative decision. The implementation of digital health technologies was different from one country to another, ‘‘countries that have quickly deployed digital technologies to facilitate planning, surveillance, testing, contact tracing, quarantine, and clinical management have remained front-runners in managing disease burden’’ (Van Spall et al., 2020).

During the Covid-19 pandemic, the access to digital technologies has emerged as a fundamental determinant of access to health and care (Evangelatos et al., 2020b). Across the Globe, health systems and societies are in different stages of digital transformation, and those health systems already advanced in terms of digital health technologies implementation have had a prompt and efficient response. The risk of digital divide in the context

of Covid-19 should be discussed even more in relationship to omics and other emerging technologies (digital twins, machine learning).

Omics is impacting nearly all dimensions of the response to the Covid-19 pandemic, including single-cell multiomics analysis of the immune response in COVID-19 (Stephenson et al., 2021), large-scale multiomics analysis of COVID-19 severity, and multiomics approach in the identification of potential therapeutic biomolecule for COVID-19 (Singh et al., 2021).

From the public health perspective, the most important benefits of the omics systems science are related to the implementation of genomic surveillance of the SARS-CoV-2 aiming to early identify potential variants of interest or variants of concern of the virus. From the medical practice perspective, one of the most challenging but promising applications of the omics sciences is the real-time phenotype capture by Internet of Things followed by data analytics using machine learning and artificial intelligence.

The measurement of the phenotype for Covid-19 by multilevel proteomics analysis reveals the perturbations produced in the human body through the interaction of SARS-CoV-2 proteins and the host proteome, profiling the interactome of the virus as well as the influence of the SARS-CoV-2 on the transcriptome, proteome, ubiquitinome and phosphoproteome of a lung-derived human cell line (Stukalov et al., 2021).

Assessment of the proteome (a measure of phenome, the interaction between the human genome and the environment, including the viral proteins) of the Covid-19 patients and the use of machine learning led to the identification of disease progression and prognostic biomarkers and risk adapted treatment strategies, as well as a map linking clinical parameters usually used for diagnostic to proteome and their dynamics in an infectious disease (Demichev et al., 2021).

Real-world data are meant to capture in real-time different sets of characteristics of an individual (phenotypes). Traditionally, the real-world data included registries, electronic medical records, and self-reported individual data, but the current understanding of the real-world data incorporate more real-time objective measurement provided by wearable devices and other Internet of Things as well. The challenge is to transform the real-world real-time data in real-world realtime evidence and knowledge.

Digital transformation in omics began with the exponential development of next generation sequencing and other emerging technologies, which will increasingly be incorporated into Internet of Things (IoT) devices, allowing real-time and precise monitoring of parameters that currently can be measured rather in hospitals and specialized clinics.

‘‘Digital transformation is impacting every facet of science and society, not least because there is a growing need for digital services and products with the COVID-19 pandemic.’’ (Lin and Wu, 2021).

Digital transformation in omics has the potential to transform the doctor-patient relationship and to increase the inequalities. The response to the Covid-19 pandemic was and is still different from continent to continent, from country to country, and sometimes within the same country. Lack of a common response and the limited use of innovative tool such as omics and digital technologies contributed to the poor control of the pandemic

around the globe. Digital determinants of health have emerged over the past 2 years in particular as important pillars of preparedness for current and future ecological threats (Moon and Kickbusch, 2021; Özdemir, 2021).

Planetary Health in the Age of Pandemic

Planetary health is a concept focused on the interconnections and interdependence of human health, animal health, and environment (Haines, 2016). Planetary health has a marked emphasis on ecological determinants of health (Seltenrich, 2018).

Climate change is a trademark of the 21st century. For infectious diseases, climate change is a threat multiplier (Chan, 2017). Climate change expands the distribution of the existing infectious disease pathogens and creates the conditions for other pathogens to emerge. Disruption of forests, rapid urbanization, and population growth are driving zoonotic events simply by increasing close contact between people and animals (Waugh et al., 2020).

At least 30 new infectious agents affecting human population have emerged over the last 40 years, with 44% involving RNA viruses (Nii-Trebi, 2017). Between 631,000 and 827,000 unknown viruses might be zoonotic and thus have the potential to infect humans after spill over from host animal populations (Jonas and Seifman, 2019). Most of these new pathogens are zoonotic, significantly correlated to socioeconomic, environmental, and ecological factors. Some of these new pathogens, apart from the current SARS-CoV-2 virus causing Covid-19 pandemic, have demonstrated in the last 20 years the capacity to spread, infect large populations, overwhelm health systems, create economic difficulties, and increase morbidity and mortality.

In 2003, SARS (the coronavirus causing acute respiratory syndrome) emerged in China, spread to around 30 countries, and caused more than 8000 cases and around 774 deaths, with an economic impact of more than 40 billion dollars worldwide in only 6 months (Council on Foreign Relations, 2020). In 2012, Middle East respiratory syndrome coronavirus (MERS-CoV) (the virus causing Middle East Respiratory Syndrome) emerged in Saudi Arabia and spread to 27 countries in the region, caused 2519 cases and 866 deaths, with a very high mortality rate (33%) (National Institute of Allergy and Infectious Diseases, 2020).

Data suggested that both SARS-CoV and MERS-CoV originated in bats. SARS-CoV then spread from civets to people, while MERS-CoV spread from dromedary camels to humans.

Research indicated for many years that bats host a significantly higher proportion of dangerous viruses than other mammals (Olival et al., 2017). This scientific data should have already led to the development of extensive viral surveillance programs of bats populations in the hotspot areas with the objective to contribute to the development of predictive models and to prevent future emergence of the virus. Established before the emergence of Covid-19 pandemic, the initiatives in biodiversity genomics as Bioscan, Earth Biogenome Project, and Global Virome Project had a limited role for Covid-19 prediction but could reach the potential of genomic and data-driven surveillance to prevent future pandemics.

Pandemic Preparedness

Established in 2005 by the World Health Organization, the International Health Regulations (IHR) provide the legal framework that defines the countries' rights and obligations in the context of public health emergencies such as pandemics. The IHR are an instrument of international law in the field of health. One hundred ninety-six countries, including 194 WHO Member States, are part of the IHR. WHO has the coordinating role in the implementation of IHR and help countries to build capacities to have the adequate response if a public health crisis arrives. States are responsible for implementing the IHR at national level. The IHR has required all the countries to build the capacities able to detect (surveillance systems), assess, report, and respond in a timely matter to any public health emergency of international concern. The goal of country implementation of IHR is to limit the spread of health risks to other countries and to prevent travel and trade restrictions.

WHO recommended that all Member States update their pandemic preparedness plans based on the lessons learnt from 2009 pandemic (H1N1), new evidence on the effectiveness of public health measures that has become available in the meantime, and an ongoing risk assessment (World Health Organization, 2019a). At the beginning of Covid-19 pandemic, only 15 out of 53 countries from the European Region of WHO have published revised national pandemic plans. None of the 15 revised national pandemic plans refers to the role of genomics in public health surveillance. The only ''omic'' mentioned in some pandemic plans was ''economic.'' (World Health Organization, 2019b).

Published jointly by WHO and European Centre for Disease Prevention and Control (ECDC), the document ''Key changes to pandemic plans by Member States of the WHO European Region based on lessons learnt from the 2009 pandemic'' include no reference to the genomics or other omics as a potential tool to be used in the pandemics control.

The IHR Review Committee declared in 2011: ''The world is ill-prepared to respond to a severe influenza pandemic or to any similarly global, sustained and threatening public-health emergency.'' According to World Health Organization, ''pandemic preparedness is a continuous process of planning, exercising, revising, and translating into action national and subnational pandemic preparedness and response plans. A pandemic plan is thus a living document, which is reviewed regularly and revised if necessary, for example, based on the lessons learnt from outbreaks or a pandemic, or from a simulation exercise.'' (World Health Organisation, 2019a).

The ''Interim progress report of the Review Committee on the functioning of the IHR (2005) during the COVID-19 response'' recognize that ''the COVID-19 pandemic has revealed significant gaps in pandemic preparedness in countries across the world, including in the areas of: surveillance, health systems, equipment and training, essential public health functions, and the role of national IHR focal points, emergency legislation, risk communication, and coordination.'' (World Health Organization, 2021b).

The Interim Progress report recognize the role of the genomics for pandemic response and public health surveillance: ''In order to engage the global scientific community in these response efforts, it is critical that pathogens, their genomic sequence, and relevant clinical samples be rapidly made available to the global medical research community.'' (World Health Organization [2021a]).

The Use of Genomics to Control the Infectious Diseases Outbreaks

SARS-CoV-2 represent the third emergence of a coronavirus outbreak in the last 20 years, after SARS-CoV and MERS-CoV, both have been known to cause lung disease in humans. To control the previous pandemics and to prevent future outbreaks with SARS-CoV and MERS, WHO provided recommendations on the surveillance systems both for animal and human health.

In 2018, the Interim Guidance of WHO on MERS-CoV specifies: ‘‘Specimens testing positive for MERS-CoV should be genetically sequenced, and the data uploaded to publicly accessible databases. If the laboratory doing the initial test does not have the capacity for genetic sequencing, an aliquot of the specimen should be forwarded to a reference centre. Such centres should attempt to isolate viruses from all cases so that whole genome sequencing can be performed, either in the national or international reference laboratory.’’ (World Health Organization, 2018).

Genomic sequencing is used to determine if the dromedary camels from North Africa and the Middle East, who are the natural reservoirs of MERS-CoV, are infected with the virus and enable a more precise approach on understanding the epidemiology and the viral dynamics (Kandeil et al., 2019). The use of genomic data could help to have a better understanding of the transmission history, of the hosts (animals and humans), natural reservoirs, and viral transmission enabling effective public health interventions, and prevention of future outbreaks.

Africa has to face a complicated landscape when it comes to infectious diseases (World Health Organization, 2020). Over the past decade, Africa experienced two Ebola virus epidemics (Delamou et al., 2017), and overall more than 140 infectious diseases outbreaks are reported every year (WHO, 2018). In this context, in some African countries, genomic surveillance systems for Lassa fever (Nigeria—Siddle et al., 2018) or Ebola virus (Democratic Republic of Congo— Mbala-Kingebeni et al., 2019) were implemented. For the implementation of the genomic surveillance of the HIV and subsequent data-driven strategies, the Phylogenetic and Networks for Generalized Epidemics in Africa (PANGEA— Abeler-Dörner et al., 2019) aiming to ‘‘guide targeted prevention efforts through characterizing transmission dynamics and identifying unrecognized clusters and untreated individuals who are probable drivers of HIV transmission’’ (Inzaule et al., 2021) was built.

The Role of Genomics in the Covid-19 Pandemic

According to the European Centre for Disease Prevention and Control, ‘‘sequencing of (partial) genes and whole genomes (WGS) has been proven as a powerful method to investigate viral pathogen genomes, understand outbreak transmission dynamics and spill-over events and screen for mutations that potentially have an impact on transmissibility, pathogenicity, and/or countermeasures (e.g., diagnostics, antiviral drugs and vaccines). The results are key to informing outbreak control decisions in public health’’ (European Center for Disease Control, 2021).

In the Covid-19 pandemics, genomics has been used so far for rapid identification and initial characterization of the virus, for the development of diagnostic tests and the development of the vaccines (Table 1). The

genomic surveillance leads also to the identification of new concerning strains of SARS-CoV-2 such as B.1.1.7, B.1.351, or P.1. Within a year of the initial identification of SARS-CoV-2, more than 450,000 full genome sequences have been shared through public database GISAID (van Noorden, 2021). The world leader in the area of genomic surveillance of SARS-CoV-2 is United Kingdom with a total of more than 350,000 samples sequenced (COG Consortium UK).

Table 1: European Centre for Disease Prevention and Control—Objectives of SARS-CoV-2 Genomic Sequencing

Main objectives	Early detection and characterization of emerging variant viruses to define if they are of particular concern; Assessing the impact of genetic and antigenic variant viruses for the pandemic and monitoring them over time to guide public health action.
Specific objectives	Investigating virus transmission dynamics and introductions of novel genetic variants; Modeling the antigenic properties of the virus to assess the risk of vaccine escape; Selecting viruses for vaccine composition; Assessing the impact of mutations on the performance of molecular diagnostic, antigen characterization and serological methods; Investigating the relationship between clades/lineages and epidemiological data such as transmissibility and disease severity or risk groups; Understanding the impact of response measures on the virus population; Assessing relatedness of viral strains within epidemiological clusters and supporting contact tracing and other public health interventions; Assessing and confirming reinfections; Monitoring emerging lineages within wild/domestic/farmed animal populations that may impact human health; Prompting further basic research investigation to confirm the relevance of observed mutations in the pathogenesis of the disease (e.g., infectivity, receptors binding); Assessing the impact of mutations on the performance of antiviral drugs; Assessing the potential incidence of vaccine-derived virus infections and transmissions should live SARS-CoV-2 vaccines become available.

SARS-CoV-2, severe acute respiratory syndrome coronavirus 2.

The Use of Genomic Information to Inform Public Health Decisions at EU Level

Published in 2012 as part of the Public Health Genomics Network Europe II Project (Brand, 2012). The *European Best Practice Guidelines for Quality Assurance, Provision, and Use of Genome-based Information and Technologies (GBIT)* provided concrete recommendations to the European Commission and Member States for the best use of Genome-Based Information and Technologies (Table 2).

Table 2: The European Best Practice Guidelines for Quality Assurance, Provision, and Use of Genome-Based Information and Technologies (Part of the Recommendations)

Develop new strategies for the use of GBIT within the public domain.
Identify and assess new GBIT for proactive and timely decision-making.
Maintain an infrastructure for the sharing of GBIT data in public health settings.
Improve the implementation of promising GBIT under conditions that safeguard the wellbeing of citizens
Support health policies related to GBIT based on good governance and trust.
Integrate GBIT into the professional training and life-long learning curricula of health professionals

GBIT, Genome-based Information and Technologies.

Despite this conceptual framework, the use of genomics in the Covid-19 pandemic differs from one Member State to another, with Denmark as European leader. Most of the EU Member States did not implement genomic

surveillance systems for infectious diseases. Denmark increased the surveillance and analytics capacity to sequence all the samples at the beginning of 2021 when the new variant B.117 threatened to become dominant.

As a consequence, both in Denmark and in United Kingdom, the genomic data informed in real-time the public health interventions implemented by the Governments to control the outbreak at the beginning of 2021.

To control the Covid-19 pandemic, European Union must be ready and prepared for the possibility of future variants being more or fully resistant to existing vaccines. European Commission announced the launch of ‘‘HERA Incubator: Anticipating together the threat of COVID-19 variants,’’ ‘‘a new bio-defence preparedness plan, to access and mobilize all means and resources necessary to prevent, mitigate, and respond to the potential impact of variants.’’ HERA Incubator ‘‘will serve as the vanguard for the European Health Emergency Preparedness and Response Authority. HERA Incubator will work closely with the European Centre for Disease Prevention and Control (ECDC) to ensure that Member States have sufficient sequencing capacities and access to sequencing support services. HERA Incubator and ECDC will standardize sequencing procedures so that the data are comparable.’’ (European Commission, 2021a).

SARS-CoV-2 genome sequencing, data sharing, and analytics should become a priority for public health policies at EU level and beyond. The free movement of citizens in the European Union is followed by the free movement of different variants of SARS-CoV-2. In this context, a common approach of all EU Member States on the genomic surveillance of SARS-CoV-2 could contribute to a better control and finally to a shortening of the Covid-19 pandemic.

The European Health Data Space, one of the top priorities of the European Commission for 2019-2025, should include a *Centre for SARS-CoV-2 Genomics Data and Analytics* to collect, standardize and analyze sequencing data from all the EU Member States and to provide regular public health reports aiming to inform the decision-makers on the evolution of the Covid-19 pandemic. This pan-EU data-driven public health initiative will pave the way to more precise and coordinated public health interventions and could shorten the course of the Covid-19 pandemic. In addition to the sequencing of the samples collected from the Covid-19 patients, the regular surveillance of wastewater could also add value.

Public health genomics, as we move forward for pandemic preparedness, will likely expand toward planetary health genomics, signalling a broadening in scope of public health genomics from public to planetary health, and emphasis on ecological determinants of health.

Digital Twin of the Earth to Prevent Future Pandemics

DestinE will be launched in 2021 by the European Commission in the context of Green Deal and Digital Strategy aiming to ‘‘unlock the potential of digital modelling of the Earth’s physical resources and related phenomena such as climate change, water/marine environments, polar areas and the cryosphere, etc. on a global scale to speed up the green transition and help plan for major environmental degradation and disasters.

At the heart of Destination Earth will be a federated cloud-based modelling and simulation platform, providing access to data, advanced computing infrastructure (including high-performance computing), software, AI

applications, and analytics. It will integrate digital twins—digital replicas of various aspects of the Earth system, such as weather forecasting and climate change, food and water security, global ocean circulation and the biogeochemistry of the oceans, and more– giving users access to thematic information, services, models, scenarios, simulations, forecasts, and visualizations. The platform will enable application development and the integration of users' own data'' (European Commission, 2021b).

A high-precision digital model of the Earth is aimed to be developed by the DestinE to monitor and simulate natural and human activity. A digital twin is a digital replica of a living or nonliving physical entity. Digital twins use multiple data sources and rely on the integration of continuous observation, modeling and high-performance simulation, resulting in highly accurate predictions of future developments (Fig. 2).

''The digital twins created in DestinE will give expert and non-expert users tailored access to high-quality information, services, models, scenarios, forecasts and visualisations. This includes models of the climate, weather forecasting, hurricane evolution and more. Digital twins rely on the integration of continuous observation, modelling and high-performance simulation, resulting in highly accurate predictions of future developments.'' (European Commission, 2021b).

From the planetary health, pandemic preparedness, and omics perspective, a planetary digital twin based on omics data generated through the continuous monitoring of high-risk animals (e.g., bats) and human populations, and also other types of real-world real-time data, could enable the prediction of future events, early warnings, and precision planetary health interventions. Big data in both the public domain and the health care industry are growing rapidly, for example, with broad availability of next-generation sequencing and large-scale phenomics datasets on patient reported outcomes (Tanwar et al., 2021).

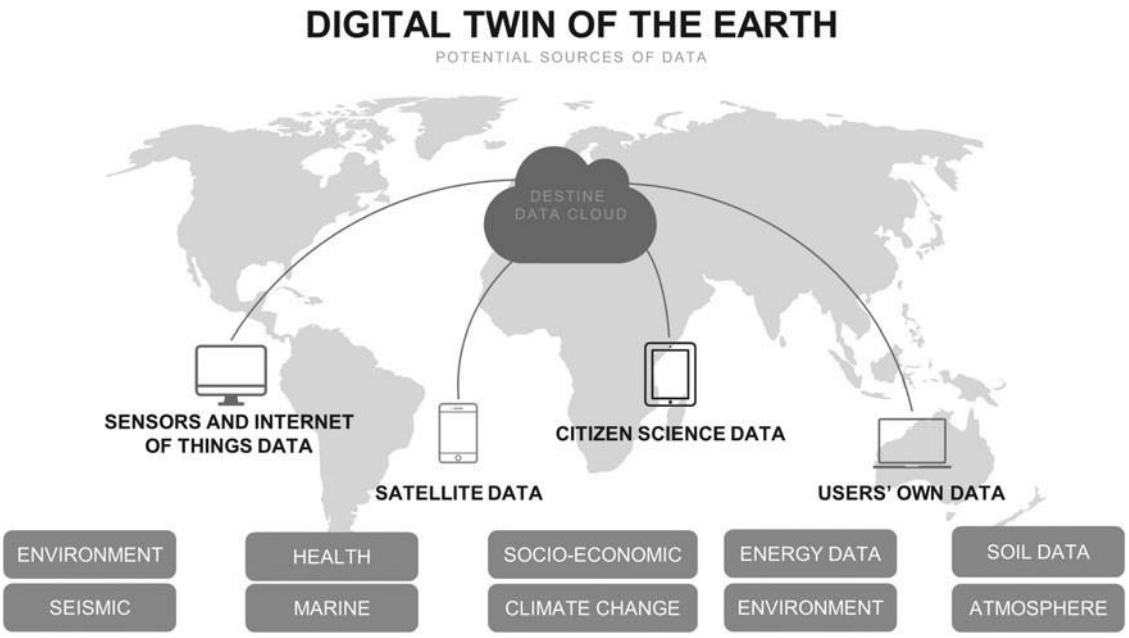


Figure 2: Digital twin of the Earth—potential sources of data. Source of information: European Commission, 2021.

Conclusions and Outlook

The Covid-19 pandemic is a planetary health issue and requires a global action based on data and new and efficient technologies. Planetary health needs to adopt a ‘‘one health’’ type approach that moves beyond an anthropocentric framework and considers the interdependencies among human, animal, and plant health. Genomic surveillance of the SARS-CoV-2 should become a key part of the pandemic response and planetary health genomics allowing the tracking of the new variants and implementation of precision planetary health interventions. For the future, the planetary digital twin could increase the resolution of the pandemic preparedness process and the capacity to predict and react to the next pandemics.

Acknowledgments

The authors thank Manipal Academy of Higher Education (MAHE) and especially the Manipal School of Life Sciences. This work would have not been possible without the support provided through the prestigious Dr. TMA Pai Endowment Chair of MAHE and the Scheme for Promotion of Academic and Research Collaboration (SPARC) project no. P1457. The authors thank the United Nations University—Maastricht Economic and Social Research Institute on Innovation and Technology (UNU-MERIT) and the Faculty of Health, Medicine, and Life Sciences (FHML) at Maastricht University for infrastructure and support.

Author Disclosure Statement

The authors declare they have no conflicting financial interests.

Funding Information

No funding was received for this article.

References:

Abeler-Dörner L, Grabowski MK, Rambaut A, Pillay D, Fraser C; and on behalf of the PANGEA consortium. (2019). PANGEA-HIV 2: Phylogenetics and networks for generalised epidemics in Africa. Curr Opin HIV AIDS 14, 173–180.

Brand A. (2012). Public health genomics and personalized healthcare: A pipeline from cell to society. Drug Metab Drug Interact 27, 121–123.

Brand A, Lal JA; and for the Public Health Genomics European Network (PHGEN II). (2012). European best practice guidelines for quality assurance, provision and use of genome based information and technologies: The 2012 declaration of Rome. Drug Metab Drug Interact 27, 177–182.

Chan M. (2017). Ten years in public health 2007-2017. https://apps.who.int/iris/bitstream/handle/10665/255355/9789241512442-eng.pdf;jsessionid=D214AD883B7A83249B596B4C0D120CEE? sequence=1. Accessed September 25, 2021.

Council on Foreign Relations. (2020). Independent task force report no. 78 improving pandemic preparedness lessons from COVID-19. https://www.cfr.org/report/pandemic-preparednesslessons- COVID-19/pdf/TFR_Pandemic_Preparedness.pdf. Accessed September 25, 2021.

Delamou A, Delvaux T, El Ayadi AM, et al. (2017). Public health impact of the 2014-2015 Ebola outbreak in West Africa: Seizing opportunities for the future. BMJ Glob Health 2, e000202.

Demichev, Tober-Lau P, Lemke O, et al. (2021). A timeresolved proteomic and prognostic map of COVID-19. Cell Syst 12, 780–794.

European Center for Disease Control. (2021). Sequencing of SARS-CoV-2—Technical guidance, 2021. https://www.ecdc .europa.eu/sites/default/files/documents/Sequencing-of-SARSCoV- 2-first-update.pdf. Accessed September 25, 2021.

European Commission. (2021a). Communication from the Commission to the European Parliament, the European Council and the Council; HERA Incubator: Anticipating together the threat of COVID-19 variants, 2021. https://ec.europa.eu/info/sites/ info/files/communication-hera-incubator-anticipating-threatcovid- 19-variants_en.pdf. Accessed September 25, 2021.

European Commission. (2021b). Destination Earth. https:// digital-strategy.ec.europa.eu/en/policies/destination-earth. Accessed September 25, 2021.

Evangelatos N, Özdemir V, and Brand A. (2020b). Blockchain for digital health: Prospects and challenges. OMICS 24, 237– 240.

Evangelatos N, Upadya S, Venne J, et al. (2020a). Digital transformation and governance innovation for public biobanks and free/Libre open source software using a blockchain technology. OMICS 24, 278–285.

Haines A. (2016). Addressing challenges to human health in the Anthropocene epoch-an overview of the findings of the Rockefeller/ Lancet Commission on planetary health. Public Health Rev 37, 14.

Inzaule S, Tessema S, Kebede Y, et al. (2021). Genomic informed pathogen surveillance in Africa: Opportunities and challenges. Lancet Infect Dis 21, E281–E289.

Jonas O, and Seifman R. (2019). Do we need a global virome project? Lancet 7, DOI: 10.1016/S2214-109X(19)30335-3.

Kandeil A, Gomaa M, Nageh A, et al. (2019). Middle east respiratory syndrome coronavirus (MERS-CoV) in dromedary camels in Africa and Middle East. Viruses 11, 717.

Lin B, and Wu S. (2021). Digital transformation in personalized medicine with artificial intelligence and the internet of medical things. OMICS [Online ahead of print]. DOI: 10.1089/ omi.2021.0037.

Mbala-Kingebeni P, Aziza A, Di Paola N, et al. (2019). Medical countermeasures during the 2018 Ebola virus disease outbreak in the North Kivu and Ituri Provinces of the Democratic Republic of the Congo: A rapid genomic assessment. Lancet Infect Dis 19, 648–657.

Moon S, and Kickbusch I. (2021). A pandemic treaty for a fragmented global polity. Lancet Public Health 6, e355–e356. National Institute of Allergy and Infectious Diseases. (2020). COVID-19, MERS & SARS. https://www.niaid.nih.gov/ diseases-conditions/covid-19. Accessed September 25, 2021.

Nii-Trebi N. (2017). Emerging and neglected infectious diseases: Insights, advances, and challenges. BioMed Res Int 2017, 5245021.

Olival K, Hosseini P, Zambrana-Torrelio C, et al. (2017). Host and viral traits predict zoonotic spillover from mammals. Nature 546, 646–650.

Özdemir V. (2021). Digital is political: Why we need a feminist conceptual lens on determinants of digital health. OMICS 25, 249–254.

Seltenrich N. (2018). Down to earth: The emerging field of planetary health. Environ Health Perspect 126, 072001.

Siddle KJ, Eromon P, Barnes KG, et al. (2018). Genomic analysis of Lassa virus during an increase in cases in Nigeria in 2018. N Engl J Med 379, 1745–1753.

Singh R, Singh PK, Kumar R, et al. (2021). Multi-omics approach in the identification of potential therapeutic biomolecule for COVID-19. Front Pharmacol 12, 652335.

Stephenson E, Reynolds G, Botting RA, et al. (2021). Singlecell multi-omics analysis of the immune response in COVID- 19. Nat Med 27, 904–916.

Stukalov A, Girault V, Grass V, et al. (2021). Multilevel proteomics reveals host perturbations by SARS-CoV-2 and SARS-CoV. Nature 594, 246–252.

Tanwar AS, Evangelatos N, Venne J, et al. (2021). Global open health data cooperatives cloud in an era of covid-19 and planetary health. OMICS 25, 169–175.

Van Noorden R. (2021). Scientists call for fully open sharing of coronavirus genome data, February 2021, Nature 590, 195–196.

Van Spall H, Topol E, Mamas M, et al. (2020). Applications of digital technology in COVID-19 pandemic planning and response. Lancet Digital Health 2, e435–e440.

Waugh C, Lam SS, and Sonne C. (2020). One health or planetary health for pandemic prevention? Lancet 396, P1882.

World Health Organisation. (2018). Investigation of cases of human infection with Middle East respiratory syndrome coronavirus (MERS-CoV) Interim guidance. https://apps.who .int/iris/bitstream/handle/10665/178252/WHO_MERS_SUR_ 15.2_eng.pdf?sequence=1. Accessed September 25, 2021.

World Health Organisation. (2019a). Pandemic Preparedness. https://www.euro.who.int/en/health-topics/communicablediseases/ influenza/pandemic-influenza/pandemic-preparedness. Accessed September 25, 2021.

World Health Organisation. (2019b). Publicly available plans prepared after 2009 pandemic. https://www.euro.who.int/en/ health-topics/communicable-diseases/influenza/pandemic-influenza/ pandemic-preparedness/national-preparedness-plans/ publicly-available-plans-prepared-after-2009-pandemic. Accessed September 25, 2021.

World Health Organisation. (2020). Emergency operations: Annual report. Saving lives and reducing suffering: WHO's work in emergency response operations in the WHO African Region in 2018. https://reliefweb.int/sites/reliefweb.int/files/ resources/WHO-AF-WHE-EMO-01-2020.pdf. Accessed September 25, 2021.

World Health Organisation. (2021a). Coronavirus dashboard. https://covid19.who.int/?gclid=CjwKCAiAkJKCBhAyEiw AKQBCkkC_NB-aTHyGLMKozjDGe_T_oICDlLopUuoPu VFrpTlJZ8PWC_TaOhoCPxMQAvD_BwE. Accessed September 25, 2021.

World Health Organisation. (2021b). Strengthening preparedness for health emergencies: Implementation of the International Health Regulations (2005). https://cdn.who.int/media/ docs/default-source/emergency-preparedness/b148_19-en .pdf?sfvrsn=c96756bc_1&download=true. Accessed September 25, 2021.

Chapter

8

Global Health Data Cloud: Laying New Directions for Collaborative Science

Tanwar, Ankit Singh, Nikolaos Evangelatos, Julien Venne, Lesley Ann Ogilvie, Kapaettu Satyamoorthy, and Angela Brand. "Global open health data cooperatives cloud in an era of COVID-19 and planetary health." *OMICS: A Journal of Integrative Biology* 25, no. 3 (2021): 169-175.

DOI: 10.1089/omi.2020.0134; **IF 3.9 (2021)**

Global Open Health Data Cooperatives Cloud in an Era of COVID-19 and Planetary Health

Abstract

Big data in both the public domain and the health care industry are growing rapidly, for example, with broad availability of next-generation sequencing and large-scale phenomics datasets on patient-reported outcomes. In parallel, we are witnessing new research approaches that demand sharing of data for the benefit of planetary society. Health data cooperatives (HDCs) is one such approach, where health data are owned and governed collectively by citizens who take part in the HDCs. Data stored in HDCs should remain readily available for translation to public health practice but at the same time, governed in a critically informed manner to ensure data integrity, veracity, and privacy, to name a few pressing concerns. As a solution, we suggest that data generated from high-throughput omics research and phenomics can be stored in an open cloud platform so that researchers around the globe can share health data and work collaboratively. We describe here the Global Open Health Data Cooperatives Cloud (GOHDCC) as a proposed cloud platform-based model for the sharing of health data between different HDCCs around the globe. GOHDCC's main objective is to share health data on a global scale for robust and responsible global science, research, and development. GOHDCC is a citizen-oriented model cooperatively governed by citizens. The model essentially represents a global sharing platform that could benefit all stakeholders along the health care value chain.

Keywords:

cloud computing, big data, health data cooperatives, health data, health data cooperatives cloud

Introduction

The availability of health data due to the emergence of high-throughput technologies is growing rapidly, and the proper management of health data becomes ever more important. Moreover, data sharing is becoming a vital element in the healthcare industry, and it is important that health data are shared across different geographical areas for better research outputs (Taichman et al., 2016). On the other hand, the storage and processing of such a huge amount of data with variability in data formats pose many challenges. A computer model that can store, run and analyze the massive amount of data simultaneously is required to overcome these technical challenges. It has been advocated by many that an open-access health data platform on a global scale is required to store, manage and share citizen-owned health data by taking a transparent and protected approach.

Big Data and its Management

When a massive amount of heterogeneous data is collected, stored, processed and analyzed at high speed, it can be referred as big data. Big data in any field is identified by its attributes such as Volume, Velocity, Variety, Variability, Veracity and Value (Assunção et al., 2014; Andreu-Perez et al., 2015; De Mauro et al., 2015). Big data in healthcare can be defined as data generated from hospitals, clinics, research centers, public sector, diagnostic labs as well as the healthcare-related industries, such as the biopharmaceutical industry. Hospitals and clinics generate data related to the diagnosis and treatment of patients also named patient-generated health

data (PGHD). PGHD contains data such as patient information, electronic health records (EHRs), diagnosis, prescribed medicines and on-going treatment methods. Research centers generate a large amount of data especially due to the emergence of advanced high-throughput techniques, such as Next Generation Sequencing (NGS), Microarrays, Whole Exome Sequencing and RNA sequencing. Advances in high-throughput techniques provided a faster approach to analyze biological samples, and it also enables the extraction of information at different levels (DNA, RNA and Protein) of the '-omics' cascade between the genome and the phenome. Indeed, a major contribution of big data in healthcare is from multi-omics studies, which generate and integrate data from fields such as genomics, proteomics, transcriptomics, metabolomics, metagenomics and epigenomics (Evangelatos et al., 2016; Mählmann et al., 2018). Healthcare or pharmaceutical industries generate a large amount of data related to novel drug discoveries and drug trials. Adaption of big data by the healthcare systems is a slow process because of the involvement of many stakeholders, providers, facilities and, importantly, their inability to share data due to different data formats and the restrictions posed by concerns on legal issues such as data protection and privacy (Mählmann et al., 2018).

Cloud Computing

Cloud computing is a modern way of current computing, which enables users to perform heavy computational tasks or store a large amount of data using a low configuration machine. To access data and run tasks on the cloud, the user needs a device (e.g., mobile phone, laptop, computer or tablet) connected to the internet. According to the National Institute of Standards and Technology (NIST), cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction. With numerous applications in different fields, a cloud model is a computing structure with three service models, four deployment models and five key characteristics, as displayed in figure 1 (Mell and Grance, 2012). Out of its three-service models, Software as a Service (SaaS) is the most common service model used by consumers around the world for applications such as e-mails, processing of word documents, social media apps, and data storage. The Platform as a Service (PaaS) model allows consumers to deploy and control applications created with the use of programming languages, tools and services supported by the provider. Whereas PaaS restricts control of underlying cloud infrastructure, the Infrastructure as a Service (IaaS) model enables consumers to have control over storage, operating systems, installed applications and certain networking components (Mell and Grance, 2012; Weber, 2013).

Among the top cloud service providers currently around the globe are the Amazon web service, Microsoft Azure, IBM Cloud, Salesforce, Google Cloud, Oracle Cloud and VMware (Armbrust et al., 2010; Zhang et al., 2010; Evans, 2017). Cloud computing platforms provide optimal solutions for the storage and processing of big data. Along with providing huge storage capacity, cloud computing also supports platforms such as Hadoop, MapReduce, Hive, Zookeeper and HBase for big data analytics (Raghupathi and Raghupathi, 2014; Ware et al., 2017).

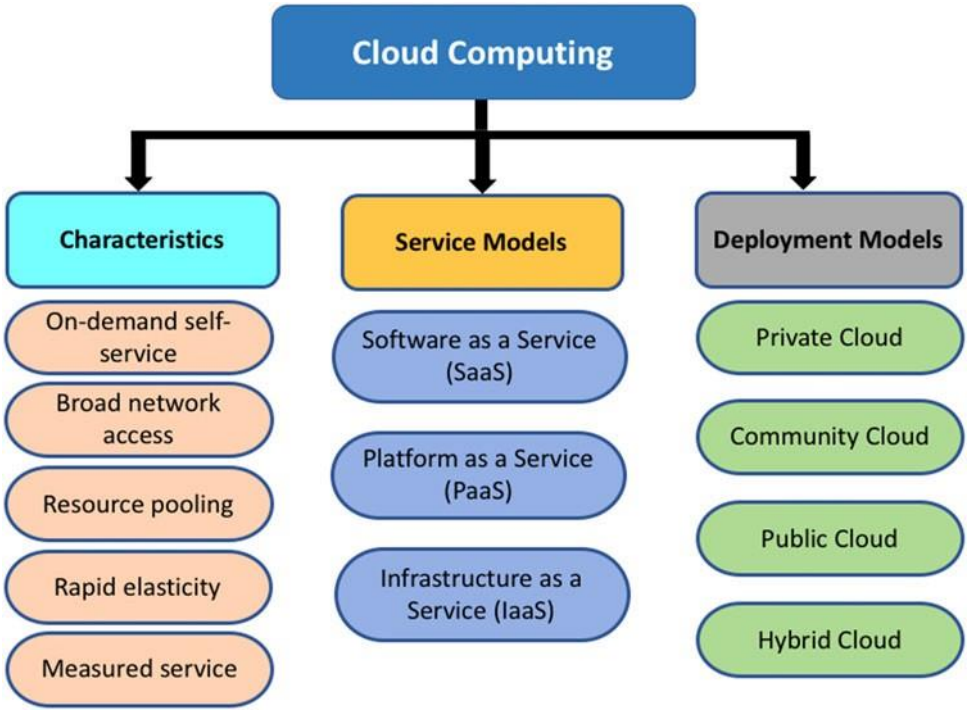


Figure 1: Structure of cloud computing. IaaS, Infrastructure as a Service; PaaS, Platform as a Service; SaaS, Software as a Service.

Challenges in Cloud Computing

Security issues and challenges in a cloud platform can arise at different levels, and these challenges, if not addressed properly and in time, can lead to dire consequences such as data theft, application run errors, system crash, low storage capacity and loss of data. A few of these major challenges are briefly described below.

Security Issues: security issues in a cloud platform can be divided into six sub-categories which include the need to (a) develop safety mechanisms to monitor the activities of the cloud server, (b) keep sensitive data confidential, (c) avoid illegal operations inside the cloud server, (d) avoid service hijacking, (e) develop protocols to prevent attackers from gaining full access to the host system, (f) provide statements of appropriate law, and implement legal jurisdiction which can help users in case of any exploitation by providers (Sun et al., 2011; Singh et al., 2016).

Application Issues: application issues can arise at different levels such as (a) attacks on user front end and back end applications, (b) compatibility and security of application on launched platform, (c) on framework level, (d) licensing of applications, as unauthorized pirated software is more prone to cyber-attacks, (e) service availability and scalability, (f) system optimization to run parallel applications (Sun et al., 2011; Singh et al., 2016).

Privacy Issues: includes (a) user control over data stored and processed in cloud and avoidance of unauthorized access to personal data, (b) replication of data to multiple location without data loss and identification of any unauthorized modification and fabrication of data, (c) level of control of cloud sub-contractors on sensitive or personal information (Sun et al., 2011; Singh et al., 2016).

Data Storage: Along with its capacity to store such huge amounts of data, any cloud platform should be capable of maintaining anonymity, 24x7 data availability (access data from any device at any time), to provide security and maintenance of data warehouse, and to prevent data loss and leakage (Puthal et al., 2015; Singh et al., 2016).

A few more challenges can be added to the above list, such as operating system compatibility issues between the cloud platform and the access devices, client management issues, and cluster computing.

Open Science Clouds

The amount of health data generated worldwide is huge and requires proper handling, storage, processing, and security. Cloud computing has the potential to store and process huge datasets generated globally in different scientific disciplines. It is important to keep access to these scientific datasets open, and the security of datasets should be a priority.

Open Science Data Cloud (OSDC): OSDC first started in 2010. It is a cloud platform, designed to store, analyze, manage and share scientific data. The OSDC is a hosted cloud platform operated by a single entity, the Open Cloud Consortium (OCC). OSDC is different from existing cloud resource in the way it is designed. OSDC architecture enables this platform to provide long term persistent storage of medium to very large scientific datasets. The OSDC also utilizes high-performance research networks, which enable data sharing over wide areas. OSDC supports a balanced architecture that utilize data locality to provide efficient execution of submitted queries and analysis (Grossman et al., 2010, 2012).

European Open Science Cloud (EOSC): The EOSC provides a federated, globally accessible environment where researchers, innovators, companies and citizens can publish, find and re-use each other's data and tools for research, innovation and educational purposes. Within the perspective of the digital single market, the EOSC aims to accelerate the process of an open science cloud and render it more effective by removing legislative and technical barriers. Furthermore, to accelerate its open science cloud platform, EOSC provides support access to systems and services and allows smooth flow of digital data across social, disciplinary and geographical borders (Ayris et al., 2016).

Health Data Cooperatives

A Health Data Cooperatives (HDC) is an ecosystem, which combines heterogeneous health data of citizens/patients with knowledge available in databases. These heterogeneous health data are used to create an integrated cloud-based analytical platform, which aims to study three major analytical models: descriptive, predictive and prescriptive (Mählmann et al., 2018). HDCs are primarily based on metadata and structured in a way that keeps citizens at the center of the stage. As a health data governance model, HDCs offer a trusted framework for overcoming ethical, societal, political, and technical challenges.

The HDCs represent unified data systems that promote data access to and linkage of heterogeneous data from a variety of sources within and outside a health domain. Heterogeneous data components include data sources

such as '-omics' data, data from electronic health records and non-electronic systems, patient m-health data, repositories of biological samples, drug data, environmental data, insurance record data, social media data, etc.

HDCs aim to provide full control of health data to citizens, who should be the main beneficiaries of this health data integrated framework. This framework should be based on trust, which implies that data processing should be executed in a secure and transparent manner. A not-for-profit HDC model is required to make the potential benefits of HDC available to platform services and research projects that serve a common good and benefit society.

The applications of HDCs are manifold including, but not limited to, (a) data-driven and evidence based policymaking, (b) provision of prevention-oriented and cost-effective healthcare, since HDCs can evaluate patient pathway management systems at primary care level, (c) a surveillance system, which combines information of big data and advanced simulation methods for early detection of public health problems, (d) evidence-based prevention strategies to evaluate the effectiveness and efficiency of implemented strategies, (e) a real-time view of current state of health of citizens that helps to evaluate individual risk estimation (risk-based preventive interventions), (f) development of innovative approaches, such as personalized medicine, to overcome the 'one-size-fits-all' treatments, (g) serving as a basis to facilitate cross-border cooperation, which will allow data sharing and access across borders, and (h) reduction of the burden of collecting and managing raw data for analysis (Mählmann et al., 2018).

A HDC on a global scale could be an excellent citizen health monitoring system, which would integrate heterogeneous health data to provide a better understanding of disease mechanisms, thus guiding healthcare and public health policies towards preventive healthcare systems. A global level HDC would require a huge amount of computational power and resources. Cloud computing strikes as a promising tool to take a HDC on a global scale. Indeed, cloud computing offers high-performance computing coupled with techniques such as parallel and cluster computing that can easily overcome the computational challenges associated with evidence-based datasets analysis.

Integration of big data to form an HDC and use of cloud computing resources to take a HDC on a global scale would provide a global health data governance system that could advance science to the benefit of both the citizens, who control their health data and other stakeholders, who make use of citizen's health data, in a transparent and secure way (Hafen et al., 2014; Mählmann et al., 2018).

A Health Data Cooperatives Cloud (HDCC) model shown in figure 2 consists of health data sources, users, and cooperatives. Health data sources like healthcare providers (hospitals and clinics), research centers, diagnostic labs, industries etc., generate health data. Generated health data are provided back to citizens, citizens store or upload all their data on HDCC, where they have full control rights to manage (read and write permissions) their health data.

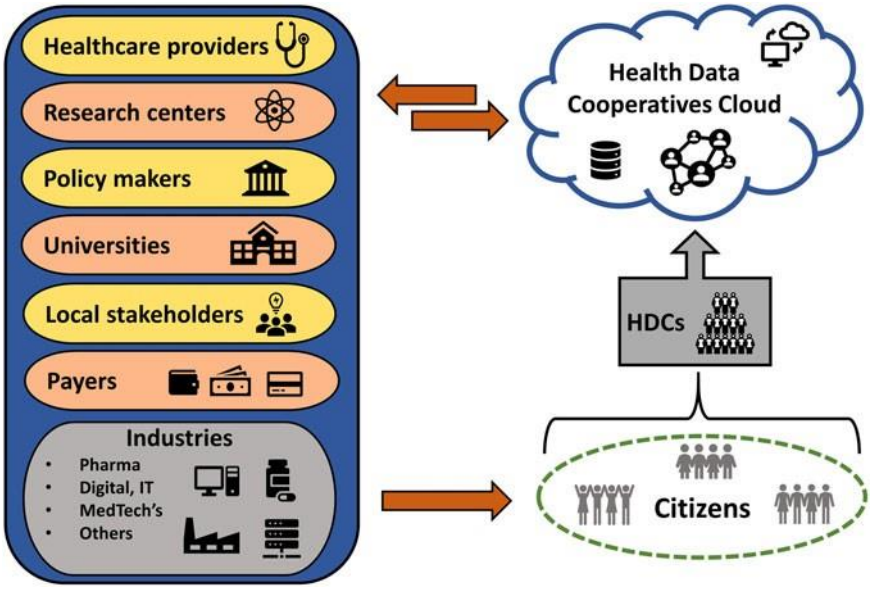


Figure 2: HDC cloud. HDCs, health data cooperatives.

When health data users need health data, they can approach an HDC. After obtaining individual consent from the members of a HDC, the health data stored in the cloud are shared with these specific users. In return for gained access to citizen's health data, health data users provide or offer benefits/incentives/services to the members of the HDC. Citizens or HDC members can decide whether they want to share only a part of their health data or all health data. HDC members decide who can access their data under what conditions, for how long (definitive period), for what purposes, and these decisions/choices are made each time the purpose of use of their personal data changes. They are fully informed about how the data will be stored and used. They also have the right to remove data from the HDCC.

The type of health data stored in the HDCC can be divided into two categories: 1) personal data and 2) environmental/ecosystem data. For example, personal data include data generated from fitness trackers and health monitoring apps. Environmental/ecosystem data include data generated from healthcare services (hospitals, clinics, primary care centers), public health services, economy and sensor (housing, education, energy, taxes, etc.), sport/fitness/physical activity, nutrition and diet, social services and private healthcare centers.

The control of cloud service providers over the data should be very limited. Cloud services should only be concerned with the data structure, data type, data size and data properties, but data content should always be confidential. The cloud service's job is to maintain data integrity, manage space and provide application support to process data. The government also plays a supportive role in this model by ensuring things run smoothly and by encouraging citizens to become members of HDCs.

Global Open Health Data Cooperatives Cloud (GOHDCC)

Many different cloud models already exist, and they all aim to store health data. Different entities regulate these health data cloud models; for example, private or governmental organizations. Furthermore, the cloud models also vary in terms of standards. While health data collection, storage and utilization are essential

components of the clouds, the health data cloud models widely differ in executing the key steps involved in health data collection, storage and utilization.

Big corporations or companies like Google, Samsung, Microsoft, and Apple are collecting citizen health data using health apps provided on electronic devices. For instance, Google launched the Google Fit program to track the fitness of its service users. Samsung uses Sami, a biometric data platform that obtains health data from electronic devices and apps. Microsoft launched a web-based platform in 2010 called HealthVault to store health and fitness information. HealthVault was officially shut down in November 2019 and deleted all user data stored on the platform.

Apple is using its HealthKit framework and its default Health app provided in all its mobile devices (Olson and Spence, 2014). All these above-mentioned models acquire and accumulate data from fitness trackers and health associated apps. These models raise concerns over privacy and how to process information as sensitive as health data. They all lack data standards and strict regulations on data collection, processing and utilization (Olson and Spence, 2014). Moreover, all these models are business driven.

In addition to the business-driven models mentioned earlier, few models with the same aim but driven by entirely different purposes and objectives exist. Two such examples already mentioned above are the European Open Science Cloud (EOSC) (Ayris et al., 2016) and the Open Science Data Cloud (OSDC) (Grossman et al., 2010, 2012).

These two models aim to collect and utilize health data for scientific purposes only; hence they are science driven. In November 2018 the European Commission announced the launch of a cloud for research data, the 'research open science cloud' (i.e. the EOSC). To support European science in its global leading role, EOSC aims to establish a trusted environment for hosting and processing research data. Both, the EOSC as well as the OSDC, aim to go global so that health data can be shared on a wider scale for better science and research. On a global scale, these models can merge to form Global Open Science Clouds (GOSC), a platform that collects and stores data related to scientific fields. One of the major drawbacks of these open science clouds is that they lack good governance of all kinds of health data.

As a solution to this problem, we propose the Global Open Health Data Cooperatives Cloud (GOHDCC) (figure 3), a model that can overcome the above-mentioned limitations regarding regulation and data governance. GOHDCC is a citizen-oriented model and is formed when different health data cooperatives clouds join, sharing the same core principles regarding data collection, storage, processing and distribution. GOHDCC's core element is the health data cooperative, which empowers citizens by providing full authority of their health data in their own hands. Citizens control how their health data should be shared with other entities. In other words, citizens hold the right to choose whether they would like to share their personal health data with private industries, hospitals, clinics, research centers, health policymakers, for clinical trial and academic purposes.

The GOHDCC overlaps with other science clouds like EOSC and OSDC in the field of health sciences. Thus, the GOHDCC can also provide health data to the already existing science clouds. Also, the GOHDCC's primary focus is health sciences, but it is not only limited to citizens' health data, but it also goes broader in terms of storing data related to non-scientific fields e.g., insurance, income, expenses, taxes etc. The GOHDCC provides a secure ecosystem build on trust and transparency as citizens control how the data are shared and as their personal data are always secure in their own hands.

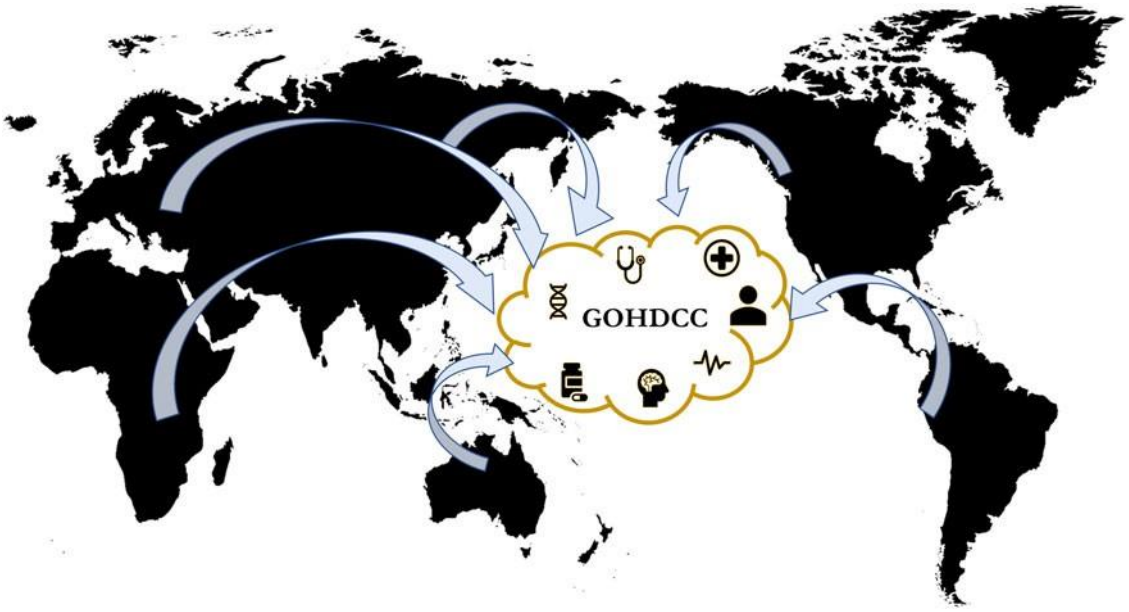


Figure 3: GOHDCC model. GOHDCC, global open health data cooperatives cloud.

GOHDCC's Role in a Pandemic and Future Health

A newly discovered coronavirus is causing the COVID-19 infectious disease that has already infected more than 23.7 million people and caused more than 814,000 deaths globally (https://www.coronatracker.com). The COVID-19 pandemic is a public health problem, and to tackle it countries are combining resources to develop a vaccine that is safe and efficient. Scientists, researchers, and physicians are generating and collecting a vast amount of data, including viral genome data, gene and protein expression data of both viral and host genomes, data from clinical trials, and literature data. COVID-19 databases and data portals are being set up around the globe to combine all the information and to provide researchers with easy access to these datasets.

Examples of COVID-19 resources are the COVID-19 database by the World Health Organization (WHO) (https://www.who.int/emergencies/diseases/novel-coronavirus-2019/global-research-on-novel-coronavirus-2019-ncov) and Centers for Disease Control and Prevention (https://www.cdc.gov), the COVID-19 data portal (https://www.covid19 dataportal.org) by the European Bioinformatics Institute (EMBL-EBI), and the COVID-evidence database (https://covid-evidence.org) by the University of Basel, to name but a few.

The GOHDCC can provide a similar infrastructure that facilitates global action by combining local inputs. As the GOHDCC is not regulated or governed by any authority and is citizen-oriented, citizens have a right to

share their health data, such as symptoms, medication course, and their immune response to treatments. Citizens participating in any ongoing clinical trials could upload their data, depending on the terms of agreement of the trial, which can lead to a better understanding of viral response in different populations.

Due to genomic differences, the mechanism by which the virus infects one's body varies in such a heterogeneous population. As a result, to develop a safe and effective vaccine, it is essential to understand the viral mechanism in different communities. The GOHDCC can function as a data platform to provide heterogeneous health data from different geographical locations to extract vital clinical features that can help to develop a safer and effective vaccine for all.

The COVID-19 pandemic qualifies as a planetary health issue, as it is a public health problem, and it also impacts our natural systems. In simple terms, planetary health is the health of human civilization and the state of the biological systems on which it depends (Whitmee et al., 2015). Human society faces environmental threats that require vital and transformative actions to safeguard present and future generations. COVID-19 is a threat to human health and the natural systems that human life depends on, as it destabilizes crucial ecological pathways (Whitmee et al., 2015).

The interaction between humans and nature shifted significantly during the past few months. The COVID-19 pandemic has caused a series of lockdowns in different parts of the world, which have resulted in the minimization of human activities and have bought our ecosystem some time to stabilize to a certain extent.

Planetary health aims at promoting health, at preventing disease and disability, at eradicating conditions that harm human health, and at fostering resilience and adaptation (Horton et al., 2014). Planetary health demands a broader governance frame that recognizes and respects other sentient life on the planet (Özdemir, 2019). To deliver planetary health and to support sustainable human development, it urges to start a social movement based on collective action at every level of society (Horton et al., 2014).

The GOHDCC can serve as an excellent communication platform to update citizens with planetary health issues; a forum to educate citizens with necessary measures to be taken to conserve, sustain, and make resilient the planetary and human systems on which health depends. Planetary health is future health; if social activities continue to stay on the current path, it will cause a collapse of human civilization. The overconsumption patterns and harm to our planet's biodiversity are unsustainable and a threat to human existence as species (Horton et al., 2014). The GOHDCC, as a global platform, can do its part by educating people on these crucial facts and assist in developing an improved understanding connection between human health and natural systems.

Conclusion

The amount of data in the healthcare industry is growing rapidly and needs to be properly stored, processed and shared across the globe. Global open health data sharing platforms provide researchers with access to citizen's health data to enable high-quality research. Open health data sharing platforms are exposed to different kind of threats, require proper security and guidelines to process data which can be shared across different

open health data platforms. Cloud computing serves as a perfect platform to build an open health data cooperatives cloud. Global Open Health Data Cooperatives Cloud (GOHDCC) is a model that allows the sharing of health data between health data cooperatives clouds and other stakeholders such as global open science clouds, private industries, healthcare and research centers. GOHDCC integrated with the HDCs ecosystem provides citizens with full control over their health data.

The COVID-19 pandemic is a public health problem that concerns all around the globe and qualifies as a planetary health issue. Under the umbrella of the WHO, countries are trying to combine resources to develop a vaccine. A global action is required in such a pandemic and the GOHDCC can provide an infrastructure that facilitates global actions by combining health data information.

Acknowledgment

The authors would like to thank Manipal Academy of Higher Education (MAHE) and especially the Manipal School of Life Sciences and the Prasanna School of Public Health. This work would have not been possible without the support provided through the prestigious Dr. TMA Pai Endowment Chairs of MAHE. The authors also would like to thank the United Nations University – Maastricht Economic and Social Research Institute on Innovation and Technology (UNUMERIT) and Maastricht University for infrastructure and support.

Author Disclosure Statement

The authors declare they have no competing financial interests.

References:

Andreu-Perez J, Poon CCY, Merrifield RD, Wong STC, and Yang GZ. (2015). Big Data for health. IEEE J Biomed Heal Informatics 19, 1193–1208.

Armbrust M, Fox A, Griffith R, et al. (2010). A view of cloud computing. Commun ACM 53, 50–58.

Assunção MD, Calheiros RN, Bianchi S, Netto MAS, and Buyya R. (2014). Big Data computing and clouds: Trends and future directions. J Parallel Distrib Comput 79–80, 3–15. Ayris P, Berthou JY, Bruce R, et al. (2016). Realising the European Open Science Cloud. DOI: 10.2777/940154. Accessed May 22, 2019.

De Mauro A, Greco M, and Grimaldi M. (2015). What is big data? A consensual definition and a review of key research topics. AIP Conf Proc 1644, 97–104.

Evangelatos N, Reumann M, Lehrach H, and Brand A. (2016). Clinical trial data as public goods: Fair trade and the virtual knowledge bank as a solution to the free rider problem—A framework for the promotion of innovation by facilitation of clinical trial data sharing among biopharmaceutical companies in the era of omics and big data. Public Health Genomics 19, 211–219.

Evans B. (2017). The Top 5 Cloud-Computing Vendors: #1 Microsoft, #2 Amazon, #3 IBM, #4 Salesforce, #5 SAP. Forbes Contrib. https://www.forbes.com/sites/bobevans1/2017/ 11/07/the-top-5-cloud-computing-vendors-1-microsoft-2-ama zon-3-ibm-4-salesforce-5-sap/#622ac8a76f2e. Accessed May 22, 2019.

Grossman RL, Greenway M, Heath AP, et al. (2012). The design of a community science cloud: The open science data cloud perspective. 2012 SC Companion: High Performance Computing, Networking Storage and Analysis, Salt Lake City, UT, November 10–16, 2012, pp. 1051–1057. IEEE.

Grossman RL, Gu Y, Mambretti J, Sabala M, Szalay A, and White K. (2010). An overview of the Open Science Data Cloud. Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing— HPDC ’10, 2010, pp. 377–384.

Hafen E, Kossmann D, and Brand A. (2014). Health data cooperatives—Citizen empowerment. Methods Inf Med 53, 82–86.

Horton R, Beaglehole R, Bonita R, Raeburn J, McKee M, and Wall S. (2014). From public to planetary health: A manifesto. Lancet 383, 847.

Mählmann L, Reumann M, Evangelatos N, and Brand A. (2018). Big Data for public health policy-making: Policy empowerment. Public Health Genomics 20, 312–320.

Mell P, and Grance T. (2012). The NIST definition of cloud computing: Recommendations of the National Institute of Standards and Technology. Public Cloud Comput Secur Priv Guidel 97–101.

Olson P, and Spence E. (2014). Google wants to collect your health data with ‘‘Google Fit.’’ Forbes. https://www.forbes .com/sites/parmyolson/2014/06/12/exclusive-google-to-launchhealth- service-google-fit-at-developers-conference/#6f72518f 400f. Accessed October 26, 2019.

Özdemir V. (2019). Innovating governance for planetary health with three critically informed frames. Omi A J Integr Biol 23, 623–630.

Puthal D, Sahoo BPS, Mishra S, and Swain S. (2015). Cloud computing features, issues, and challenges: A big picture. Proceedings—1st International Conference on Computational Intelligence and Networks, CINE 2015, pp. 116–123.

Raghupathi W, and Raghupathi V. (2014). Big data analytics in healthcare: Promise and potential. Heal Inf Sci Syst 2, 3. Singh S, Jeong YS, and Park JH. (2016). A survey on cloud computing security: Issues, threats, and solutions. J Netw Comput Appl 75, 200–222.

Sun D, Chang G, Sun L, and Wang X. (2011). Surveying and analyzing security, privacy and trust issues in cloud computing environments. Procedia Eng 2852–2856.

Taichman DB, Backus J, Baethge C, et al. (2016). Sharing clinical trial data: A proposal from the International Committee of Medical Journal Editors. Chin Med J (Engl) 129, 127–128. Ware A, Janvale G, Shaikh F, and

Harke S. (2017). HADOOP: Solution for big data challenges in bioinformatics and its prospective in India. IOSR Journal of Computer Engineering. 51–54.

Weber AS. (2013). Cloud computing in education. Ubiquitous Mob Learn Digit Age 19–36.

Whitmee S, Haines A, Beyrer C, et al. (2015). Safeguarding human health in the Anthropocene epoch: Report of the Rockefeller Foundation-Lancet Commission on planetary health. Lancet 386, 1973–2028.

Zhang Q, Cheng L, and Boutaba R. (2010). Cloud computing: State-of-the-art and research challenges. J Internet Serv Appl 1, 7–18.

Chapter 9

Discussion

The thesis starts with the description of human microbiome and microbial diversity in human body which plays a vital role in building and regulating human health. The balance in microbiome diversity is important to maintain quality health and emergence of pathogens can break this crucial balance. The thesis also highlights the relationship between human microbiome and associated cancers. Microbiome in cancers play a dual role and it is important to understand their mechanism to differentiate pathogens and non-pathogens to develop targeted therapeutic strategies. Application of omics, such as in field of microbial genomics can be utilized to screen pathogens and determine their virulence potential. It can assist in understanding the underlying mechanisms of antibiotic resistance and aid in identifying new therapeutic targets to combat antimicrobial resistance.

Amount of multiomics data generated from microbial studies can provide a genomic surveillance system to identify emerging pathogens and predict future pandemics. Anticipating pandemics can provide governments sufficient time to take necessary precautions and equip themselves with the resources they need to deal with these circumstances. Applications of multiomics studies can also significantly contribute towards planetary health genomics.

The thesis explains the significance and need of a global health data platform to collaborate and unify research to support planetary health. A cloud platform to share health data on a global scale for robust and responsible global science, research, and development.

Chapter 2 (Human Microbiome): This study was conducted to investigate the impact of Ayurvedic *prakriti* phenotypes on the diversity of the gut/oral microbiome in healthy individuals. The aim was to understand human-microbe interactions and identify microbial signatures that can serve as biomarkers for personalized and community health. The gut and oral microbiome of individuals grouped into three *prakriti* categories were analyzed for microbial diversity. Microbial signatures and composition are unique to each individual and can be impacted by factors such as age, diet, lifestyle, stress, and environment. The study found that overall species diversity was significantly higher in older individuals for both the gut and oral microbiome.

We found that *Prevotella*, *Bacteroides*, and *Dialister* were the predominant genera in the gut microbiome across all prakriti types and genders. We discovered a negative correlation between the relative abundance of *Prevotella* and *Bacteroides*, suggesting that diet plays a role in the enrichment of certain bacterial species. In the oral microbiome, the major genera identified were *Streptococcus* and *Neisseria*, which is in agreement with previous studies conducted on the Indian population (Chaudhari *et al*. 2019).

Most of the organisms in both oral and gut microbiomes were found to be gram-negative. A higher proportion of biofilm-forming microbes were identified in the oral microbiome. Although some species were identified as potential pathogens in both microbiomes, the least amount of potentially pathogenic organisms were present in individuals of *pitta prakriti* phenotype in both cases.

Overall aim was to examine the connection between *prakriti*, the foundation of personalized medicine in *Ayurveda*, and gut microbiome, which is increasingly considered as a reliable indicator of an individual's

health. To achieve this goal, we analyzed the bacterial metagenomes in saliva and stool samples, which represent the oral and gut microbiome respectively, from 272 healthy individuals, and studied their relationship with *prakriti* to see if there is a correlation between microbial diversity and *prakriti* identity.

Chapter 3 (Cancer): This review focuses on the markers that regulate epithelial-mesenchymal transition (EMT) and the complex network of molecular mechanisms involved in the pathogenesis of oral submucous fibrosis (OSF) and oral squamous cell carcinoma (OSCC). Epithelial-mesenchymal transition (EMT) plays a crucial role in various physiological and pathological events, including embryonic development, wound healing, organ fibrosis, and the development of cancer. Recent studies have shown that EMT also plays a significant role in the invasion and metastasis of cancer cells. Additionally, EMT's involvement in the onset of oral submucous fibrosis (OSF) and its malignant transformation, as well as the inflammatory reaction leading to fibrosis, has not been extensively explored.

There is evidence to suggest that different signalling pathways engage in crosstalk, and some studies indicate that the inhibition of a single transcription factor is sufficient to block EMT. The detrimental role of EMT in fibrosis progression and cancer metastasis highlights the need to identify suitable targets for preventing EMT induction, particularly given the poor prognosis of oral cancer and the development of drug resistance.

This review centres on the signalling pathways and mechanisms that induce changes in gene expression signatures, resulting in EMT in both OSF and OSCC. The discovery of signature genes that affect EMT could reveal new pathways that play a crucial role in the advancement of oral cancer. These EMT markers might serve as effective targets for controlling the disease's spread and enhancing the prognosis of OSCC. In this study, a systematic analysis of gene-disease associations has demonstrated their involvement in OSF and SCC. A pathway analysis has also revealed the involvement of upregulated and downregulated genes in various EMT regulating pathways.

Chapter 4 and 5 (Pathogens and AMR): Two comparative genome analysis studies were done to identify antimicrobial resistance and virulence potential of studied strains. The presence of an array of virulence and antimicrobial resistance genes in studied strains suggests their potential role as emerging pathogens with significant impact on public health.

The first research focuses on analyzing the genomes of three *Clostridia* strains. The study compared two *Clostridium* strains that have been sequenced (*C. butyricum* and *C. tertium*) to the sequenced pathogenic strain *Clostridioides difficile* (CD). CD is a Gram-positive bacterium that can cause illness and is linked to diseases such as sepsis, pseudomembranous colitis, and colorectal cancer. *C. difficile* infections (CDI), typically following antibiotic exposure that result in dysbiosis of the gut microbiome, is one of the leading causes of diarrhoea in the elderly population. In the study, three isolates (*C. tertium* [CT MALS001], *C. butyricum* [CB MALS002], and *C. difficile* [CD MALS003]) were sequenced and evaluated for their antimicrobial, cytotoxic, antiproliferative, genomic, and proteomic profiles.

The goal of the current work was to study the genomes of CT MALS001, CB MALS002 and compare it with that of CD MALS003 for genomic surveillance to uncover virulence-associated factors and their potential role as emerging pathogens. We found a total of 66 different toxin genes in the three isolates, with CD MALS003 harbouring maximum number of toxin-coding genes (54) while CB MALS002 (13) and CT MALS001 (14) showed fewer toxin-coding genes. We identified major exotoxin genes (*tcdA* and *tcdB*) along with binary toxins *cdtA* and *cdtB* in CD MALS003 genome. In CT MALS001 one toxin gene each from category exoenzyme and hemolysin were predicted while in CB MALS002, two hemolysins and one exotoxin was predicted. A maximum number of flagellar and cell adhesion genes were predicted for CD MALS003 (31 genes) followed by CT MALS001 (6 genes) and CB MALS002 (3 genes).

The second study focuses on the examination of *Staphylococcus aureus* (*S. aureus*), which is a highly virulent and widely prevalent pathogen in diabetic foot ulcer infections. In this comparative genome analysis study, we compared four *S. aureus* strains (MUF168, MUF256, MUM270, and MUM475) in terms of the presence of genes involved in antibiotic resistance and virulence potential, biofilm formation, and variants of potential drug targets that may contribute to antibiotic resistance development. Our *in-silico* analysis revealed that strains MUM270 and MUM475 had a higher number of antibiotic resistance genes (ARGs) compared to the MUF strains. However, strain MUM270 was found to be sensitive to all nine antibiotics tested, indicating that the presence of ARGs does not necessarily result in increased antibiotic resistance. Strain MUF256 was found to have the highest number of virulence genes, followed by strains MUM475, MUF168, and MUM270. Additionally, strain MUM475 was determined to be a high biofilm producer. Finally, our analysis showed that the MUM strains had a higher number of deleterious variants compared to the MUF strains.

In summary, our study based on whole-genome sequence analysis was aimed at identifying potential virulence factors, antimicrobial resistance genes, mobile genetic elements, biofilm-forming capabilities and sporulation factors, which contribute to pathogenic potential of microbes. Further studies using omics approach can provide critical clues on key regulators of microbial virulence and factors that contribute to antimicrobial resistance in clinically relevant pathogenic isolates.

Chapter 6 (Microbial Genomics): The content provided in this chapter serves as a vital reference for selecting the most pertinent and appropriate computational tool(s) for genome assembly and genetic signature identification. The realm of genomics has brought about a transformative approach to genome assembly, annotation and drug discovery. Through the utilization of high-throughput sequencing technologies, scientists can now assemble complete genome sequences and pinpoint potential drug targets contained within them. Progress made in the field of genomics has led to a deeper comprehension of the genome, unveiling fresh prospects.

Key stages in this process encompass genome assembly, annotation, and the identification of AMR resistance, virulence effects, and drug-target interactions. Fortunately, numerous resources exist to aid researchers in these endeavours, despite the intricate and challenging nature caused by the genome's abundance of repetitive sequences. A multitude of software/tools are available at researchers' disposal, each possessing its own

strengths and weaknesses. In this review, our objective was to emphasize the widely recognized and extensively utilized resources for different aspects of microbial research, including genome assembly and annotation, profiling antibiotic genes, identifying virulence factors, and studying drug interactions.

Chapter 7 (Genomic Surveillance): The Covid-19 pandemic is a critical global health crisis, and one of the three most significant outbreaks of infectious diseases in the early decades of the current century. It has spurred advances not only in the field of infectious diseases, but also in digital technology to enhance the capability to monitor, predict, and address planetary and ecological hazards. The field of omics systems science holds great promise for monitoring emerging pathogens through genomic surveillance, including new zoonotic threats. From a public health perspective, the most significant advantage of this field is the ability to implement genomic surveillance of SARS-CoV-2, which can help detect potential variants of interest or concern early on.

The transformation of omics through digital means started with the rapid advancements in next generation sequencing and other cutting-edge technologies. These developments will be increasingly integrated into Internet of Things (IoT) devices, enabling real-time and accurate monitoring of parameters that are currently only measured in hospitals and specialized clinics. The European Centre for Disease Prevention and Control states that sequencing partial genes and complete genomes (WGS) is an effective approach for studying the genomes of viral pathogens, comprehending the spread of outbreaks and spillover events, and identifying mutations that might affect transmissibility, pathogenicity, and countermeasures such as diagnostics, antiviral drugs, and vaccines. The results are key to informing outbreak control decisions in public health'' (European Center for Disease Control, 2021).

Chapter 8 (Global Health Data Cloud): We propose the Global Open Health Data Cooperatives Cloud (GOHDCC), a global platform for sharing health data that can support global research and development. This platform is citizen-led and cooperatively governed, making it beneficial for all stakeholders involved in the healthcare system. The model integrates cloud computing to manage large amounts of data and highlights the significance of big data management. It also highlights existing cloud-based health data platforms such as the Open Science Data Cloud (OSDC) and European Open Science Cloud (EOSC).

The GOHDCC model is centred around the Health Data Cooperatives ecosystem, which combines diverse health data from citizens/patients and databases to form an integrated cloud-based analytical platform. This platform is designed to study three major analytical models: descriptive, predictive, and prescriptive (Mählmann et al., 2018). The goal of the HDCs is to give citizens full control over their health data and make them the primary beneficiaries of the integrated framework.

A global HDC has the potential to serve as an excellent citizen health monitoring system by integrating diverse health data to enhance understanding of disease mechanisms. This, in turn, could steer healthcare and public health policies towards preventive systems. GOHDCC is a citizen-focused model, created by the collaboration of various health data cooperative clouds that share common principles for data collection, storage, processing, and distribution.

In pandemics like COVID-19, GOHDCC can act as a data platform to gather heterogeneous health data from various geographical locations and identify crucial clinical features to create a safer and effective vaccine for all. The COVID-19 pandemic is a public health problem that concerns all around the globe and qualifies as a planetary health issue. A global action is required in such a pandemic and GOHDCC provides the infrastructure needed to facilitate global collaboration by integrating health data information.

I S A

Impact Paragraph

Understanding the relationship between host and microbes is crucial for maintaining human health, and it is imperative to investigate this intricate association in greater depth. Microbes that reside within the human body can have both positive and negative effects, making it vital to differentiate between pathogenic and beneficial microbes in disease conditions. With the advancement of high-throughput technologies, our understanding of host-microbe interactions is improving each day. By analysing the vast amounts of data generated from microbial genomics research, we can better comprehend these interactions, manage diseases and infections, create vaccines and therapeutic targets, and promote the health of both host and microbes.

The main objective of this thesis was to delve into the intricate dynamics of the host-microbe relationship and shed light on the significance of the human microbiome. In Chapter 2, a comprehensive examination is conducted to examine the diverse microbial signatures and compositions observed among individuals. It becomes evident that these microbial profiles are susceptible to various influencing factors, including age, diet, lifestyle, stress, and environment. By studying these fluctuations, it is possible to identify microbial signatures that hold great potential as biomarkers for personalized health management and the overall well-being of communities. These insights pave the way for a deeper understanding of the intricate interplay between the human body and its microbial inhabitants, opening up new avenues for personalized healthcare approaches.

Chapters 4 and 5 of this thesis focus on the use of comparative genomics to assess the virulence potential of sequenced strains. Utilizing genomics data and computational resources, these chapters identify factors that contribute to virulence and antimicrobial resistance. Comparative genomic studies are a valuable tool in the fight against antibiotic resistance, as they can detect emerging pathogens and help determine the pathogenicity of closely related strains within a species. The evolution of emerging pathogens is a significant global health concern, and monitoring this evolution allows for the identification of potential molecular drug targets to combat future infections. Furthermore, comparative genomic studies can reveal novel regions encoded in microbial genomes that allow them to adapt and survive in exposed environment.

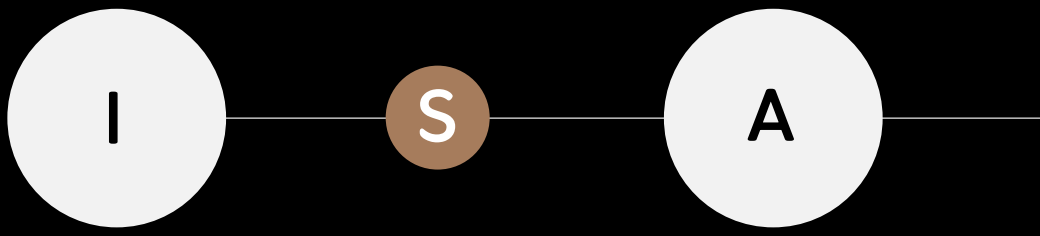
Chapter 6 of this thesis lists all major computational tools and resources that play a crucial role in the field of microbial genomics, aiding researchers in the analysis and interpretation of genomics data. This chapter compile most important tools together to provide a resource guide for key steps involved in microbial research. In this chapter tools and databases are listed for genome assembly, genome annotation, metagenomics, antibiotic resistance prediction, virulence factor and potential drug target identification. The purpose of this chapter is to function as a valuable reference guide for researchers facing challenges in finding and selecting the most suitable tools for their analysis.

Chapter 7 of the thesis emphasizes on the significance of genomic surveillance, using the COVID-19 pandemic as a prime example of why monitoring emerging pathogens is crucial. Outbreaks of infectious diseases lead to significant losses in human life and cause socio-economic damages. Modern technologies can produce copious amounts of data that enable us to prepare for and control such outbreaks. However, merely generating data is insufficient; it is vital to store this data in global repositories to facilitate research and report new scientific findings.

Chapter 8 of the thesis provides an exemplary instance of a global cloud repository for storing and sharing health data worldwide. The model proposed in this chapter is founded on a health data cooperative ecosystem that is primarily citizen-oriented ("For the people by the people"). This model represents an exceptional platform for predictive studies in various scientific fields. The availability of data enables its usability, which, in turn, provides significance to the raw data. The platform aims to integrate data from existing science clouds to facilitate global scientific collaboration and provide solutions for planetary health concerns.

Using the research discussed above, I have endeavoured to contribute novel information to the realm of human microbiome. However, additional experimental validations are necessary to make definitive assertions regarding the host-microbe interactions, as the *in-silico* studies are limited in their scope. It is crucial to investigate the molecular level of human microbiome relationships beyond the microbial abundance to gain a deeper understanding of this complex system. While the comparative genomic analysis conducted in this thesis provides valuable insights, experimental validation is still required to ensure that the phenotypic characteristics of the sequenced strains align with the genomic data. It is important to note that the mere presence of antimicrobial resistance genes or virulence factors does not necessarily indicate a pathogenic phenotype of a particular strain.

Like any other scientific work, it is impossible to fully encapsulate the complexity of a topic within a limited number of pages. Nonetheless, I believe the results of my research, which was carried out as part of this dissertation, have made a valuable contribution to the scientific community.



Summary

The human microbiome is a complex ecosystem that resides within our bodies. It consists of trillions of microorganisms that inhabit various niches, such as the gut, skin, oral, and reproductive organs. The human microbiome has been extensively studied in recent years and has been found to influence numerous aspects of our health, including digestion, metabolism, immune function, and even mental health. Imbalances or disruptions in the microbiome have been linked to a wide range of diseases and conditions, such as obesity, autoimmune disorders, allergies, and mood disorders.

Beyond the human microbiome, the microbiomes of other organisms, such as plants, animals, and even the environment itself, also play critical roles in maintaining planetary health. For example, plant-associated microbiomes help plants extract nutrients from the soil, protect them from pathogens, and enhance their resilience to environmental stresses. In turn, healthy plants contribute to the overall stability and productivity of ecosystems, including food production, carbon sequestration, and biodiversity conservation.

Similarly, animal microbiomes are essential for their digestion, immune function, and overall health. In natural ecosystems, animals interact with their environments and exchange microbes, influencing the microbial diversity and dynamics at a broader scale. This interconnectedness between animals, plants, and the environment forms a complex web of interactions, with the microbiome at its core.

The planetary health perspective emphasizes the interdependence of human health, animal health, and the environment. By recognizing the importance of the microbiome in maintaining these connections, we can develop strategies to promote global health and sustainability. For instance, understanding the microbiome's role in nutrient cycling and disease resistance can lead to more sustainable agricultural practices, reduced dependence on chemical fertilizers and pesticides, and improved soil health.

Furthermore, the microbiome has implications for infectious disease prevention and control. Research on microbial communities can help us better understand the transmission dynamics of pathogens and develop targeted interventions. By manipulating the microbiome, we may be able to enhance disease resistance and reduce the spread of infections.

Additionally, the microbiome has the potential to revolutionize medicine and healthcare. Advancements in microbiome research have led to the development of novel therapies, such as fecal microbiota transplantation (FMT), which involves transferring healthy microbial communities to restore the balance in individuals with disrupted microbiomes. Furthermore, probiotics and prebiotics are being explored for their potential in promoting a healthy microbiome and preventing or treating various diseases.

In conclusion, the microbiome plays a crucial role in shaping planetary health. From influencing human health to maintaining the balance of ecosystems, the microbial communities that surround us have a profound impact on the well-being of our planet. By understanding and harnessing the power of the microbiome, we can work towards a more sustainable, resilient, and healthier future for ourselves and the planet.

Part 1

First part of this dissertation, chapter 2, 3, 4 and 5 outline the host-microbe relationship and its crucial role in maintaining human health. Host-microbe interactions are complex and multifaced and can have both positive and negative impacts on human body. It is vital to distinguish between pathogenic (disease-causing) and beneficial microbes and with the advent of high-throughput technologies, our understanding of host-microbe interactions is constantly improving. These technologies allow us to generate vast amounts of data and by analyzing this wealth of data, we can gain a better comprehension of how microbes interact with their hosts at the molecular, cellular, and systemic levels.

Understanding the mechanisms by which pathogenic microbes interact with the human body can aid in the development of targeted therapies and the design of vaccines to prevent infections. Additionally, studying host-microbe interactions can help identify beneficial microbes that promote health and well-being. Furthermore, investigating host-microbe interactions can also provide insights into the development of microbial drug resistance. By studying how microbes adapt and evolve in response to therapeutic interventions, we can design more effective strategies to combat drug-resistant infections.

Part 2

The second part of this dissertation delves into the field of microbial genomics, specifically focusing on the significant amount of multiomics data generated from various techniques. The dissertation briefly explores the wide range of applications of microbial genomics. These applications can include studying microbial diversity, investigating the role of microorganisms in various ecosystems, understanding their interactions with hosts (such as in human microbiota research), and exploring their potential in biotechnology, agriculture, and medicine.

One crucial aspect discussed in this part of the dissertation is the proper storage of the vast amount of healthcare-associated multiomics data in cloud platforms. Cloud platforms provide scalable and cost-effective solutions for managing and storing large volumes of multiomics data, allowing researchers to efficiently store, analyze, and share their data with the scientific community. Managing and preserving this data is essential because it ensures that the information is readily available to the scientific community for further exploration and analysis. Availability of data is crucial for researchers to investigate key areas of research and make important discoveries.

To assist researchers in their microbial genome studies, this part of the dissertation highlights key resources that are available for microbial genome research. These resources can include databases, software/tools, computational platforms, and other relevant sources of information. By providing a guide to these resources, the dissertation helps researchers select the most appropriate computational tools and resources that best suit their microbial study.

Additionally, the dissertation emphasizes the importance of genomic surveillance, particularly in the context of detecting emerging pathogens. Genomic surveillance involves the systematic monitoring and analysis of

pathogen genomes to identify and track the spread of infectious diseases. By detecting and studying the genomic variations of pathogens, scientists can gain insights into their transmission patterns, virulence factors, and potential treatment strategies. This part of the dissertation underscores the significance of genomic surveillance as a proactive approach to public health, allowing for early detection and response to emerging infectious diseases.

Overall, this section of the dissertation provides an overview of microbial genomics, emphasizes the need for open health data clouds for proper data storage and availability, underline microbial research associated computational resources, and highlights the importance of genomic surveillance in detecting and addressing emerging pathogens.

Samenvatting

Het menselijk microbioom is een complex ecosysteem dat zich in ons lichaam bevindt. Het bestaat uit triljoenen micro-organismen die in verschillende niches leven, zoals de darmen, de huid, de mond en de voortplantingsorganen. Het menselijk microbioom is de afgelopen jaren uitgebreid bestudeerd en blijkt van invloed te zijn op tal van aspecten van onze gezondheid, zoals spijsvertering, stofwisseling, immuunfunctie en zelfs geestelijke gezondheid. Onevenwichtigheden of verstoringen in het microbioom zijn in verband gebracht met een groot aantal ziekten en aandoeningen, zoals obesitas, auto-immuunziekten, allergieën en stemmingsstoornissen.

Naast het menselijke microbioom speelt ook het microbioom van andere organismen, zoals planten, dieren en zelfs het milieu zelf, een cruciale rol bij het behoud van de gezondheid op aarde. Het plantgebonden microbioom helpt planten bijvoorbeeld om voedingsstoffen uit de bodem te halen, beschermt ze tegen ziekteverwekkers en vergroot hun veerkracht bij stress in het milieu. Op hun beurt dragen gezonde planten bij aan de algehele stabiliteit en productiviteit van ecosystemen, waaronder voedselproductie, koolstofvastlegging en behoud van biodiversiteit.

Ook dierlijke microbiomen zijn essentieel voor hun spijsvertering, immuunfunctie en algehele gezondheid. In natuurlijke ecosystemen interageren dieren met hun omgeving en wisselen ze microben uit, waardoor ze de microbiële diversiteit en dynamiek op grotere schaal beïnvloeden. Deze onderlinge verbondenheid tussen dieren, planten en het milieu vormt een complex web van interacties, met het microbioom als kern.

Het planetaire gezondheidsperspectief benadrukt de onderlinge afhankelijkheid van menselijke gezondheid, diergezondheid en het milieu. Door het belang van het microbioom in het onderhouden van deze verbanden te erkennen, kunnen we strategieën ontwikkelen om wereldwijde gezondheid en duurzaamheid te bevorderen. Zo kan inzicht in de rol van het microbioom in de nutriëntencyclus en ziekteresistentie leiden tot duurzamere landbouwpraktijken, minder afhankelijkheid van kunstmest en pesticiden en een gezondere bodem.

Bovendien heeft het microbioom implicaties voor de preventie en bestrijding van infectieziekten. Onderzoek naar microbiële gemeenschappen kan ons helpen de transmissiedynamiek van ziekteverwekkers beter te begrijpen en gerichte interventies te ontwikkelen. Door het microbioom te manipuleren, kunnen we mogelijk de weerstand tegen ziekten verhogen en de verspreiding van infecties verminderen.

Bovendien heeft het microbioom het potentieel om een revolutie teweeg te brengen in de geneeskunde en de gezondheidszorg. Vooruitgang in het microbioomonderzoek heeft geleid tot de ontwikkeling van nieuwe therapieën, zoals fecale microbiotatransplantatie (FMT), waarbij gezonde microbiële gemeenschappen worden overgebracht om het evenwicht te herstellen bij personen met een verstoord microbioom. Bovendien worden probiotica en prebiotica onderzocht op hun potentieel om een gezond microbioom te bevorderen en verschillende ziekten te voorkomen of te behandelen.

Concluderend kan worden gesteld dat het microbioom een cruciale rol speelt bij het vormgeven van de gezondheid van onze planeet. Van het beïnvloeden van de menselijke gezondheid tot het in stand houden van

het evenwicht van ecosystemen, de microbiële gemeenschappen om ons heen hebben een diepgaande invloed op het welzijn van onze planeet. Door de kracht van het microbioom te begrijpen en te benutten, kunnen we werken aan een duurzamere, veerkrachtigere en gezondere toekomst voor onszelf en onze planeet.

Deel 1

Het eerste deel van dit proefschrift, hoofdstuk 2, 3, 4 en 5, schetst de relatie tussen gastheer en microbe en de cruciale rol ervan bij het behoud van de gezondheid van de mens. Gastheer-microbiële interacties zijn complex en veelzijdig en kunnen zowel positieve als negatieve gevolgen hebben voor het menselijk lichaam. Het is van vitaal belang om onderscheid te maken tussen pathogene (ziekteveroorzakende) en nuttige microben en met de komst van high-throughput technologieën wordt ons begrip van gastheer-microbiële interacties steeds beter. Met deze technologieën kunnen we enorme hoeveelheden gegevens genereren en door deze schat aan gegevens te analyseren, kunnen we beter begrijpen hoe microben op moleculair, cellulair en systemisch niveau met hun gastheer interageren.

Inzicht in de mechanismen waarmee pathogene microben interageren met het menselijk lichaam kan helpen bij de ontwikkeling van doelgerichte therapieën en het ontwerp van vaccins om infecties te voorkomen. Daarnaast kan het bestuderen van gastheer-microbiële interacties helpen bij het identificeren van nuttige microben die de gezondheid en het welzijn bevorderen. Verder kan het bestuderen van gastheer-microbiële interacties ook inzicht verschaffen in de ontwikkeling van microbiële geneesmiddelenresistentie. Door te bestuderen hoe microben zich aanpassen en evolueren in reactie op therapeutische interventies, kunnen we effectievere strategieën ontwerpen om geneesmiddelresistente infecties te bestrijden.

Deel 2

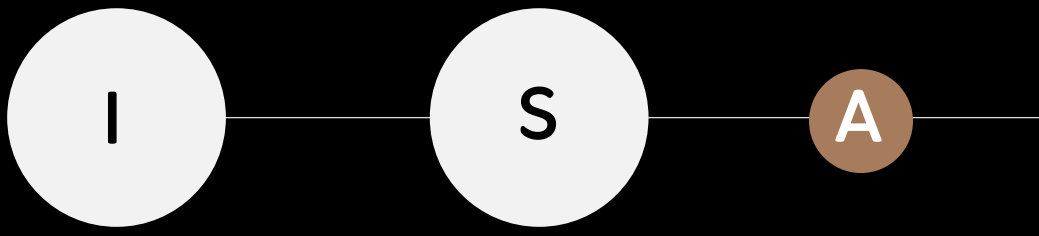
Het tweede deel van dit proefschrift gaat in op het gebied van microbiële genomics, waarbij specifiek wordt ingegaan op de significante hoeveelheid multiomics data die door verschillende technieken wordt gegenereerd. Het proefschrift verkent kort het brede scala aan toepassingen van microbiële genomics. Deze toepassingen kunnen bestaan uit het bestuderen van microbiële diversiteit, het onderzoeken van de rol van micro-organismen in verschillende ecosystemen, het begrijpen van hun interacties met gastheren (zoals in onderzoek naar de menselijke microbiota) en het verkennen van hun potentieel in de biotechnologie, landbouw en geneeskunde.

Een cruciaal aspect dat in dit deel van het proefschrift wordt besproken, is de juiste opslag van de enorme hoeveelheid multiomics-gegevens uit de gezondheidszorg in cloudplatforms. Cloudplatforms bieden schaalbare en kosteneffectieve oplossingen voor het beheren en opslaan van grote hoeveelheden multiomics-gegevens, waardoor onderzoekers hun gegevens efficiënt kunnen opslaan, analyseren en delen met de wetenschappelijke gemeenschap. Het beheren en bewaren van deze gegevens is essentieel omdat het ervoor zorgt dat de informatie direct beschikbaar is voor de wetenschappelijke gemeenschap voor verdere exploratie en analyse. De beschikbaarheid van gegevens is cruciaal voor onderzoekers om belangrijke onderzoeksgebieden te onderzoeken en belangrijke ontdekkingen te doen.

Om onderzoekers te helpen bij hun onderzoek naar het microbiële genoom, belicht dit deel van het proefschrift de belangrijkste bronnen die beschikbaar zijn voor onderzoek naar het microbiële genoom. Deze bronnen kunnen databases, software/tools, computationele platforms en andere relevante informatiebronnen omvatten. Door een gids te bieden voor deze bronnen, helpt het proefschrift onderzoekers bij het selecteren van de meest geschikte computationele tools en bronnen die het beste passen bij hun microbiële studie.

Daarnaast benadrukt het proefschrift het belang van genomische surveillance, vooral in de context van het opsporen van opkomende pathogenen. Genomische surveillance omvat het systematisch monitoren en analyseren van genomen van ziekteverwekkers om de verspreiding van infectieziekten te identificeren en te volgen. Door de genoomvariaties van ziekteverwekkers te detecteren en te bestuderen, kunnen wetenschappers inzicht krijgen in hun transmissiepatronen, virulentiefactoren en mogelijke behandelingsstrategieën. Dit deel van het proefschrift onderstreept het belang van genomische surveillance als een proactieve benadering van de volksgezondheid, die vroegtijdige detectie en reactie op opkomende infectieziekten mogelijk maakt.

Over het algemeen geeft dit deel van het proefschrift een overzicht van microbiële genomica, benadrukt het de noodzaak van open gezondheidsgegevenswolken voor goede gegevensopslag en beschikbaarheid, onderstreept het microbiële onderzoek geassocieerde computationele bronnen, en benadrukt het het belang van genomische surveillance bij het opsporen en aanpakken van opkomende ziekteverwekkers.



Acknowledgements

"*If you work on something a little bit every day, you end up with something that is massive.*"

-Kenneth Goldsmith

The path to a PhD is hard but we don't have to walk alone. In the final chapter of my thesis, I would like to thank and acknowledge all the personalities who have supported me in my PhD journey.

First and the main character of my PhD, my Pilot/ Superhero/ Supervisor, Dr. Angela Brand. Words are not enough to appreciate or express the kind of support and love you showed me throughout my journey. I am fortunate to have a supervisor like you who cared for me like a mother, I think that is why everyone in Manipal felt that you are my second mother. I will always remember our non-ending sessions which brought us so close that I could share things with you so comfortably. You were always so kind and open to discuss, to clear my doubts regarding any difficulties I faced. I learned so much from your actions and your way of treating other people. I will be always grateful to have you as my PhD supervisor. A big thank you for your support, guidance, care and kindness in my journey. Thank you so much!

The next person I would like to thank is my Co-Supervisor/ Co-Pilot, Dr. Kapaettu Satyamoorthy. Sir, you taught me some of the best qualities a PhD candidate should have. You taught me to approach a problem with multiple solutions. You showed me that it is not always necessary to take a long way instead choose a smart one. You always made me triple-check my results and taught me to be thorough with the literature background. Your critical reviews and suggestions always turned our projects into sensible ones. I will always remember our long conference room meetings. I thank you for your great support and guidance and for being a critical part of my PhD journey. I will never forget it.

Everyone deserves a break and sometimes it's important to go out, have drinks and have a random conversation to bring a new perspective in life. I would like to thank Dr. Helmut Brand for being that person who always brought new perspectives to my life. Thank you for all the stories you shared during our dinner time. You brought a lot of fun to the table with your jolly nature. It was nice to have you around and your support was vital to me. You are a great person and a big thank you for being so supportive.

I would like to thank all my international project collaborators, Dr. Hans Lehrach, Dr. Nikolaos Evangelatos (Nikos), Dr. Lesley Ann Ogilvie, Julien Venn and Marius Geanta. I have learned so much from them while working on our publications. Comments and suggestions from all my collaborators were vital for successful publications. A special thanks to Dr. Hans Lehrach for allowing me to work in his lab in MPI, Berlin.

The next people I want to thank are my project collaborators in Manipal, Shruptha Padival, Apoorva Jnana, Dr. Murali and Dr. Bobby Paul. Thank you for the collaborations and discussions we had. I learned a lot from each of you and I enjoyed working with all of you. Best wishes to all of you.

I would like to thank the most important part of my PhD journey, my friends/ my boys/ Manipal Warriors. Guys thank you so much for those countless memories and moments we all had. Manipal was only a better

place but you guys made it great. Thank you so much for those long trips, hikes, late-night beach visits, amazing food joints, and those long conversations in the canteen. Thank you for making conferences fun and thank you for covering up for me. Thank you so much for those laughs we all had together. The list is just too big and it's hard to cover all of it in a page. So, my boys, Sathvik, Dinesh, Pradyumna, Jishnu, Akshay and Prasanth, thank you for such an amazing time. I will always miss our time together. Love you all. Marvel Forever!

I would like to thank my best friends, Dhruv and Sumit for always being so supportive and guiding me in the right direction whenever I was going off track. Lucky to have you both in my life. Thank you so much.

I have spent a great amount of time (11 years) in the Manipal School of Life Sciences (MSLS), from my bachelor's to PhD. I have seen a lot of people coming and going and some constants too. For my entire time in MSLS, I would like to thank all the directors, faculties, technicians, non-teaching staff, maintenance staff and security team for their contribution to my Manipal life. Learned a lot about MSLS and it provided me with such great opportunities to transform myself into a better version. Thank you all.

To the new addition to my family, my wife, Priya. I would like to thank you for patiently waiting for me while I was busy writing my thesis. Thank you for your support and time. Most importantly thank you so much for understanding. THANK YOU.

I would like to thank my biggest and constant supporters in life, my parents and my sister. Thank you for giving me an opportunity to learn and explore my interests. Thank you for being there for me always. I can’t thank you enough for all your sacrifices just to fulfil my wishes. A big thank you to my sister Ashi, for always motivating me. Thank you, mom and dad, for believing in me and I hope you guys feel proud of me (Means a lot to me). Thanks to you, I am standing here today and defending my thesis to become a PhD. Thank you for everything.

Cheers to all of you and best wishes

Curriculum Vitae

Ankit Singh Tanwar

Mobile: +91 8123000722

ankittanwar12@gmail.com | LinkedIn ID | Google Scholar | ResearchGate

Summary

Bioinformatician with 6+ years of experience in high-throughput data analysis, metagenomics, single-cell transcriptomics, microbial genomics, and proteomics. Worked extensively with 16S rRNA sequencing data, particularly in analysing gut, oral, and cervical microbiomes. Also, well-versed in handling whole genome datasets obtained from different sequencing platforms such as Illumina, 10xGenomics, Ion-torrent, and Oxford Nanopore.

Professional Experience

❖ **Ph.D. fellow (Faculty of Health, Medicine and Life Sciences (FHML), Maastricht University, Maastricht, The Netherlands)**
[Apr 2020 – Present]

- ✓ Conducted research activities in the field of bioinformatics and public health genomics. Worked on microbial genomics-oriented research projects to understand host-microbe interactions. Communicated research findings through publications in peer-reviewed journals.

❖ **Research fellow (Manipal School of Life Sciences, MAHE, Karnataka, India)**
[Oct 2017 – Sept 2023]

- ✓ Developed expertise in various types of sequencing data analysis (RNA-Seq, ChIP-Seq, long read sequencing and miRNA-Seq) and performed robust data quality control and validation. Routine maintenance and updating in-house server for data analysis.
- ✓ Supported research staff with NGS data analysis and visualization and trained junior research fellows, bachelor's and master's students in bioinformatics-oriented research projects.
- ✓ Organized and demonstrated NGS data analysis in various hands-on workshops. Contributed to conferences, webinars and symposiums by presenting research outputs. Published collaborated research outputs in peer-reviewed journals.

❖ **Intern – Bioinformatics (Max Planck Institute for Molecular Genetics, Berlin, Germany)**
[Sep – Oct 2018]

- ✓ Performed single cell RNA-seq data analysis of melanoma patients. Worked on DigiTwin project to model biological pathways and reactions using PyBioS and libSBML (Matlab libraries).

Research Projects

- Metagenomic analysis of gut and oral microbiome of individuals with specific ayurveda prakriti.
- Metagenomic analysis of cervical cancer microbiome in different cancer stages associated with HPV infection.
- Whole genome sequencing and analysis of *C. butyricum*, *C. difficile* and *C. tertium* strains and exploring their pathogenic role in *C. difficile* infections (CDI).

- Whole genome comparison of four clinical strains of *S. aureus* involved in diabetic foot ulcers. To explore differences and similarities between their antimicrobial and virulence properties.
- Single-cell transcriptomics data analysis of melanoma patients to identify intra and inter-tumor heterogeneity.
- In-silico motif prediction for zinc-finger genes and their role in target gene expression.
- Optimization of diagnostic testing strategy for Triplet Repeat Expansion in HTT, FMR1 genes causing Huntington disease and Fragile X syndrome.
- Building a linux-based tool for estimating coding density from whole genome sequences.

Key Skills

- **Sequencing datasets:** DNA-seq, RNA-seq, ChIP-seq, single cell RNA-seq and WGS.
- **Analysis:** Metagenomic, gene expression, mutational, phylogenetic, network and pathway, protein modelling, docking and simulation.
- **Visualization:** ggplot2 and Matplotlib.
- **Others:** Docker, GitHub, Gitpod.
- **Workflow manager:** Nextflow and nf-core.
- **Programming languages:** Shell scripting, R, Python, PERL and MATLAB.
- **Microsoft tools:** Power BI, Excel, Word and PowerPoint.
- **Operating systems:** Linux, Windows and MacOS.
- **Web and database:** HTML, CSS, XML and MySQL.
- **Design and editing:** Adobe illustrator, Inkscape and Canva.
- **Lab skills:** PCR, cell culture, DNA isolation and flow cytometry.

Education

Course	University	Passed Year	Marks (%)
MSc Bioinformatics	Manipal School of Life Sciences, MAHE	2017	78
BSc Biotechnology	Manipal School of Life Sciences, MAHE	2015	67

Presentation, Conferences and Courses

- Demonstrated NGS data analysis at the "High-End workshop on Next Generation Sequencing Data Analysis using Workflows with Nextflow and nf-core". (July 2022)
- Participated in one-month online workshop (hands-on) "Introduction to Computational Drug Design" co-organized by Schrödinger and Pharmacy Council of India. (21st Sept – 23rd Oct 2020)
- Presented research poster at the international symposium titled "Genome instability: from bench to bedside". (January 2020)
- Demonstrated data analysis at workshop on "Data Science for Genomics". (February 2019)
- Presented research poster at the international conference on "Advances in Cellular, Genomic and Epigenomic Insights on Environmental Mutagenesis and Health", EMSI. (January 2017)
- Attended summer training program organized by Manipal University Student Research Forum (MUSRF) in Manipal. (July 2016)

Extracurricular activities and Achievements

- Won best outgoing sportsmen award for the year 2015.
- Captain of the intra-college football team secured first prize in year 2015.
- Won best football player award in year 2015.
- Secured multiple gold, silver and bronze medals in athletic events held during the annual sports day between the year 2012-2018.
- Secured third prize in fashion show competition of 'Utsav' (University level) for the year 2012 and 2014.
- Secured second and third prize in western dance competition for the year 2013 and 2014.

Publications

Publications

1. **Tanwar, A. S.**, Shruptha, P., Jnana, A., Brand, A., Ballal, M., Satyamoorthy, K., & Murali, T. S. (2023). Emerging pathogens in planetary health and lessons from comparative genome analyses of three *Clostridia* species. OMICS: A Journal of Integrative Biology, 27(6), 247-259.

2. **Tanwar, A. S.**, Shruptha, P., Paul, B., Murali, T. S., Brand, A., & Satyamoorthy, K. (2023). How can omics inform diabetic foot ulcer clinical management? a whole genome comparison of four clinical strains of *Staphylococcus aureus*. OMICS: A Journal of Integrative Biology, 27(2), 51-61.

3. **Tanwar, A. S.**, Evangelatos, N., Venne, J., Ogilvie, L. A., Satyamoorthy, K., & Brand, A. (2021). Global open health data cooperatives cloud in an era of COVID-19 and planetary health. OMICS: A Journal of Integrative Biology, 25(3), 169-175.

4. Samantray, D., **Tanwar, A. S.**, Murali, T. S., Brand, A., Satyamoorthy, K., & Paul, B. (2023). A Comprehensive Bioinformatics Resource Guide for Genome-Based Antimicrobial Resistance Studies. OMICS: A Journal of Integrative Biology, 27(10), 445-460.

5. Shalini, T. V., Jnana, A., Sriranjini, S. J., **Tanwar, A. S.**, Brand, A., Murali, T. S., ... & Gangadharan, G. G. (2021). Exploring the signature gut and oral microbiome in individuals of specific *Ayurveda prakriti*. Journal of Biosciences, 46(3), 1-20.

6. Geanta, M., **Tanwar, A. S.**, Lehrach, H., Satyamoorthy, K., & Brand, A. (2022). Horizon scanning: Rise of planetary health genomics and digital twins for pandemic preparedness. OMICS: A Journal of Integrative Biology, 26(2), 93-100.

7. Shetty, S. S., Sharma, M., Fonseca, F. P., Jayaram, P., **Tanwar, A. S.**, Kabekkodu, S. P., ... & Radhakrishnan, R. (2020). Signaling pathways promoting epithelial mesenchymal transition in oral submucous fibrosis and oral squamous cell carcinoma. Japanese Dental Science Review, 56(1), 97-108.

8. Shruptha, P., **Tanwar, A. S.**, Jayaram, P., Brand, A., & Satyamoorthy, K. Taxonomic diversity and functional profiling of cervical microbiota associated with cancer progression. (Journal submitted to: npj Biofilms Microbiomes)