MASTER THESIS

Course code: BIO5011 Name / Candidate no.: **Soumitra Chowdhury** / 1

Whole genome sequencing of a bacterium and a yeast isolated from the intestine of Atlantic salmon

---

Date: 01.09.2020 Total number of pages: 78

NORD University

www.nord.no

Acknowledgements

I express my deep gratitude to my main supervisor Professor Kiron Viswanath for his guidance, valuable suggestions, moral support, and motivation throughout the course of my master study. Without his continuous support I would not have completed the thesis. It has been a pleasure to work with you and learn from you.

My heartfelt thanks to PhD scholar Yousri Abdelmutalab Ahmed Abdelhafiz, co-supervisor, for his support including training on molecular techniques, DNA sequencing, bioinformatics, and providing valuable suggestions. He always made himself available whenever I approached him.

I would like to thank post-doctoral scholar Dr. Juline Marta Walter for teaching and assisting me in microbial culture and DNA extraction techniques as well as for giving constructive comments.

My sincere gratitude goes to former laboratory engineer Ghana Kalerammana Vasanth for providing me help as well as necessary laboratory knowledge even when she was busy.

I am very grateful to Bisa Saraswathy for her kind suggestions and remarks during the preparation of my thesis.

I am thankful to the Faculty of Bioscience and Aquaculture, Nord University for providing the necessary support that facilitated the conduct of this project.

No words are enough to complement my parents, my brother, and my lovely wife for their unconditional love and moral support, without which I would not have been able to achieve this goal.

Soumitra Chowdhury

1st September 2020, Bodø

Abstract

Gut microbes are integral components of vertebrate hosts, and they play important roles in host development, growth, and health. Fish intestine also harbors microbes such as bacteria, yeasts, and unicellular protists. As the aquaculture industry is growing every day to fulfill the protein demands of the growing world population, in-depth understanding of the resident microbes in the host, both at individual and community level, is necessary for improved health management of the farmed animals. Whole Genome Sequencing technology facilitated by the continuous advancements in sequencing techniques is now more feasible than ever and can be utilized in the studies of commensal microbes. In the present study, one bacterium (NU1901-B013) and one yeast (NU1901-Y022) were selected from the microbial biobank of the Faculty of Biosciences and Aquaculture, Nord University, Bodø, Norway to sequence their genomes. These organisms were previously isolated from the intestinal samples of Atlantic salmon (*Salmo salar*). The DNA extracted from pure cultures of the microorganisms were sequenced employing illumina MiSeq platform. The whole genome sequences obtained from the platform were assembled for downstream analyses. The bacterium genome is identified up to the species level as *Kocuria rhizophila* while the yeast genome is identified only up to the genus level as *Rhodotorula.* The bacterium, *K. rhizophila* NU1901-B013 has a genome size of 2.67 million bases (Mb) with a GC content of 71.2%. The genome possesses 2,450 protein coding genes as well as 56 RNA coding genes. On the other hand, the genome size of *Rhodotorula* sp. NU1901-Y022 is 22.77 Mb with a GC content of 57.3%, 7,294 protein coding genes and 186 RNA coding genes. The functional potential of both the genomes were also studied and compared with the related genomes. Both the genomes contain genes related to the functional pathways that are capable of utilizing the nutrients in the gut microenvironment and producing the associated metabolites, some of which can be beneficial to the host fish, salmon. In addition, the genes in *K. rhizophila* NU1901-B013 genome are linked to the production of vitamins (such as B1, B6, B9, H, and K). *Rhodotorula* sp. NU1901-Y022 genome can synthesize carotenoids. The metabolic potential of the two microorganisms revealed through this study could lead to further research on their precise roles in nutrient utilization and metabolite production.

Table of Contents

List of Tables

List of Figures

List of Supplementary Tables

List of Supplementary Figures

1. Introduction

Study of microbes started in the 17th century with the invention of the microscope. Later, scientists started to isolate microbes from infected hosts to understand their involvement in causing diseases. However, as we all know now, microorganisms are present everywhere, on ocean floors, icebergs, volcanos, and even in places that are exposed to poisonous chemicals (Prins et al., 1990). Each of these environments bears a microbial community signature. Microbes in a particular community are interdependent and the term "microbial soup" was introduced by Massaquoi and Guillemin (2018). This "soup" is home to a microscopic community that contains mostly bacteria (~99%), which exist along with other microbes such as yeasts, archaea, and other unicellular organisms known as protists (Robinson et al., 2010).

1.1 Microbes and their niches within host

Throughout the course of evolution, all the multicellular eukaryotes have had a continuous, habitual, and functional association with microbes. The "hologenome theory of evolution", which considers hosts and associated microorganisms as a single unit of selection, was introduced to understand the influence of the latter on the co-evolution of both organisms (Zilber-Rosenberg and Rosenberg, 2008). Microorganisms inhabit the outer surface (skin) of and/or colonize organs in mammals, namely the gastrointestinal tract (GIT/gut), and lungs, thus occupying various ecological niches (Butt & Volkoff, 2019). Each of the ecological niches on (skin) or within (gut, lungs) host's body is known to have their specific conditions (optimum temperature, pH, oxygen, salinity and nutrients) that are conducive to growth and reproduction of specific kinds of microbes. For instance, i) obligatory or facultative anaerobic microorganisms that reside in the gut of a host where there is little or no oxygen, ii) halotolerant microbes on the skin that can tolerate high-level of salinity and iii) acidophiles that reside in the stomach (Ikeda-Ohtsubo et al., 2018). The resident microorganisms obtain their nutrition and dwelling place. On the other hand, the host relies on these microbes for functions that are not encoded by the host genome. Thus, higher organisms, live in a symbiotic or mutualistic relationship with trillions of microbes that belong to thousands of different species.

1.2 Gut microbiota

Microbes that live in a defined environment or biome are collectively called microbiota, for example the skin microbiota or gut microbiota (Marchesi and Ravel, 2015). Autochthonous

(those that colonize the host surfaces) and allochthonous (transient type that enter the host surfaces from diet or other administrative strategies) are considered as components of the host microbiota. There are more autochthonous or indigenous bacteria on mucosal surfaces than the total cells of a host organism (Tlaskalová-Hogenová et al., 2004). Furthermore, gut is the most densely populated mucosal organ among all microbial niches within a vertebrate host. It is presumed that these resident gut microbes are comprised of more than a thousand bacterial species, with at least 160 (~10%) common species in different human individuals (Qin et al., 2010). This fact indicates both the diversity and similarity of the indigenous microbiome associated with individuals. The microbial composition within the hosts is governed by many factors such as diet, environment, and host genetics is crucial among them (Perez et al., 2010). Variation in the bacterial or fungal community within individuals is found mainly at the lower taxonomic levels, and the composition at the higher taxonomic levels is highly conserved (Robinson et al., 2010). Although the microbial compositional differences related to health or disease conditions have been established through different studies, genetics as well as metabolic pathways underlying these conditions are yet to be revealed through in-depth studies (Ghanbari et al., 2015). Hence it is important to understand the interactions between the host and the microbial community it harbors.

Although gut microbiota consists of both prokaryotes and eukaryotes, very few studies have been conducted to understand the gut protists as well as fungal community compared to bacterial studies. To distinguish the gut fungi from the gut bacteria Ghannoum et al. (2010) coined a term “mycobiota”, and small-scale, culture dependent/independent mycobiota studies were conducted during the last decade (Forbes et al., 2019). However, to obtain in-depth understanding of host-microbiota interactions, more studies must be conducted in this relatively new but exciting field of mycobiota.

1.3 Beneficial effects of gut microbes on host

The influences of microbes on host health are historically thought as harmful to the host health because only the adverse impacts of these microbes, especially on human health (for instance to cause inflammatory diseases), had been extensively studied until late 20th century. Such partial understanding created a widespread fear towards the microbial community. However, numerous studies during the last two decades provide evidence that microbiota is an essential part of host and associated with improved health conditions, thereby benefiting the host (Butt

and Volkoff, 2019; Feng et al., 2018). It is now clear that there always exists a barrier between microbiota and host epithelium - while the skin surface has a barrier of multi-layered epithelial tissue, a thick mucus layer is the primary barrier at the mucosal surfaces. The dense and nutrient-rich mucus not only prevents the microbes from entering the epithelial cells but also provides nutrition from the host diet to the commensal ones (Schroeder, 2019; Sommer and Bäckhed, 2013). New sequencing techniques have revealed the influence of microbiota on host physiology. Now it is recognized that microbiota is a beneficial cohabitant that positively impacts the functions of hosts rather than an immunological threat to the host (Sommer and Bäckhed, 2013).

Host-microbiota interactions affect many physiological conditions within the host body, ranging from metabolic activity to physiological homeostasis (Rawls et al., 2004). The gut microbiota is sometimes referred to as "an extra organ" because it acts like another physiological unit of an organism (Butt and Volkoff, 2019; Feng et al., 2018).

1.3.1 Beneficial effects of gut bacteria on host

Various studies that employed mammalian models have indicated that the gut bacteria are able to regulate appetite, digestion, and metabolism in hosts (Bliss and Whiteside, 2018; Duca et al., 2012; Fetissov, 2017; Read and Holmes, 2017). Moreover, some specific bacteria can produce essential vitamins, such as vitamin K and B, whereas others produce a multitude of natural metabolites, the functions of which are yet to be revealed (Durack and Lynch, 2019). The short chain fatty acids (SCFAs) such as butyrate, propionate, and acetate that are produced during the fermentation of dietary carbohydrates, serve as the major energy source of the gut epithelia. These SCFAs also play important roles in maintaining the gut health. Butyrate, for example, promotes barrier function and immune tolerance by regulating T cell development and intestinal macrophage functions (Zhang and Davies, 2016). In addition, the gut microbiota has a significant role in "colonization resistance" that wards off pathogenic microbes; this phenomenon prevents potential pathogens from proliferating and creating new niches in the gut (Sorbara and Pamer, 2019). The dietary substrates, consumed by the hosts, are nutritional sources for both the host and its microbes. The gut bacteria also have effective machinery to metabolize dietary carbohydrates, lipids, and proteins (Jandhyala et al., 2015; Yu et al., 2019). Yet another interesting aspect of host-bacterial interaction is the strong bidirectional communication between the host brain and gut microbiome; the route of this communication

is known as “gut-brain axis” (Cryan and O’Mahony, 2011). This dynamic interaction is facilitated through the ability of the bacteria to produce mammalian neurotransmitters such as serotonin, dopamine, noradrenaline, gamma-aminobutyric acid (GABA), acetylcholine, histamine (Strandwitz, 2018). These neurotransmitters are microbial endogenous chemicals that take part in transmitting neural signals from one neuron to another. Thus, by involving in a multitude of functions at the neural, hormonal, and immunological levels and in digestion, gut microbiota confers different physiological homeostasis in hosts.

1.3.2 Beneficial effects of gut mycobiota on host

In comparison with the large number of articles and review papers on the beneficial effects of gut bacteria, benefits of gut mycobiome are largely unknown. Nevertheless, one of the studies on the effects of gut fungi has indicated that the probiotic yeast *Saccharomyces boulardii* helps in the prevention of antibiotic-associated diarrhea and inflammatory bowel disease (Zanello et al., 2009). Weiler and Schmitt (2003) suggested that probiotic yeasts, those that can produce disease-fighting proteins, may be effective against pathogenic yeasts such as *Candida* spp. Another study showed that monocolonization with either *Candida albicans* or *Saccharomyces cerevisiae* can prevent the development of dextran sulfate sodium (DSS)-induced colitis in commensal bacteria depleted mice (Jiang et al., 2017). Moreover, Budden et al. (2017) explained the communication between gut and lungs (“gut-lung axis”) and showed that immune cells and bacterial metabolites such as SCFAs or fungal cell components like mannan migrates from the gut to lungs and confer protection. These results indicate that mycobiota is not only a key component to maintain intestinal homeostasis, but also important for other host organs such as lungs. So far, fungal involvement in host digestion, growth, and development is largely unknown. However, the gradually emerging mycobiome studies may delineate host-gut mycobiome interactions.

1.4 Microbes and their adverse effects during unfavorable conditions

Some of the microbes can harm the host, either directly by disrupting epithelial barrier or indirectly by producing metabolites including toxins that initiate inflammatory responses that damage tissues or cause a physiological imbalance (Perez et al., 2010). These microbes are termed pathogens. On the other hand, opportunistic pathogens (either a resident or transient) are those that do not usually cause diseases but infect hosts with underlying disease conditions

(Sedghizadeh et al., 2017). Stress is one of the conditions that can lead to alteration in immune responses and create opportunities for the opportunistic pathogens to invade host epithelia and/or cause infection by colonization (Sandrini et al., 2015). Moreover, loss of microbial diversity and abundance, especially of a specific microbe, can lead to immune intolerance; which is suggested as the main reason behind chronic diseases (such as autism), obesity, and metabolic syndromes (Durack and Lynch, 2019).

1.5 Fish associated microbiota

Like other vertebrates, fish also hosts microbiota inside and outside of their body. These microbes are involved in their development and in various physiological conditions. Fish live in an aquatic environment, and hence continuously deal with a more complex environmental microbiota compared to terrestrial organisms, and these organisms may have both positive and negative effects on fish health and growth (Butt and Volkoff, 2019). Hence, understanding microbial community is a necessity to effectively manage aquaculture practices. On an average, fish gut contains 10^7-10^8 microbial cells per gram of mucus (Austin, 2002; Gomez and Balcazar, 2008). As noted earlier, most of them are bacteria and in the intestine of a healthy marine fish, genera of *Aeromonas, Alcaligenes, Acinetobacter, Cytophaga, Alteromonas, Carnobacterium, Flavobacterium, Micrococcus, Moraxella, Pseudomonas* and *Vibrio* are predominant (Egerton et al., 2018). On the other hand, the core and/ or predominant fungal community of marine fish intestine has not yet been clearly defined because the results of the small-scale studies regarding gut-mycobiota are not consistent. Understanding the fish associated microbial community more precisely, especially their genome, transcriptome and their metabolic potential, is a necessity because the knowledge could be employed by the aquaculture industry in disease management, enhancing growth, and producing high quality fish fillets.

1.6 Limitations and current methods in microbiota studies

Until the late 20th century, microbiologists depended on culture techniques and microscopy to study microorganisms associated with different environments. Interestingly, a phenomenon known as “The Great Plate Count Anomaly” was soon perceived after understanding that more microbes can be observed under a microscope compared to those in a culture media. Although it is frequently reported that the percentage of bacteria or yeasts that can be cultured from the

samples of host microbiota ranges from 1% to maximum 50%, some recent studies suggest that it is possible to grow up to 90% of the gut microbes if several culture conditions are applied (Medvecky et al., 2018). Recently, advanced omics (genomics, transcriptomics, proteomics, metabolomics etc.) technology–that are employed to analyze microbiome data revealed through culture independent methods–have facilitated us to identify and study a large percentage of microorganisms and their metabolic capabilities (Lagier et al., 2018). Moreover, while whole genome metagenomics can reveal all the genes and metabolic potential of a microbial community. For better accuracy in identification and understanding the functional capacity of a specific microbe, a suitable approach is to prepare high-quality genomes from pure cultures of the organisms of interest. Culture-dependent methods should also be advanced in the same way as the culture-independent methods to understand individual microbes.

1.7 DNA sequencing technologies

Rapid advancement in DNA sequencing technologies in the last two decades has influenced the field of microbiology significantly. The methods have allowed scientists to understand the complicated microbial communities associated with different environments, including those in fish gut.

Sanger method was originally developed to sequence DNA. However, this sequencing technique is time-consuming and not suitable for large-scale projects such as sequencing a whole genome as it can sequence only one DNA fragment at a time. Later, after the introduction of shotgun sequencing during Human Genome Project (HMP), the entire human genome was sequenced faster than expected (Zhang et al., 2011). This technique facilitated the launch of massive parallel sequencing approach which is used in the Next Generation Sequencing (NGS) technologies since 2005

Initially, high throughput NGS technologies had their inherent challenges. The accuracy of sequences generated using NGS could hardly compete with the “Gold Standard” -Sanger sequencing. In addition, the amplification step of the NGS technique was found to be biased; due to PCR bias from adapters, formation of chimeric sequences and secondary structures related issues (Shokralla et al., 2012). The main disadvantage of NGS technologies, especially Illumina systems, was the phasing-caused optical signal decay, which limits the length of high-quality reads. This demerit limited the application of these technologies in situations where no reference sequence was available to align, assign and annotate the short sequences (Zhou et al.,

2010). However, new sequencing chemistry, ever enriching online databases and bioinformatic advancements are continuously solving many of the limitations such as combining long reads from the third-generation sequencing with short reads from NGS.

1.7.1 illumina sequencing

The illumina Genome Analyzer (illumina Inc.) is currently the most widely used DNA sequencing system. It uses a 'sequencing by synthesis' chemistry, in which all four deoxyribonucleotides (dATP, dTTP, dGTP, dCTP) with DNA polymerase are added simultaneously into flow cell channels. The sequencing steps are illustrated in Figure 1.1. At present, the illumina MiSeq can produce reads of 2*300 bp (paired end reads), and the new chemistry of MiSeq reagent kit (version 3) can generate up to 25 million reads which is equal to 15 giga bp of sequences per run. In addition, illumina sequencing technology has a sequencing accuracy rate of 99.9% which makes it a reliable sequencing technology.

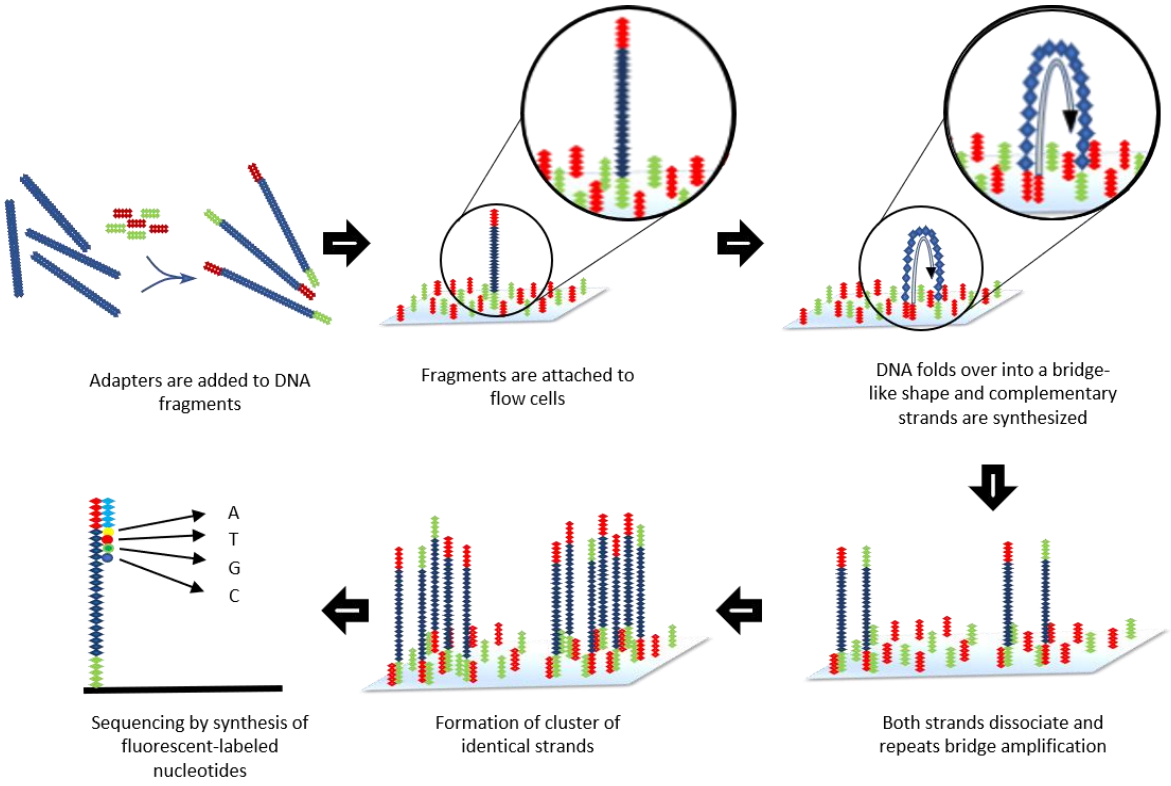


Figure 1 Schematic representation of 'sequencing by synthesis' of fluorescent-labeled nucleotides. *First adapters are attached to the DNA fragments. Next, these fragments attach to the flow cell. Cluster strands of identical fragments created by bridge amplification are primed and all four fluorescently labeled, 3 -OH blocked nucleotides are added with DNA polymerase to the flow cell. The cluster strands are extended by one nucleotide at a time, then during the base calling step, the final signals are converted into DNA sequences with their associated quality score, which can be downloaded as FASTQ files. The diagram was drawn based on the information in illumina sequencing protocol.*

1.7.2 NGS application in Whole Genome Sequencing (WGS)

Whole genome sequencing started in 1976 with the sequencing of bacteriophage MS2 RNA genome (Fiers et al., 1976). The term "genomics" was used for the first time in 1986 and defined by Thomas Roderick as "encompassed the structure and function of genes, and comparative genomics elucidated the hereditary relationships and evolution within and between different species" (Kuska, 1998). However, since the introduction of NGS, the field of "genomics" includes only genome mapping and organization, *de novo* sequence differentiations, and resequencing of genomes as well as exonic or targeted sequences within metagenomes (Kulski, 2016). Whole Genome Sequencing (WGS) using NGS technology played a crucial role to overcome the obstacles associated with Genome Wide Association Studies (GWAS), namely insufficient sample size, limitation of arrays for certain genetic variation, and/ or heterogeneity in phenotype. In fact, the most comprehensive approach in genomics is WGS. Rapid cost reduction in sequencing and the capability of NGS technologies to produce large amount of data make WGS an effective tool for genomic research. WGS is generally employed for human genome sequencing, however, its flexibility makes it valuable for sequencing any species. Moreover, WGS is efficient in unraveling bacterial and viral genomes. It has become progressively easier, faster, and cheaper, and hundreds of genomes that are available in online genome databases can be used for comparative genomics researches (Kulski, 2016).

1.7.3 NGS data analysis using bioinformatics

Bioinformatics is essential to store, analyze and interpret the data generated by NGS technology (Land et al., 2015). There are at least three steps in sequence analysis. Firstly, an integrated software within the sequencing instrument generates sequence reads (A, T, G, and C) from raw signals and associate the sequences with quality scores. The second step is to assemble the raw short reads into contigs, scaffolds, and whole chromosome (if possible) using a software with specific alignment/mapping or assembly algorithm. Finally, various bioinformatics software are used to annotate, integrate and visualize different aspects of assembled genome (Kulski, 2016).

1.8 Importance of WGS in microbiome studies

Rapid cost-reduction in high-throughput sequencing has allowed us to access the genomes of microorganisms from pure cultures or samples. WGS can be employed to understand the metabolic pathways of individual microbial species and predict the roles of microorganisms in maintaining host health status. The available information about the sequenced genome in the National Center for Biotechnology Information (NCBI) database, Ensembl Genome Browser and other databases aid in gene prediction and annotation of new sequenced genomes. WGS allows scientists not only to study protein encoding genes but also to understand the regions within the DNA that influences the regulation of genes. Moreover, identification of microbes using whole genomes, through phylogenetic and similarity-based analyses, is more precise than the existing identification procedures that target 16S (bacteria), or 18S genes or internal transcribed spacer (ITS) regions (eukaryotes) as well as any morphological or chemical feature-based identification (Paul et al., 2019). Unlike targeted sequencing or single nucleotide polymorphism (SNP) arrays which captures only a part of DNA, WGS includes every single code within a whole microbial genome. As the existing genome databases are being updated every day, origin, evolutionary history, structure, and functional potential of a particular microbe can also be described through comparative genomics. Thus, WGS provides us with a comprehensive understanding of the genome, and it is accepted as a crucial method in pathogen detection, epidemiological typing, and antibiotic/ drug susceptibility or resistance (Gautam et al., 2019). Even though resistance/ virulence genes detected via WGS might not be expressed in *in vitro* phenotypic tests due to variable test conditions, the information can be useful in cases where the organism in question develops pathogenicity.

Microbiota-based therapeutics is directed at manipulation of microbial composition to maintain the health of host. WGS can be employed to provide precise information about the functional potential of the specific commensal organisms for use as therapeutics. The whole genome with functional annotation in the genbank can be used by industries or scientists who aim to exploit the metabolic potential of some microbes or find solutions to reduce adverse effects of others. For instance, the increased use of antibiotics has led to the emergence of new antibiotic resistant strains in farmed animals. This is a problem that must be tackled sooner than later. The most rational substitute of antibiotics is probiotics (beneficial gut microbes) (de Bruijn et al., 2018). Whole genomic information of commensal microbes is a pre-requisite to produce immune-friendly commercial probiotics that is aimed at microbiota-based therapeutics. Comprehensive understanding of the functional capabilities–i.e. underlying mechanisms or pathways utilized

by the microbes to benefit the host–could be obtained through WGS. On the other hand, whole genome information about potential pathogens can enlighten our knowledge regarding the molecular differences between a beneficial and a harmful microbe. Thus, whole genome sequencing along with phylogenetic analysis, function prediction and comparative genomic analysis can shed light on many aspects essential for the different industries. Hence in the present master thesis the following 4 objectives were considered to provide comprehensive information about two microbes for the industries and future gut-microbiota studies.

1.9 Objectives of this study

1. To sequence the whole genome of two microorganisms isolated from salmon intestine.

2. To identify the taxonomic ranks of the two microbes.

3. To compare the genomic features and identify potential metabolic features of the two microbes.

2. Methods and Materials

2.1 Microbial samples

Samples for the present study were obtained from the microbial biobank of the Faculty of Biosciences and Aquaculture, Nord University, Bodø, Norway. These microbes were previously isolated from the intestine of Atlantic salmon. They were then identified using Biolog (Biolog Inc., Hayward, CA, USA) system, and preserved at -80°C. From these isolated microbes we chose one bacterium (NU1901-B013) and one yeast species (NU1901-Y022) because they were found in different intestine samples from salmon. These were obtained from experiments that were conducted at our faculty.

2.2 Experimental design

The workflow that was adopted in this study is illustrated in Figure 2. DNA from the selected microbes was extracted and DNA libraries were prepared to sequence the whole genomes of both the microbes. After obtaining the raw short reads from the sequencer, bioinformatics approaches were adopted to assemble the genomes. Next, the assembled genomes and predicted rRNA genes were employed to confirm the species level identification of the microbes. Thereafter, functional annotation was performed to understand various aspects of the genomes and to identify the associated pathways. In addition, comparative genomics was performed by comparing the sequenced genomes of each microbe with related genomes; to reveal the similarities and dissimilarities in the basic features.

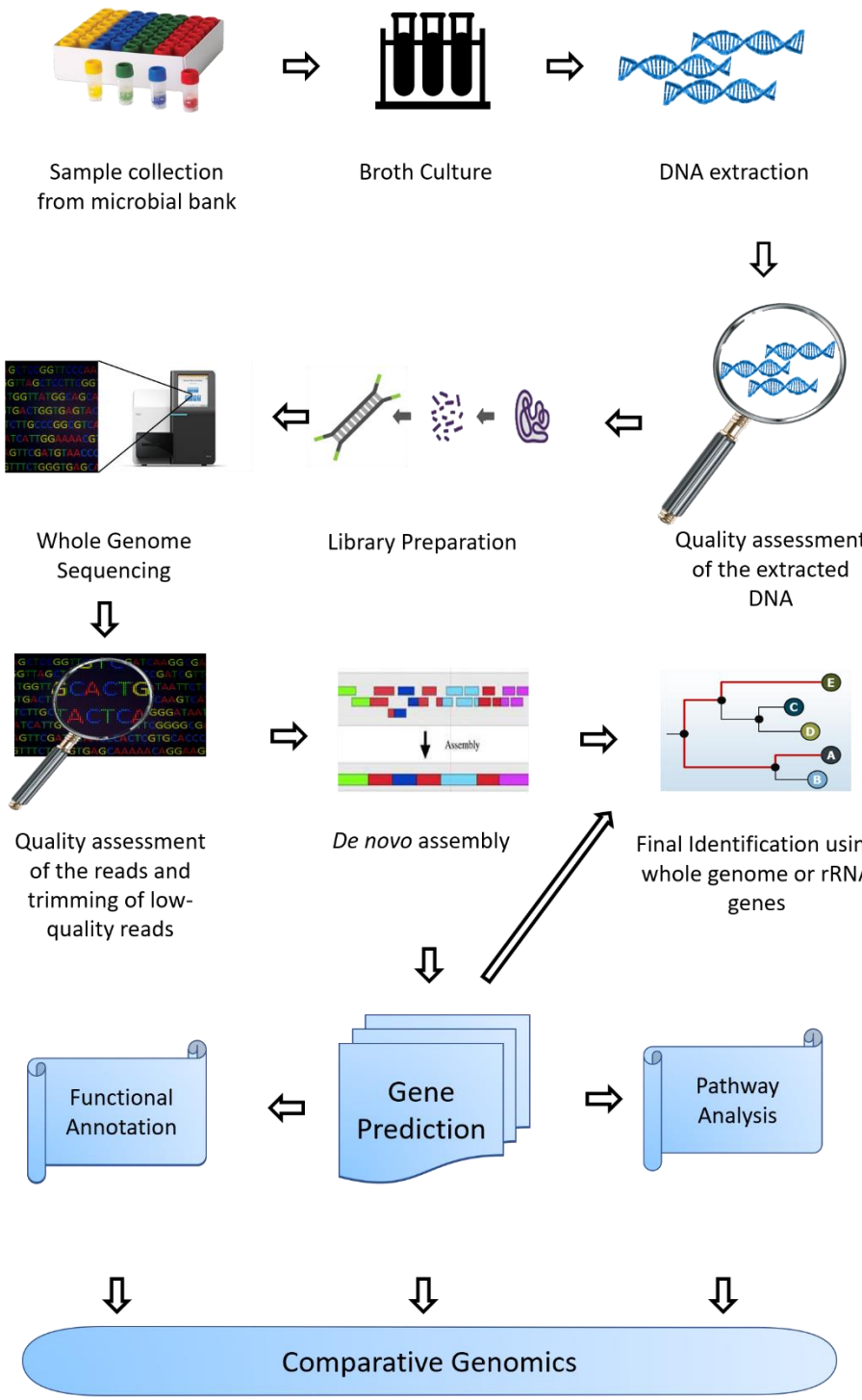


Figure 2 Overview of the study workflow. Samples were prepared for whole genome sequencing. *Raw sequences from the sequencing instrument were then assembled and analyzed using various bioinformatics software.*

2.3 DNA extraction, library preparation and whole genome sequencing

2.3.1 Bacterium and yeast cultivation

The isolates that were preserved on beads in cryotubes (Microbank™, Pro-Lab Diagnostics, Round Rock, Texas, USA) at -80 °C were used for preparing the broth culture. Tryptic Soy Broth or Trypticase Soy Broth (TSB) (Sigma-Aldrich, St. Louis, Missouri, USA) medium was prepared following the manufacturer's guidelines. For this, the broth powder was dissolved in distilled water (30 mg/L) in a conical flask and autoclaved, after which 8 ml of the broth medium was poured into three 15 ml centrifuge tubes. One bead, from each of the cryotubes containing the stock microorganism, was added to the first two tubes and the third tube was kept as a control to check contamination. The tubes were then sealed, transferred to a shaking incubator, and maintained at 23°C for ~48 h, that is until the required cell growth for DNA extraction was achieved.

2.3.2 DNA extraction

Whole DNA from both the microbes was extracted using QIAamp Fast DNA Stool Mini Kit (Qiagen Inc., Valencia, California, USA), following the manufacturer's instructions, but with minor modifications. The microbial cultures were centrifuged at 14,000 x *g*, and the resulting cell pellets were collected after discarding the supernatant. The cells were then resuspended in 2 ml of inhibitEX buffer and incubated for 2 min. Briefly, 1.3 ml of the mixture was transferred to new 2 ml tubes that contained 0.5 mm glass beads (Bertin Technologies, Aix-en-Provence, France) and 1.4 mm ceramic beads (Bertin Technologies). The mixtures were first homogenized for 30 s, in 3 cycles, and then centrifuged at 14000 rpm for 5 min to get rid of the foam that formed during homogenization. Thereafter, all the lysates, obtained from the previous step, were mixed in a vortex and transferred to new low bind 2 ml tubes. Then these lysates were incubated for 15 min at 70° C. The incubated samples were immediately mixed using a vortex for 1 min, then centrifuged at 14,000 rpm for 1 min. Next, 600 µl of supernatant was transferred to new 2 ml tubes containing 25 µl of proteinase K that hydrolyzes the peptide bonds. Thereafter, 600 µl of lysis buffer AL was added to the above solution, which was mixed immediately using vortex and incubated at 70° C for 15 min. After incubation, 600 µl of ethanol (96%) was added to the mixture, which was subjected to a quick spin. Then, 600 µl of the lysate was added to QIAamp spin column, which was placed in new 2 ml collection tubes. These tubes were centrifuged for 1 min at 14000 rpm, and the collection tubes containing the filtrate were discarded. This process was repeated until all the lysates were loaded onto the column.

Thereafter, 3 cleaning steps, employing 2 wash buffers (AW1 once and AW2 twice), were carried out to ensure that only the DNA remains in the column. Next, 60 µl of Tris EDTA (ATE) buffer was added to each column and incubated for 5 min at room temperature. The column was then placed in new tubes and centrifuged for 2 min. Finally, the purified DNA samples was suspended in ATE buffer and then stored at -20° C for further use.

2.3.3 Quality and quantity check of the extracted DNA

The DNA was quantified using Qubit™ dsDNA High Sensitivity (HS) assay kit (Invitrogen, Thermo Fisher Scientific, Eugene, OR, USA). First, the Qubit™ fluorometer (Invitrogen, Thermo Fisher Scientific) was standardized with the two standards. Then, the concentration of each DNA sample was checked.

In addition, to assess the quality of the extracted DNA, NanoDrop™ 1000 spectrophotometer (Thermo Fisher Scientific, Wilmington, DE, U.S.A) was used. The blank measurement was performed using ATE buffer, which was used in the final step of the DNA extraction process to store DNA molecules as a suspension. Next, option "DNA-50" was selected in the NanoDrop software and 1 µl of DNA solution, from each sample, was loaded onto the instrument to measure the nucleic acid absorbance at 260 nm and 280 nm. A 260/280 ratio of ~1.8 is considered as 'contamination free'.

2.3.4 Library preparation for illumina sequencing

After ensuring the quality and quantity of the DNA samples, a DNA library was constructed for sequencing the whole genome of each microbes. Nextera DNA Flex Library preparation kit (Illumina Inc., San Diego, CA, USA) was used for the library preparation, according to the manufacturer's protocol. The DNA libraries were prepared to obtain a library with an average insert size of 500 bp. Prior to library construction, a sample sheet was prepared in the 'Illumina Experiment Manager' software; for future reference of the samples and the index adaptors during the sequencing. Next, a 96-well PCR plate was labeled and ~100 ng DNA solution from each sample was pipetted into specific wells of the plate. In a low binding tube 10 µl of Tagmentation Buffer 1 (TB1) and 10 µl of Bead-linked Transposome (BLT) were added to prepare the tagmentation master mix, which was mixed vigorously using vortex for 10 s. Next, 20 µl of the tagmentation master mix was added to the DNA sample in each well. The resulting solution was mixed thoroughly by pipetting to resuspend the beads. The PCR plate was then

sealed and placed on a pre-programed thermal cycler for incubation; first an incubation at 55°C for 15 min with a pre-heated lid at 100°C (volume set to 50 µl), and then a 10°C hold. As soon as the temperature of the sample reached 10°C, 10 µl Tagmentation Stop Buffer (TSB) was added to each well. The solutions were then pipetted gently to resuspend the beads. Again, the plate was sealed with a Microseal B (Bio-rad, California, USA) and incubated at 37°C for 15 min with a pre-heated lid at 100°C. After incubation, the plate was spun quickly and placed on a magnetic plate for ~3 min until the beads formed a tight pellet. The supernatant was first discarded, and then the plate was removed from the magnetic plate. Next, 100 µl of Tagment Wash Buffer (TWB) was added in 3 cycles to each sample; in each cycle, the beads were resuspended through gentle pipetting. Then the PCR plate was placed back on the magnetic plate for 3 min, and lastly supernatant was removed. At the end of the third cycle, the TWB buffer was retained in the wells to avoid drying. Next, PCR master mix was prepared in a separate low binding tube; by adding 20 µl of Enhanced PCR Mix (EPM) and 20 µl of molecular grade water. After preparing the master mix, the TWB in the wells was removed carefully and the plate was removed from the magnetic plate. Immediately after this step, 40 µl PCR master mix was added to each well and the pellet was resuspended by pipetting. Then, 10 µl of appropriate index pairs was added to each well, and the solution was mixed well by pipetting 10 times carefully. The plate was again sealed and put on the thermal cycler with pre-programmed settings (step 1: 68°C for 3 min; step 2: 98°C for 3 min; step 3: 5 cycles of 98°C for 45 s, 62°C for 30 s and 68°C for 2 min; step 4: 68°C for 1 min; step 5: hold at 10°C; total volume: 50 µl). After the thermal cycler run was completed, the plate was centrifuged at 280 x *g* for 1 min. The amplified library was finally cleaned in different steps; the plate was first placed on the magnetic plate to separate the beads from the supernatant that contained DNA, and then 45 µl of the supernatant was transferred to new wells and labelled accordingly. After adding 85 µl of Sample Purification Beads (SPB) to the new wells, they were mixed by pipetting 10 times. The plate was incubated for 5 min at room temperature, and then it was placed on magnetic plate so that the beads form tight pellets. Next, 125 µl of supernatant was transferred to new wells. After adding 15 µl of SPB into the supernatant, the solution was mixed well and incubated for 5 min. At this stage, the DNA is bound to the magnetic beads and the supernatant was discarded. Finally, the beads were washed 2 times by adding 170 µl of 80% fresh ethanol. The cleaned beads that contained DNA were resuspended by adding 32 µl of Resuspension Buffer (RSB) to transfer the DNA from the beads to RSB. The suspension was incubated for 3 min at room temperature. The PCR plate was then placed on a magnetic

plate so that the beads form a tight pellet. Lastly, 30 µl of supernatant, from each well, was transferred to new low bind tubes and labeled as our final library.

2.3.5 Library quantification and pooling

To visualize the size of the DNA fragments of each library, Agilent 2200 TapeStation system (Agilent Technologies, Waldbronn, Germany) along with Agilent High Sensitivity D1000 ScreenTape and D1000 reagents (Agilent Technologies) were used, based on the instruction of the manufacturer. Briefly, 1 µl of D1000 Ladder or DNA library solution was added to 3 µl of D1000 sample buffer in PCR tubes and was mixed in a vortex for 1 min. The samples were then loaded into the TapeStation, and then D1000 ScreenTape was placed into the machine Subsequently, the size of the DNA fragments, within each DNA library, was obtained from the TapeStation and then visualized.

Each library was then quantified using Roche LightCycler^{TM} 96 Real-Time PCR system (Roche Diagnostics, Basel, Switzerland) along with KAPA Library Quantification Kit (Roche Diagnostics). The concentrations of the bacterium and yeast libraries were 33.25 nM and 21.53 nM. All libraries were diluted and pooled in equimolar concentrations and the concentration of the pool was checked again.

Genome of the selected yeast was assumed to be approximately 7 times larger in length. Therefore, both the yeast and bacterium libraries were pooled together in the ratio of 7:1.

2.3.6 Sequencing

The pooled libraries were sequenced by illumina MiSeq sequencing platform (illumina, San Diego, CA, U.S.A) using MiSeq reagent V3 kit (illumina, San Diego, CA, U.S.A), which is capable of generating 25 million reads (with a paired end read length of 300*2 bp) per run. Following the manufacturer's protocol, the libraries were denatured in NaOH 0.2N for 5 min, and then they were diluted with hybridization buffer (HT1) to the starting concentration, which was 4nM. Phix control (~15%) was added to the diluted pooled-library, which was loaded onto the reagent cartridge. Finally, both the reagent cartridge and the flow cell were loaded into the sequencer to start the sequencing process.

2.4 Bioinformatic analysis

2.4.1 Quality assessment of RAW data

After obtaining the raw data from the sequencer, the quality of the raw reads were checked by FastQC 0.11.9 (Andrews, 2010). In addition, Trimmomatic 0.39 software (Bolger et al., 2014) was used to trim sequences with a quality Phred score below 20 and to remove the sequencing adapters. The quality of the processed data was checked again using FastQC software.

2.4.2 *De novo* assembly

High quality reads from both the genomes were assembled using the procedures described below. Clean sequences were then subjected to a *de novo* assembly approach. A combination of two assemblers that uses de Bruijn graph algorithm were used for the genome assembly. These software are SPAdes 3.14.0 (Bankevich et al., 2012) and Velvet 1.2.09 (Zerbino and Birney, 2008). First SPAdes (k-mer 21, 33, 55, 77, 99 and 127) and velvet (k-mer 149) were used separately and finally SPAdes was used to combine the assemblies employing the parameter "—trusted-contigs". Contigs from the assembly were further used to create scaffolds using SSPACE-standard 3.0 (Boetzer and Pirovano, 2012) which utilises the information from the raw reads. Subsequently, GapFiller 1.10 (Boetzer and Pirovano, 2012) was used to close the gaps in the assembled genomes. The bacterial genome was further refined by the software G-Finisher 1.4 (Guizelini et al., 2016), wherein a reference genome (*Kocuria rhizophila* DC2201) was used for the identification of the assembly errors by checking the GC Skew bias, and then the assembly was refined according to the reference. After finishing the genome assembly, completeness of the assemblies was assessed using BUSCO (Seppey et al., 2019). This software searches for a set of core genes within an assembled genome and calculates the completeness of an assembly.

The circular genome map of bacterium was constructed using the server (https://www.patricbrc.org/) maintained by the Pathosystems Resource Integration Center (PATRIC). Genomic islands (the mobile genetic elements associated with Horizontal Gene Transfer, HGT) within the bacterium genome were predicted using Island Viewer 4 (Bertelli et al., 2017).

2.4.3 Gene prediction and functional annotation

Different pipelines were used for gene prediction and functional annotation of each microbe. The assembled bacterial genome was uploaded into Rapid Annotation using Subsystem Technology (RAST) web server (https://rast.nmpdr.org/). Gene prediction was performed using the server pipeline. The pipeline also includes the mapping of the genes to their subsystems that predicts the metabolic potential (Aziz et al., 2008; Overbeek et al., 2014).

The yeast genome, on the other hand, was structurally and functionally annotated by GenSAS (https://www.gensas.org/) (Humann et al., 2019), a step-by-step annotation pipeline. It started with the prediction and masking of repeat sequences by RepeatMasker (Smit et al., 2019) for further analysis. GeneMark-ES (Ter-Hovhannisyan et al., 2008) was used for ab initio gene prediction. RNAmmer 1.2 (Lagesen et al., 2007) within the GenSAS pipeline was used for rRNA prediction and tRNAscan-SE 2.0 (Lowe and Eddy, 1997), also in the GenSAS pipeline, was used for tRNA prediction. Finally, the predicted genes were assigned a particular gene name using a combination of DIAMOND software (Buchfink et al., 2015) and BLASTp (protein alignment).

All the genes within the annotated genomes (both bacterium and yeast) were then clustered to functional categories through orthology assignment by eggNOG-mapper 2.0 (Huerta-Cepas et al., 2017). The number of genes that are related to specific ontologies/ pathways were obtained from the output data sheet generated by the software. Moreover, the amino acid sequence file, obtained from the annotation of the yeast genome, was submitted to BlastKOALA (Kanehisa et al., 2016) tool, within the KEGG website http://www.kegg.jp/blastkoala. The software then reconstructed the KEGG pathways based on the individual gene functions assigned through KEGG Orthology (KO).

2.4.4 Comparative genomics and phylogenetic analysis of the isolates

Various aspects of the sequenced genomes, such as genome similarity, GC percentage, predicted genes and key functional features of the genes obtained from the previous steps were compared to the downloaded reference genomes (Supplementary Table 7 and 8) from the genome database of the National Center for Biotechnology Information (NCBI) website (https://www.ncbi.nlm.nih.gov/). Orthologous Average nucleotide identity Tool version 0.93.1 (OAT) (Lee et al., 2016) was employed to calculate both original ANI (Average Nucleotide Identity) and orthologous ANI (OrthoANI). Genome to Genome Distance Calculator, GGDC

2.1 (https://ggdc.dsmz.de) (Auch et al., 2010) was used for species delimitation by considering *in silico* DNA-DNA hybridization (DDH) between the query and reference genomes (Meier-Kolthoff et al., 2013). This software was also used to calculate GC percentage differences and genome to genome distances.

Phylogenetic tree based on target regions: To find the evolutionary relationships between the sequenced isolates and their close relatives− by using specific target regions (16S for the bacterium and both 18S and Internal Transcribed Spacer for yeast)− I followed four steps: i) selected specific target regions within the sequenced genomes ii) blasted the target sequences against NCBI database to find similar organisms iii) downloaded the related sequences from the database and iv) constructed phylogenetic tree. The 16S and 18S rRNA genes in the genome were predicted by RNAmmer 1.2., and to obtain the sequence of the ITS region from the sequenced yeast whole genome, one of the *Rhodotorula* ITS sequences (*Rhodotorula mucilaginosa* CBS 316) from the database was downloaded and blasted against the sequenced yeast whole genome. The aligned portion of the sequence, within the sequenced yeast genome, was downloaded and used further in phylogeny construction. These sequences (16S, 18S and ITS) were then blasted against the rRNA/ITS database of NCBI (https://www.ncbi.nlm.nih.gov/). Here, only the type strains are considered in the BLAST search. Next, sequences of the closely related and one/ two distantly related type strains (outgroups) (supplementary table 4, 5, and 6) were downloaded from the BLAST result. Average nucleotide identity (ANI) results indicated that some of the downloaded target sequences of both *K. rhizophila* and *Rhodotorula* sp. non-type strains were more closely related to the sequenced genomes than the type strain, that was found to be most closely related in the BLAST search. However, only some of these non-type sequences were retrieved from the nucleotide database of NCBI, and some of them were predicted using RNAmmer 1.2 from the whole genomes (supplementary table 6 and 7). Next, each set (16S, 18S and ITS) of sequences were aligned using MUSCLE algorithm employing MEGA X (Molecular Evolutionary Genetics Analysis) software (Kumar et al., 2018). The phylogenetic tree based on 16S (bacterium) gene was constructed using Neighbor-joining method and the recommended model (Maximum Composite Likelihood). The same software was used to infer 18S and ITS based trees employing Maximum Likelihood method and the model predicted as best fit to each data set (Tamura-3 and General Time Reversal-parameter, respectively for 18S and ITS).

Phylogenetic tree based on whole genome: Two methods were employed to construct phylogenetic trees of the bacterium. The Type (Strain) Genome Server (TYGS)

(https://tygs.dsmz.de) was employed only for the genome of the bacterium. The Reference sequence Alignment based Phylogeny builder (REALPHY; https://realphy.unibas.ch) was used for the genomes of both bacterium and yeast.

The TYGS undertakes the whole genome based taxonomic analysis (Meier-Kolthoff and Göker, 2019). The whole genome sequence of the bacterium was first uploaded into the server to compare–via the MASH algorithm (a fast approximation of intergenomic relatedness)–against all type strain genomes available in the TYGS database (Ondov et al., 2016), and the ten type strains with the smallest MASH distances were chosen for phylogenetic analysis. The Genome BLAST Distance Phylogeny approach (GBDP) was employed to calculate the precise distances within the genomes (Meier-Kolthoff et al., 2013) and a balanced minimum evolutionary tree was inferred based on the resulting intergenomic distances.

The REALPHY needs both a query genome along with their corresponding reference genomes to generate the phylogenetic tree. The server then performed multiple sequence alignments using bowtie2. Next, employing phyML tree construction method the server generated the phylogenetic trees (Bertels et al., 2014).

3. Results

In this section, I will provide the information regarding the assembled bacterium and yeast genomes and the description about genome completeness side by side. Thereafter, the key features, phylogenetic analysis, and the predicted functions of the assembled bacterial genome, and then the corresponding details about yeast genome will be described.

3.1 Quantity and quality of the extracted DNA

The DNA concentrations for bacterium and yeast were 39.3 µl/ml and 8.97 µl/ml, respectively. while the absorbance ratios at 260 nm and 280 nm were 1.89 for bacterium and 2.01 in the case of yeast. In addition, the corresponding absorbance ratios at 260 nm and 230 nm were 1.94 and 2.44, respectively (Supplementary table 1). The DNA fragments were in the range 200 - 1300 bp, and the samples intensity peaks were at 457 bp and 456 bp for the bacterium and yeast, respectively (Supplementary Figure 1).

3.2 Characteristics of the preprocessed and processed whole genome sequences

A total of 614,258 paired end reads, with a length of 301 bp and an approximate coverage of 69x, were generated for the bacterium. On the other hand, a total of 4,469,040 paired end reads were generated in the case of yeast, with an approximate coverage of 59x. After quality trimming all the reads have a Phred quality score of 20 or higher (Supplementary Figure 2 and 3), where, 50% and 49% of the base pairs were maintained for further analysis of the yeast and bacteria, respectively (Supplementary table 3).

3.3 Genome assembly reports

3.3.1 Genome assembly of NU1901-B013

SPAdes generated 48 contigs with an N50 of 246,953 bp, and the obtained largest contig was of size 623,328 bp (Table 1). Velvet-SPAdes combination−generated 46 contigs with an N50 value of 265,960 bp, and the maximum length was still 623,328 bp. After filling the gaps within the scaffolds, i.e. 100 Ns in 1 gap with known bases, the scaffolds were sorted using G-Finisher software, based on a reference genome. The sorting step generated 13 scaffolds with an N50 value of 335,683 bp and the maximum scaffold length was 652,348 bp (Table 1). The assembled genome was employed for the downstream analyses.

Table 1 Characteristics of theNU1901-B013 genome assembled using different software

Software	Number of consensus sequences	Maximum size (bp)	N50 (bp)	Number of Ns
Contigs				
SPAdes	48	623328	246953	100
Velvet	1716	7125	1251	N/A
Combined	46	623328	246808	100
Scaffolds				
SSPACE-Standard	46	623328	246953	100
GapFiller	46	623328	246953	0
G-Finisher	13	652348	335683	0

3.3.2 Genome assembly of NU1901-Y022

SPAdes and Velvet generated 1,017 and 17,442 contigs, respectively (Table 2). While SPAdes assembly had an N50 value of 62,030 bp and the largest contig size was 470,784 bp, Velvet assembly provided an N50 value of 1085 bp and the largest contig size was 9,491 bp. SPAdes-Velvet combination generated 963 contigs. The scaffolding process generated 958 contigs. The N50 value of 72,323 bp and the largest contig length of 500,506 bp were identical for the assemblies from both SPAdes-Velvet combination and scaffolding. After filling 3 out of 12 gaps employing GapFiller, i.e. by substituting 672 Ns (out of 848) with known bases 958 contigs were obtained for further analysis (Table 2**).**

3.3.3 Genome completeness

The assembled bacterium genome had a 97% BUSCO score. Here, 39 complete and 1 partial core genes were detected in the 40 query core genes. On the other hand, 270 complete and 10 partial core genes were found in the yeast genome when queried against 290 core genes, and the BUSCO score of the assembled yeast genome was 93.1% (Table 3). After performing this quality check using BUSCO, the assembled genomes were used for the next step.

Table 2 Characteristics of the NU1901-Y022 genome assembled using different software

Software	Number of contigs / scaffolds	Maximum size (bp)	N50 (bp)	Number of Ns
Contigs				
SPAdes	1017	470784	62030	977
Velvet	17442	9491	1085	N/A
Combined	963	500506	72323	848
Scaffolds				
SSPACE-Standard	958	500506	72323	848
GapFiller	958	500506	72323	176

Table 3 Completeness scores of the genome assemblies

	Bacterium	**Yeast**
Total number of core genes queried	40	290
Core genes (complete) detected	39 (97.50%)	270 (93.10%)
Core genes (complete+partial) detected	40 (100%)	280 (96.55%)
Number of missing core genes	0	10 (3.45%)
Average number of orthologs per core genes	1.03	1.00
% of detected core genes that have more than 1 ortholog	2.56	0.00
BUSCO score	97%	93.1%

3.4 Genomic features, identification, and functional analysis of the NU1901-B013 genome.

3.4.1 Structural features of the assembled genome

The assembled bacterium (NU1901-B013) genome had a total of 2,676,931 bp with an average GC content of 71.16% (Table 4). A total of 2,506 coding genes and 12 pseudogenes were predicted within the genome. In addition, 56 RNA genes are present in the genome, including 10 ribosomal RNAs (rRNAs) and 46 transfer RNAs (tRNAs). No noncoding RNAs (ncRNAs) were detected in the assembled genome (Table 4). The assembled genome also has antimicrobial resistant genes in them (Figure 3).

Table 4 Key features of NU1901-B013 genome

Feature	Genome
Genome size	2,676,931 bp
GC content	71.16%
Predicted genes (total)	2,518
Coding sequences (CDS)	2,506
Protein coding genes	2,450
RNA genes	56
rRNA	10
5S rRNA	3
16S rRNA	4
23S rRNA	3
tRNA	46
Pseudo genes	12
repeats	0

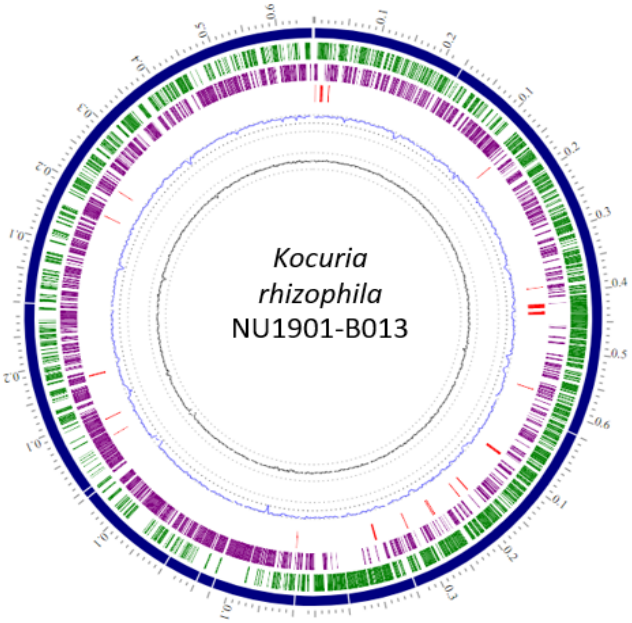


Figure 3 Circular map showing the key features of the genome NU1901-B013.
The outer most navy-blue concentric circle denotes the size of the contigs. The green and purple concentric circles represent the predicted protein-coding genes, present in the forward (green circle) and reverse (purple circle) strands of the genome. The red lines within the next inner concentric circle indicates the presence of antimicrobial resistant genes within certain loci of the genome. The two innermost concentric circles represent GC content (blue circle) and GC skew (black circle), respectively.

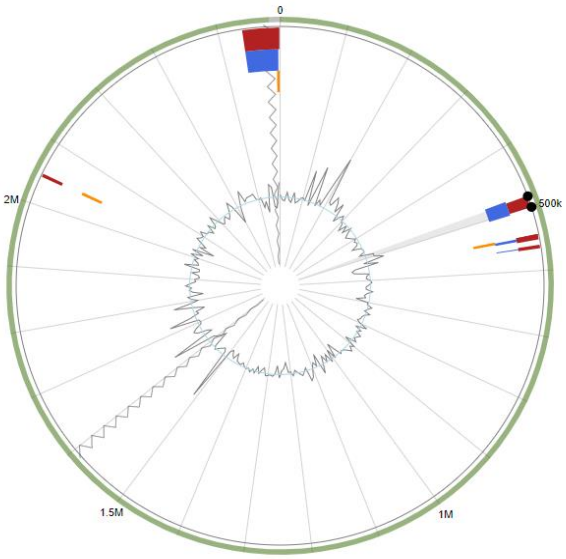


Figure 4 Genomic islands in the genome of NU1901-B013.
The predicted genomic islands in red were generated using an integrated approach, that employed IslandPath-DIMOB (blue) and SIGI-HMM (yellow).

3.4.2 Genomic islands

A total of 7 genomic islands were found in NU1901-B013 genome (Figure 4). These islands contain 264 genes. Three of these islands were predicted by IslandPath-DIAMOND and 4 islands were predicted through SIGI-HMM method.

3.4.3 Phylogenetic analysis

Three phylogenetic trees were constructed from the bacterial genome; one employing the software MEGA X, two employing the servers TYGS, and REALPHY. Based on 16S gene similarity, the bacterium was found within the clade of *Kocuria rhizophila* strains. The most closely related strains were found to be *K. rhizophila* NCTC8340 and *K. rhizophila* FDAARGOS_302 (Figure 5A). REALPHY server provided another phylogeny (Figure 5B), based on the mapping of reference and query genomes. Moreover, TYGS server also provided a phylogenetic tree that suggests that the type genome (within TYGS database) of *Kocuria rhizophila* is the most closely related species of the sequenced genome NU1901-B013 (Figure 5C).

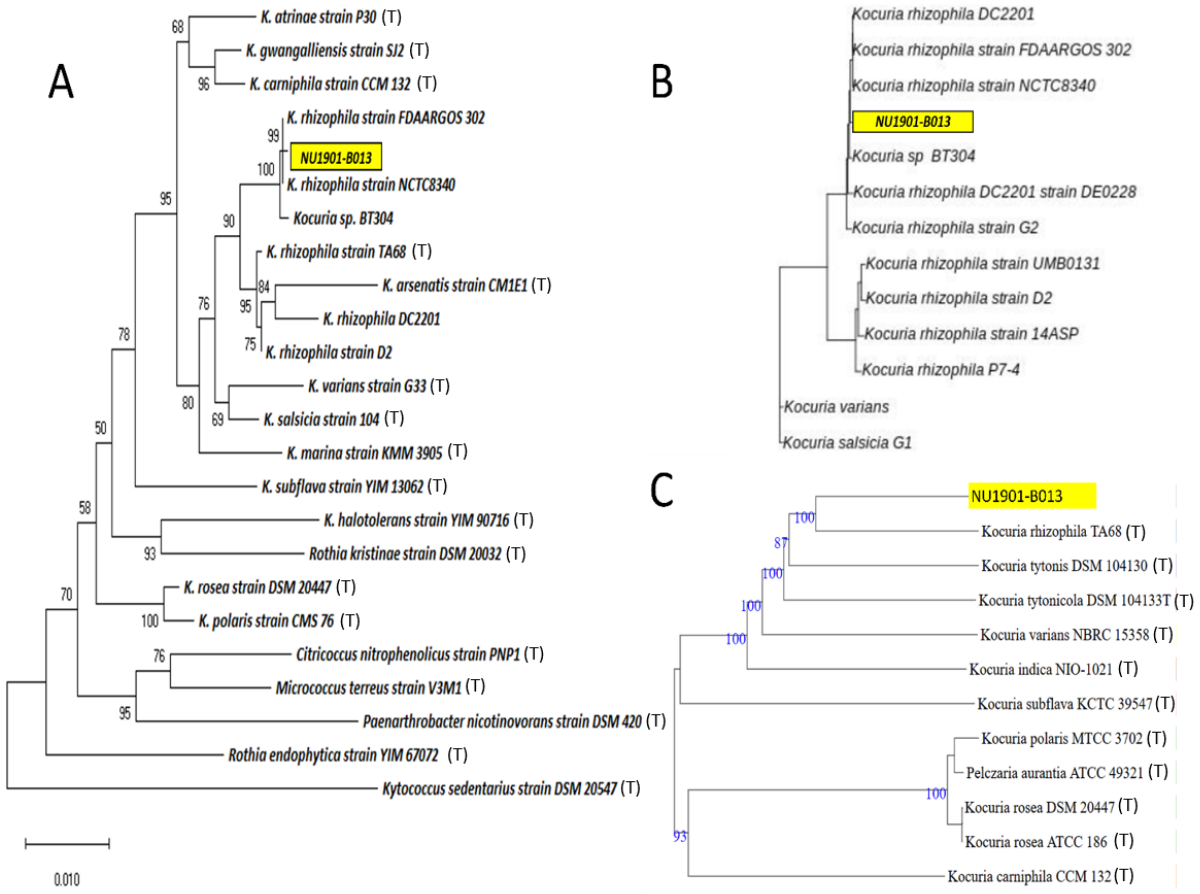


Figure 5 Phylogenetic tree constructed for the assembled NU1901-B013 genome.
A) tree based on 16S rRNA genes using neighbor-joining method and Maximum Composite Likelihood model. The tree was constructed using the 16S rRNA gene of NU1901-B013 and the respective reference sequences from NCBI. The scale indicates 0.01 nucleotide substitutions per nucleotide site. B) tree based on alignments of whole genomes (query and references), constructed by REALPHY server. C) tree, provided by TYGS server, constructed with FastME 2.1.6.1 (Lefort et al., 2015) based on GBDP distances calculated within the genome sequences. The tree was generated with an average branch support of 86.3 % with 100 bootstrap replications and the branch lengths are scaled according to the Genome BLAST Distance Phylogeny (GBDP) distance. T: type strain.

3.4.4 Predicted functions of the genes in the genome of the bacterium

Functional annotation of the bacterial genome was conducted using both RAST and EggNOG-mapper. The 2,450 protein-coding genes, within the genome, could be classified under 245 subsystems within 27 categories, based on the annotations by the RAST server (Figure 6). “Amino acids and derivatives” was the largest category and are associated with 219 genes, followed by “carbohydrates” (155 genes), “protein metabolism” (161 genes), “cofactors, vitamins, prosthetic groups, pigment” (121 genes), “nucleosides and nucleotides” (75 genes), and “fatty acids, lipids, and isoprenoids” (59 genes). Within the category of “amino acids and

derivatives", 48 genes are related to a sub-category "branched-chain amino acid", including "branched-chain amino acid biosynthesis" (10 genes).

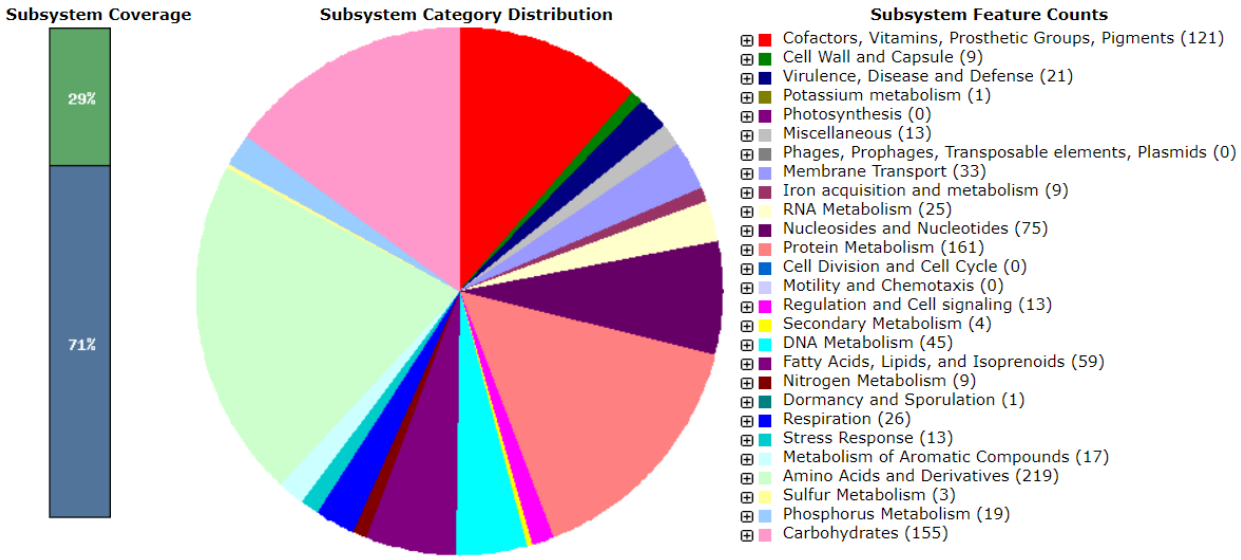


Figure 6 Functional potential of the genes in the genome of NU1901-B013.
The genome was classified into 27 categories and 245 subsystems employing the RAST server. Bar chart indicates the percentage of proteins (green shaded portion) aligned to the subsystem database. The pie chart in the middle represents percentage distribution of the categories and the corresponding gene counts are shown on the right.

A summary of the functional categories and subcategories in the EggNOG database for the annotated genes in the genome is provided in Table 5. A total of 2,108 protein-coding genes in the bacterium genome were found to be orthologous to other species in the EggNOG database. "information storage and processing" seems to be performed by 484 genes, 1013 genes are found to be involved in "metabolism", 347 genes represented "cellular processes and signalling", and 457 genes with unknown functions. In the sub-categories, there are 164 genes that are related to "translation, ribosomal structure, and biogenesis", 161 transcriptional genes, and 158 genes are involved in "DNA replication, recombination and repair" within the category of "information storage and processing". Moreover, in the "metabolism" section 48 genes are engaged in "secondary metabolites biosynthesis, transport, and catabolism".

Table 5 Functional potential of the predicted genes in the genome of NU1901-B013.

Functional categories	No. of genes
INFORMATION STORAGE AND PROCESSING	**484**
Translation, ribosomal structure, and biogenesis	164
Transcription	161
Replication, recombination, and repair	158
Chromatin structure and dynamics	1
CELLULAR PROCESSES AND SIGNALING	**347**
Cell cycle control, cell division, chromosome partitioning	33
Defence mechanisms	31
Signal transduction mechanisms	62
Cell wall/membrane/envelope biogenesis	99
Cell motility	6
Intracellular trafficking, secretion, and vesicular transport	36
Posttranslational modification, protein turnover, chaperones	80
METABOLISM	**1013**
Energy production and conversion	133
Carbohydrate transport and metabolism	137
Amino acid transport and metabolism	249
Nucleotide transport and metabolism	83
Coenzyme transport and metabolism	100
Lipid transport and metabolism	112
Inorganic ion transport and metabolism	151
Secondary metabolites biosynthesis, transport, and catabolism	48
POORLY CHARACTERIZED	**457**
Function unknown	457

3.4.5 Relatedness of NU1901-B013 to the closely related reference genomes

The results of genome comparisons are presented in tables 6 and 7. Average Nucleotide Identity (OrthoANI and ANI) was determined to understand the genetic relatedness of the whole genome of the bacterium and the corresponding references. Although the highest OrthoANI score was obtained for the comparison with the genome of Kocuria rhizophila DC2201

(99.013%) (Figure 7), the highest ANI was found for the comparison with the genome of *Kocuria rhizophila* strain NCTC8340. Six of the reference genomes had a DDH (in silico DNA-DNA Hybridization) value of =>70%. NU1901-B013 genome was found to contain features that are present in the other strains of *Kocuria rhizophila*. The functional aspects, of the bacterium, include vitamin biosynthesis, branched chain amino acid (BCAA) production, butanol and butyrate biosynthesis etc.

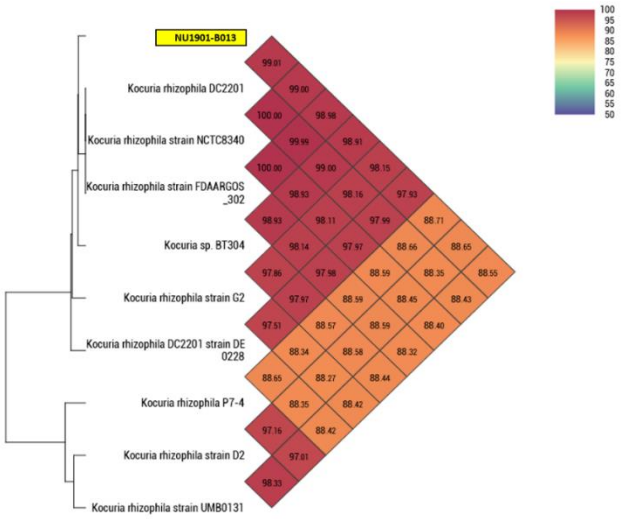


Figure 7 Heatmap showing the similarities between the NU1901-B013 and the reference genomes. *The OrthoANI values calculated by the OAT software were employed to generate the heatmap.*

Table 6 Comparison of structural and functional features of NU1901-B013 with those of related genomes

		NU1901-B013	*Kocuria* sp. BT304	*K. rhizophila* DC2201	*K. rhizophila* FDAARGOS_302	*K. rhizophila* G2	*K. rhizophila* D2	*K. rhizophila* 14ASP	*K. rhizophila* P7-4	*K. rhizophila* NCTC8340	*K. rhizophila* DE0228
Structural features	**Contigs/scaffolds**	13	1	1	1	87	34	183	54	1	62
	Size bp	2,676,931	2,763,150	2,697,540	2,697,877	2,881,857	2,636,961	2,698,103	2,820,331	2,697,831	2,869,143
	GC %	71.2	71.2	71.2	71.2	70.8	70.8	70.8	70.5	71.2	70.9
	Source	Salmon gut	Bovine gut	soil	N/A	Wall in slaughterhouse	Human feces	Soil	Fish gut	N/A	Soil
	Number of Coding Sequences	2450	2488	2450	2465	2682	2343	2732	2540	2446	2665
	Number of RNAs	56	55	55	55	51	49	52	50	55	50
	Number of subsystems (RAST)	245	247	247	247	244	247	247	253	247	249
Number of genes involved in RAST subsystem	**Virulence**	0	0	0	0	0	0	0	0	0	0
	Resistance to antibiotics and toxic compounds	9	10	10	10	11	14	21	14	10	18
	Resistance to fluoroquinolones	2	2	2	2	2	2	4	2	2	2
	Biotin biosynthesis	14	14	14	14	14	14	16	16	14	14
	Menaquinone and Phylloquinone Biosynthesis	12	12	12	12	12	12	14	12	12	12
	Thiamin biosynthesis	7	7	7	7	7	7	8	8	7	7
	Pyridoxin (Vitamin B6) Biosynthesis	6	6	6	6	6	6	6	6	6	6
	NAD and NADP cofactor biosynthesis	8	8	8	8	9	8	7	8	8	8

Folate biosynthesis and cluster	23	23	23	23	23	23	11	23	23	24
Coenzyme biosynthesis and cluster	14	14	14	14	14	14	14	14	14	14
BCAA biosynthesis	10	9	9	9	10	9	11	9	9	11
Glutamine, Glutamate, Aspartate and Asparagine Biosynthesis	15	15	16	16	17	16	25	15	16	15
Lactate utilization	8	8	8	8	8	8	10	8	8	8
Butanol Biosynthesis	10	10	10	10	10	8	9	9	10	10
Acetyl-CoA fermentation to Butyrate	15	15	15	15	15	14	17	15	15	15
Toxin-antitoxin replicon stabilization systems	4	2	0	0	2	0	2	0	0	2

Table 7 Parameters indicating the genetic relatedness of NU1901-B013 with its closely related genomes

Query	References	DDH (%)	Probability DDH >= 70%	OrthoANI value (%)	Original ANI value (%)	GGDC distance	Difference in GC percentage
NU1901-B013	*Kocuria rhizophila* DC2201	91.5	96.26	99.0135	98.9977	0.0105	0.01
NU1901-B013	*Kocuria rhizophila* strain NCTC8340	91.5	96.26	98.9964	99.001	0.0105	0.01
NU1901-B013	*Kocuria rhizophila* strain FDAARGOS_302	91.5	96.26	98.9779	98.9797	0.0105	0.01
NU1901-B013	*Kocuria* sp. BT304	90.9	96.09	98.9099	98.863	0.0112	0.05
NU1901-B013	*Kocuria rhizophila* strain G2	84	93.22	98.1484	98.1203	0.0187	0.34
NU1901-B013	*Kocuria rhizophila* DC2201 strain DE0228	82.5	92.32	97.9349	98.024	0.0204	0.27
NU1901-B013	*Kocuria rhizophila* P7-4	35.4	0.74	88.7056	88.2678	0.1166	0.63
NU1901-B013	*Kocuria rhizophila* strain D2	35.2	0.7	88.6522	88.3013	0.1173	0.32
NU1901-B013	*Kocuria rhizophila* strain 14ASP	35.1	0.67	88.506	88.2301	0.1179	0.4

3.5 Genomic features, identification, and functional analysis of the NU1901-Y022 genome.

3.5.1 Structural features of the genome

The genome size of the yeast (NU1901-Y022) was estimated to be 22,772,963 bp with an average GC content of 57.3% (Table 8). A total of 10,613 genes were predicted, among which 3,251 genes were found to be pseudogenes. Rest of the 7,294 predicted genes are protein-coding and 186 are RNA genes. Among the RNA genes, 10 are rRNA genes (seven 8S rRNAs, one 18S rRNA and two 28S rRNAs) and 176 are tRNA genes.

Table 8 Key features of the NU1901-Y022 genome

Feature	Genome
Genome size	22,772,963 bp
GC content	57.3%
Predicted genes (total)	10,613
CDS	7,362
Protein-coding Genes	7,294
RNA genes	186
rRNA	10
8S rRNA	7
18S rRNA	1
28S rRNA	2
tRNA	176
Pseudo genes	3,251
repeats	13,544

3.5.2 Phylogenetic analysis

Three phylogenetic trees were constructed from the yeast genome; two employing MEGA X and one with REALPHY. The 18S rRNA based phylogenetic tree indicates that the yeast NU1901-Y022 is more related to a clade containing three different yeast species [*Rhodotorula evergladenensis, Rhodotorula mucilaginosa (with the highest BLAST alignment score of 99.34%),* and *Rhodotorula alboruscens*] than a clade of other *R. mucilaginosa* strains (Figure 8A). ITS-based tree indicates that the yeast is most closely related with *Rhodotorula mucilaginosa* strain PY_32 (Figure 8B). On the other hand, whole-genome based phylogenetic tree, provided by REALPHY, identifies *Rhodotorula* sp. JG-1B as the closest relative to the

sequenced yeast genome. *R. mucilaginosa* strains have only a distant relationship with the sequenced genome (Figure 8C).

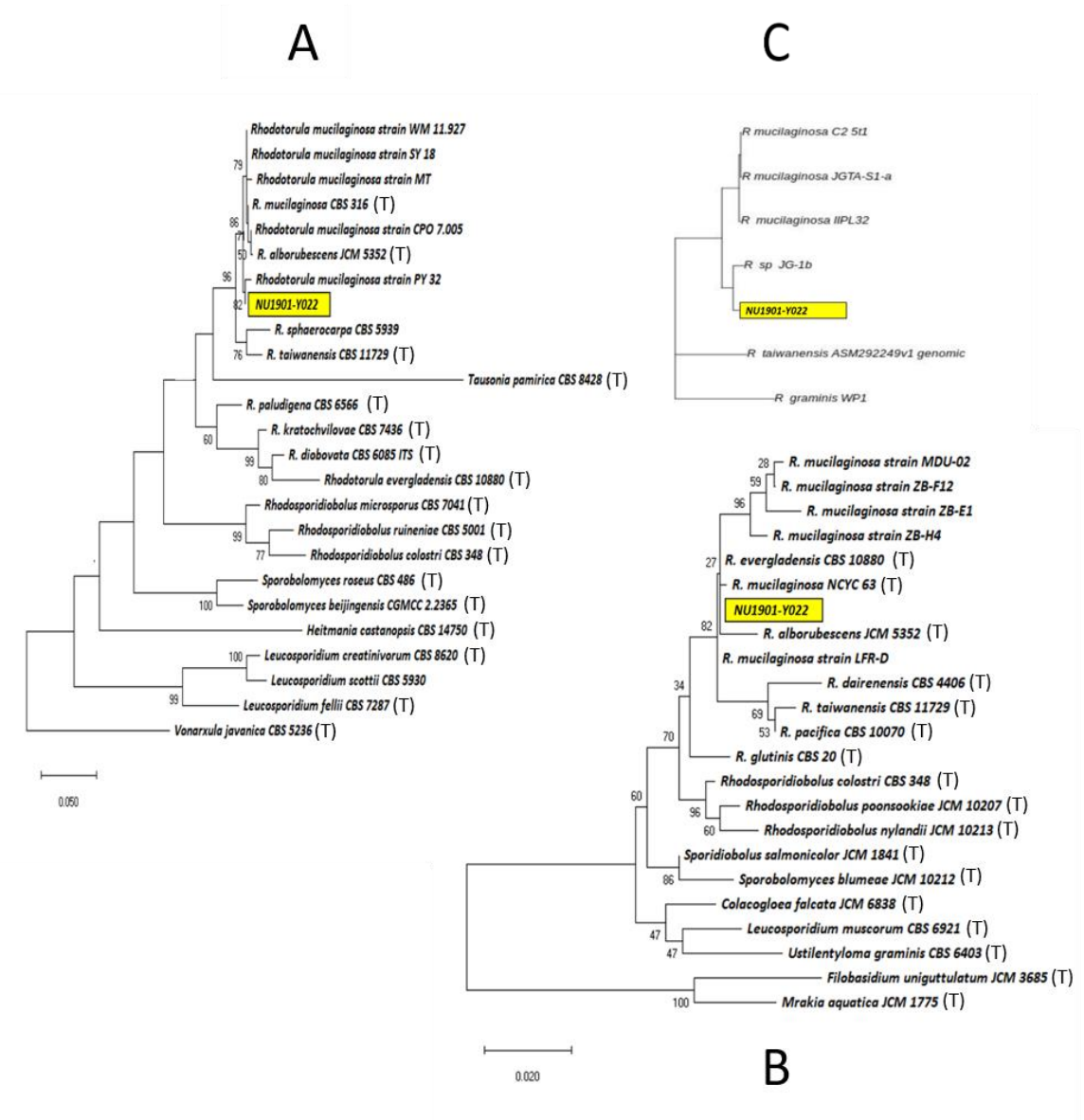


Figure 8 Phylogenetic tree constructed for the assembled NU1901-Y022 genome.
A) Phylogenetic tree generated using the ITS regions and the respective reference sequences from NCBI. The tree was constructed employing Maximum Likelihood method and General Time Reversal model. The scale indicates 0.05 nucleotide substitutions per nucleotide site. B) 18S rRNA based phylogenetic tree. The tree was constructed employing Maximum Likelihood method and Tamura-3 model. The scale indicates 0.02 nucleotide substitutions per nucleotide site. C) Phylogenetic tree based on alignments of whole genomes (query and references), constructed using REALPHY. T: type strain.

3.5.3 Functional features of the genome

In the case of yeast genome, EggNOG-mapper and BlastKOALA were employed to understand its functional potential. First, the annotated genes, obtained from the GenSAS pipeline, were categorized through orthology assignment in EggNOG-mapper. A total of 3,494 genes were found to be orthologous to other species in the EggNOG database, among which 796 genes are of unknown function. Out of the genes with known functions, 812 genes are found to be involved in "information storage and processing", 1070 genes in "cellular processes and signaling" and 1109 genes in "metabolism". Under "metabolism" category, among other sub-categories, there are genes involved in "carbohydrate transport and metabolism" (205 genes), "amino acid transport and metabolism" (221 genes), "lipid transport and metabolism" (175 genes), "inorganic ion transport and metabolism" (103 genes) and "secondary metabolites biosynthesis, transport, and catabolism" (115 genes). Table 9 provides the number of genes, within the yeast genome, that are involved in various functional categories. Moreover, BlastKOALA was used for KEGG pathway annotation of the genome. The software was able to reconstruct complete pathways within the functional categories from a total of 3096 genes (43.9% of the protein-coding genes) (Figure 9). Here, almost half of the genes of the annotated genes were assigned to the pathways related to the category, genetic information processing (1370 genes). There were also genes that were involved in carbohydrate metabolism (224 genes), amino acid (125 genes), and lipid (113 genes) metabolisms. The overview of the main functional categories is summarized in Table 10.

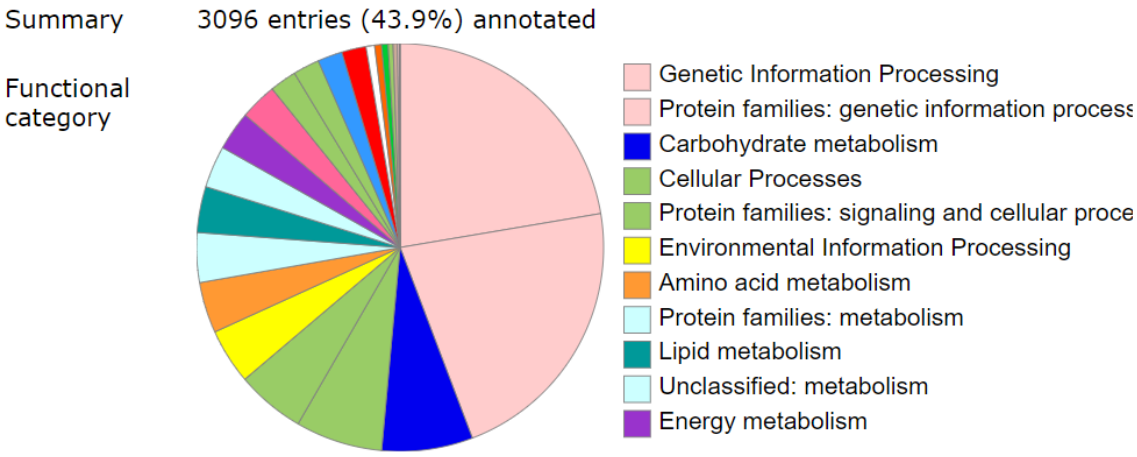


Figure 9 A pie chart showing the percentage of genes involved in the functional categories, annotated by BlastKOALA.

Table 9 Functional annotation of the predicted genes of NU1901-Y022 using EggNOG-mapper

Functional categories	No. of genes
INFORMATION STORAGE AND PROCESSING	**812**
RNA processing and modification	180
Translation, ribosomal structure, and biogenesis	226
Transcription	165
Replication, recombination, and repair	173
Chromatin structure and dynamics	68
CELLULAR PROCESSES AND SIGNALING	**1070**
Cell cycle control, cell division, chromosome partitioning	106
Defence mechanisms	36
Signal transduction mechanisms	259
Cell wall/membrane/envelope biogenesis	51
Cell motility	3
Intracellular trafficking, secretion, and vesicular transport	225
Posttranslational modification, protein turnover, chaperones	297
Extracellular structures	5
Nuclear structure	16
Cytoskeleton	72
METABOLISM	**1109**
Energy production and conversion	138
Carbohydrate transport and metabolism	205
Amino acid transport and metabolism	221
Nucleotide transport and metabolism	55
Coenzyme transport and metabolism	97
Lipid transport and metabolism	175
Inorganic ion transport and metabolism	103
Secondary metabolites biosynthesis, transport, and catabolism	115
POORLY CHARACTERIZED	**796**
Function unknown	796

Table 10 Important functional pathways in NU1901-Y022 and the number of genes connected to them

Functional categories	**No. of genes**
Metabolic pathways	632
Biosynthesis of secondary metabolites	257
Microbial metabolism in diverse environments	131
Carbohydrate metabolism	224
Amino acid metabolism	125
Lipid metabolism	113
Metabolism of cofactors and vitamins	112
Pantothenate and CoA biosynthesis	15
Ubiquinone and other terpenoid-quinone biosynthesis	10
Metabolism of terpenoids and polyketides	25
Carotenoid biosynthesis	2
Zeatin biosynthesis	1
Biosynthesis of ansamycins	1
Biosynthesis of other secondary metabolites	38
Betalain biosynthesis	1
Penicillin and cephalosporin biosynthesis	1
Carbapenem biosynthesis	2
Streptomycin biosynthesis	4
Neomycin, kanamycin, and gentamicin biosynthesis	1
Novobiocin biosynthesis	3
Xenobiotics biodegradation and metabolism	54
Toluene degradation	1
Dioxin degradation	1
Naphthalene degradation	3

3.5.4 Relatedness of NU1901-Y022 to the closely related reference genomes

The OrthoANI, original ANI and the DDH values also indicate that *Rhodotorula* sp. JG-1B is the closest relative with genome similarity of 93.15%, 92.69% and 48.4%, respectively. However, none of the DDH values were ≥ 70% (Table 10). In the heatmap, generated based on OrthoANI values, none of the refence genomes has a genome similarity ≥ 95% (Figure 9).

Basic structural and functional features of NU1901-Y022 were compared with five *Rhodotorula* genomes. The NU1901-Y022 has the largest genome size, and the number of predicted genes in the genome is more compared to the other analysed genomes. However, the assembled genome has fewer orthologous genes in EggNOG database than the reference genomes used in this study (Table 11). In addition, it has the least GC percentage among the considered genomes. Genes involved in "information storage and processing" and "metabolism" categories are more in *Rhodotorula mucilaginosa* JGTA-S1, while the number of genes related to "cellular processes and signalling" category is more in *Rhodotorula graminis* WP1.

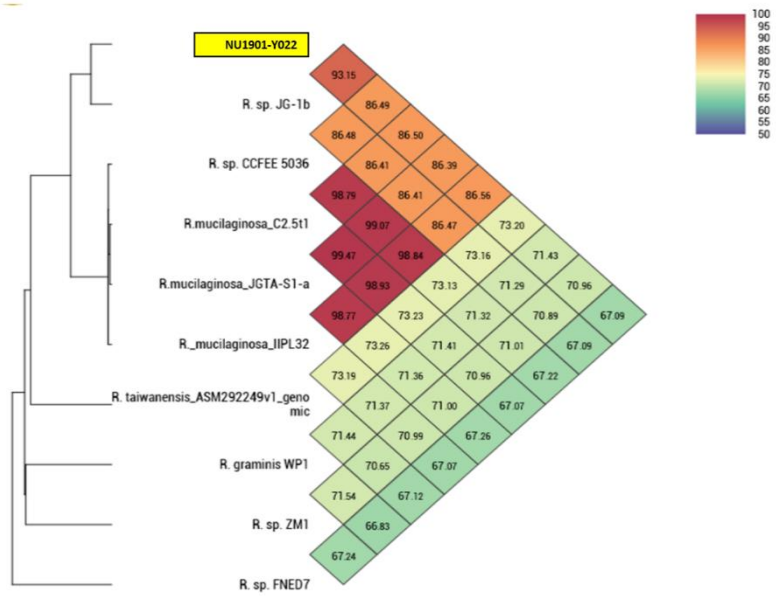


Figure 10 Heatmap showing the similarity of NU1901-Y022 with its close relatives.

Table 11 Parameters indicating the genetic relatedness of NU1901-Y022 with its closely related genomes

Query Genome	Reference Genome	DDH (%)	Probability DDH >= 70%	OrthoANI value (%)	Orginal ANI value (%)	GGDC distance	Difference in GC percentage
NU1901-Y022	*Rhodotorula.* sp. JG-1b	48.4	15.06	93.15	92.69	0.0754	3.27
NU1901-Y022	*R. mucilaginosa* IIPL32	29.7	0.09	86.56	85.77	0.1438	3.21
NU1901-Y022	*R. mucilaginosa* C2.5t1	29.6	0.09	86.50	85.70	0.1443	3.19
NU1901-Y022	*Rhodotorula.* sp. CCFEE 5036	29.5	0.09	86.49	85.70	0.1446	3.23
NU1901-Y022	*R. mucilaginosa* JGTA-S1-a	29.5	0.04	86.39	85.66	0.1451	3.17
NU1901-Y022	*R. graminis* WP1	19.4	0	71.43	70.95	0.2260	10.42
NU1901-Y022	*R. taiwanensis* ASM292249v1	19.1	0	73.20	72.26	0.2307	4.34
NU1901-Y022	*Rhodotorula.* sp. ZM1	19	0	70.96	70.37	0.2309	3.82
NU1901-Y022	*Rhodotorula.* sp. FNED7	18.6	0	67.09	66.74	0.2363	7.88

Table 12 Comparison of the structural and functional features of NU1901-Y022 and reference genomes

Features	**NU1901-Y022**	***Rhodotorula* sp. JG-1B**	***Rhodotorula* sp. CCFEE 5036**	***Rhodotorula mucilaginosa* JGTA-S1**	***Rhodotorula graminis* WP1**	***Rhodotorula taiwanensis***
Genome size (Mb)	22.7729	20.014	19.3934	20.072	19.0726	19.6067
GC percentage	57.3	67.5	60.6	59.9	60.5	61.6
Contigs/Scaffolds	958	26	156	46	155	181
Protein-coding genes	7294	7225	6681	7110	6434	7014
Genes found orthologous in egg-NOG database	3494	5196	4925	5385	5432	5321
Information storage and processing (genes)	812	1249	1181	1295	1242	1240
Cellular processes and signalling (genes)	1070	1558	1447	1581	1601	1571
Metabolism (genes)	1109	1558	1485	2182	1666	1593
Function unknown (genes)	796	1243	1182	1341	1353	1334

4. Discussion

The present study describes the whole genomes of a bacterium and a yeast isolated from the intestine of Atlantic salmon and provides information about their taxonomic identity using genomic information. Based on this information, their structural features and potential functions that could benefit the host will be elaborated in the following sections.

Whole genome sequencing of gut microbes helps to reveal the important characteristics of microbes that can be exploited by appropriate industrial sectors. Such information will be useful either to develop feed probiotic organisms or for using them as biocontrol agents in aquaculture or livestock industries. Furthermore, the information is expected to support the collaborative efforts by scientists who are dedicated to provide scientific evidence on natural substances including probiotics (Swanson et al., 2020).

4.1 High quality DNA and high-quality reads for reliable genome assembly

The purity and the quantity of the extracted DNA was acceptable for library preparation and sequencing. The sequences from the sequencer were pre-processed for reliable genome assembly. Bases with a Phred score of 20 or higher were retained and used for the *de novo* assembly, indicating a base-calling accuracy of 99%. For assembly of a draft genome, a Phred score of 20 or higher is acceptable, according to Lee (2020). During the sequencing process, the sequencer determines the probability of base-calling accuracy, which is known as Phred quality score (Del Angel et al., 2018). Low-quality raw reads from the sequencer were trimmed to obtain a reliable genome assembly.

4.2 SPAdes-Velvet combination improved the assembly of the genomes

SPAdes assembler was used for the de novo assembly. Del Angel et al. (2018) suggested SPAdes as the best assembly tool for smaller genomes of microorganisms. In addition, the contig file obtained from velvet assembler (with k-mer 149) was used in SPAdes employing a parameter "—trusted-contig", which is used to improve the assembly through graph construction, gap closure and repeat resolution from the additional assembly, as suggested by Prjibelski et al. (2020). In the present study, this combined approach increased both N50 and the largest contig size as well as reduced the number of contigs. These parameters are considered primarily to assess the quality and completeness of an assembly (Carneiro et al.,

2012). The values obtained here indicates the improvement in quality of the assemblies of the sequenced genomes.

4.3 GFinisher improved the bacterium assembly remarkably

GFinisher was used to refine and finalize the bacterium genome assembly. This software finds probable assembly error by pointing out GC skew bias, orders contigs based on reference genome and closes the gaps (Kremer et al., 2017). Other studies on whole genome sequencing of bacterium species have employed GFinisher to improve the genome assembly (Da Costa et al., 2017; Haubert et al., 2018; Palmeiro et al., 2019). In the present study, the bacterium genome was finally reordered and joined into 13 scaffolds (from 46 scaffolds, which was obtained using the scaffolding software, SSPACE-Standard). The lesser the number of scaffolds the more assembled will be the genome (Boetzer and Pirovano, 2012). Generally, assembly with short reads rarely generate fewer scaffolds as reported here, suggesting that genome for the bacterium is nearly complete.

Gfinisher can only be used to polish prokaryote genomes. The available software that can perform similar functions in eukaryotes is RaGOO (Alonge et al., 2019b). However, to generate a reference guided refinement of a *de novo* assembly as for the present bacterium, there should be an appropriate reference genome. However, employing the same procedures adopted for the bacterium, I was not able to identify the yeast up to the species level. To select an appropriate reference genome, the sequenced genome should be assigned to a species that could be used by an assembly-polishing software for a reliable refinement (Alonge et al., 2019a; Silva et al., 2013).

4.4 The assemblies of the genomes were of high quality

BUSCO, a well-established genomic tool was used to assess genome completeness based on core gene mapping (Seppey et al., 2019). A BUSCO score of >90% is accepted by researchers, in the field of whole genome sequencing, as high degree of completeness (Fletcher et al., 2018; Johnson et al., 2020; Manni et al., 2020), and I obtained 97% for the bacterium and 93.1% for the yeast. Moreover, low percentages of duplicated core genes (2.6% and 0% in bacterium and yeast genome) also confirm the high quality of the assemblies (Seppey et al., 2019).

4.6 NU1901-B013 is *Kocuria rhizophila*

NU1901-B013 was identified as *K. rhizophila*; based on multiple phylogenetic analyses and genome similarity-based assessments. 16S rRNA gene is the gold-standard marker gene that is employed for taxonomic assignment of unknown bacterium as it is the most conserved gene found in the genome of all the species under the domain bacteria (Lagier et al., 2018). In the present study, the phylogenetic tree constructed based on 16S rRNA of the sequenced genome clearly indicates that NU1901-B013 shares a common ancestor with other *K. rhizophila* strains. The tree also revealed the closest relatives of the sequenced bacterium as *K. rhizophila* FDAARGOS_302 and NCTC8340 strains. A study of the whole genome of *Kocuria* sp. BT304 presented a similar 16S-based tree (Whon et al., 2018); they used only 16S genes of type strains and hence the target strain was positioned in an outer branch with a common ancestor of *K. rhizophila* TA68 and *K. arsenatis* CM1E. It should be noted that Whon et al. (2018) were able to identify their sequenced bacterium only up to the genus level. In the present study, non-type strains of *K. rhizophila* were included which clarified the fact that the *K. arsenatis* CM1E is an unexpected occurrence within the clade of *K. rhizophila* strains. The whole genome-based tree reported here, employing multiple alignments of query and reference genomes, provided identical findings to that of 16S-based taxonomical position of the sequenced bacterium NU1901-B013. The other whole genome-based tree in the present study also indicated that the genome of *K. rhizophila* TA68 is closely related to NU1901-B013 among all the type strains of *Kocuria* sp. within TYGS database. Furthermore, both the ANI and OrthoANI values were ~98-99% for the comparisons of NU1901-B013 with *K. rhizophila* DC2201, *K. rhizophila* NCTC8340, *K. rhizophila* FDAARGOS_302, *K. rhizophila* DE0228, *Kocuria* sp. BT304, and *K. rhizophila* G2. The recommended threshold of 70% DNA-DNA Hybridization (DDH) score for species delineation (Wayne et al., 1987) correlates with the ANI and OrthANI value of >95-96% (Goris et al., 2007; Richter and Rosselló-Móra, 2009). The DDH scores were also found to be within the range of 82.5-91.5% when compared to the above-mentioned six genomes of *K. rhizophila* strains, which confirms that NU1901-B013 belongs to the species, *K. rhizophila*. On the other hand, comparison of NU1901-B013 with other strains, namely *K. rhizophila* P7-4, *K. rhizophila* D2, and *K. rhizophila* 14ASP gave ANI and OrthoANI values of ~88% and the DDH scores were ~35%. The misidentification of the deposited strains at the species level cannot be ruled out. These three strains are clustered together in a clade (of the whole genome-based phylogenetic tree generated in this study), which is positioned away from other *K. rhizophila* strains in the whole genome-based tree (REALPHY). However, genomic features

of NU1901-B013 were also found to be similar to those of the *Kocuria rhizophila* strains selected from the database. Since all the reference genomes were annotated with the same annotation pipeline (RAST) the comparisons can be considered as reliable. Structural and functional features of the sequenced bacterium genome and those of the reference genomes were comparable, even though all the strains were isolated from different sources. All the evidences confirm that NU1901-B013 is *K. rhizophila*.

4.7 NU1901-Y022 is a probable novel yeast species belonging to the genus *Rhodotorula*

Phylogenetic tree constructed using 18S rRNA genes positioned the yeast (NU1901-Y022) within a clade that is comprised of three different *Rhodotorula* spp. i.e., *R. mucilaginosa* NCYC_63, *R. alborubescence* JCM 5352, and *R. evergladensis* CBS 10880. ITS-based tree indicates that the yeast has the most recent common ancestor with *R. mucilaginosa* PY 32, and both of them are related to the sister clade of other *R. mucilaginosa* strains. Though, ITS region is approved by the Consortium for Barcode of Life as official fungal barcode marker, there is no universal cut-off value for species delineation (Raja et al., 2017). However, the whole genome-based tree, constructed with the probable reference genomes, found *Rhodotorula* sp. JG-1b as the closely related genome, which itself is not defined up to the species level. However, together with NU1901-Y022, *Rhodotorula* sp. JG-1b shares a common ancestor with the sister clade of three other *R. mucilaginosa* genomes. The genome-relatedness-based analyses also suggested that the NU1901-Y022 strain can be delineated only up to the genus level (*Rhodotorula*), based on the existing database. Although, the highest ANI and OrthoANI values (93.2% and 92.7%, respectively) were obtained when NU1901-Y022 and *Rhodotorula* sp. JG-1b (which itself is not identified up to the species level) genomes were compared, the values are not within the recommended threshold to consider them as belonging to the same species, as suggested by Richter and Rosselló-Móra (2009). In addition, the DDH score was only 48.5%, which is far below the recommended score (70%) for species delineation (Wayne et al., 1987). These low values indicate that these two genomes do not belong to the same species. The aforementioned comparisons with *Rhodotorula* sp. JG-1b gave the highest ANI, OrthoANI and low DDH scores than when compared to the other reference genomes considered in this study. Hence, a taxonomic classification up to the species level as that of the compared genomes cannot be given to NU1901-Y022. However, these *Rhodotorula* sp. strains showed 85-86% genomic relatedness with NU1901-Y022 in terms of ANI and OrthoANI, which is >85% and indicates that they belong to at least the genus *Rhodotorula*, as suggested

by Lee et al. (2016). Furthermore, the comparative genomic analyses showed a distant relationship of NU1901-Y022 with other closely related genomes in the existing database. The genomic comparison can be trusted as all the genomes were annotated with the same annotation pipeline. All the structural features (such as genome size, GC content, coding gene number) and the functional categories (annotated by eggNOG-mapper) of the sequenced yeast varied from those of the reference genomes. Hence, the genomes selected from the database do not seem to belong to the same species as the sequenced yeast. Thus, NU1901-Y022 could be a potential novel species of *Rhodotorula*, which should be confirmed through future studies.

4.8 NU1901-B013 has some potential benefits to the host

The *K. rhizophila* NU1901-B013 strain was isolated from the distal intestine of Atlantic salmon, where the gut microbiota metabolizes the host's dietary nutrients (such as carbohydrates, lipids, and amino acids) (Butt and Volkoff, 2019; Vatsos, 2017). The genome of the bacterium contains 219 genes that are related to the subsystem category "amino acids and derivatives". All the studied genomes of *K. rhizophila* strains, including the sequenced bacterium genome contain almost the same number of genes (216-231 genes), under this category. When annotated with eggNOG-mapper 249 genes were found under the orthologous category of "amino acid transport and metabolism". This indicates the capability of dietary amino acid utilization and metabolism by *K. rhizophila* strains, as suggested by Whon et al. (2018). NU1901-B013 contains 48 genes in the sub-category "branched chain amino acid (BCAA)" under the category "amino acids and derivatives"; ten genes within this sub-category are related to BCAA biosynthesis. BCAAs are the molecules responsible for lipogenesis in adipocytes (Green et al., 2016). This capacity of bacteria can be further examined in nutritional approaches to improve the flesh quality of fish. The NU1901-B013 genome also contains genes (155 genes by RAST subsystem, 137 by eggNOG-mapper) which are related to carbohydrate utilization and metabolism. Among them, there are genes that are associated with lactate utilization (8 genes). Lactate found in the gut ecosystem can be either from diet or from lactate-producing bacteria such as Lactobacilli, Bifidobacteria, which can cross feed butyrate producing bacteria (Belenguer et al., 2006; Moens et al., 2017). Interestingly, NU1901-B013 is also a butyrate-producing bacterium, which contains at least 15 genes related to "acetyl-CoA fermentation to butyrate" sub-category of RAST subsystem. Butyrate is one of the most important short chain fatty acids produced by the gut microbiota, which is primarily known as the energy provider for the gut epithelia and contributor to host defense in terms of inducing

gut barrier function by promoting "physiological hypoxia" in epithelium cells (Zhang and Davies, 2016). NU1901-B013 genome also possesses some genes that are responsible for the biosynthesis of vitamins, cofactors, and prosthetic groups (121 genes). The genome has 14 genes related to biotin biosynthesis. Biotin is a vitamin (also known as vitamin H), which is mainly known for its ability to convert nutrients into energy (Said, 2008). Biotin is also involved in the regulation of, among others, cytokine genes and glucose metabolism-linked genes (Rodriguez-Melendez and Zempleni, 2003). In addition, the genome contains genes linked to menaquinone and phylloquinone biosynthesis (12), also known as vitamins K1 and K2. These vitamins have well-established and important physiological roles such as blood-coagulation and Ca^{2+} binding to bones and tissues (Kaneki et al., 2006). The genome also contains thiamin- (7 genes) and pyridoxin- (6) producing genes. Pyridoxine is also known as vitamin B6, which has several beneficiary roles such as hemoglobin production, immune function and dietary metabolism (Shils and Shike, 2006). Thiamin, known as vitamin B1, also helps in dietary metabolism, and is an essential micronutrient belonging to vitamin B complex family (Coates et al., 2010). The genome further contains genes for "NAD and NADP cofactor biosynthesis" (8 genes), "coenzyme biosynthesis" (14 genes) and "folate biosynthesis" (also known as vitamin B9, needed for DNA synthesis and cell division) (23 genes). In summary, the NU1901-B013 has the capacity to utilize diet-derived compounds within salmon intestine and may produce beneficial molecules to help the host in digestion, growth, and health. However, a critical question is whether all of these functions, that are encoded by the bacterium, help the bacterium itself or other members of the gut microbial community or the host salmon. This question needs to be answered by conducting in depth studies. It has been reported that if certain members of fish microbiota are capable of producing a vitamin then the animals do not require the vitamin from diet (Ramírez et al., 2018). Furthermore, Whon et al. (2018) suggested that the *Kocuria rhizophila* BT304 they sequenced can be a probable probiotic organism for bovine animals, based on the BCAA production capacity. They also compared the number of virulence related factors (under the subsystem category of "virulence, disease and defense") present in other commercially used probiotics such as *Bifidobacterium animali*s subsp. lactis BL03 (22 genes), *Bifidobacterium animalis* subsp. lactis BI04 (22 genes), *Lactobacillus helveticu*s BD08 (47 genes) etc. The NU1901-B013 strain contains fewer genes (21 genes) under the category "virulence, disease and defense" than any of the abovementioned probiotics. Therefore, the *Kocuria rhizophila* NU1901-B013 can be a suitable probiotic candidate.

4.9 NU1901-Y022 has some potential benefits to the host

Members of the genus *Rhodotorula* is predominantly found in the intestine of many fish species, and they belong to the phylum *Basidiomycota* (Bogusławska-Wąs et al., 2019). *Rhodotorula* is a saprophytic fungus mainly known for its carotenoid producing ability (Gan et al., 2017). *Rhodotorula* NU1901-Y022 was isolated from the intestine of Atlantic salmon. According to the orthology analysis of eggNOG-mapper, the genome contains genes that can participate in carbohydrate (205 genes), lipid (175 genes), and amino acid (221 genes) transport and metabolism. KEGG pathways also indicated that the genome is capable of utilizing and metabolizing carbohydrates, amino acids, and lipids. For the microbe to survive in the gut ecosystem, it must obtain nutrients from host diet-derived products. This ability of the strain in utilizing host dietary products, metabolizing them, and transporting them points to the survivability of NU1901-Y022 as an indigenous microbe. Bogusławska-Wąs et al. (2019) also suggested that *Rhodotorula* is a permanent resident of fish intestine. Several KEGG pathways connected to biosynthesis of beneficial metabolites are found in the genome. Like the sequenced bacterium, NU1901-B013, this yeast also has the ability to biosynthesize biotin and some other vitamins and cofactors. For instance, fifteen genes were involved in the pathway "pantothenate and CoA biosynthesis". Pantothenate is also known as vitamin B5, which is necessary for all animals to synthesize coenzyme A (Leonardi and Jackowski, 2007). Moreover, the genome also can produce carotenoids, and other studies confirmed this ability of several *Rhodotorula* sp. (Aksu and Eren, 2005; Buzzini et al., 2005; Zhao et al., 2019). Carotenoid is mainly known as organic pigments. Interestingly, carotenoid deposition is responsible for the red flesh color of salmon (Shahidi and Brown, 1998). The flesh color is one of the most important parameters that determines the quality of salmon. Because of the consumers perception of a redder flesh as a better-flavored, fresh, and high-quality fish diet, the flesh-color plays a decisive role in the business of farmed salmon (Anderson, 2001). Apart from the pigmentation, carotenoids have antioxidant properties, which can protect humans from head and neck cancer (Leoncini et al., 2015), prostate cancer (Soares et al., 2015), Parkinson's disease (Takeda et al., 2014), and breast cancer (Chajès and Romieu, 2014). So, consumption of carotenoid-rich diet, such as salmon, is beneficial to humans. Furthermore, if the carotenoids that are produced by *Rhodotorula* sp. can be incorporated as additives in salmon feeds or the organism can be supplemented as probiotics, the benefits are multifaceted; the fillet with high nutritive value could be preserved for a longer time. In addition, NU1901-Y022 possesses the genes related to betalain biosynthesis, which is also a pigment (red or yellow),

which has antioxidant properties (Escribano et al., 1998), and it pharmacological property is exploited by the food industry (Choo, 2017).

NU1901-Y022 is also capable of the biosynthesis of ansamycins, which are secondary metabolites with antimicrobial activity (Wehrli and Staehelin, 1971), as well as antiviral activity against bacteriophages. The strain is also capable of producing a range of antibiotics such as penicillin, cephalosporin, carbapenem, streptomycin, neomycin, kanamycin, gentamicin, and novobiocin. In-depth studies must be performed to check, if these antibiotics, produced by the yeast, are harmful to the indigenous bacterial community of the intestine of Atlantic salmon or they provide protection against pathogenic bacteria.

The yeast genome also contains genes those are involved in the pathways of xenobiotic degradation. Pathways related to the degradation of toxic compounds such as toluene, napthalene, and dioxin, indicates the ability of NU1901-Y022 in toxicity-management.

To summarize, NU1901-Y022 is a strain of the genus *Rhodotorula*, which is capable of producing multitude of important metabolites. Nevertheless, the safety for the administered fish and humans should be confirmed through in-depth studies as recommended by the International Scientific Association for Probiotics and Prebiotics (ISAPP) (Swanson et al., 2020).

5. Data availability

All the data related to the present study will be deposited in European Nucleotide Archive.

6. Conclusion

This thesis describes the whole genome sequences of two microbes that commonly occur in the intestine of salmon. The structural features were uncovered, and phylogenetic analyses as well as genome similarity-based analyses confirmed their taxonomic identity. The bacterium isolate NU1901-B013 was identified up to the species level as *Kocuria rhizophila*, while the identity of the yeast isolate NU1901-Y022 was delineated only up to the genus level as *Rhodotorula* sp. Furthermore, all the genes of both the genomes were functionally annotated to reveal the potential metabolic features of the microbes. Both the genomes have the ability to utilize host-dietary components and produce metabolites, some of which are likely to be beneficial for host-physiology. Although, future studies are needed for further clarification about the overall role of these two microbes, this research opens the possibility of using the genome sequence information for improving our understanding on the roles of gut microorganisms in fish.

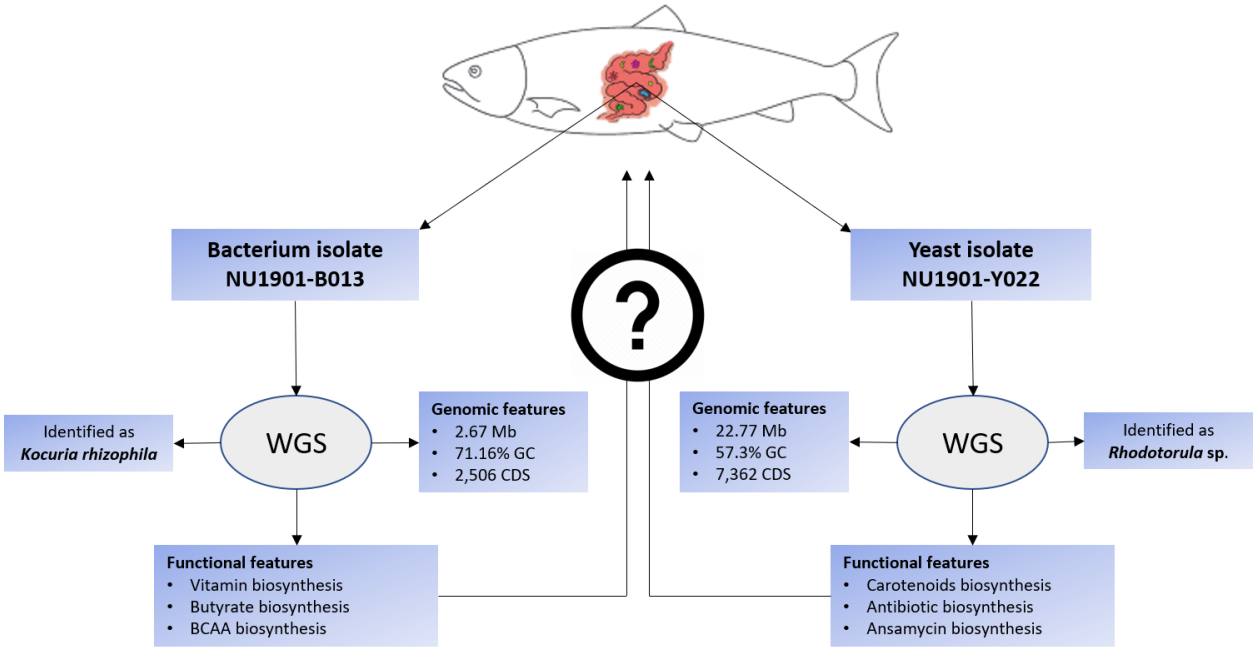


Figure 11 Summary of the present study.

7. Limitations and future perspectives

The sequenced genomes could not be assembled up to the chromosome level. Assembling up to chromosome level is possible with the help of third-generation sequencing techniques such as PacBio (Pacific Bioscience, California, USA) and Nanopore (Oxford Nanopore Technologies, Oxford, UK) sequencing platforms, which are capable of generating long reads. The present study was conducted using the available sequencer (illumina MiSeq) at the genomics facility of Nord University. Only short reads can be generated employing this sequencer. This limitation makes it difficult for the assembler to identify the consensus sequence of a genome up to the chromosome level, especially those that have long repeats, as in the case of eukaryotic genomes.

Future studies must be conducted to confirm that the metabolites encoded by the sequenced microbial genomes are of functional value for the host. Virulence of the genomes should be explored through further studies. In-depth studies can plausibly reveal the unknown functional potential of the genome. In addition, due to the continuous curation, upgrading, and updating of the subsystem (RAST) or KEGG pathway databases, the sequenced genomes will reveal more functions in future. When the metabolic features of these organisms are experimentally demonstrated, it could eventually be established as probiotics in aquaculture or other industries.

8. References

Aksu, Z., Eren, A.T., 2005. Carotenoids production by the yeast *Rhodotorula mucilaginosa*: use of agricultural wastes as a carbon source. Process Biochemistry 40, 2985-2991.

Alonge, M., Soyk, S., Ramakrishnan, S., Wang, X., Goodwin, S., Sedlazeck, F.J., Lippman, Z.B., Schatz, M.C., 2019a. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. Genome Biology 20, 1-17.

Alonge, M., Soyk, S., Ramakrishnan, S., Wang, X., Goodwin, S., Sedlazeck, F.J., Lippman, Z.B., Schatz, M.C., 2019b. RaGOO: fast and accurate reference-guided scaffolding of draft genomes. Genome Biology 20, 224.

Anderson, S., 2001. Salmon color and the consumer. International Institute of Fisheries Economics and Trade, http://localhost/files/9s1616848.

Andrews, S., 2010. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.

Auch, A.F., von Jan, M., Klenk, H.-P., Göker, M., 2010. Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. Standards in Genomic Sciences 2, 117-134.

Austin, B., 2002. The bacterial microflora of fish. Scientific World Journal 2, 558-572.

Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., Formsma, K., Gerdes, S., Glass, E.M., Kubal, M., Meyer, F., Olsen, G.J., Olson, R., Osterman, A.L., Overbeek, R.A., McNeil, L.K., Paarmann, D., Paczian, T., Parrello, B., Pusch, G.D., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A., Zagnitko, O., 2008. The RAST Server: rapid annotations using subsystems technology. BioMed Central Genomics 9, 75.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A.A., Dvorkin, M., Kulikov, A.S., Lesin, V.M., Nikolenko, S.I., Pham, S., Prjibelski, A.D., 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of Computational Biology 19, 455-477.

Belenguer, A., Duncan, S.H., Calder, A.G., Holtrop, G., Louis, P., Lobley, G.E., Flint, H.J., 2006. Two routes of metabolic cross-feeding between *Bifidobacterium adolescentis* and

butyrate-producing anaerobes from the human gut. Applied and Environmental Microbiology 72, 3593-3599.

Bertelli, C., Laird, M.R., Williams, K.P., Simon Fraser University Research Computing Group, Lau, B.Y., Hoad, G., Winsor, G.L., Brinkman, F.S., 2017. IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. Nucleic Acids Research 45, W30-W35.

Bertels, F., Silander, O.K., Pachkov, M., Rainey, P.B., van Nimwegen, E., 2014. Automated reconstruction of whole-genome phylogenies from short-sequence reads. Molecular Biology and Evolution 31, 1077-1088.

Bliss, E.S., Whiteside, E., 2018. The gut-brain axis, the human gut microbiota and their integration in the development of obesity. Frontiers in Physiology 9.

Boetzer, M., Pirovano, W., 2012. Toward almost closed genomes with GapFiller. Genome Biology 13, R56.

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114-2120.

Buchfink, B., Xie, C., Huson, D.H., 2015. Fast and sensitive protein alignment using DIAMOND. Nature Methods 12, 59-60.

Budden, K.F., Gellatly, S.L., Wood, D.L.A., Cooper, M.A., Morrison, M., Hugenholtz, P., Hansbro, P.M., 2017. Emerging pathogenic links between microbiota and the gut–lung axis. Nature Reviews Microbiology 15, 55-63.

Butt, R.L., Volkoff, H., 2019. Gut microbiota and energy homeostasis in fish. Frontiers in Endocrinology 10.

Buzzini, P., Martini, A., Gaetani, M., Turchetti, B., Pagnoni, U.M., Davoli, P., 2005. Optimization of carotenoid production by *Rhodotorula graminis* DBVPG 7021 as a function of trace element concentration by means of response surface analysis. Enzyme and Microbial Technology 36, 687-692.

Carneiro, A.R., Ramos, R.T.J., Barbosa, H.P.M., Schneider, M.P.C., Barh, D., Azevedo, V., Silva, A., 2012. Quality of prokaryote genome assembly: indispensable issues of factors affecting prokaryote genome assembly quality. Gene 505, 365-367.

Chajès, V., Romieu, I., 2014. Nutrition and breast cancer. Maturitas 77, 7-11.

Choo, W.S., 2017. Betalains: application in functional foods, in: Mérillon, J.-M., Ramawat, K.G. (Eds.), Bioactive Molecules in Food. Springer International Publishing, Cham, pp. 1-28.

Coates, P.M., Betz, J.M., Blackman, M.R., Cragg, G.M., Levine, M., Moss, J., White, J.D., 2010. Encyclopedia of dietary supplements. CRC Press.

Cryan, J.F., O'Mahony, S.M., 2011. The microbiome-gut-brain axis: from bowel to behavior. Neurogastroenterology and Motility 23, 187-192.

Da Costa, E.M., Guimarães, A.A., Vicentin, R.P., de Almeida Ribeiro, P.R., Leão, A.C.R., Balsanelli, E., Lebbe, L., Aerts, M., Willems, A., de Souza Moreira, F.M., 2017. *Bradyrhizobium brasilense* sp. nov., a symbiotic nitrogen-fixing bacterium isolated from Brazilian tropical soils. Archives of Microbiology 199, 1211-1221.

de Bruijn, I., Liu, Y., Wiegertjes, G.F., Raaijmakers, J.M., 2018. Exploring fish microbial communities to mitigate emerging diseases in aquaculture. FEMS Microbiology Ecology 94, fix161.

Del Angel, V.D., Hjerde, E., Sterck, L., Capella-Gutierrez, S., Notredame, C., Pettersson, O.V., Amselem, J., Bouri, L., Bocs, S., Klopp, C., 2018. Ten steps to get started in Genome Assembly and Annotation. F1000Research 7.

Duca, F.A., Swartz, T.D., Sakar, Y., Covasa, M., 2012. Increased oral detection, but decreased intestinal signaling for fats in mice lacking gut microbiota. PloS One 7, e39748.

Durack, J., Lynch, S.V., 2019. The gut microbiome: Relationships with disease and opportunities for therapy. The Journal of Experimental Medicine 216, 20-40.

Egerton, S., Culloty, S., Whooley, J., Stanton, C., Ross, R.P., 2018. The gut microbiota of marine fish. Frontiers in Microbiology 9, 873-873.

Escribano, J., Pedreño, M.A., García-Carmona, F., Muñoz, R., 1998. Characterization of the antiradical activity of betalains from *Beta vulgaris* L. roots. Phytochemical Analysis: An International Journal of Plant Chemical and Biochemical Techniques 9, 124-127.

Feng, Q., Chen, W.-D., Wang, Y.-D., 2018. Gut microbiota: An integral moderator in health and disease. Frontiers in Microbiology 9.

Fetissov, S.O., 2017. Role of the gut microbiota in host appetite control: bacterial growth to animal feeding behaviour. Nature Reviews Endocrinology 13, 11-25.

Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Jou, W.M., Molemans, F., Raeymaekers, A., Van den Berghe, A., 1976. Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. Nature 260, 500-507.

Fletcher, K., Klosterman, S.J., Derevnina, L., Martin, F., Bertier, L.D., Koike, S., Reyes-Chin-Wo, S., Mou, B., Michelmore, R., 2018. Comparative genomics of downy mildews reveals potential adaptations to biotrophy. BMC Genomics 19, 851.

Forbes, J.D., Bernstein, C.N., Tremlett, H., Van Domselaar, G., Knox, N.C., 2019. A fungal world: could the gut mycobiome be involved in neurological disease? Frontiers in Microbiology 9, 3249.

Gautam, S.S., KC, R., Leong, K.W., Mac Aogáin, M., O'Toole, R.F., 2019. A step-by-step beginner's protocol for whole genome sequencing of human bacterial pathogens. Journal of Biological Methods.

Ghanbari, M., Kneifel, W., Domig, K.J., 2015. A new view of the fish gut microbiome: advances from next-generation sequencing. Aquaculture 448, 464-475.

Ghannoum, M.A., Jurevic, R.J., Mukherjee, P.K., Cui, F., Sikaroodi, M., Naqvi, A., Gillevet, P.M., 2010. Characterization of the oral fungal microbiome (mycobiome) in healthy individuals. PLoS Pathogens 6, e1000713.

Gomez, G.D., Balcazar, J.L., 2008. A review on the interactions between gut microbiota and innate immunity of fish. FEMS Immunology and Medical Microbiology 52, 145-154.

Goris, J., Konstantinidis, K.T., Klappenbach, J.A., Coenye, T., Vandamme, P., Tiedje, J.M., 2007. DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. International Journal of Systematic and Evolutionary Microbiology 57, 81-91.

Green, C.R., Wallace, M., Divakaruni, A.S., Phillips, S.A., Murphy, A.N., Ciaraldi, T.P., Metallo, C.M., 2016. Branched-chain amino acid catabolism fuels adipocyte differentiation and lipogenesis. Nature Chemical Biology 12, 15-21.

Guizelini, D., Raittz, R.T., Cruz, L.M., Souza, E.M., Steffens, M.B., Pedrosa, F.O., 2016. GFinisher: a new strategy to refine and finish bacterial genome assemblies. Nature Scientific Reports 6, 34963.

Haubert, L., Kremer, F.S., da Silva, W.P., 2018. Whole-genome sequencing identification of a multidrug-resistant *Listeria monocytogenes* serotype 1/2a isolated from fresh mixed sausage in southern Brazil. Infection, Genetics and Evolution 65, 127-130.

Huerta-Cepas, J., Forslund, K., Coelho, L.P., Szklarczyk, D., Jensen, L.J., Von Mering, C., Bork, P., 2017. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. Molecular Biology and Evolution 34, 2115-2122.

Humann, J.L., Lee, T., Ficklin, S., Main, D., 2019. Structural and functional annotation of eukaryotic genomes with GenSAS. Methods in Molecular Biology 1962, 29-51.

Ikeda-Ohtsubo, W., Brugman, S., Warden, C.H., Rebel, J.M.J., Folkerts, G., Pieterse, C.M.J., 2018. How can we define “optimal microbiota?”: A comparative review of structure and functions of microbiota of animals, fish, and plants in agriculture. Frontiers in Nutrition 5.

Jandhyala, S.M., Talukdar, R., Subramanyam, C., Vuyyuru, H., Sasikala, M., Nageshwar Reddy, D., 2015. Role of the normal gut microbiota. World Journal of Gastroenterology 21, 8787-8803.

Jiang, T.T., Shao, T.-Y., Ang, W.G., Kinder, J.M., Turner, L.H., Pham, G., Whitt, J., Alenghat, T., Way, S.S., 2017. Commensal fungi recapitulate the protective benefits of intestinal bacteria. Cell Host & Microbe 22, 809-816. e804.

Johnson, L.K., Sahasrabudhe, R., Gill, J.A., Roach, J.L., Froenicke, L., Brown, C.T., Whitehead, A., 2020. Draft genome assemblies using sequencing reads from Oxford

Nanopore Technology and Illumina platforms for four species of North American Fundulus killifish. GigaScience 9, giaa067.

Kaneki, M., Hosoi, T., Ouchi, Y., Orimo, H., 2006. Pleiotropic actions of vitamin K: protector of bone health and beyond? Nutrition 22, 845-852.

Kremer, F.S., McBride, A.J.A., Pinto, L.d.S., 2017. Approaches for in silico finishing of microbial genome sequences. Genetics and Molecular Biology 40, 553-576.

Kulski, J.K., 2016. Next-generation sequencing—an overview of the history, tools, and "Omic" applications. Next Generation Sequencing–Advances, Applications and Challenges, 3-60.

Kumar, S., Stecher, G., Li, M., Knyaz, C., Tamura, K., 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. Molecular Biology and Evolution 35, 1547-1549.

Kuska, B., 1998. Beer, Bethesda, and biology: how "genomics" came into being. Oxford University Press.

Lagesen, K., Hallin, P., Rødland, E.A., Stærfeldt, H.-H., Rognes, T., Ussery, D.W., 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Research 35, 3100-3108.

Lagier, J.-C., Dubourg, G., Million, M., Cadoret, F., Bilen, M., Fenollar, F., Levasseur, A., Rolain, J.-M., Fournier, P.-E., Raoult, D., 2018. Culturing the human microbiota and culturomics. Nature Reviews Microbiology 16, 540-550.

Land, M., Hauser, L., Jun, S.-R., Nookaew, I., Leuze, M.R., Ahn, T.-H., Karpinets, T., Lund, O., Kora, G., Wassenaar, T., 2015. Insights from 20 years of bacterial genome sequencing. Functional & Integrative Genomics 15, 141-161.

Lee, D.-H., 2020. Complete genome sequencing of influenza a viruses using next-generation sequencing, in: Spackman, E. (Ed.), Animal Influenza Virus: Methods and Protocols. Springer US, New York, NY, pp. 69-79.

Lee, I., Ouk Kim, Y., Park, S.-C., Chun, J., 2016. OrthoANI: an improved algorithm and software for calculating average nucleotide identity. International Journal of Systematic and Evolutionary Microbiology 66, 1100-1103.

Leonardi, R., Jackowski, S., 2007. Biosynthesis of pantothenic acid and coenzyme A. EcoSal Plus.

Leoncini, E., Nedovic, D., Panic, N., Pastorino, R., Edefonti, V., Boccia, S., 2015. Carotenoid intake from natural sources and head and neck cancer: A systematic review and meta-analysis of epidemiological studies. Cancer Epidemiology Biomarkers & Prevention 24, 1003-1011.

Lowe, T.M., Eddy, S.R., 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Research 25, 955-964.

Manni, M., Simao, F.A., Robertson, H.M., Gabaglio, M.A., Waterhouse, R.M., Misof, B., Niehuis, O., Szucsich, N.U., Zdobnov, E.M., 2020. The genome of the blind soil-dwelling and ancestrally wingless dipluran *Campodea augens*: A key reference hexapod for studying the emergence of insect innovations. Genome Biology and Evolution 12, 3534-3549.

Marchesi, J.R., Ravel, J., 2015. The vocabulary of microbiome research: a proposal. Microbiome 3, 31.

Massaquoi, M.S., Guillemin, K., 2018. Evolving in a microbial soup: you are what they eat. Developmental Cell 47, 682-683.

Medvecky, M., Cejkova, D., Polansky, O., Karasova, D., Kubasova, T., Cizek, A., Rychlik, I., 2018. Whole genome sequencing and function prediction of 133 gut anaerobes isolated from chicken caecum in pure cultures. BMC Genomics 19.

Meier-Kolthoff, J.P., Auch, A.F., Klenk, H.-P., Göker, M., 2013. Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinformatics 14, 60.

Meier-Kolthoff, J.P., Göker, M., 2019. TYGS is an automated high-throughput platform for state-of-the-art genome-based taxonomy. Nature Communications 10, 1-10.

Moens, F., Verce, M., De Vuyst, L., 2017. Lactate-and acetate-based cross-feeding interactions between selected strains of lactobacilli, bifidobacteria and colon bacteria in the presence of inulin-type fructans. International Journal of Food Microbiology 241, 225-236.

Ondov, B.D., Treangen, T.J., Melsted, P., Mallonee, A.B., Bergman, N.H., Koren, S., Phillippy, A.M., 2016. Mash: fast genome and metagenome distance estimation using MinHash. Genome Biology 17, 132.

Overbeek, R., Olson, R., Pusch, G.D., Olsen, G.J., Davis, J.J., Disz, T., Edwards, R.A., Gerdes, S., Parrello, B., Shukla, M., Vonstein, V., Wattam, A.R., Xia, F., Stevens, R., 2014. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Research 42, D206-214.

Palmeiro, J.K., De Souza, R.F., Schörner, M.A., Araujo, H.P., Grazziotin, A.L., Vidal, N.M., Venancio, T.M., Dalla Costa, L.M., 2019. Molecular epidemiology of multidrug-resistant *Klebsiella pneumoniae* isolates in a Brazilian tertiary hospital. Frontiers in Microbiology 10, 1669.

Paul, B., Dixit, G., Murali, T.S., Satyamoorthy, K., 2019. Genome-based taxonomic classification. Genome 62, 45-52.

Perez, T., Balcazar, J.L., Ruiz-Zarzuela, I., Halaihel, N., Vendrell, D., de Blas, I., Muzquiz, J.L., 2010. Host-microbiota interactions within the fish intestinal ecosystem. Mucosal Immunology 3, 355-360.

Prins, R.A., de Vrij, W., Gottschal, J.C., Hansen, T.A., 1990. Adaptation of microorganisms to extreme environments. FEMS Microbiology Reviews 6, 103-104.

Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., Korobeynikov, A., 2020. Using spades de novo assembler. Current Protocols in Bioinformatics 70, e102.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K.S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D.R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J.-M., Hansen, T., Le Paslier, D., Linneberg, A., Nielsen, H.B., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Li, S., Jian, M., Zhou, Y., Li, Y., Zhang, X., Li, S., Qin, N., Yang, H., Wang, J., Brunak, S., Doré, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill,

J., Weissenbach, J., Meta, H.I.T.C., Bork, P., Ehrlich, S.D., Wang, J., 2010. A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464, 59-65.

Raja, H.A., Miller, A.N., Pearce, C.J., Oberlies, N.H., 2017. Fungal identification using molecular tools: a primer for the natural products research community. Journal of Natural Products 80, 756-770.

Ramírez, C., Coronado, J., Silva, A., Romero, J., 2018. Cetobacterium is a major component of the microbiome of giant amazonian fish (*Arapaima gigas*) in ecuador. Animals : An Open Access Journal From MDPI 8, 189.

Rawls, J.F., Samuel, B.S., Gordon, J.I., 2004. Gnotobiotic zebrafish reveal evolutionarily conserved responses to the gut microbiota. Proceedings of the National Academy of Science of the United States of America 101, 4596-4601.

Read, M.N., Holmes, A.J., 2017. Towards an integrative understanding of diet–host–gut microbiome interactions. Frontiers in Immunology 8.

Richter, M., Rosselló-Móra, R., 2009. Shifting the genomic gold standard for the prokaryotic species definition. Proceedings of the National Academy of Sciences 106, 19126-19131.

Robinson, C.J., Bohannan, B.J.M., Young, V.B., 2010. From structure to function: The ecology of host-associated microbial communities. American Society for Microbiology 74, 453-476.

Rodriguez-Melendez, R., Zempleni, J., 2003. Regulation of gene expression by biotin. The Journal of Nutritional Biochemistry 14, 680-690.

Said, H.M., 2008. Cell and molecular aspects of human intestinal biotin absorption. The Journal of Nutrition 139, 158-162.

Sandrini, S., Aldriwesh, M., Alruways, M., Freestone, P., 2015. Microbial endocrinology: host–bacteria communication within the gut microbiome. Journal of Endocrinology 225, R21-R34.

Schroeder, B.O., 2019. Fight them or feed them: how the intestinal mucus layer manages the gut microbiota. Gastroenterology Report 7, 3-12.

Sedghizadeh, P.P., Mahabady, S., Allen, C.M., 2017. Opportunistic oral infections. Dental Clinics of North America 61, 389-400.

Seppey, M., Manni, M., Zdobnov, E.M., 2019. BUSCO: Assessing genome assembly and annotation completeness. Methods in Molecular Biology 1962, 227-245.

Shahidi, F., Brown, J.A., 1998. Carotenoid pigments in seafoods and aquaculture. Critical Reveiws in Food Science 38, 1-67.

Shils, M.E., Shike, M., 2006. Modern nutrition in health and disease. Lippincott Williams & Wilkins.

Shokralla, S., Spall, J.L., Gibson, J.F., Hajibabaei, M., 2012. Next-generation sequencing technologies for environmental DNA research. Molecular Ecology 21, 1794-1805.

Silva, G.G., Dutilh, B.E., Matthews, T.D., Elkins, K., Schmieder, R., Dinsdale, E.A., Edwards, R.A., 2013. Combining de novo and reference-guided assembly with scaffold builder. Source Code for Biology and Medicine 8, 1-5.

Smit, A., Hubley, R., Green, P., 2019. 2013–2015. RepeatMasker Open-4.0. http://www.repeatmasker.org.

Soares, C.N., Teodoro, A.J., Lotsch, P.F., Granjeiro, J.M., Borojevic, R., 2015. Anticancer properties of carotenoids in prostate cancer. A review. Histology and Histopathology 30, 1143-1154.

Sommer, F., Bäckhed, F., 2013. The gut microbiota — masters of host development and physiology. Nature Reviews Microbiology 11, 227-238.

Sorbara, M.T., Pamer, E.G., 2019. Interbacterial mechanisms of colonization resistance and the strategies pathogens use to overcome them. Mucosal Immunology 12, 1-9.

Strandwitz, P., 2018. Neurotransmitter modulation by the gut microbiota. Brain Research 1693, 128-133.

Swanson, K.S., Gibson, G.R., Hutkins, R., Reimer, R.A., Reid, G., Verbeke, K., Scott, K.P., Holscher, H.D., Azad, M.B., Delzenne, N.M., Sanders, M.E., 2020. The International

Scientific Association for Probiotics and Prebiotics (ISAPP) consensus statement on the definition and scope of synbiotics. Nature Reviews Gastroenterology & Hepatology.

Takeda, A., Nyssen, O.P., Syed, A., Jansen, E., Bueno-de-Mesquita, B., Gallo, V., 2014. Vitamin A and carotenoids and the risk of parkinson's disease: a systematic review and meta-analysis. Neuroepidemiology 42, 25-38.

Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y.O., Borodovsky, M., 2008. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome Research 18, 1979-1990.

Tlaskalová-Hogenová, H., Štěpánková, R., Hudcovic, T., Tučková, L., Cukrowska, B., Lodinová-Žádnıková, R., Kozáková, H., Rossmann, P., Bártová, J., Sokol, D., 2004. Commensal bacteria (normal microflora), mucosal immunity and chronic inflammatory and autoimmune diseases. Immunology Letters 93, 97-108.

Vatsos, I., 2017. Standardizing the microbiota of fish used in research. Laboratory Animals 51, 353-364.

Wayne, L., Brenner, D., Colwell, R., Grimont, P., Kandler, O., Krichevsky, M., Moore, L., Moore, W., Murray, R., Stackebrandt, E., 1987. Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. International Journal of Systematic and Evolutionary Microbiology 37, 463-464.

Wehrli, W., Staehelin, M., 1971. Actions of the rifamycins. Bacteriological Reviews 35, 290-309.

Weiler, F., Schmitt, M.J., 2003. Zygocin, a secreted antifungal toxin of the yeast *Zygosaccharomyces bailii*, and its effect on sensitive fungal cells. FEMS Yeast Research 3, 69-76.

Whon, T.W., Kim, H.S., Bae, J.-W., 2018. Complete genome sequence of *Kocuria rhizophila* BT304, isolated from the small intestine of castrated beef cattle. Gut Pathogens 10, 42.

Yu, Y., Raka, F., Adeli, K., 2019. The role of the gut microbiota in lipid and lipoprotein metabolism. Journal of Clinical Medicine 8, 2227.

Zanello, G., Meurens, F., Berri, M., Salmon, H., 2009. *Saccharomyces boulardii* effects on gastrointestinal diseases. Current Issues in Molecular Biology 11, 47.

Zerbino, D.R., Birney, E., 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Research 18, 821-829.

Zhang, J., Chiodini, R., Badr, A., Zhang, G., 2011. The impact of next-generation sequencing on genomics. Journal of Genetics and Genomics 38, 95-109.

Zhang, L.S., Davies, S.S., 2016. Microbial metabolism of dietary components to bioactive metabolites: opportunities for new therapeutic interventions. Genome Medicine 8.

Zhao, Y., Guo, L., Xia, Y., Zhuang, X., Chu, W., 2019. Isolation, identification of carotenoid-producing *Rhodotorula* sp. from marine environment and optimization for carotenoid production. Marine Drugs 17, 161.

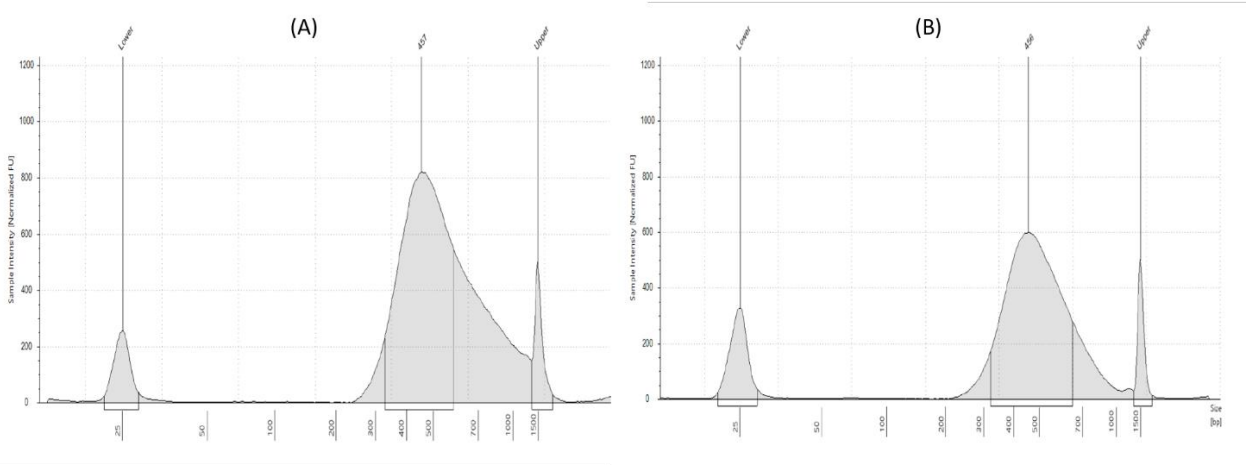
Zhou, X., Ren, L., Li, Y., Zhang, M., Yu, Y., Yu, J., 2010. The next-generation sequencing technology: a technology review and future perspective. Science China Life Sciences 53, 44-57.

Zilber-Rosenberg, I., Rosenberg, E., 2008. Role of microorganisms in the evolution of animals and plants: the hologenome theory of evolution. FEMS Microbiology Reviews 32, 723-735.

9. Supplementary material

Supplementary Table 1 *Concentration and absorbance ratios of the extracted DNA*

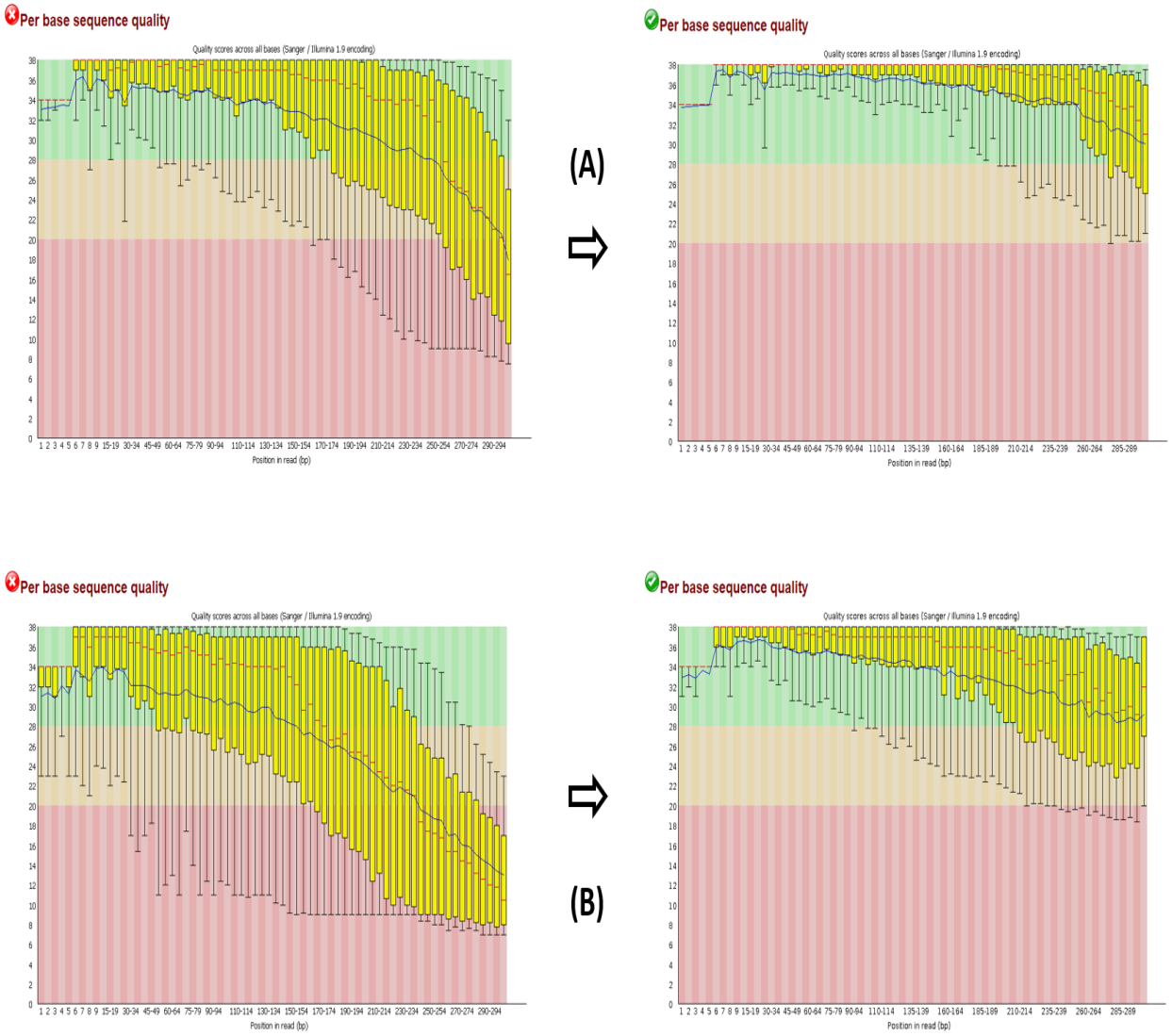
	Concentration	A 260	A 280	260/280	260/230
Bacterium DNA	39.3 µl/ml	1.254	0.663	1.89	1.94
Yeast DNA	8.97 µl/ml	1.152	0.573	2.01	1.46



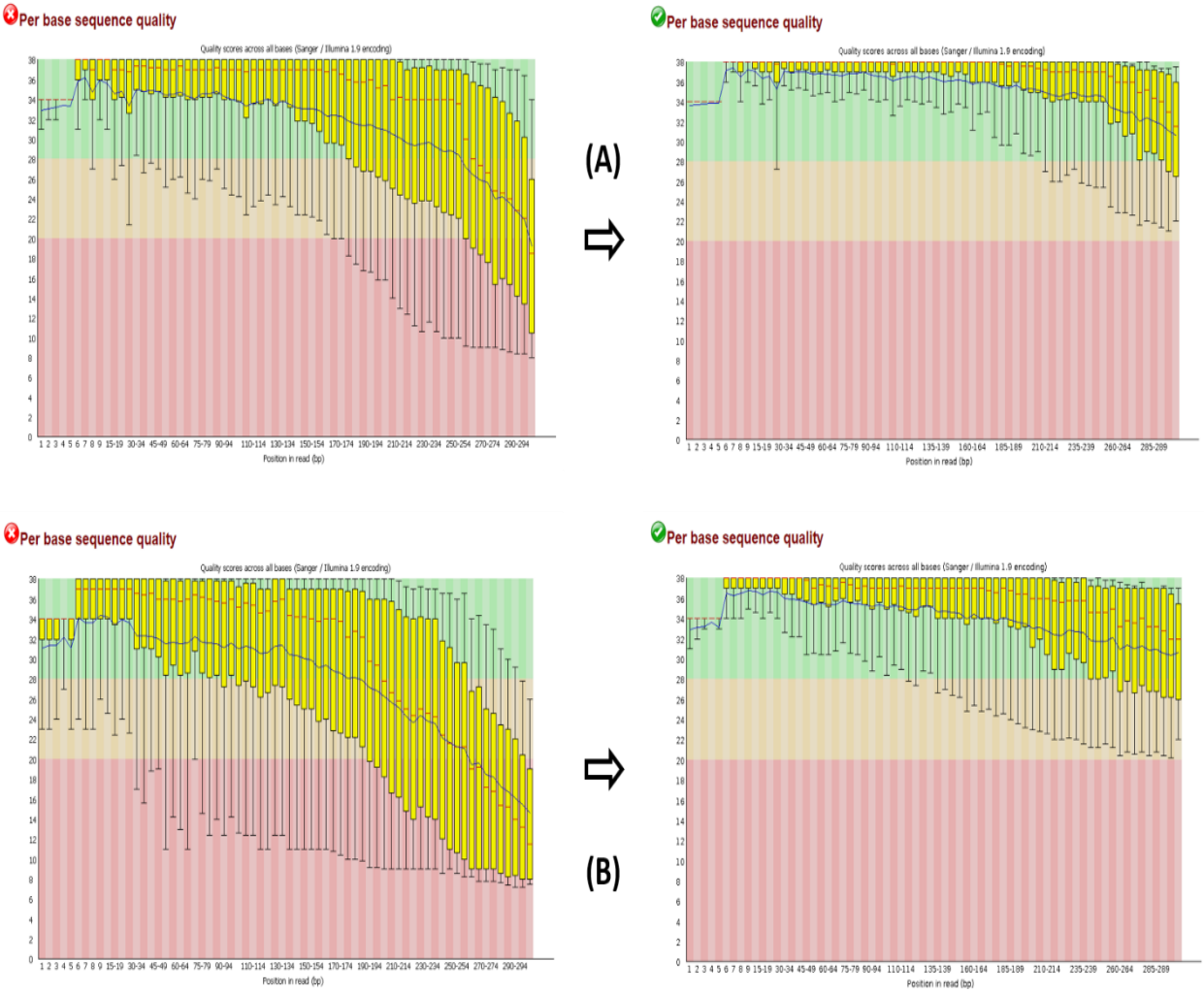
Supplementary Figure 1 *Electropherogram curve showing the size distribution of DNA fragments within the libraries of (A) bacterium and (B) yeast.*

Supplementary Table 2 *Concentration of the libraries before and after dilution*

	PRIMARY QUANTIFICATION	DILUTION	QUANTIFICATION AFTER DILUTION
Bac	33.23	2/3 x	22.84
Yeast	21.53	1 x	22.46



Supplementary Figure *2 Per base sequence quality of bacterium DNA reads. (A) Forward reads before (left) and after (right) quality trimming, (B) Reverse reads before (left) and after (right) trimming.*



Supplementary Figure 3 *Per base sequence quality of yeast DNA reads. (A) Forward reads before (left) and after (right) quality trimming, (B) Reverse reads before (left) and after (right) trimming.*

Supplementary Table 3 *Number of raw reads obtained from each sample with depth coverage and number of reads after trimming*

	No. of Raw reads	Depth x	No. of reads after trimming	Percent survived
Bacteria	614,258	69	300,949	49
Yeast	4,469,040	59	2,250,801	50

Supplementary Table 4 *List and accession number of 16S genes used in this study.*

Ref 16S	Bioproject	Accession No.
Kocuria rhizophila strain TA68	PRJNA33175	NR_026452
Kocuria carniphila strain CCM 132	PRJNA33175	NR_027193
Kocuria rhizophila DC2201	N/A	KM460939
Kocuria sp. BT304	N/A	KT368978
Kocuria rhizophila strain 14asp	N/A	KF875448
Kocuria rhizophila strain D2	N/A	MH005095
Kocuria rhizophila strain NCTC8340	N/A	Extracted from WG
Kocuria rhizophila FDAARGOS_302	N/A	Extracted from WG
Kocuria varians strain G33	PRJNA33175	NR_029297
Kocuria gwangalliensis strain SJ2	PRJNA33175	NR_116266
Kocuria subflava strain YIM 13062	PRJNA33175	NR_144586
Kocuria salsicia strain 104	PRJNA33175	NR_117299
Kocuria atrinae strain P30	PRJNA33175	NR_116744
Kocuria marina strain KMM 3905	PRJNA33175	NR_025723
Kocuria rosea strain DSM 20447	PRJNA33175	NR_044871
Kocuria polaris strain CMS 76	PRJNA33175	NR_028924
Kocuria halotolerans strain YIM 90716	PRJNA33175	NR_044025
Kocuria arsenatis strain CM1E1	PRJNA33175	NR_148610
Citricoccus nitrophenolicus strain PNP1	PRJNA33175	NR_117546
Micrococcus terreus strain V3M1	PRJNA33175	NR_116649
Paenarthrobacter nicotinovorans strain DSM 420	PRJNA33175	NR_026194
Rothia kristinae strain DSM 20032**	PRJNA33175	NR_026199
Kytococcus sedentarius strain DSM 20547**	PRJNA33175	NR_074714

**Outgroups

Supplementary Table 5 *List and accession number of 18S genes used in this study.*

Ref 18S	Bioproject	Accession no.
Rhodotorula mucilaginosa NCYC 63	PRJNA39195	NG_065157
Rhodotorula mucilaginosa strain MDU-02	N/A	KT000655
Rhodotorula mucilaginosa strain LFR-D	N/A	MH644858
Rhodotorula mucilaginosa strain ZB-H4	N/A	FJ538169
Rhodotorula mucilaginosa strain ZB-F12	N/A	FJ538168
Rhodotorula mucilaginosa strain ZB-E1	N/A	FJ538166
Rhodotorula glutinis CBS 20	PRJNA39195	NG_062726
Sporidiobolus salmonicolor JCM 1841	PRJNA39195	NG_063452
Rhodosporidiobolus poonsookiae JCM 10207	PRJNA39195	NG_062132
Rhodosporidiobolus ruineniae JCM 1839	PRJNA39195	NG_062129
Rhodosporidiobolus nylandii JCM	PRJNA39195	NG_060980
Sporobolomyces blumeae JCM 10212	PRJNA39195	NG_063456
Colacogloea falcata JCM 6838	PRJNA39195	NG_065485
Rhodotorula dairenensis CBS 4406	PRJNA39195	NG_063019
Leucosporidium muscorum CBS 6921	PRJNA39195	NG_062181
Rhodotorula alborubescens JCM 5352	PRJNA39195	NG_063540
Rhodotorula evergladensis CBS 10880	PRJNA39195	NG_063017
Ustilentyloma graminis CBS 6403	PRJNA39195	NG_062670
Rhodotorula taiwanensis CBS 11729	PRJNA39195	NG_063018
Rhodotorula pacifica CBS 10070	PRJNA39195	NG_063016
Rhodosporidiobolus colostri CBS 348	PRJNA39195	NG_062179
Filobasidium uniguttulatum JCM 3685**	PRJNA39195	NG_063470
Mrakia aquatica JCM 1775**	PRJNA39195	NG_063458

****Outgroup**

Supplementary Table 6 *List and accession number of ITS genes used in this study.*

Ref ITS	Bioproject	Accession No.
Rhodotorula mucilaginosa CBS 316	PRJNA177353	NR_073296
Rhodotorula mucilaginosa strain MT	N/A	AF128797
Rhodotorula mucilaginosa strain PY 32	N/A	KX525688
Rhodotorula mucilaginosa strain SY 18	N/A	KX525685
Rhodotorula mucilaginosa strain CPO 7.005	N/A	KU688203
Rhodotorula mucilaginosa strain WM 11.927	N/A	KP132588
Rhodotorula alborubescens JCM 5352	PRJNA177353	NR_153197
Rhodotorula evergladensis CBS 10880	PRJNA177353	NR_137709
Rhodotorula taiwanensis CBS 11729	PRJNA177353	NR_157462
Rhodotorula sphaerocarpa CBS 5939	PRJNA177353	NR_073269
Rhodotorula paludigena CBS 6566	PRJNA177353	NR_073265
Rhodotorula kratochvilovae CBS 7436	PRJNA177353	NR_073282
Rhodotorula diobovata CBS 6085	PRJNA177353	NR_073271
Rhodosporidiobolus ruineniae CBS 5001	PRJNA177353	NR_155707
Rhodosporidiobolus microsporus CBS 7041	PRJNA177353	NR_073290
Rhodosporidiobolus colostri CBS 348	PRJNA177353	NR_155730
Sporobolomyces roseus CBS 486	PRJNA177353	NR_155845
Sporobolomyces beijingensis CGMCC 2.2365	PRJNA177353	NR_137663
Leucosporidium creatinivorum CBS 8620	PRJNA177353	NR_073329
Leucosporidium fellii CBS 7287	PRJNA177353	NR_073276
Heitmania castanopsis CBS 14750	PRJNA177353	NR_160333
Leucosporidium scottii CBS 5930	PRJNA177353	NR_073267
Vonarxula javanica CBS 5236	PRJNA177353	NR_111079
Tausonia pamirica CBS 8428**	PRJNA177353	NR_154490

****Outgroup**

Supplementary Table 7 *List and accession number of bacterial reference genomes used in this study*

Reference genome	Bioproject	Assembly accession no.
Kocuria rhizophila DC2201	PRJDA27833	GCA_000010285.1
Kocuria rhizophila strain NCTC8340	PRJEB6403	GCA_900637835.1
Kocuria rhizophila strain FDAARGOS_302	PRJNA231221	GCA_002208685.2
Kocuria sp. BT304	PRJNA475186	GCA_003290245.1
Kocuria rhizophila strain G2	PRJEB9947	GCA_001499775.1
Kocuria rhizophila DC2201 strain DE0228	PRJNA543692	GCA_007677595.1
Kocuria rhizophila P7-4	PRJNA66631	GCA_000214115.2
Kocuria rhizophila strain D2	PRJNA428934	GCA_002879775.1
Kocuria rhizophila strain 14ASP	PRJNA286912	GCA_001038535.1

Supplementary Table 8 *List and accession number of yeast reference genome used in this study*

Reference genome	Bioproject	Assembly accession no.
Rhodotorula graminis WP1*	PRJNA342700	GCA_001329695.1
Rhodotorula. sp. CCFEE 5036*	PRJNA342238	GCA_005059875.1
Rhodotorula. sp. FNED7-22	PRJNA354502	GCA_001914285.1
Rhodotorula. sp. JG-1b*	PRJNA195770	GCA_001541205.1
Rhodotorula. sp. ZM1	PRJNA486254	GCA_009806315.1
Rhodotorula taiwanensis ASM292249v1*	PRJNA352283	GCA_002922495.1
Rhodotorula mucilaginosa IIPL32	PRJNA387690	GCA_002806785.1
Rhodotorula mucilaginosa C2.5t1	PRJNA270792	GCA_000931965.1
Rhodotorula mucilaginosa JGTA-S1-a*	PRJNA393004	GCA_003055205.1

*Genomes used for comparative genomics