Microbiome Surfaceome and Secretome Atlas: Building a dictionary of putative surface and secreted proteins on the Human Microbiome

A Project Report

submitted by

LAVANYA CB

in partial fulfilment of the requirements
for the award of the degree of

MASTER OF TECHNOLOGY

COMPUTATIONAL BIOLOGY
INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY DELHI

NEW DELHI- 110020

18-05-2024

THESIS CERTIFICATE

This is to certify that the thesis titled **Microbiome Surfacome and Secretome Atlas: Building a dictionary of putative surface and secreted proteins on the Human Microbiome** , submitted by **LAVANYA CB**, to the Indian Indraprastha Institute of Information Technology Delhi (IIIT-Delhi) for the award of the degree of **MASTER OF TECHNOLOGY**, is a bona fide record of the research work done by her under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Dr. Tarini Shankar Ghosh
Thesis Supervisor
Assistant Professor
Dept. of Computational Biology
IIIT Delhi, 110020

Place: New Delhi

Date: 19th January 2024

ACKNOWLEDGEMENTS

I extend my sincere gratitude to Dr. Tarini Shankar Ghosh for his invaluable guidance, and to our dedicated team for their technical assistance and collaboration. Special thanks to the institution for providing the resources and environment necessary for this project's success. With their support, I've embarked on a journey of discovery, aiming to make a meaningful impact in our field.

Abstract

Microbial effectors like genotoxins have been shown to play a crucial role in the pathogenesis of diseases like colorectal cancer. These effectors are agents capable of causing DNA damage within cells and are pivotal in CRC development. Our study consists of two parts. The first part begins by aiming to discern genotoxins produced by microbes and elucidate their distribution patterns. Microbes, like bacteria, highly contribute significantly to genotoxin production due to their metabolic versatility and environmental interactions. By identifying and mapping microbial genotoxins, we can grasp their role in disease parthenogenesis and evaluate their similarities across diverse microbial species.

In the second part, initiated to identify novel peptides or effectors with genotoxic potential, we extend our investigation to identify transmembrane proteins and signal peptides across all microbial species is paramount in understanding the intricate mechanisms underlying microbial physiology and pathogenesis. Utilizing reference proteomes from 169 species, we have created a proto-type of an atlas of the putative surfaceome and secretome of the human gut microbiome. This atlas is organized in clusters generated using similarities of various protein properties.

To demonstrate the application of this atlas, we have leveraged this database to map to identify homologs of different genotoxin proteins and proteins showing similarity to different human surfacome and secretome proteins, focusing on a core set of 169 species. This investigation unearthed numerous novel proteins and elucidated their potential roles in microbial pathogenesis and host-microbe interactions. Our analysis has revealed significant overlaps between microbial proteins and disease-producing human proteins, putatively shedding light on potential mechanisms of microbial pathogenicity and disease progression.

Keywords: Genotoxins, colorectal cancer, transmembrane proteins, signal peptides, microbial species, database creation, bioinformatics tools, protein prediction, clustering, microbial pathogenicity, disease progression, host-microbe interactions, infectious diseases, global health.

Contents

List of Figures

Chapter 1

Introduction

Understanding the impact of genotoxins on the human gut microbiome is vital for comprehending their implications for health and disease, particularly in the context of colorectal cancer (CRC). The gut microbiome, consisting of trillions of microorganisms, plays a pivotal role in human physiology, affecting digestion, immune function, and mental well-being. However, genotoxins pose a significant threat to the stability and function of this microbial community. Genotoxins encompass various substances capable of causing DNA or chromosomal damage, with potential implications for heritable changes and cancer initiation, including CRC. While extensive testing methods have been developed to detect genotoxicity across different cellular contexts, understanding how genotoxins interact with the gut microbiome, particularly in the development of CRC, remains a challenge.Exploring microbial peptides that correlate with genotoxins offers significant potential in deciphering the complex dynamics within the gut microbiome. By pinpointing peptides that interact with genotoxic agents, researchers gain insights into the pathways through which these toxins impact microbial communities. This understanding can pave the way for strategies aimed at minimizing genotoxin-induced harm and preserving microbiome equilibrium. Furthermore, the identification of microbial peptides associated with genotoxins opens avenues for the discovery of novel biomarkers, crucial for assessing genotoxicity levels and predicting individual susceptibility to diseases like colorectal cancer. Moreover, the elucidation of specific microbial peptides influencing genotoxin activity holds promise for the development of tailored interventions aimed at safeguarding against genotoxic damage and lowering the risk of related illnesses. This investigation into the interplay between microbial peptides and genotoxins not only deepens our comprehension of environmental influences on the gut microbiome but also offers avenues for targeted therapeutic interventions to uphold gut health and overall well-being.[1]

Moreover, The transmembrane and secretome systems of microbes are crucial for their survival and interaction with the environment, forming the foundation for innovative applications like bacterial biosensors. Microbes utilize their transmembrane receptors and signaling pathways to sense and respond to various stimuli, enabling them to adapt their behavior accordingly. This capability allows for the engineering of bacteria to detect toxins, monitor process parameters, and even combat infections by producing targeted therapeutic agents. Additionally, the secretome, comprising proteins and molecules released by microbes, further enhances their ability to interact with their surroundings and modulate host-microbe interactions. Leveraging these natural systems in bacterial biosensors holds promise for revolutionizing diagnostics, therapeutics, and environmental monitoring, offering cost-effective, portable solutions for diverse applications in medicine and beyond[2].

Until now, there has been a conspicuous absence of comprehensive data on the transmembrane receptors and secretome constituents of the human gut microbiome. This pioneering initiative seeks to fill this critical void, embarking on an ambitious endeavor to map and characterize these molecular landscapes. By delving into uncharted territory, we aim to revolutionize our understanding of microbial-host interactions within the gut ecosystem. This groundbreaking atlas holds immense promise for unveiling novel insights into microbial communication networks, nutrient acquisition strategies, and immunomodulatory functions. Moreover, it lays the foundation for precision medicine approaches aimed at modulating gut microbiome function to alleviate a spectrum of gastrointestinal and systemic disorders.

Our study aimed to identify novel genotoxic peptides within the human gut microbiome, expanding our understanding of their potential implications for colorectal cancer (CRC) and related diseases. This endeavor led to the development of the first prototype atlas of the human gut surfacome and secretome, providing intense insights into microbial-host interactions. This atlas not only brighten our understanding of gut microbiome motility but also holds promise for advancing diagnostics, therapeutics, and precision medicine approaches. Through collaborative efforts, it serves as a valuable resource for interdisciplinary research and paves the way for transformative discoveries in microbiology and healthcare.

Chapter 2

METHODS

2.1 Literature review and protein identifications

In our investigation of genotoxins, we conducted an extensive literature search that led in the identification of genotoxic impacts on disease pathogenesis. Our focus involved in understanding the intricate roles played by genotoxins and their associated pathways, thereby broadening our understanding of the underlying molecular mechanisms. We utilized targeted search terms to locate genotoxin proteins, employing keywords like "genotoxin," "DNA damage", "microbiome"," "mutagen," gut microbiome," and "genotoxicity and microbes"." By using these specific terms, we aimed to efficiently identify relevant literature and databases containing information on proteins involved in genotoxic processes.

2.2 Methods

Our research had two primary objectives. Firstly, we sought to identify novel genotoxin-associated peptides present across a variety of microbial species, employing sophisticated bioinformatics methodologies. Secondly, we aimed to develop a foundational dictionary elucidating the complexities of the human gut microbiome. Additionally, we intended to integrate this atlas with the human surfacome and secretome database. This section will provide a detailed methodology of the two main objectives.

2.2.1 Identifying Novel Peptides or Effectors with Genotoxic Potential

Identifying and mapping the microbial genotoxins helps to uncover harmful substance and also to investigate the role of unknown microbes in producing them.

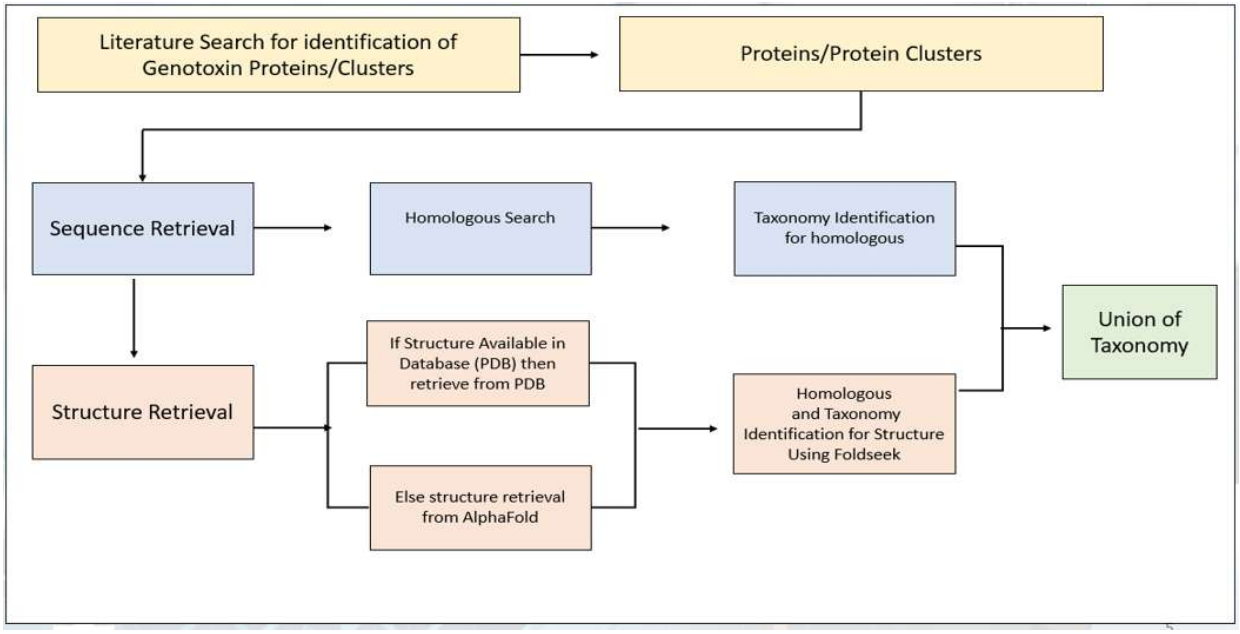

Figure 2.1: Workflow of Genotoxin

An extensive search to identify genotoxic proteins and their associated counterparts, revealing a multitude of proteins and enzymes with carcinogenic potential. Following thorough validation, the scope was narrowed down to 22 proteins, where after our focus shifted towards delineating their associated proteins. The rationale for focusing on clustered proteins stemmed from the fundamental understanding that proteins seldom operate in isolation; rather, they intricately interact with a network of associated proteins to execute their biological functions effectively. This interconnectedness underscores the importance of studying protein clusters, as they often collaborate synergistically to orchestrate complex cellular processes. For instance, colibactin is a genotoxin produced by certain strains of Escherichia coli, notably those possessing the polyketide synthase (pks) genomic island. Within this island, the clbB gene encodes an enzyme involved in colibactin biosynthesis. Studies have shown that colibactin-producing E. coli strains

are associated with an increased risk of colorectal cancer, highlighting the significance of understanding its mechanisms of action and associated proteins in pathogenesis [3]. Leveraging the diamond tool, conducted a sequence homology search, applying rigorous standards, with 80% standard query coverage and 90% mapping identity threshold for the homologous searches. These searches were conducted across two extensive metagenomic sequence databases, namely MGnify and EggNOG, ensuring comprehensive exploration of genetic data. Subsequently, the structural sequences of these 22 proteins were meticulously identified. In cases where three-dimensional structures were unavailable, it was harnessed by the predictive capabilities of the Alphafold tool to model their intricate three-dimensional architectures, ensuring a comprehensive understanding of their functional dynamics and molecular interactions. Structural homology analysis of PDB structures was conducted using Foldseek., a bioinformatics tool utilized for predicting protein. A structural similarity search was conducted, employing

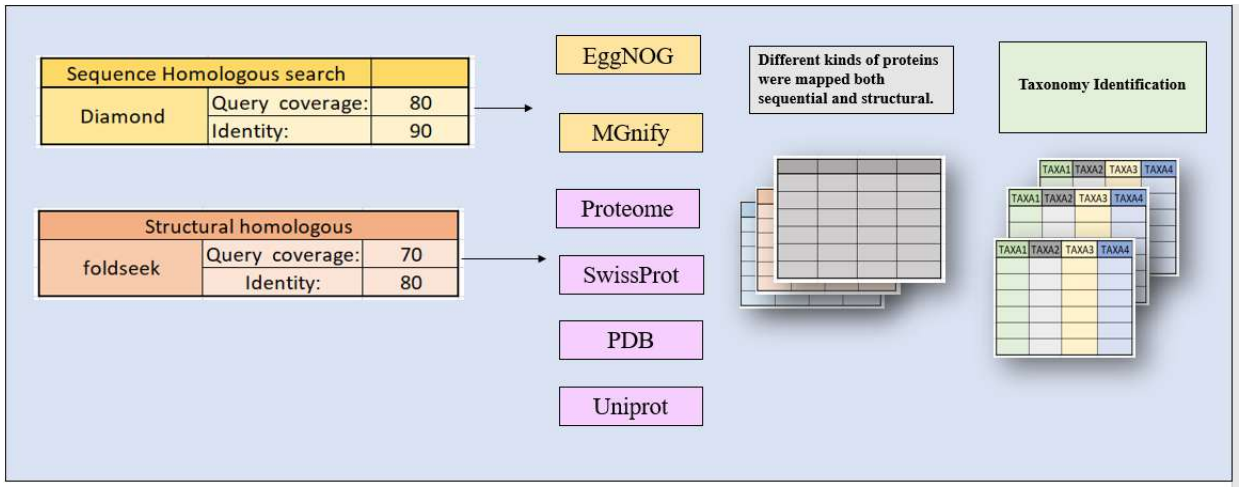

Figure 2.2: Homologous identification and Taxonomy Identification

a blast algorithm to compare three-dimensional protein structures against four distinct databases: Proteome, SwissProt, PDB (Protein Data Bank), and UniProt. The Proteome database encompasses the entirety of an organism's encoded proteins, offering insights into its functional landscape. SwissProt provides a meticulously curated collection of protein sequences, ensuring data reliability and accuracy. PDB serves as a global repository for experimentally determined protein structures, supporting struc-

tural analysis and drug discovery endeavors. UniProt integrates diverse data sources, including SwissProt, TrEMBL, and PDB, offering a comprehensive platform for protein sequence and functional information. Each database contributes uniquely to structural bioinformatics, collectively enhancing our understanding of protein structure, function, and evolution [Figure:2.2].

Further, the homologous results are carried out for taxonomy identification. This step is vital as it allows the classification of the identified proteins within their evolutionary lineages, providing insights into their evolutionary relationships and functional implications. The abundance of proteins retrieved through homologous searches underscores the importance of taxonomy identification in discerning common ancestry and evolutionary conservation among these proteins, thus enriching our understanding of their biological roles and significance in cellular processes. During the investigation of genotoxin peptides, a notable observation emerged regarding the involvement of numerous proteins in cell membrane interactions and signal transduction pathways. This prompted a strategic shift in focus towards these domains, aiming to enhance the understanding of microbial behavior through the exploration of trans-membrane and signal peptides.

2.2.2 Creation of the initial prototype for a Surfacome and Secretome Atlas dedicated to the human gut microbiome.

Identifying trans-membrane and signal peptides of microbes serves as a fundamental endeavor for several compelling reasons. Firstly, these peptides are pivotal in orchestrating the intricate interplay between microbial cells and their surrounding milieu, which includes crucial interactions with host cells in pathogenic contexts. By comprehensively understanding the composition and attributes of these peptides, we gain invaluable insights into the mechanisms underpinning microbial colonization, invasion, and host-cell modulation. Secondly, trans-membrane and signal peptides often serve as conduits for the transport of various molecules across cellular membranes, encompassing toxins, virulence factors, and other bioactive compounds. By unraveling the identities and functionalities of these peptides, we unravel the intricate tactics employed by microbes to subvert host defenses and propagate infection. Moreover, the discernment of trans-membrane and signal peptides holds promise in the realm of therapeutic interventions and vaccine development, as they represent potential targets for disrupting microbial pathogenicity and virulence. In essence, the identification of trans-membrane and signal peptides stands as a cornerstone in unraveling microbial behavior, pathogenesis, and host-microbe interactions, while also offering avenues for the innovation of novel therapeutic modalities [2].

To enable the creation of the expansive human gut Surfacome and Secretome Atlas, it was imperative to access robust and comprehensive databases. This pivotal requirement was adeptly fulfilled through the utilization of esteemed microbial proteome database resources, prominently featuring MGnify and UNINA Segata. Notably, MGnify boasts an extensive repository comprising approximately 204,938 genomes originating from 4,744 prokaryotic organisms, thereby providing an extensive taxonomic spectrum. Impressively, taxonomic information at the species level is available for a substantial subset, encompassing 3,577 entries [4]. Likewise, the UNINA Segata database encompasses a diverse array of metagenomic assembly genomes, collectively spanning nearly 154,723 sequences. This remarkable dataset curation was meticulously conducted, incorporating rigorous criteria such as the establishment of a sequence threshold to exclude sequences with less than 5% genome contamination and ensuring over 90% completeness [5].

A comprehensive pipeline was developed to facilitate the creation of a database encompassing trans-membrane and Secretome proteins sourced from a vast collection of microbial proteomes. This pipeline entailed the collection and storage of all proteomic data in the FASTA format. Integral to this process was the utilization of TMHMM 2.0, an advanced trans-membrane prediction tool. Operating on a probabilistic framework utilizing hidden Markov models, TMHMM 2.0 analyzes protein sequences to predict their placement relative to the cellular membrane. Outputs from this tool delineate whether the sequence resides within the inner loop, trans-membrane loop, or outer loop. This approach was meticulously orchestrated to ensure the creation of a robust database while adhering to professional standards and ethical guidelines [6].

The output of the TMHMM tool yields three essential files: the ".annotation" file, which outlines the orientation of fasta sequences concerning inner, outer, and trans-membrane regions; the ".summary" file, providing a detailed summary of amino acid residues within these regions; and the ".plot" file, offering a visual representation of the results. Leveraging data from these files, a filtration process is applied to isolate protein sequences predicted as surfacome, enabling subsequent analysis. The residue summary data from the ".summary" file is then utilized to filter the outside residue which highly contributes as secretome and applied SignalP6.0 tool that works like a detective for proteins. It looks at the sequence of amino acids in a protein and tries to find special patterns that signal where the protein needs to go in the cell. These patterns are called signal peptides. Once it finds them, it predicts where the signal peptide gets cut off, which is important for the protein to function correctly. It's like finding the right address on an envelope and predicting where to tear it open to get to the letter inside which helped us predict secretome proteins, thereby enhancing the understanding of cellular secretion mechanisms [7]. Subsequent to these analyses, a comprehensive

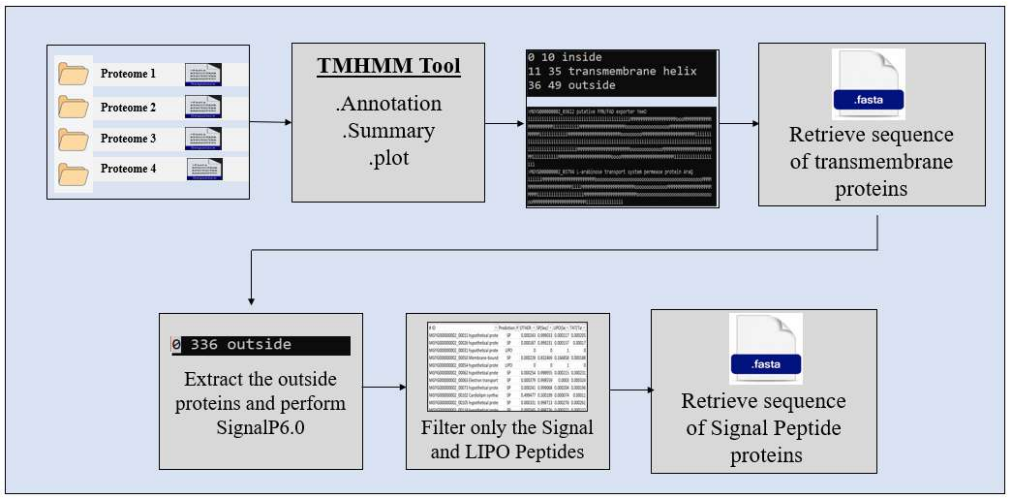

Figure 2.3: Pipeline for surfacome and secretome prediction

atlas is constructed, housing protein sequence information earmarked as surfacome and secretome constituents. This atlas serves as a pivotal resource, facilitating in-depth exploration and investigation into cellular surface and secretory processes [Figure:2.3].

The pipeline was executed specifically for 169 gut-associated species, representing globally prevalent taxa observed across numerous individual samples. This targeted approach ensured a focused analysis on organisms commonly found in diverse microbiomes, thereby providing insights into shared features and functionalities across various gut microbiota compositions. Out of the 169 gut-associated species examined, 130 were identified through filtering from UNINA data, while the remaining 39 were extracted from MGnify datasets. This meticulous selection process ensured a comprehensive representation of gut microbiota species from diverse sources, facilitating robust analyses and interpretations.

2.2.3 Deciphering the Taxonomic Landscape: Clustering Surfacome and Secretome Proteins of the Human Gut Microbiome

Now that the atlas of the surfacome and secretome of the human gut microbiome has been developed, next stage is to build the dictionary inorder to make sense of the data. As part of this, carried out the clustering of the surfacome and scretome proteins using MMseqs2, a bioinformatics tool used for fast and sensitive sequence searching and clustering. It works by employing advanced indexing and search algorithms to

quickly compare large sets of sequences [8]. Linclust utilizes linear time and memory complexity algorithms for fast clustering of massive sequence datasets. This module is particularly efficient for clustering protein sequences into families or clusters based on sequence similarity. It employs a threshold-based approach to identify clusters, where sequences are grouped together if they share a certain degree of similarity. Linclust is a highly regarded tool within the bioinformatics community, esteemed for its efficacy in diverse applications such as protein family classification, functional annotation, and evolutionary analysis. Its popularity stems from its notable attributes including speed, scalability, and accuracy. Leveraging a threshold-based approach, Linclust efficiently clusters sequences, making it particularly adept for organizing large datasets of protein sequences into meaningful groups. In our study, we employed Linclust to cluster all surfacome and secretome proteins, a task pivotal for unraveling their functional roles and evolutionary relationships. This strategic utilization underscores the tool's significance in elucidating complex biological phenomena.

The post-processing of clustered results derived from MMseqs2 involved extracting representative clusters and their corresponding cluster members, followed by the identification of each cluster member's species origin. This step was essential for elucidating the taxonomic composition within each cluster and understanding the evolutionary relationships among the clustered sequences. By concatenating species names to individual protein sequences prior to clustering, we simplified the process of associating sequences with their respective species. This facilitated a comprehensive analysis of the taxonomic distribution within each cluster, enabling us to discern patterns of species co-occurrence and identify conserved features across related taxa [Figure:2.4]. This detailed taxonomic annotation is crucial for inferring functional properties and ecological roles associated with specific microbial communities, thus enhancing our understanding of the complex dynamics within gut-associated microbiomes.

In our endeavor to ensure robust cluster representation, we removed the prevalent occurrence of singleton sequences within the representative clusters, prompting a critical reassessment. To address this potential source of bias, established a prudent threshold criterion, restricting the focus to representative clusters harboring five or more cluster members. This judicious refinement not only bolstered the integrity of subsequent analyses but also augmented the reliability of the findings. Subsequent to this curation, embarked on the construction of a species abundance matrix across surfacome and secretome proteins, documenting their distribution across the diverse microbial species inhabiting the intricate landscape of the gut microbiome. This matrix, a fundamental cornerstone of the investigation, served as a quantitative depiction of protein abundance dynamics across microbial taxa, offering valuable insights into the interplay between microbial communities and their functional attributes. To further unravel the

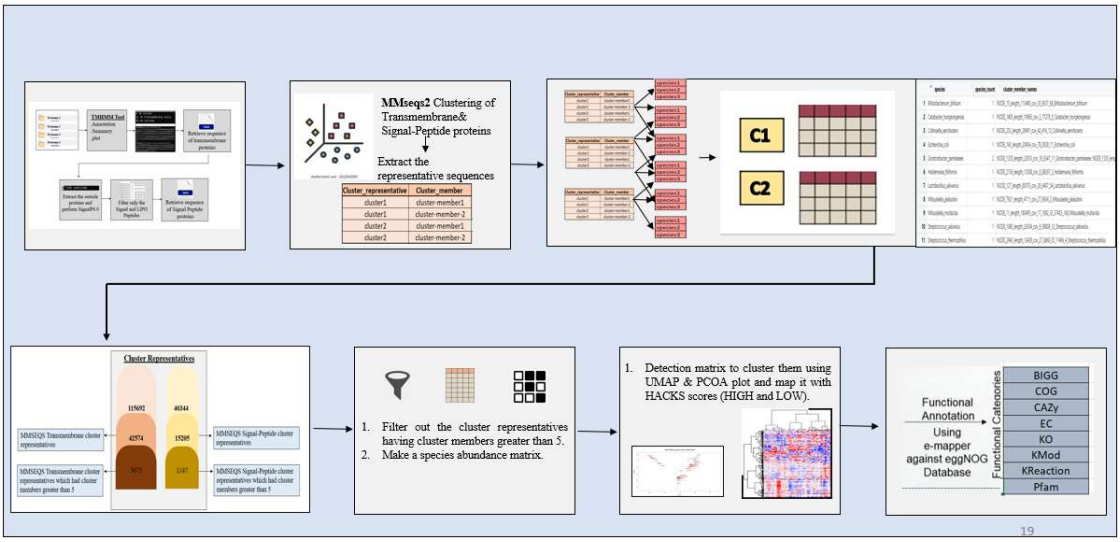

Figure 2.4: Unveiling Insights: Clustering and Interpretation of Gut Microbiome Protein Data

underlying patterns within this complex web of protein abundances, employed the Principal Coordinate Analysis (PCoA), a powerful multivariate statistical technique. PCoA holds exceptional utility in discerning relationships among multidimensional datasets, allowing for the exploration of similarities and dissimilarities between samples based on their abundance profiles. Unlike traditional Principal Component Analysis (PCA), it transforms high-dimensional data into a reduced set of orthogonal axes (principal coordinates), enabling visualization and interpretation of complex relationships within the data. By employing PCoA, we aimed to unravel the underlying structure of our dataset, elucidating new patterns and associations that may not be discernible through traditional approaches. Ultimately, the integration of PCoA into the analytical framework served as a cornerstone for uncovering the intricate interrelationships among clustered proteins and their abundances across microbial taxa, thus advancing our understanding of the ecological dynamics within the gut microbiome. [Figure:2.4].

Functional annotation of proteins is essential for understanding their roles in biological processes. It enables the prediction of protein function, guiding biomedical research and biotechnological applications. Additionally, functional annotation aids in comparative genomics, facilitating the study of evolutionary relationships across species. Overall, it provides crucial insights into cellular functions and informs advancements in various fields, including medicine and biotechnology. A comprehensive functional an-

notation was conducted for the entire atlas of surfacome and secretome proteins generated thus far. This annotation utilized the e-mapper tool, integrated with the eggNOG database, to provide insights into the functional classification of proteins. The annotation yielded information about various functional categories, including: Bacterial Integrated Gene Catalog (BIGG), Clusters of Orthologous Groups (COG), Carbohydrate-Active enZymes (CAZY), Enzyme Commission (EC) numbers, KEGG Orthology (KO), KEGG Modulation, KEGG Reaction, and Protein family (Pfam) domains. These annotations offer valuable details regarding protein function, evolutionary relationships, metabolic pathways, and protein domain architecture, enhancing our understanding of the biological roles and regulatory mechanisms governing surfacome and secretome proteins [Figure:2.4].

2.2.4 Assessing Cluster Validity: Intra and Inter-Cluster Similarity Analysis

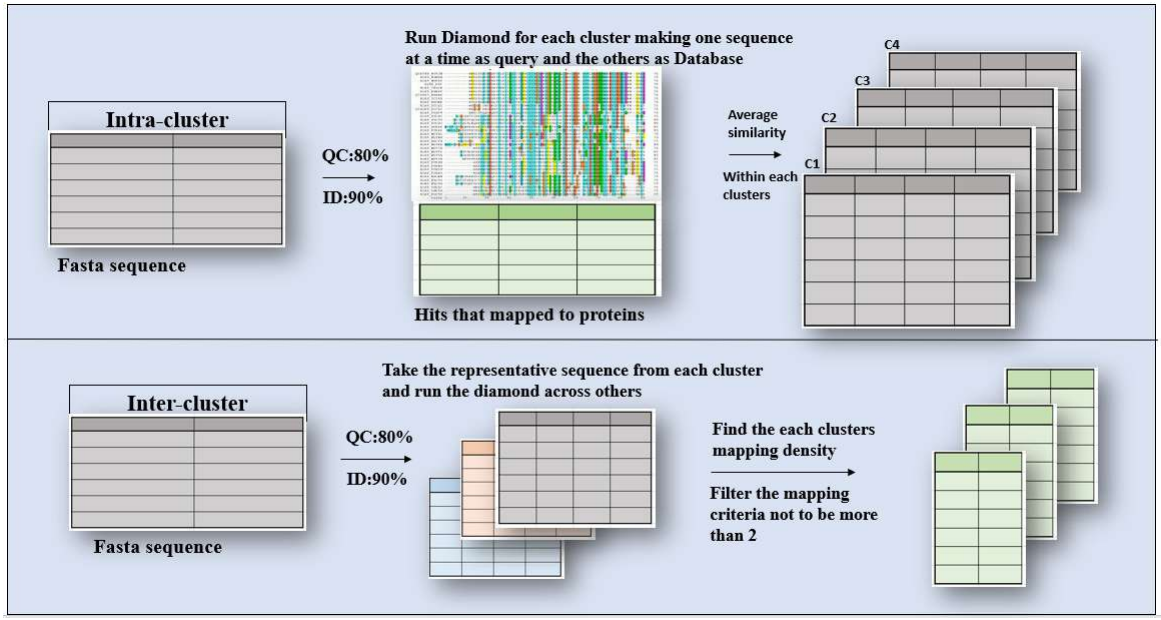

Figure 2.5: Validation workflow for Surfacome and Secretome clusters of Human gut Micorbiome

Validating the clusters helps ensure they accurately represent the data and are

biologically meaningful. Ultimately, validation enhances our confidence in the reliability and interpretability of the clustering outcomes. The validation of surfacome and secretome clusters from the human gut microbiome involved assessing the intracluster similarity to evaluate protein cohesion within each cluster, as well as measuring inter-cluster dissimilarity to discern differences between proteins across clusters. This comprehensive approach ensured the robustness and accuracy of the clustering results, providing valuable insights into the functional and taxonomic relationships among gut microbiome proteins.

The assessment of intracluster similarity involved conducting similarity searches within each surfacome and secretome protein cluster. This process utilized one protein from the cluster as a query sequence against the remaining proteins within the same cluster, with a query coverage of 80% and an identity threshold of 90%. The average similarity was then calculated within each cluster. Similarly, for intercluster similarity analysis, representative sequences from each cluster were compared using the same criteria against representatives from other clusters. To ensure the quality of clustering, it was crucial to examine the mapping density of each representative against others, with a predetermined threshold of no more than two mapping hits per representative. This rigorous approach provided a reliable assessment of both intra- and intercluster similarities, thus validating the clustering outcomes.

2.2.5 Optimizing Analysis: Mapping Human Gut Microbiome Secretome and Surfacome Proteins to Human Counterparts

The final objective of the study is to explore the applications and potential utility of our data. we opted to identify the microbial species influencing the human surfacome and secretome due to their pivotal role as the initial interface where microbes interact with the host. This decision stemmed from the recognition that understanding the microbial modulation of these host compartments could offer crucial insights into host-microbe interactions and potential mechanisms of microbial colonization and pathogenesis. The Human Protein atlas is the database developed to investigate various aspects of the human proteome, including the secretome, membrane proteome, druggable proteome, cancer proteome, and metabolic functions across 32 different tissues and organs. The resulting comprehensive dataset is integrated into an interactive web-based database, facilitating exploration of individual proteins and navigation of global expression patterns across major tissues and organs in the human body[9]. Utilizing this database, we extracted a total of 19,218 human secretome proteins following the exclusion of unclas-

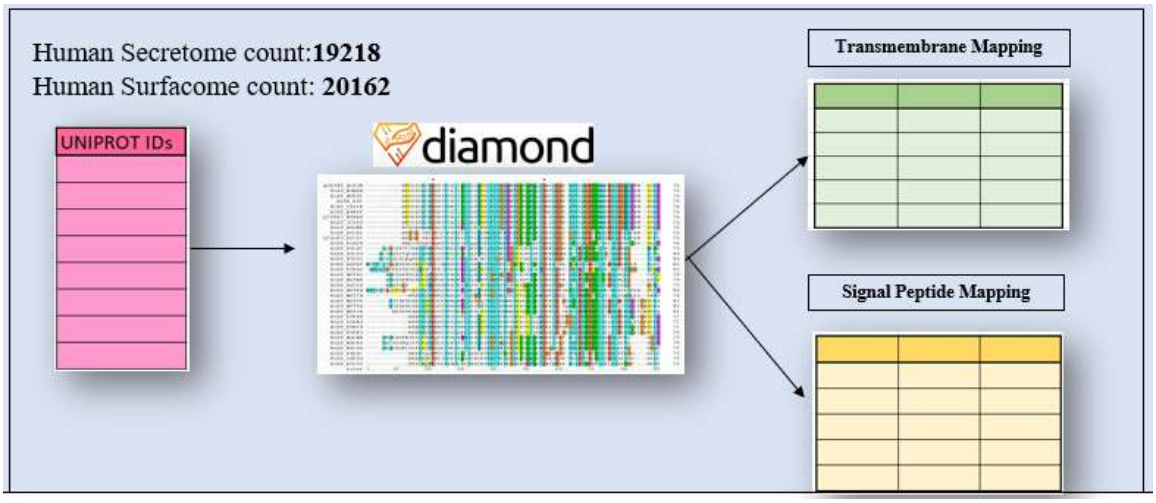

Figure 2.6: Workflow of mapping human surfacome and secretome against the microbiome atlas

sified or invalid identifiers [Figure:2.6]. Similarly, for the human surfacome, we utilized the in-silico human surfacome as our primary resource, from which we retrieved a total of 20,162 surfacome proteins. Among these, 2,216 were classified as non-surface level proteins, 2,886 as surface level proteins, and the remaining 15,091 were categorized as unclassified proteins. This information is vital for understanding the composition and distribution of proteins associated with the human cell surface, facilitating insights into cell-cell communication, signaling pathways, and potential therapeutic targets. Following the acquisition of UniProt IDs from primary resources, we retrieved corresponding FASTA sequences and conducted homologous searches using the Diamond tool. These searches were performed against our surfacome and secretome atlas of the human gut microbiome, employing standard query coverage thresholds of 80% and identity thresholds of 90%. This approach enabled the identification of potential homologous proteins within the gut microbiome atlas, facilitating the exploration of conserved protein functionalities and inter-species interactions at the protein level [Figure:2.6].

Once we identified microbial proteins similar to human ones and aimed to pinpoint which specific microbial proteins were similar to human ones. To achieve this, we gathered all the matches from our dataset and determined the origin of each microbial protein. Then, we created a detailed table showing how frequently each human protein matched with microbes from various species. This allowed us to explore and understand the interactions between human and microbial proteins in greater depth [Figure:2.7].

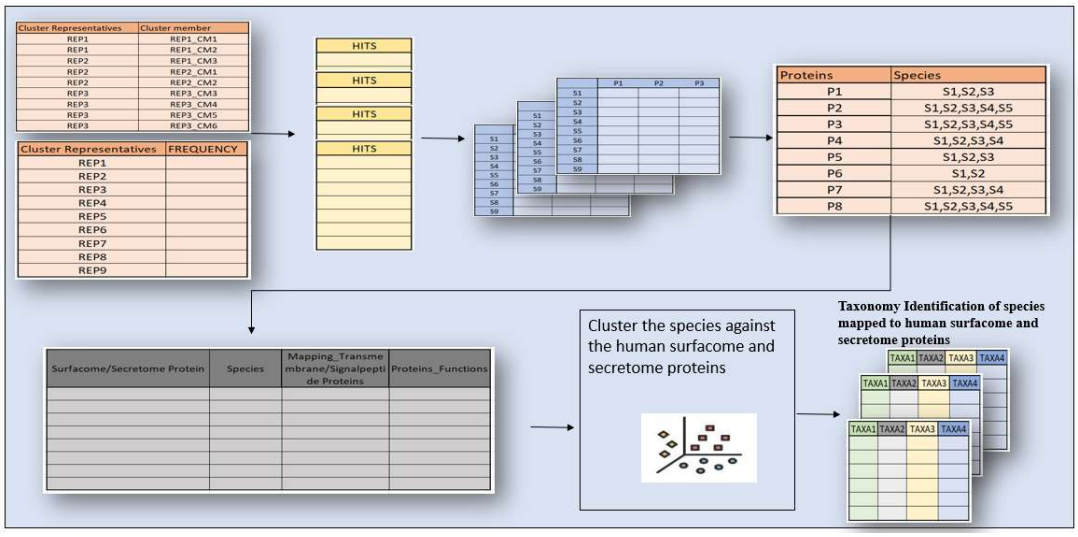

Figure 2.7: Exploring Interactions Between Human Proteins and Microbial Players

Furthermore, comprehensive functional annotations were conducted for both microbial and human proteins, offering valuable insights into their individual roles and potential implications in various biological processes. Subsequently, leveraging this annotation, we performed clustering of microbial species that were most frequently mapped to the human secretome and surfacome. This clustering approach allowed us to identify groups of microbial species with similar protein profiles, aiding in the elucidation of common functional traits or pathways. Additionally, taxonomy identification was conducted to categorize and classify these microbial species based on their evolutionary relationships and taxonomic hierarchy. Such taxonomic information is essential for understanding the diversity and evolutionary context of the microbial communities interacting with the human host. Overall, these analyses provide a deeper understanding of the functional and taxonomic landscape of the human-microbiome interaction, facilitating insights into host-microbe dynamics and potential implications for health and disease.

Chapter 3

RESULTS

3.1 Genotoxin Results:Identification of Proteins and Protein clusters

A thorough investigation of the literature yielded valuable insights into the molecular mechanisms underlying genotoxin-mediated disease progression.Additionally, integrating this information into computational models may facilitate the prediction of potential therapeutic targets or the development of novel interventions aimed at mitigating genotoxin-associated pathologies [Figure: 3.2].

Delving deeper into the homologous results, we uncovered a diverse array of microbial peptides intricately associated with the genotoxins. Scankey plot is used to display the hierarchical relationships between taxonomic categories. The identification of taxonomy in homologous results of genotoxin proteins gave the understanding of their evolutionary relationships, functional conservation, ecological associations, and for comparative studies. Taxonomy helped to trace the origins and diversification of genotoxin proteins across different taxa, predict their functional conservation across related species, explore ecological interactions between organisms harboring these proteins, and inform comparative studies across diverse taxa. Overall, taxonomy enriches our understanding of genotoxicity from evolutionary, functional, ecological, and comparative perspectives. Figure: 3.2 is an example of four proteins,Avr protein(Avr, Adenosine(ADO), Campylobacter adhesion to fibronectin(CaDF) and Azurin(AZO) there are more likely to be causing carcinogens.

The mapping of the protein CadF [Figure:3.2] with species like *Bacillus methanolicus*, *Streptococcus entericus, Carnobacterium gallinarum*, and others reveals intriguing insights into the potential distribution of genotoxic proteins across microbial taxa.

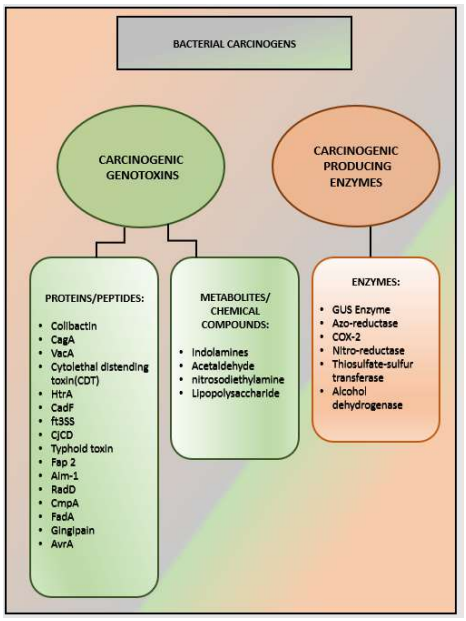

Figure 3.1: Genotoxin proteins identified by literature search

While CadF is traditionally known for its role in bacterial adhesion and colonization, its presence in diverse species raises questions about its broader functions. This observation prompts further investigation into the functional significance of CadF in different bacterial lineages and its potential contributions to genotoxicity. Understanding the distribution of genotoxic proteins across microbial species sheds light on the complex interactions between microbes and their hosts, underscoring the importance of comprehensive analyses in unraveling the molecular mechanisms underlying microbial pathogenesis.

During the investigation of genotoxin peptides, a notable observation emerged regarding the involvement of numerous proteins in cell membrane interactions and signal transduction pathways. This observation underscores the significance of our secondary objective in developing the human gut secretome and surfacome atlas. By compiling a comprehensive atlas, we aim to glean deeper insights into the presence of potential genotoxins within the surfacome and secretome regions.

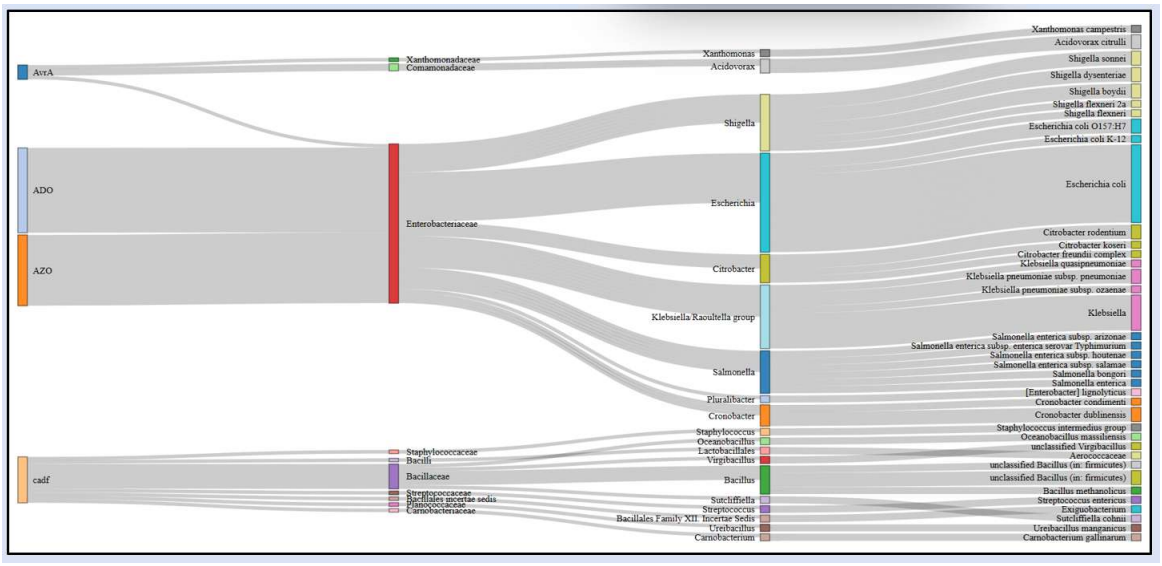

Figure 3.2: Taxonomy classification of peptides that mapped with four genotoxins

3.2 Results of Atlas Creation Pipeline: Comprehensive Overview

We have successfully generated the initial prototype of the human gut secretome and surfacome atlas, focusing on 169 prevalent gut-associated species identified from comprehensive datasets obtained from MGnify and UNINA repositories. In total, our analysis retrieved 484,503 proteins from these species. Among these, 115,692 proteins (23.8%) were predicted to be associated with the surfacome, while 40,344 proteins (8.3%) were identified as secretomes. This comprehensive dataset forms the foundation for further exploration and characterization of the human gut microbiome's secretome and surfacome profiles [Figure:3.3].

Furthermore, [Figure 3.4] provides a comprehensive visualization of the top species-wise distribution of predicted transmembrane and secretome protein counts, juxtaposed against their actual protein counts. This comparative analysis offers valuable insights into the predictive accuracy of our methodology and the distribution patterns of these functional protein categories across different species. Subsequent to the clustering of the 169 core gut-associated species using MMseqs2, a comprehensive analysis was conducted to discern patterns within the data. A total of 42,574 clusters were identified under the transmembrane category, while 150,205 clusters were identified under the

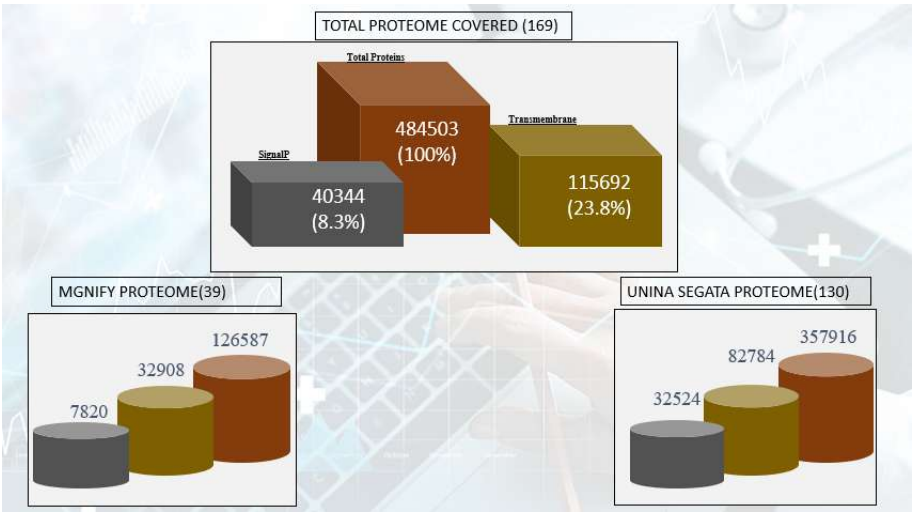

Figure 3.3: Total Number of predicted transmembrane and secretome across core microbial proteins

secretome category. Notably, a substantial portion of the representative sequences were singleton clusters, prompting the application of a threshold to consider only representative clusters with more than five members. This refinement facilitated the formation of the initial cluster hierarchy.

Thus far, significant progress has been made in the preparation of the atlas for the human gut microbiome surfacome and secretome. Subsequently, efforts were directed towards deriving meaningful insights from this atlas. Initially, the focus was on clustering the proteins and structuring the data using R functions. By mapping the cluster members with their respective species of origin, valuable information was obtained regarding the taxonomic composition of each cluster representative. This process enabled the identification of species grouped under specific representatives, elucidating the diversity within the clusters. Furthermore, the creation of a species abundance matrix facilitated a comprehensive assessment of the relative abundance of different microbial species across the samples, providing a quantitative framework for further analyses and interpretations. Overall, these steps represent essential strides toward comprehensively characterizing the human gut microbiome and extracting biologically relevant information from the generated atlas.

Subsequently, Principal Coordinate Analysis (PCoA) was employed to jaccard taxonomic-oriented patterns within the transmembrane profiles. This analysis not only highlighted health-associated core keystone species but also identified potential pathobionts that

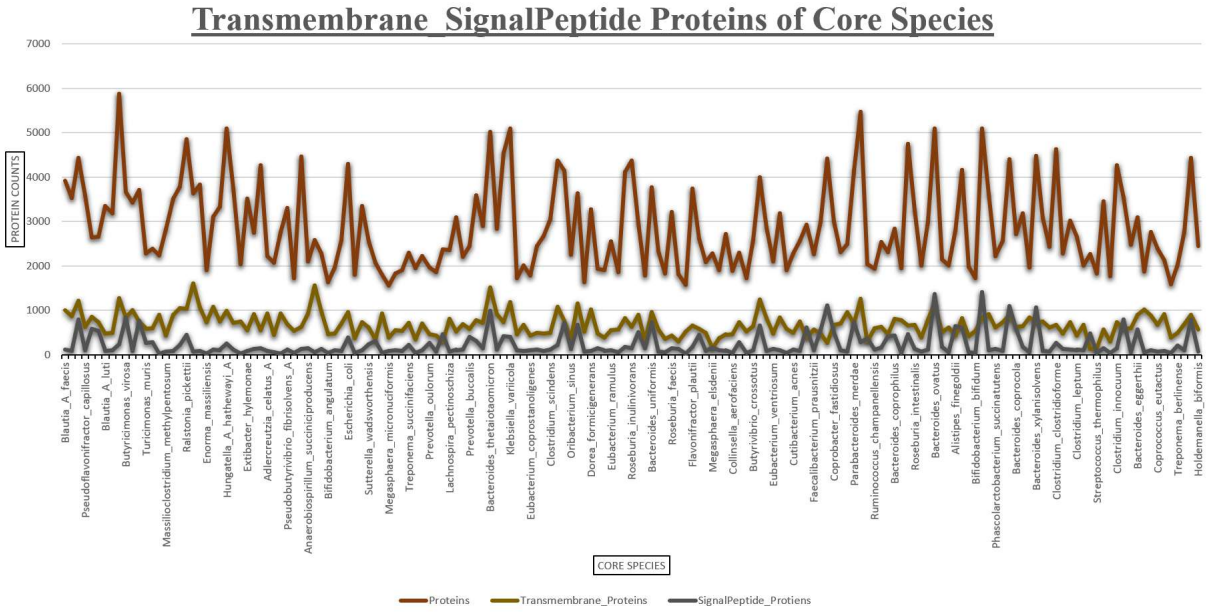

Figure 3.4: Core Specieswise distribution of predicted transmembrane and secretome proteins counts along with total protein counts

may contribute to health implications. These findings offer valuable insights into the intricate dynamics of the gut microbiome and pave the way for further investigations into the functional roles and ecological significance of these microbial communities in human health and disease

The PCoA analysis provides insights into the similarities and dissimilarities between microbial communities, highlighting key microbial taxa crucial for maintaining gut homeostasis and overall host health. By identifying shifts or dysbiosis in the gut microbiome composition associated with health conditions, PCoA aids in understanding the complex interactions between microbial communities and host health [Figure:3.5]. Ultimately, this approach facilitates targeted investigations into the mechanisms underlying gut microbiome-mediated health outcomes.Likewise, the PCoA plot reveals pathobiont-enriched microbial communities and their interactions with the host [Figure:3.6]. Ultimately, this analysis informs the development of targeted interventions to restore microbial balance and mitigate disease risk.

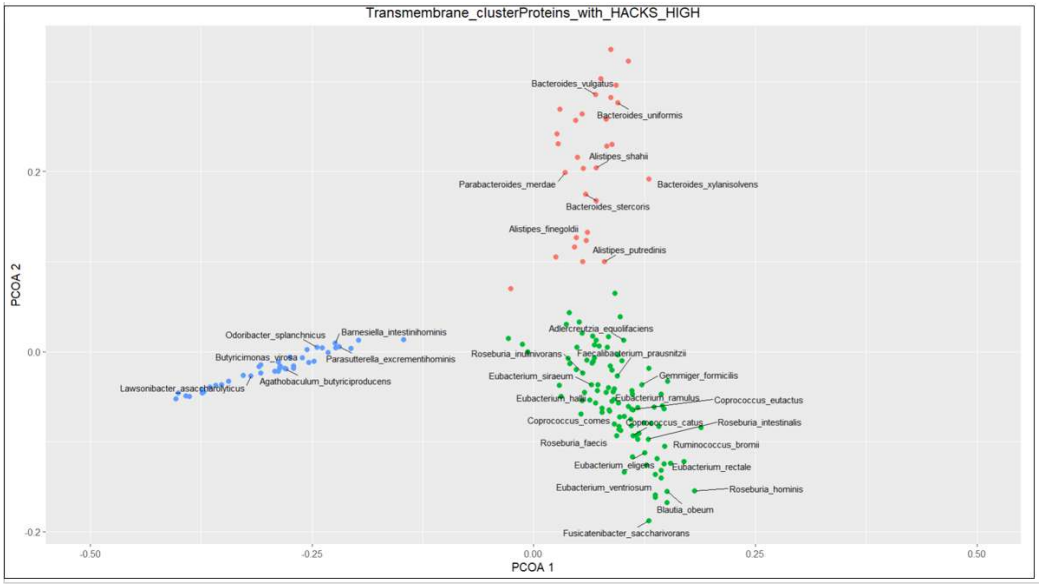

Figure 3.5: Principal Coordinate Analysis(PCoA): Trans-membrane proteins with Health associated core keystone species highlighted

3.3 Validation of the clusters: Intra-cluster similarity and inter-cluster similarity

The clustering of human gut transmembrane and secretome sequences using the MMseqs2 tool was a pivotal step in our analysis pipeline. However, ensuring the credibility of these cluster results was paramount to the integrity of our findings. To achieve this, we devised a meticulous validation process involving both intra-cluster and inter-cluster similarity assessments. Intra-cluster similarity analysis aimed to evaluate the degree of similarity among proteins within each cluster. By comparing a representative protein sequence from each cluster to the remaining proteins within the same cluster, we sought to ascertain the consistency and cohesion of the protein groupings. We established a threshold of more than 50% average similarity within each cluster and across all clusters to ensure that proteins within the same cluster exhibited substantial sequence homology, indicative of functional or structural relatedness.

On the other hand, inter-cluster similarity assessment enabled us to gauge the dissimilarity between proteins from different clusters. By comparing representative clusters against one another, we aimed to delineate distinct protein clusters and identify potential overlaps or discrepancies in their composition. We set a criterion where each

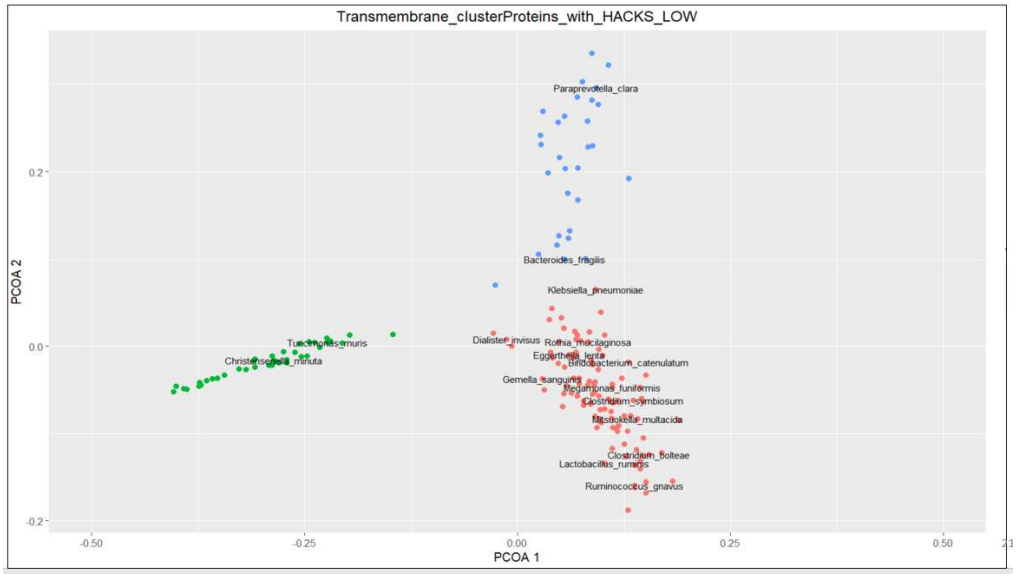

Figure 3.6: Principal Coordinate Analysis(PCoA): Trans-membrane proteins with Pathobionts

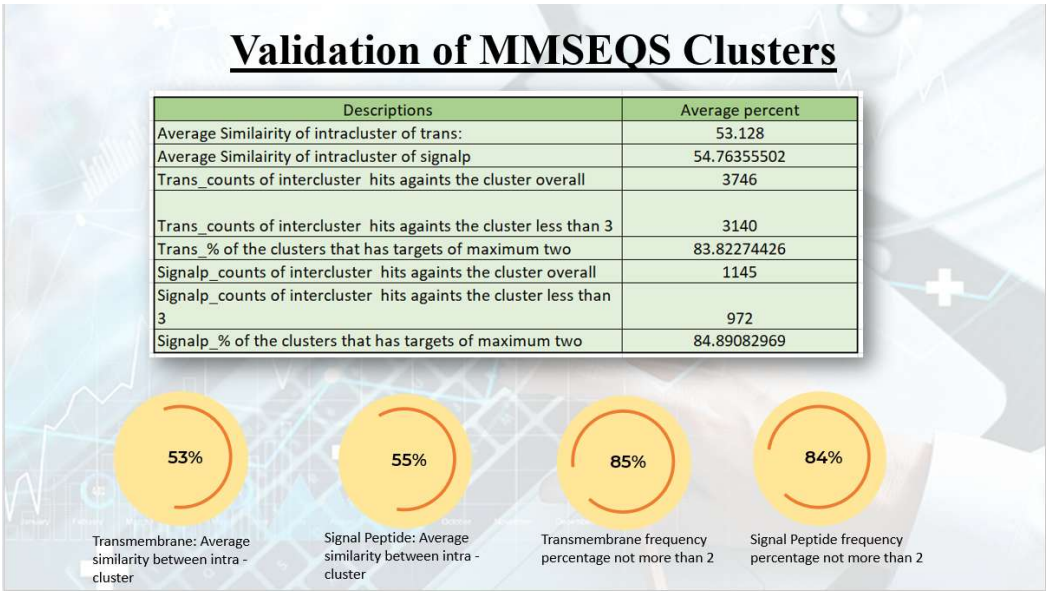

Figure 3.7: Validation Results: Intra-cluster and Inter-cluster similarity

representative should not map to more than 2 other representatives, with an expected percentage of more than 50%. This criterion ensured that representative clusters remained distinct entities, devoid of significant overlap with other clusters.

From [Figure:3.7], we can infer the results of intra-cluster similarity which revealed notable findings regarding both transmembrane and secretome proteins. Specifically, an average similarity of 53% was observed among transmembrane proteins, while secretome proteins exhibited a slightly higher average similarity of 55%. Additionally, the assessment of intra-cluster similarity highlighted promising outcomes, with 85% of transmembrane proteins and 84% of secretome proteins demonstrating representative hits not exceeding two within their respective clusters. These results underscore the robustness and cohesion of the clustered protein groups, indicating reliable clustering outcomes.

From these validation metrics, several important insights can be inferred. Firstly, the observed average similarity values suggest significant sequence homology within protein clusters, indicative of functional or structural relatedness among clustered proteins. Furthermore, the high percentage of representative hits within the specified threshold underscores the distinctiveness and integrity of individual clusters, minimizing potential overlaps between clusters.

These findings instill confidence in the quality and reliability of our clustered protein groups, affirming their suitability for downstream analyses. With validated clusters, our atlas of human gut microbiome surfacome and secretome emerges as a valuable resource for large-scale analyses and investigations. Leveraging this comprehensive atlas, researchers can delve into diverse aspects of gut microbiome biology, ranging from microbial community dynamics to host-microbe interactions, thereby advancing our understanding of gut health and disease.

3.4 Analysis of Results: Integration of Human and Microbial Secretome and Surfacome: Insights from Comparative Mapping

To utilize the large atlas of the secretome and surfacome human gut microbiome.Knowing the importance of matching secretome and surfacome atlas of the human gut microbiome with the corresponding human surfacome and secretome serves a crucial purpose in comprehensively understanding the interactions between microbial and host proteins. This comparative analysis enables the identification of shared proteins between the two datasets, shedding light on potential molecular mechanisms underlying host-microbe interactions in the gut. By elucidating commonalities and differences in protein

composition, we can uncover key players involved in maintaining gut homeostasis or contributing to disease pathogenesis.

Secretome Fasta	Fasta_sequence_count	Queries_matched	unique targets (atlas) mapped	cluster_representative_mapped_greater_5	cluster_representative_matching_with_species_and_annotations_from_representatives
Secretome_alltrans	19218	6243	3179	117	224/6243
Secretome_allsignal	19218	1257	334	9	43/1257
Percantage covered	(our target of trans+signal)(115692+40344)	39.02591321	2.251403522		
Surfacome_alltrans	20193	6297	3227	119	222/6297
Surfacome_allsignal	20193	1160	340	9	35/1160
percentage_covered	(115692+40344)	36.92863864	2.286010921		

Figure 3.8: Statitics of Human surfacome and secretome matched with the microbial proteins as well as the representatives

The [Figure 3.9] illustrates the distribution of human proteins mapped to various identity ranges, specifically between 40-60%, 60-80%, and 80-100%. This detailed breakdown provides valuable insights into the fidelity of protein mappings across different identity thresholds. Understanding the distribution of mappings within these identity ranges is crucial for assessing the reliability and accuracy of the homology-based protein mappings. By discerning how many proteins fall within each identity range, researchers can gauge the level of similarity between the human proteins and their microbial counterparts. This information is instrumental in evaluating the robustness of the protein mappings and in elucidating potential evolutionary relationships between host and microbial proteins.

Secretome_allsignal	Identity percentages in the range 40-60%: 1241
	Identity percentages in the range 60-80%: 16
	Identity percentages in the range 80-100%: 0
Secretome_alltrans	Identity percentages in the range 40-60%: 6048
	Identity percentages in the range 60-80%: 194
	Identity percentages in the range 80-100%: 1
Surfacome_alltrans	Identity percentages in the range 40-60%: 6089
	Identity percentages in the range 60-80%: 207
	Identity percentages in the range 80-100%: 1
Surfacome_allsignal	Identity percentages in the range 40-60%: 1142
	Identity percentages in the range 60-80%: 18
	Identity percentages in the range 80-100%: 0

Figure 3.9: Number of human surfacome and secretome mapped within the percentage identity thresholds

The total protein IDs retrieved comprised 20,162 from the human surfacome and 19,218 from the human secretome. These were subsequently mapped with the atlas of microbial secretome and surfacome proteins, revealing that 39% of human secretome proteins and 37% of surfacome proteins were mapped against microbial transmembrane and secretome counterparts. However, when considering the unique targets of microbial proteins mapped, only 2% were identified from the secretome, and similarly,

Furthermore, mapping microbial proteins to their human counterparts provides insights into the functional overlap or complementarity between microbial and host proteins, offering valuable clues about microbial strategies for colonization, immune modulation, or nutrient acquisition within the host environment. Ultimately, this integrative approach enhances our knowledge of the complex interplay between the gut microbiome and host physiology, paving the way for targeted therapeutic interventions and personalized approaches for managing gut-related disorders Likewise, we commenced the mapping process by procuring UniProt IDs from two primary sources: the In-silico human surfacome and the Human Protein Atlas. These selected proteins were acquired and utilized as the reference dataset for comparative analysis. Subsequently, employing a systematic approach, aligned these IDs with our comprehensive atlas database comprising microbial surfacome and secretome proteins. This meticulous procedure ensured accurate matching and facilitated the exploration of potential associations between host and microbial proteins.

Following the initial mapping process, our focus shifted towards identifying microbial species that exhibited widespread mapping across a considerable array of human surfacome and secretome proteins. This investigative endeavor revealed compelling outcomes, showcasing the prevalence of numerous commensal microbiome species and pathobionts intricately intertwined with the human surfacome and secretome. Such findings not only underscore the complex interplay between microbial inhabitants and host proteins within the gut environment but also offer promising avenues for advancing our comprehension of pathogen-host interaction mechanisms. Indeed, the presence of these mapped microbial species holds immense potential for unveiling novel insights into the dynamics of host-microbe cross-talk, shedding light on pivotal processes governing gut health and disease progression. By deciphering the intricate molecular dialogues occurring at the interface of the gut microbiome and host tissues, we stand poised to unravel the underlying mechanisms driving microbial colonization, immune modulation, and disease pathogenesis. Ultimately, these discoveries are poised to propel forward our understanding of host-microbe interactions, paving the way for innovative therapeutic interventions and personalized approaches in managing gut-related disorders. [2].

Subsequently, we constructed a species abundance matrix that complemented human proteins against the diverse microbial species identified. This analytical approach afforded us a deeper understanding of the intricate interactions between host proteins and microbial inhabitants. By systematically organizing this information, we gained valuable insights into the relative abundance patterns of microbial species associated with specific human proteins. This comprehensive matrix provided a multifaceted perspective, allowing us to discern potential associations and dependencies between host and microbial components within the gut environment. Such insights are instrumen-

tal in elucidating the complex dynamics governing host-microbe interactions and offer valuable clues towards unraveling the underlying mechanisms shaping gut health and disease.

3.4.1 Intersecting Transmembrane and Secretome Microbial Peptides with Human Surfacome and Secretome Proteins

Following the construction of the species abundance matrix, we further enriched our analysis by generating heatmaps specifically tailored for transmembrane and signal peptide proteins. These heatmaps provided visual representations of the abundance patterns of microbial species associated with these distinct protein subsets. By depicting the relative abundance levels across different microbial species and protein categories, these heatmaps offered enhanced insights into the functional dynamics of host-microbe interactions within specific cellular compartments. Such visualizations facilitated the identification of potential trends, clusters, or correlations, thereby deepening our understanding of the complex interplay between microbial communities and host proteins. Leveraging these heatmap analyses, we were able to glean valuable insights into the spatial distribution and functional implications of microbial colonization within the host secretome and surfacome, thereby advancing our knowledge of gut microbiome ecology and its impact on host physiology.

The insights gleaned from Figure 3.10 shed light on the prominent microbial species that exhibit extensive mapping with human surfacome and secretome proteins. This visualization offers a comprehensive overview of the top species involved in interactions with a multitude of human proteins at the surface level. By identifying these key microbial players, we gain valuable insights into the potential functional implications and host-microbe interactions within the complex ecosystem of the gut microbiome. Additionally, this analysis provides a foundation for further exploration into the specific roles and contributions of these microbial species to human health and disease.

Examining the heatmap results of human secretome and surfacome proteins mapped with microbial species unveils numerous protein connections with both beneficial and harmful microorganisms. For instance, proteins such as DnaJ exhibit notable interactions with various microbial species. DnaJ homolog subfamily proteins, known as heat shock proteins, are pivotal in maintaining protein stability and facilitating the degradation of damaged or unwanted proteins. Notably, certain pathobionts like *Clostridium scindes, Coprococus comes*, and *Croprococcus ectactus*, prevalent in Inflammatory Bowel Disease (IBD) patients, are mapped. This suggests a potential for interaction between microbes producing similar proteins, potentially disrupting the immune system. Con-

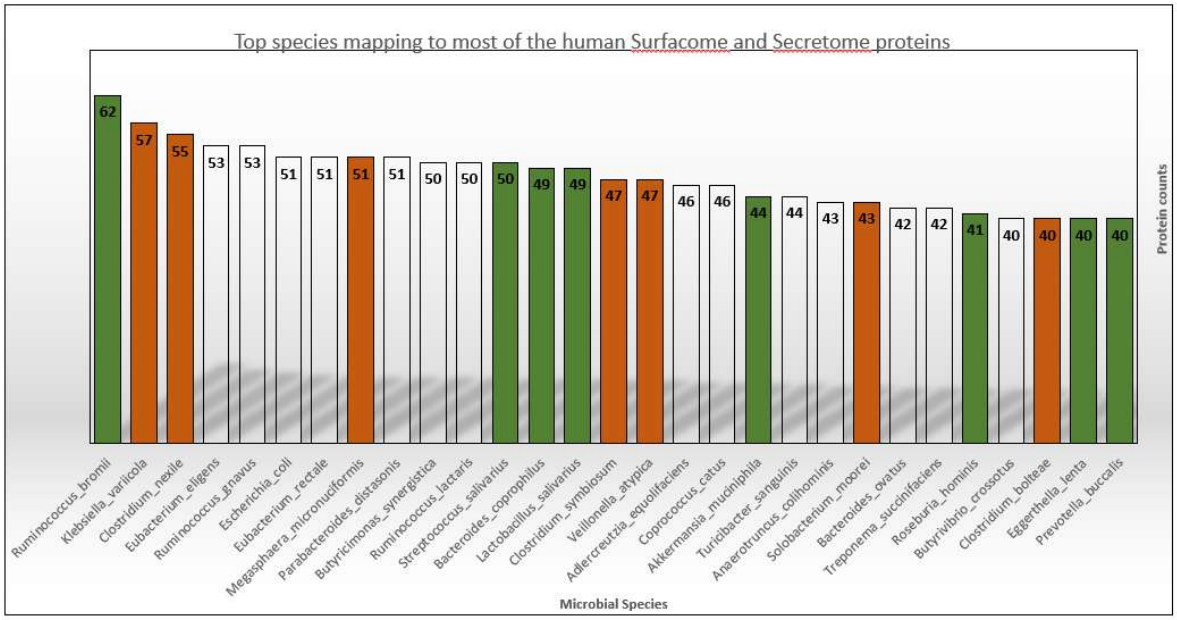

Figure 3.10: Top species that are mapped to human secretome and surfacome proteins

versely, commensal bacteria like *Faecalibacterium prausnitzii* and *Roseburia* are also identified, indicating a possible role in enhancing protein stability. This comprehensive analysis offers insights into the intricate interplay between host proteins and microbial species, shedding light on potential mechanisms underlying health and disease states [Figure:3.11 and 3.13].

In a similar vein, ATP synthase enzymes were observed to interact with commensal bacteria like *Akkermansia muciniphila*, known for their ability to degrade mucin. This interaction may play a role in maintaining mucosal barrier function, potentially offering protection against inflammatory bowel disease (IBD). Conversely, pathobionts such as *Klebsiella pneumoniae* and *Vibrio cholerae* pose significant health risks. These pathogens may target ATP-binding proteins to facilitate infection and evade host defenses, exacerbating disease conditions. This highlights the intricate interplay between microbial species and host proteins, underscoring the importance of understanding these interactions in the context of human health and disease[Figure:3.12 and 3.14].

The intricate interplay between secretome proteins and microbial species offers a fascinating insight into the dynamic relationship within the gut microbiome. Among these interactions, the connection with neurologin stands out as particularly noteworthy, given its pivotal role in facilitating communication between the gut and the brain.

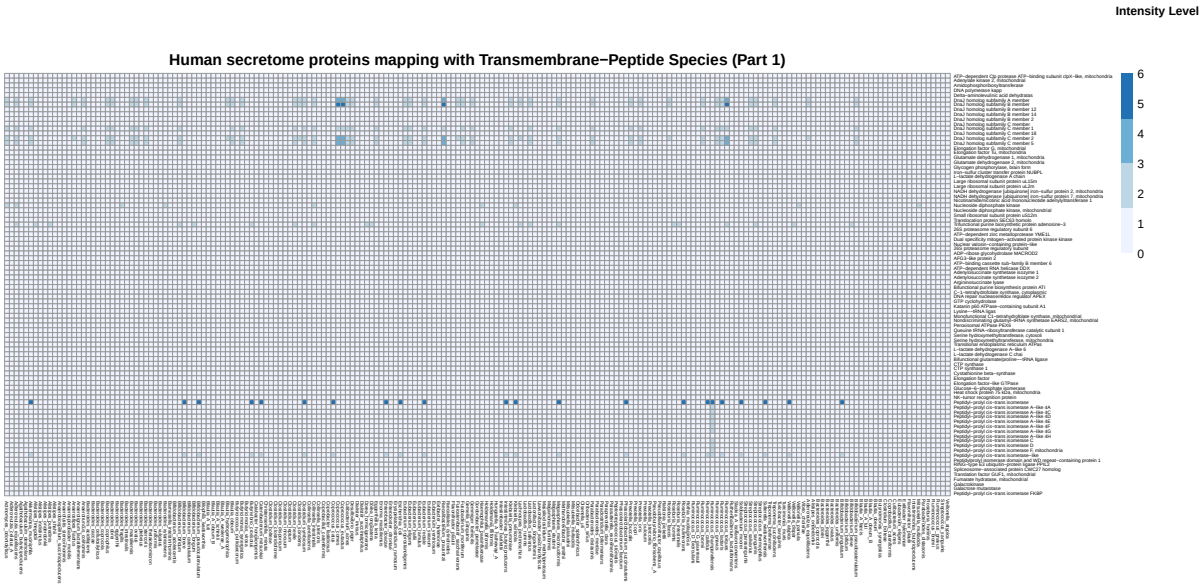

Figure 3.11: Transmembrane-Peptide microbial species mapping with Human secretome/surfacome proteins(Part1)

Neurologin serves as a vital bridge for transmitting essential messages that influence various aspects of health and well-being. What's captivating is the influence of gut microbiota on this communication pathway. By interacting with neurologin, microbes can exert a profound impact on neurotransmitter activity, modulating the delicate balance of signaling molecules that regulate mood, cognition, and stress response. This intricate modulation has significant implications for neurological health, potentially contributing to the onset or exacerbation of mood disorders and other mental health conditions. Notably, certain pathobionts notorious for their pathogenic effects, such as *Sutterella wadsworthensis* and *Megasphaera elsdenii*, may disrupt neurotransmitter function, while beneficial commensal species like *Bacteroides uniformis* and *Bacteroides coprocola* could bolster neurotransmitter activity, promoting overall neurological well-being. This dynamic interplay underscores the remarkable complexity of the gut-brain axis and highlights the profound influence of gut microbiota on neurological function and mental health [Figure:3.15 and 3.16]. In our last endeavor, the incorporation of taxonomy identification into our analysis of species mapping to human proteins represents a significant advancement in our understanding of host-microbe interactions.

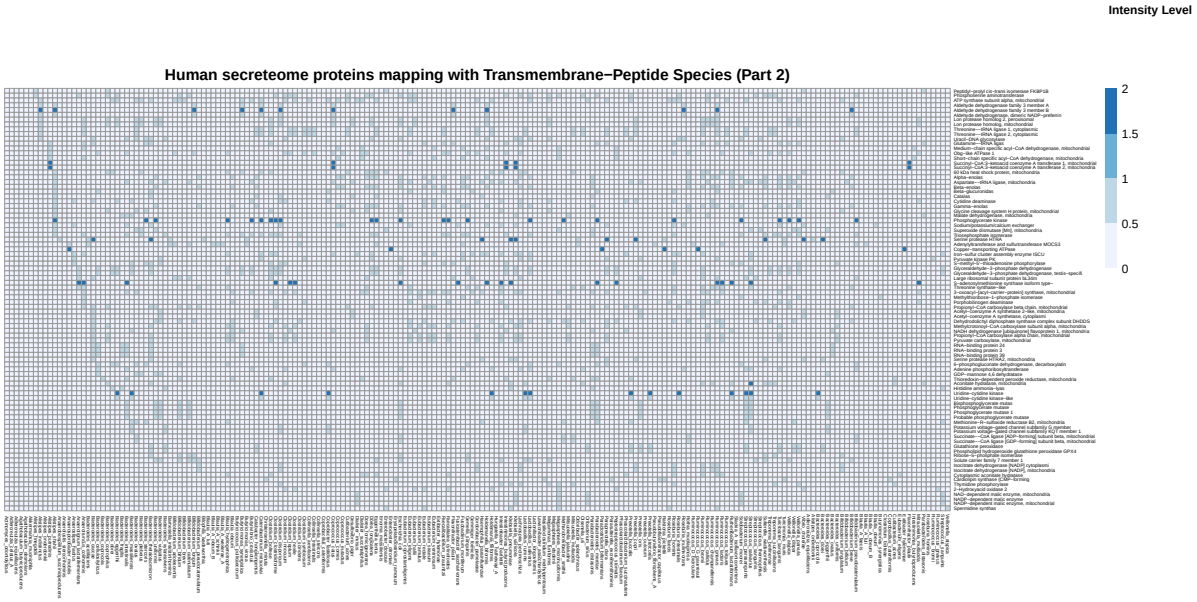

Figure 3.12: Transmembrane-Peptide microbial species mapping with Human secretome/surfacome proteins(Part2)

By unraveling the taxonomic composition of microbial species engaged in these interactions, we have gained invaluable insights into their diverse roles in shaping human physiology and health. This comprehensive approach paves the way for future research aimed at deciphering the complex interplay between microbes and their human hosts, with the ultimate goal of improving human health and well [Figure:3.17 and 3.15].

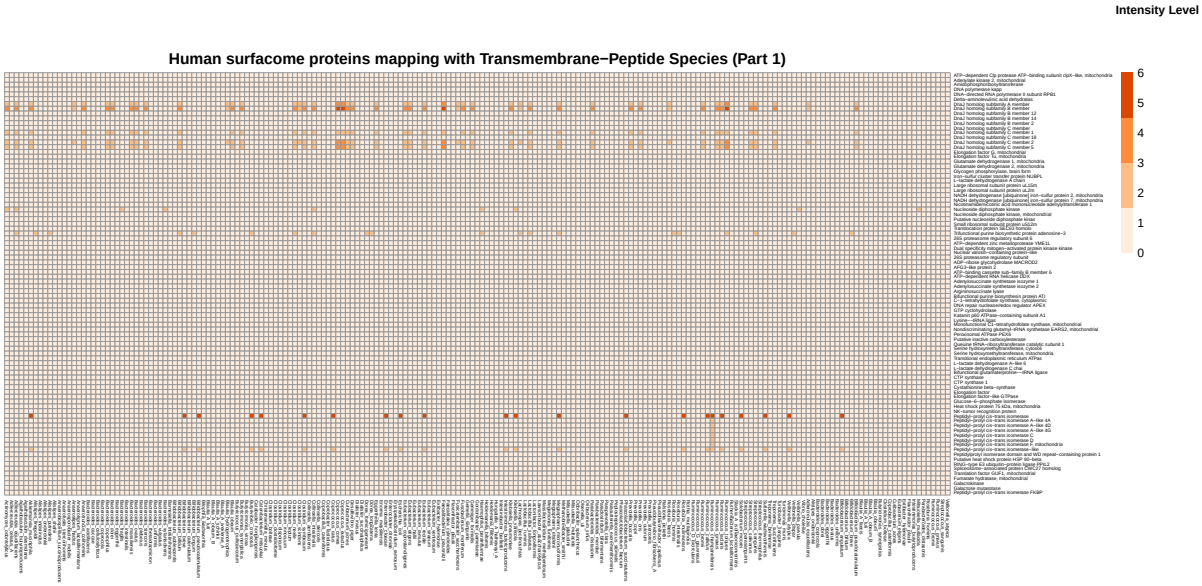

Figure 3.13: Transmembrane-Peptide microbial species mapping with Human surfacome proteins(Part1)

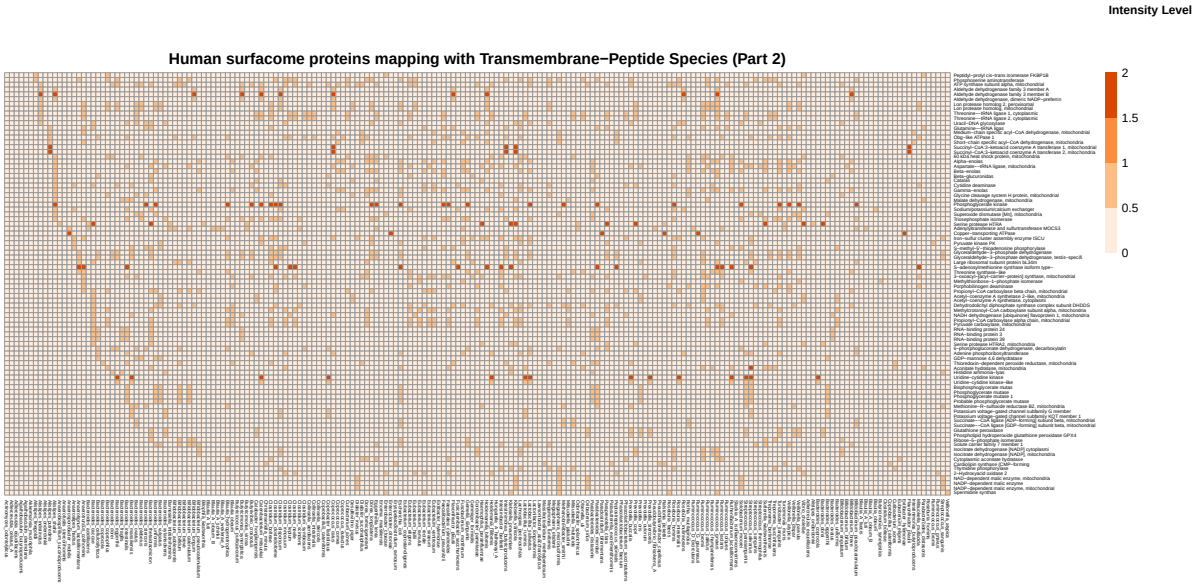

Figure 3.14: Transmembrane-Peptide microbial species mapping with Human surfacome proteins(Part2)

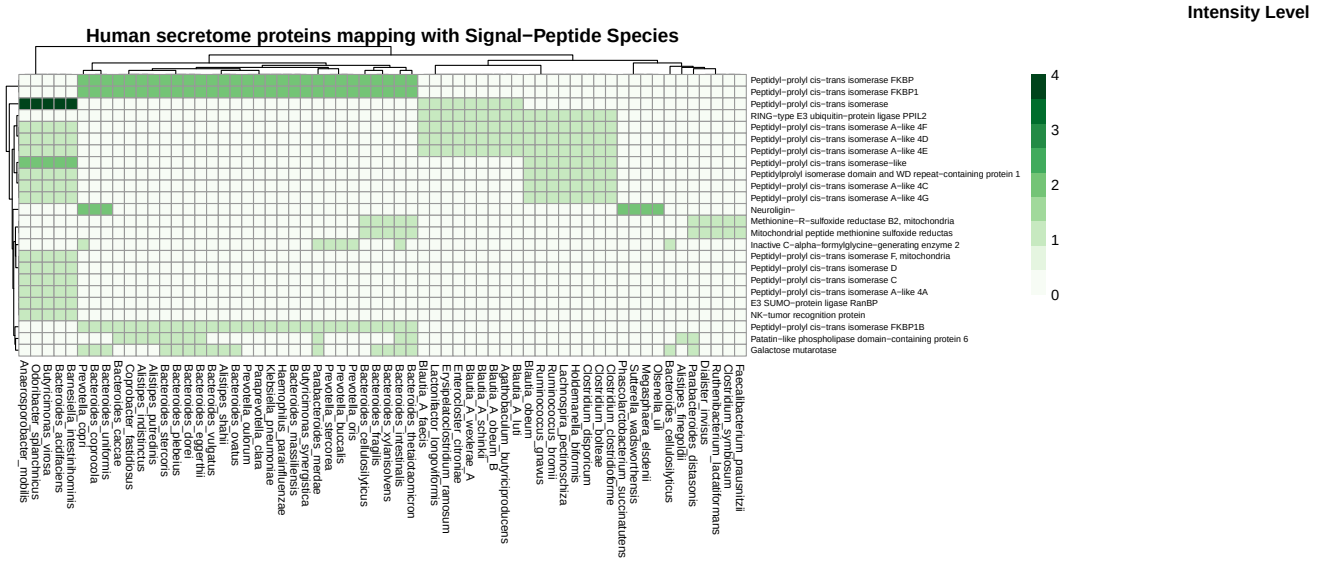

Figure 3.15: Secretome-Peptide microbial species mapping with Human surfacome/secretome proteins

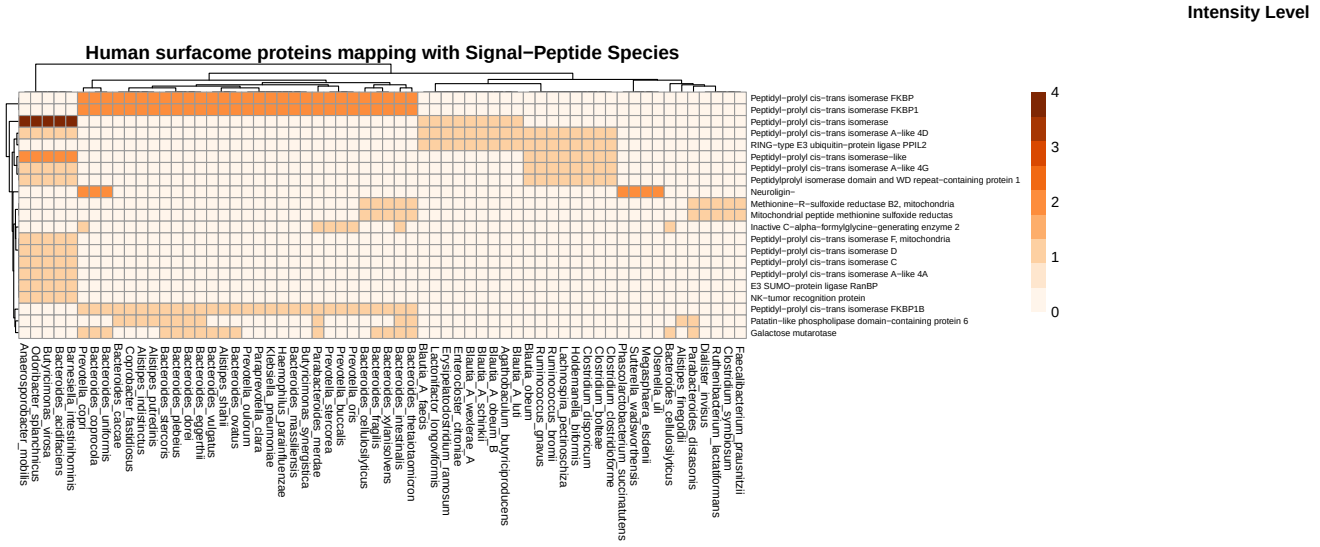

Figure 3.16: Secretome-Peptide microbial species mapping with Human /surfacome proteins

Taxonomy identification of Transmembrane Peptide species that mapped to human secretome and surfacome protein

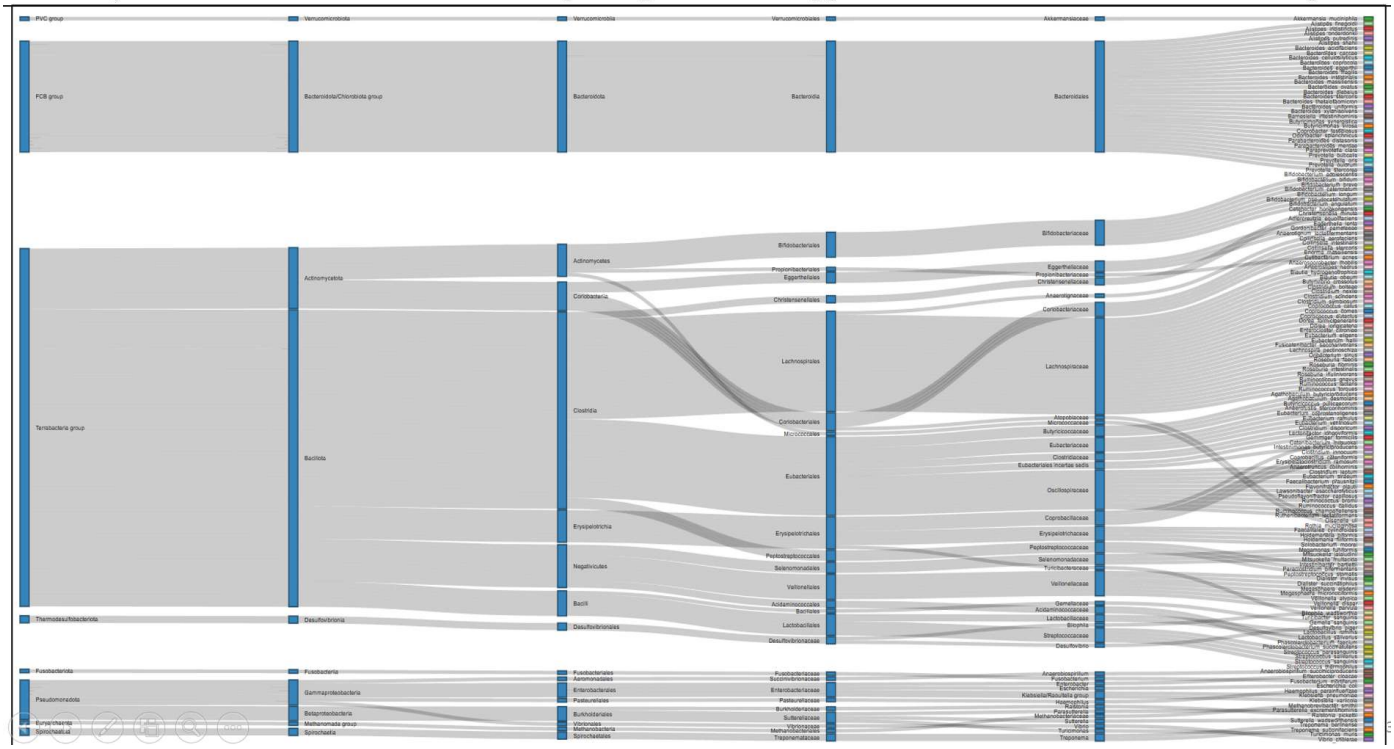

Figure 3.17: Taxonomy Identification of Transmembrane microbial species mapped with Human surfacome and secretome proteins

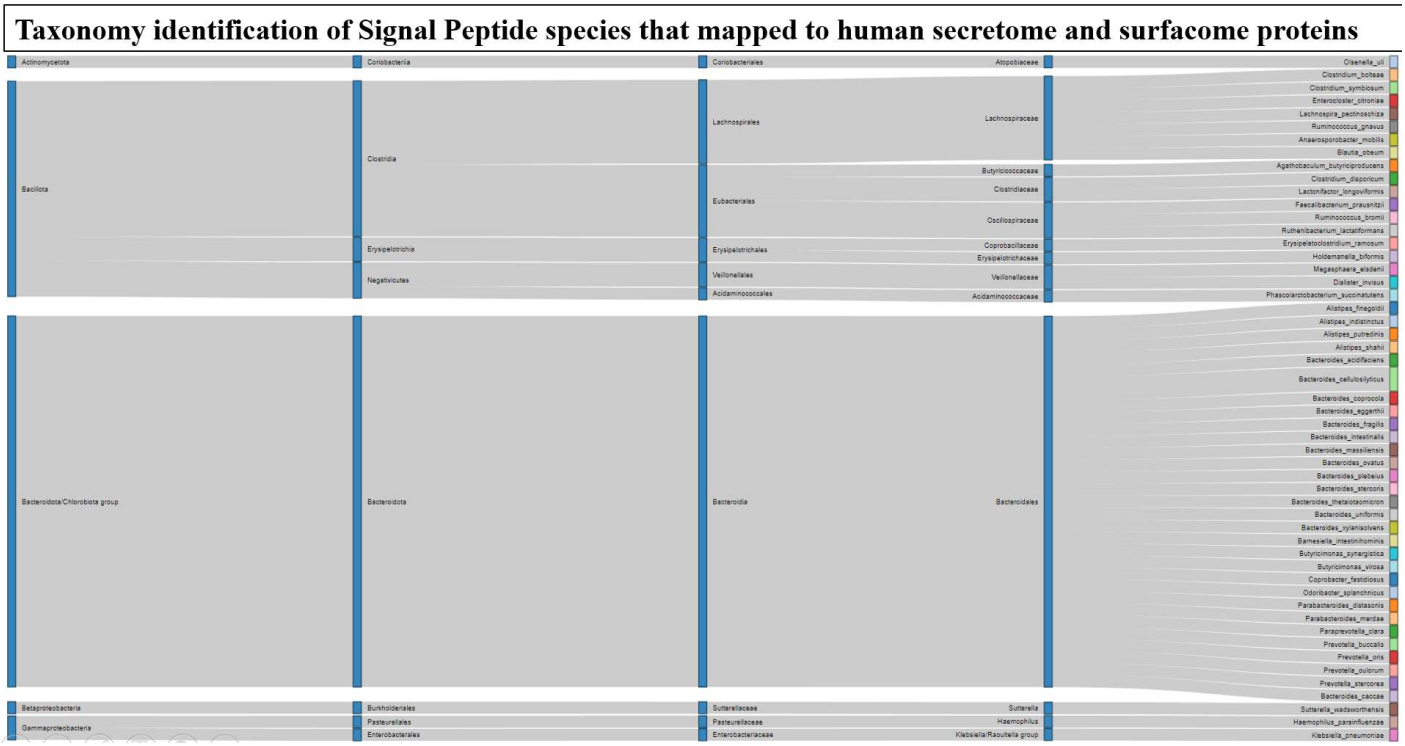

Figure 3.18: Taxonomy Identification of Secretome microbial species mapped with Human surfacome and secretome proteins

Chapter 4

Conclusion and Future Scope

4.1 Conclusion

In conclusion,my thesis study represents a comprehensive investigation into the role of microbial effectors, particularly genotoxins, in the pathogenesis of diseases such as colorectal cancer (CRC). Through a two-part approach, we aimed to elucidate the distribution patterns of genotoxins and identify novel peptides or effectors with genotoxic potential.

In the first part of our study, we focused on discerning genotoxins produced by microbes and mapping their distribution patterns. Microbes, particularly bacteria, are known to significantly contribute to genotoxin production due to their metabolic versatility and environmental interactions. By identifying and mapping microbial genotoxins, we aimed to understand their role in disease pathogenesis and evaluate their similarities across diverse microbial species.

In the second part, we extended our investigation to identify novel peptides or effectors with genotoxic potential. This involved the identification of transmembrane proteins and signal peptides across all microbial species, providing insights into the intricate mechanisms underlying microbial physiology and pathogenesis. Utilizing reference proteomes from 169 species, we developed a prototype of an atlas of the putative surfaceome and secretome of the human gut microbiome. This atlas, organized into clusters based on similarities of various protein properties, served as a valuable resource for our analysis.

Our investigation leveraged this atlas to map homologs of different genotoxin proteins and proteins showing similarity to different human surfacome and secretome proteins, focusing on a core set of 169 species. This analysis unveiled numerous novel proteins and elucidated their potential roles in microbial pathogenesis and host-microbe

interactions. Importantly, we identified significant overlaps between microbial proteins and disease-producing human proteins, providing insights into potential mechanisms of microbial pathogenicity and disease progression.

Overall, our study contributes to a deeper understanding of the intricate interplay between microbial effectors, host physiology, and disease development. By unraveling the role of microbial genotoxins and identifying novel peptides with genotoxic potential, we pave the way for future research aimed at elucidating the mechanisms underlying microbial pathogenesis and developing targeted interventions for disease prevention and treatment.

4.2 Future Perspectives

The findings presented in this thesis study is driven by the ambition to explore the intricacies of microbial biology through the prediction of transmembrane and signal peptides across a broad spectrum of microbial species. This ambitious endeavor is poised to significantly expand our repository of knowledge and catalyze the launch of an extensive database, serving as a cornerstone for researchers worldwide. By providing access to a wealth of data on protein interactions and signaling pathways, this database will empower scientists with invaluable insights, propelling biomedical and biotechnological research to new heights.

Central to our study is the development of protein language-based models, which will serve as powerful tools for delving into the rich biological dictionary of microbial peptides. These models will unlock the mysteries encoded within these peptides, offering a deeper understanding of their significance in various biological processes. Through meticulous mapping of these peptides onto crucial proteins, we aim to illuminate the intricate web of host-pathogen interactions, unraveling the underlying mechanisms of disease.

Moreover, our research endeavors to go beyond mere prediction and exploration. By uncovering the hidden connections between microbial peptides and crucial proteins, we aspire to identify novel therapeutic targets that hold the promise of revolutionizing treatment strategies for a myriad of diseases. This holistic approach not only deepens our understanding of microbial physiology but also opens new avenues for therapeutic intervention, offering hope for improved health outcomes worldwide.

In essence, our thesis study represents a significant step forward in the quest to unravel the mysteries of microbial biology. Through the expansion of our repository, the development of advanced models, and the mapping of peptides to crucial proteins, we aim to unlock a treasure trove of knowledge that will shape the future of biomedical research and pave the way for innovative solutions to global health challenges.

Bibliography

[1] R. M. Pollet, E. H. D'Agostino, W. G. Walton, Y. Xu, M. S. Little, K. A. Biernat, S. J. Pellock, L. M. Patterson, B. C. Creekmore, H. N. Isenberg, R. R. Bahethi, A. P. Bhatt, J. Liu, R. Z. Gharaibeh, and M. R. Redinbo, "An Atlas of -Glucuronidases in the Human Intestinal Microbiome," *Structure*, vol. 25, no. 7, pp. 967–977.e5, Jul. 2017. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0969212617301375

[2] K. Jung, F. Fabiani, E. Hoyer, and J. Lassak, "Bacterial transmembrane signalling systems and their engineering for biosensing," *Open Biology*, vol. 8, no. 4, p. 180023, Apr. 2018. [Online]. Available: https://royalsocietypublishing.org/doi/10.1098/rsob.180023

[3] S. Y. Lee, D. Y. Lee, J. H. Kang, J. H. Kim, J. W. Jeong, H. W. Kim, D. H. Oh, S. H. Yoon, and S. J. Hur, "Relationship between gut microbiota and colorectal cancer: Probiotics as a potential strategy for prevention," *Food Research International*, vol. 156, p. 111327, Jun. 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0963996922003842

[4] L. Richardson, B. Allen, G. Baldi, M. Beracochea, M. Bileschi, T. Burdett, J. Burgin, J. Caballero-Pérez, G. Cochrane, L. Colwell, T. Curtis, A. Escobar-Zepeda, T. Gurbich, V. Kale, A. Korobeynikov, S. Raj, A. Rogers, E. Sakharova, S. Sanchez, D. Wilkinson, and R. Finn, "MGnify: the microbiome sequence data analysis resource in 2023," *Nucleic Acids Research*, vol. 51, no. D1, pp. D753–D759, Jan. 2023. [Online]. Available: https://academic.oup.com/nar/article/51/D1/D753/6880769

[5] E. Pasolli, F. Asnicar, S. Manara, M. Zolfo, N. Karcher, F. Armanini, F. Beghini, P. Manghi, A. Tett, P. Ghensi, M. C. Collado, B. L. Rice, C. DuLong, X. C. Morgan, C. D. Golden, C. Quince, C. Huttenhower, and N. Segata, "Extensive Unexplored Human Microbiome Diversity Revealed by

Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle," *Cell*, vol. 176, no. 3, pp. 649–662.e20, Jan. 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0092867419300017

[6] A. Krogh, B. Larsson, G. v. Heijne, and E. L. L. Sonnhammer, "Predicting transmembrane protein topology with a hidden markov model: application to complete genomes11Edited by F. Cohen," *Journal of Molecular Biology*, vol. 305, no. 3, pp. 567–580, 2001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0022283600943158

[7] F. Teufel, J. J. Almagro Armenteros, A. R. Johansen, M. H. Gíslason, S. I. Pihl, K. D. Tsirigos, O. Winther, S. Brunak, G. Von Heijne, and H. Nielsen, "SignalP 6.0 predicts all five types of signal peptides using protein language models," *Nature Biotechnology*, vol. 40, no. 7, pp. 1023–1025, Jul. 2022. [Online]. Available: https://www.nature.com/articles/s41587-021-01156-3

[8] M. Mirdita, M. Steinegger, and J. Söding, "MMseqs2 desktop and local web server app for fast, interactive sequence searches," *Bioinformatics*, vol. 35, no. 16, pp. 2856–2858, Jan. 2019, _eprint: https://academic.oup.com/bioinformatics/article-pdf/35/16/2856/50719300/bty1057.pdf. [Online]. Available: https://doi.org/10.1093/bioinformatics/bty1057

[9] M. Uhlén, L. Fagerberg, B. M. Hallström, C. Lindskog, P. Oksvold, A. Mardinoglu, Sivertsson, C. Kampf, E. Sjöstedt, A. Asplund, I. Olsson, K. Edlund, E. Lundberg, S. Navani, C. A.-K. Szigyarto, J. Odeberg, D. Djureinovic, J. O. Takanen, S. Hober, T. Alm, P.-H. Edqvist, H. Berling, H. Tegel, J. Mulder, J. Rockberg, P. Nilsson, J. M. Schwenk, M. Hamsten, K. Von Feilitzen, M. Forsberg, L. Persson, F. Johansson, M. Zwahlen, G. Von Heijne, J. Nielsen, and F. Pontén, "Tissue-based map of the human proteome," *Science*, vol. 347, no. 6220, p. 1260419, Jan. 2015. [Online]. Available: https://www.science.org/doi/10.1126/science.1260419