Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Natural Sciences
Department of Biotechnology and Food Science

Amir Hamza

Constructing a Hybrid *De novo* Genome Assembly of *Aurantiochytrium* sp. T66 using MinION and Illumina reads to Improve the Quality of its Annotation

Master's thesis in MBIOT5
Supervisor: Helga Ertesvåg
Co-supervisor: Tonje Marita Bjerkan Heggeset
June 2021

NTNU
Norwegian University of
Science and Technology

Amir Hamza

Constructing a Hybrid *De novo* Genome Assembly of *Aurantiochytrium* sp. T66 using MinION and Illumina reads to Improve the Quality of its Annotation

Master's thesis in MBIOT5
Supervisor: Helga Ertesvåg
Co-supervisor: Tonje Marita Bjerkan Heggeset
June 2021

Norwegian University of Science and Technology
Faculty of Natural Sciences
Department of Biotechnology and Food Science

NTNU
Norwegian University of
Science and Technology

Preface

This master thesis is part of Auromega project. The project is collaboration between Norwegian University of Science and Technology (NTNU), and SINTEF. It was carried out at NTNU Department of Biotechnology and Food science and at SINTEF lab facilities located at Kjemiblokk 3 in Gløshaugen.

I am grateful to my supervisor Professor Helga Ertesvåg for the help all along, and also to my co-supervisor Tonje Marita Bjerkan Heggeset. I am also grateful to Snorre Sulheim and Tone Haugen for all the help during the thesis.

Abstract

Thraustochytrids are unicellular marine heterotrophs which have attention due to their ability to accumulate high amount of important nutraceuticals, in particular polyunsaturated fatty acids (PUFA). However, their biotechnological potential has not been fully utilized as the metabolism of these organisms have not been completely unraveled. One of the ways to understand metabolism of an organism to sequence and annotate its genome. *Aurantiochytrium* sp. T66 is one of the strains of thraustochytrids the genome of which has been sequenced and annotated previously. However, the genome assembly constructed for annotation had numerous gaps. Therefore, there was possibility that some genes might have missed the annotation process. In this study, a *de novo* hybrid genome assembly of *Aurantiochytrium* sp. T66 was constructed using MinION's long reads and Illumina's shotgun reads. The purpose was to build an assembly of better quality and annotate genes which might have missed the annotation process previously.

With automatic and manual functional annotation, 317 genes were identified which had not been annotated previously. PUFA synthase subunit A and C, phosphopantetheinyl transferase, and glutamine synthase are genes important for PUFA synthesis. In the previous annotation, only partial or fragmented sequences of these genes were identified. In addition to 317 genes, complete sequences of these four genes were also identified. ATP-citrate lyase, which is also an important gene for lipid synthesis and could not be detected in the published assembly of *Aurantiochytrium* sp. T66, could not be detected in this study either, suggesting that *Aurantiochytrium* sp. T66 might have some alternate gene(s) that does the task of ATP-citrate lyase, which is to prove to acetyl-CoA for lipid synthesis. Furthermore, mitochondrial genome of *Aurantiochytrium* sp. T66 was also identified and annotated.

The approach of constructing a hybrid assembly using MinION's long reads and Illumina's shotgun reads to improve the genome assembly and annotation of *Aurantiochytrium* sp. T66 has proven to be effective. However, there is still room for improvement of the assembly. It is suggested that further polishing the quality of the assembly might further help improve the annotation of *Aurantiochytrium* sp. T66.

Table of Contents

1. Introduction

1.1. Thraustochytrids

Thraustochytrids are marine unicellular protists. Nutritionally, they are obligate saprotroph, i.e., they feed on non-living organic matter. They are found all over the world feeding on detritus, from 1000m to 3000m deep sea zone to mangrove forests, playing an important role in the decomposition of dead matter(Raghukumar and Raghukumar, 1999; Raghukumar, 2002). Previously, there have been confusions regarding their taxonomical classification, and have been conventionally categorized as algae. At present, they are classified as belonging to the kingdom of stramenophiles in the phylum heterokonta and in the order t.(Morabito et al., 2019). They belong to the order of thraustochytriales. Taxonomical classification of one of the ten family members of thraustochytrids, *Aurantiochytrium,* is given in Figure 1.

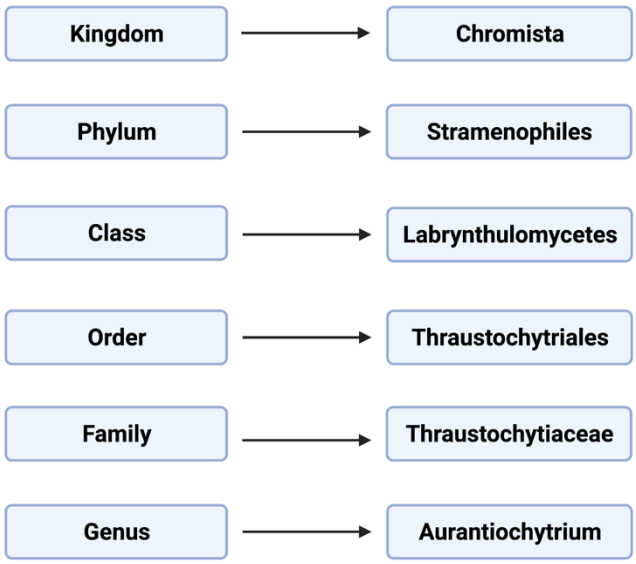


Figure 1. ***Taxonomic Classification of Aurantiochytrium.*** *(Created with BioRender.com).*

The main attractive attribute of thraustochytrids has been their ability to accumulate high lipid content, in particular omega (ω)-3 fatty acids. ω-3 fatty acids are polyunsaturated fatty acids characterized by a double bond after the third carbon in the carbon chain. They are essential for brain development and function, which is also the reason they are present in the humans'

breast milk (Juber et al., 2017). Thraustochytrids have been reported to accumulate lipids up to 50% of their biomass, with docosahexaenoic acid (DHA), a type of ω-3 fatty acid, contributing to 50% of the total lipid content (Bajpai et al., 1991). Other than ω-3 fatty acids, thraustochytrids have also shown potential for the mass production of squalene and carotenoids, two monetarily profitable nutraceuticals in pharmaceutical and cosmetic industries. (Aki et al., 2003; Park et al., 2018; Patel et al., 2019)

Though studies demonstrating the biotechnological potential of thraustochytrid have been surfacing since 1990s (Ratledge, 1993), the commercialization of the potential applications have not been at the same pace. Currently, the production of ω-3 fatty acids has been commercialized, the company Dutch State Mines (DSM) being the leader in the production of DHA oil from thraustochytrids, however, it makes very small percentage of the overall market for ω-3 fatty acids production (Barclay et al., 2010; Ratledge, 2012)**.** The primary reason for not being able to fully tap the biotechnological potential of thraustochytrids has been the insufficient knowledge about the thraustochytrids. For instance, complete lipid synthesis pathway in thraustochytrids has not been unraveled. (Heggeset et al., 2019; Aasen et al., 2016).

One of the ways to unravel metabolic pathways, and so to exploit the biotechnological potential of an organism, is to sequence and annotate its genome, and search for the relevant genes. *Aurantiochytrium* sp T66 is one of the many strains of thraustochytrids which have shown potential for the high production of ω-3 fatty acids. Its genome has been sequenced and annotated previously by Heggeset et al., (2019). Previously, the sequencing performed was with Next Generation Sequencing (NGS) technologies, Illumina HiSeq and Roche 454 FLX++. As would be explained later, the drawback of most current NGS technology platforms is that the reads generated, though very accurate, are very short. That becomes the reason for various problems which result in inaccuracies in the genome annotation. In the subsequent sub-sections, technologies for genome sequencing, and the process of genome assembly construction and its annotation would be explained briefly. Then, the process of lipid synthesis in thraustochytrids, which is the most studied pathway in thraustochytrids, would be described that will help elaborate the current state of genome annotation of *Aurantiochytrium* sp T66. It would be then explained what approach was taken in this study and how it could possibly help improve the genome annotation of *Aurantiochytrium* sp T66.

1.1. Genome Sequencing technologies

1.1.1. Next-generation sequencing (NGS)

Genome sequencing, as is evident from the term itself, is decoding of the entire genome of the organism. Genome sequencing is dominated today by next-generation sequencing (NGS) technologies. NGS technologies are characterized by their ability to perform sequencing in massively parallel manner, and is therefore also referred to as high-through put sequencing (Churko et al., 2013). There are various NGS platforms in existence today. Different NGS platforms differ on the technical details of sequencing, for instance, on the method of DNA sample preparation, or/and the method of sequencing, and each method has its own advantages and disadvantages. However, most NGS platforms have a common theme (Shendure and Ji, 2008)**.** The DNA sample is fragmented and amplified. Then, thousands of copies of identical single-stranded DNA fragments are anchored at one place. The complementary strands of these fragments are synthesized such that after incorporation of each nucleotide, the synthesis stops. The identity of the incorporated nucleotide is determined. The synthesis of the complementary strand is reinitiated, and the process is repeated until the entire fragment is sequenced (Shendure and Ji, 2008).

The largest share in the market for genome sequencing is of NGS technologies developed by Illumina, Inc (Goodwin et al., 2016; Giani et al., 2019). To explain the mechanism briefly, the sample DNA is sheared. The fragments are ligated with adapters. Using these adapters, the fragments are ligated to a slide, and then amplified, generating hundreds of identical copies of each DNA fragment. DNA fragments are then flooded with modified deoxyribonucleotide triphosphates (dNTPs). These dNTPs are fluorescently labelled, which also act as synthesis terminators. The unincorporated dNTPs are washed away. The incorporated dNTP, which is now actually deoxyribonucleotide monophosphate (dNMP), is excited with a light source, and the light emitted by the dNMP is used to determine the identity of dNMP (Metzker, 2010). Since many copies of a DNA fragment are sequenced simultaneously, the incorporated dNTPs in all the copies of a DNA fragment are used to draw consensus to determine the dNTP incorporated. The fluorescent label, which is also chain-terminating group, is cleaved, so that the cycle can be repeated. In this manner, the entire fragment is determined. Since the chain-terminating characteristic of the process is reversible, and the process is repeated to sequence a DNA fragment, the process is therefore called cyclic reversible termination method for sequencing (Metzker, 2010).

1.1.1.1. Limitations of NGS

The ability of NGS technologies to perform sequencing at such large scale at more than 99 % accuracy has massively advanced the studies of genomics. However, NGS technologies have limitations which impede the accomplishment of certain objectives in genomic studies. One of the limitations present in almost all NGS platforms is that the sequenced DNA fragments (reads) are very short, with the maximum length of reads being 1 kb (Bleidorn, 2016). As it would become clearer in the section 1.2.3, the genome assembly generated using short reads tends to be very fragmented due to the inability to resolve repetitive regions in the genome, which can negatively interfere in the process of genome annotation (Dijk et al., 2018).

Most NGS platforms require PCR amplification of the DNA sample to perform sequencing (Quail et al., 2012). GC-rich regions in the genome tend to be more thermostable, which makes them more resistant to annealing of DNA strands required for PCR amplification. Such regions, since, are more reluctant to amplification, they are relatively less covered in the sequencing process. This contributes to what is called GC-biasedness ('contributes to' as PCR is not the only factor responsible for GC-bias). Consequently, this GC-bias contributes to uneven genome coverage in the sequencing, further negatively contributing to the contiguity of the genome assembly (Teytelman et al., 2009; Aird et al., 2011).

Another problem in many NGS platforms is what is called dephasing (Kircher and Kelso, 2010). The term 'dephasing' can be explained using Illumina sequencing described above. As mentioned in earlier, hundreds of copies of a DNA fragment are sequenced simultaneously, and it is the consensus between all the identical fragments that is used to determine the identity of each incorporated dNTP. However, it often happens that dNTP is not incorporated in all identical copies of a DNA fragment in one cycle, but the process goes normal in the next cycle. It might happen in many cycles. The result is that the consensus between identical DNA fragments for each incorporated dNTP decreases, resulting in the errors in the sequencing (Metzker, 2010).

The limitations of short-reads from NGS platforms can be partly addressed by a sequencing technique called mate-pair sequencing (Miyamoto et al., 2014). In mate-pair sequencing, DNA is sheared into long fragments, in contrast to short fragments as is the case in shotgun

sequencing. The fragment size can range from 3-40 kb, depending on the platform (Van Nieuwerburgh et al., 2012; Wu et al., 2012; Berglund et al., 2011). The fragments are end-labeled with biotin and then circularized. The result of circularization is that two biotin-labeled ends of a DNA fragment, which are separated by a known number of nucleotides, become adjacent to each other (Gao and Smith, 2015). The circularized fragments are then chopped into 300-500 bp fragments. The biotin-labelled fragments are purified using streptavidin-coated magnetic beads. These purified fragments are then paired-end sequenced for 200-300 kb. Paired-end sequencing means that a DNA fragment is sequenced from both ends, as opposed to from only one end. Result is that two reads are generated, which span by known length of DNA, or as referred to in literature, by known insert size. (Hampton et al., 2017).

Mate-pair sequencing can help lessen the problem of resolving the repeat elements, and can therefore, increase the contiguity of the genome assembly. However, the fragmentation of genome assembly because of GC-biasedness, or errors because of dephasing still persist. Furthermore, with mate-pair sequencing, it is difficult to resolve homopolymer regions, which are regions in genome which have consecutive stretches of identical nucleotides (Berglund et al., 2011).

1.1.2. Third-generation sequencing (TGS)

Shortcomings of NGS platforms can be made less severe with what is called third-generation sequencing (TGS) technologies. TGS technology is characterized by its ability to generate long reads. It differs from NGS by the basic approach it adopts for sequencing. TGS technology, instead of sequencing thousands of identical templates in a parallel manner, directly sequences a single polynucleotide (Schadt et al., 2010). Nanopore sequencing is one of the TGS technologies based on nanopores. To briefly explain the primary mechanism of nanopore sequencing, nanopores, which are either engineered biological or entirely synthetic pores, are embedded on an electrically insulated membrane (Niederweis et al., 1999; Manrao et al., 2012). Electric field is applied across nanopore, and the current flow through the nanopore is measured thousand times per second by a sensor (Lu et al., 2016). Polynucleotide passes through the nanopore by electrophoresis, i.e., polynucleotide being negatively charged migrate towards the positive end of the electric field. As the polynucleotide passes through the nanopore, individual nucleotides cause disruption in the current flow, making a characteristic disruption pattern,

which is translated into the sequence of polynucleotides (Deamer et al., 2016; Lu et al., 2016; Branton et al., 2008).

In 2015, Oxford Nanopore Technology (ONT) launched the first commercially available sequencing device, called MinION, based on nanopore sequencing technology. Following sub-chapters would entail the basic mechanism of MinION sequencing and its benefits as compare to NGS.

1.1.2.1. MinION Sequencing

ONT's MinION is a hand-sized sequencing device, weighing around 90g (Figure 2). It can be connected to a computer with a standard USB-3 port. The device has a flow cell with 512 channels or sensors connected to a data processing unit called application-specific integrated circuit (ASIC) (Jain et al., 2016; Mikheyev and Tin, 2014). Each channel has four individual nanopores, adding up to 2048 nanopores in total in a flow cell. Each nanopore is embedded in a separate stable membrane immersed in an ionic solution. At any one time, only one nanopore from each channel can take part in DNA sequencing. That means that at any given time, only 512 nanopores at max are active (Lu et al., 2016; Jain et al., 2016).

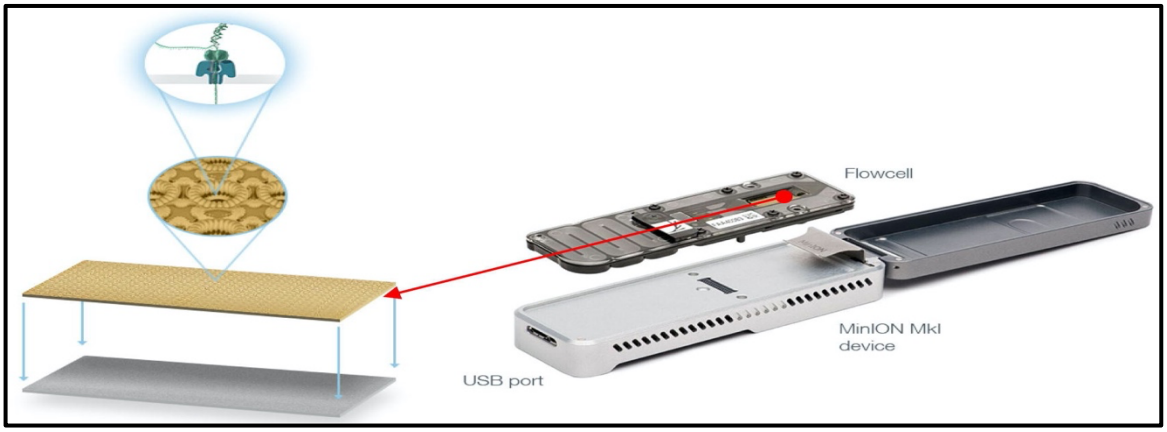


*Figure 2: **MinION sequener**. (Lu et al., 2016)*

The latest flow cell is termed as R10 flow cell. In R10 flow cell, the nanopore used is engineered version of transport channel CsgG of *Escherichia coli* (Van der Verren et al., 2020). The transport protein CsgG is found in many species of gram-negative bacteria. CsgG of *E.coli* is 262 amino acid long, 36-stranded β-barrel transmembrane protein (Goyal et al., 2014;

Loferer et al., 1997). The natural function of CsgG protein is its involvement in the secretion of curli protein, an extracellular fibrous protein involved in the formation of bacterial biofilm (Barnhart and Chapman, 2006). For ONT's nanopore sequencing, the CsgG protein of *E.coli* has been engineered to make it suitable for the sequencing purpose (Carter and Hussain, 2017).

Before sequencing, the DNA sample is prepared for sequencing. This process of preparing the DNA sample before sequencing is termed as ''library preparation" in nucleic acid sequencing terminology. In MinION sequencing, library preparation protocols differ depending on the aim of sequencing. For DNA sequencing, the basic steps are as follows: DNA is fragmented; ends of DNA fragments are repaired; here, the term repaired means repairing the damage such as oxidization of bases, deamination of cytosine, and creating blunt ends in the fragmented DNA; then, adenine bases are added at the ends of DNA fragments, a process which is termed as dA-tailing; two types of adapters are ligated at the ends of the DNA fragments: Y-adapter, which is ligated to the blunted 5'end of 'template' strand, and optional hairpin adapter, which is ligated to 3' end of the 'temple'; Y-adapter helps in the 'recognition' of the polynucleotide by the nanopore, whereas 'hairpin' adapter helps 2D sequencing (Lu et al., 2016).

Prepared DNA sample is loaded onto the sequencer. Sequencing starts with the recognition of the adapter-ligated polynucleotide by nanopore. As the nanopore recognizes the polynucleotide, it is unzipped by the motor protein such that only a single strand of the polynucleotide passes through the nanopore (Dijk et al., 2018). If the option with a hairpin adapter is utilized, the complementary strand is also passed through the nucleotide, and thus termed as 2D sequencing. If the hairpin adapter is not used, only a single strand is sequenced, and thus would be termed as 1D sequencing. The passing of nucleotides causes disruption in the current flow, making a pattern of disruption. This disruption pattern is transferred to ASIC for processing and is translated/base called into nucleotide sequence by an ONT-provided software called MinKNOW. With MinKNOW, the progress of sequencing can be visualized in real-time (Plesivkova et al., 2019).

1.1.2.2. Advantages of ONT's MinION sequencer in comparison to NGS platforms

There are numerous advantages of using MinION, some of which are attributed to their ability to generate long-reads, and some are attributed to the nanopore-sequencing technology itself. The length of reads from the MinION sequencer is significantly higher than the reads from

NGS technologies(Schadt et al., 2010). NGS platform-based sequencing, though has a high throughput and is very accurate, can only sequence a few hundred bases (Erlich et al., 2008). For instance, the read length with Illumina MiSeq, one of the NGS platforms of Illumina, is around 2 x 300 bp paired-end reads (Schirmer et al., 2016). On the other hand, it is not unusual to have read lengths of 100kbp in MinION sequencing. Long reads, as would be explained later, help resolve the problem of repeat elements and fragmentation in genome assembly. For instance, in one study, 36kb +MinION reads were used to resolve a 50kb gap in the human Xq24 reference genome (Jain et al., 2015). Since nanopore sequencing can generate long reads, full-length cDNA reads can be sequenced, providing more accuracy in RNA expression analysis (Bolisetty et al., 2015).

Another vital aspect of MinION is its independency to PCR amplification to perform sequencing. Most NGS platforms require PCR amplification of the DNA sample to perform sequencing (Quail et al., 2012). GC-rich regions in the genome tend to be more thermostable, which makes them more resistant to annealing of DNA strands required for PCR amplification. Such regions, since, are more reluctant to amplification, this contributes to what is called GC-bias (Aird et al., 2011). GC-bias contributes to uneven genome coverage (Teytelman et al., 2009). As PCR amplification is not a requirement for MinION, this lessens the GC-bias problem.

Benefits attributed to the technology itself include the ability to directly detect modification on the nucleotides, which is not possible with NGS technology-based sequencing platforms. (Wescoe et al., 2014; Laszlo et al., 2013). Time-effectiveness is another crucial factor which renders MinION superiority over NGS technology-based sequencing platforms. For instance, with HiSeq 3000, one of Illumina's NGS technology-based sequencer, a sequencing run can take about 4 days (Bleidorn, 2016). Though the default timing of MinION is 2 days, the majority of the data is generated within a day for most projects (Tyler et al., 2018). Furthermore, the portability provided by the MinION sequencer is first of its kind.

1.1.3. Genome assembly

Genome assembly construction is the process of reconstructing the genome of the organism from the sequenced reads (Foxman, 2012). The assembly can be constructed by aligning the sequenced reads to a reference genome of the organism, and thus the assembly would be termed

as reference-guided genome assembly. If reference genome of the organism is not available, for instance, in the case if the genome of the organism has never been sequenced before, then the genome has to be constructed *de novo* (i.e. from the beginning), thus the assembly would be termed as *de novo* genome assembly (Sutton, 2010)

Genome assembly construction is primarily a computational process. There are different programs available for the construction which are tailored according to: ***a)*** the type of assembly to be built, for instance, if the assembly is to be constructed *de novo* or with the help of a reference; ***b)*** the type of genome, for instance, if the genome is from prokaryotic or eukaryotic organism; ***c)*** the type of reads with which the assembly is to be constructed, for instance, if the reads are long or short. Different programs utilize different algorithms for constructing assembly. In principle, all programs rely on the overlapping sequences of the reads to construct longer sequence, and eventually assemble the longer sequences into genome (Kalyanaraman, 2011; 2010; Sutton, 2010).

Organizationally, a genome assembly is arranged into contigs and scaffolds. A contig is a contiguous sequence of DNA without any gaps, though it may have ambiguous/undetermined bases, which are represented by the letter 'N' in the sequence. It is constructed by merging series of overlapping reads, and the construction terminates as the overlapping reads for the contig finish. Contigs, in turn, are arranged into higher organizational order called scaffold (Choudhuri, 2014). A scaffold comprises of contigs with gaps of known length between them. The Lower the number of contigs and scaffolds in a genome assembly, the more contiguous, thus of better quality, the assembly is considered, and vice versa.

Contiguity of an assembly is highly dependent on the length of the reads the assembly is constructed with. Genomes tend to have repetitive sequences which can be several thousand base pairs long (Zhang et al., 2011; Tyson et al., 2018). In genome assembly construction using short reads, short reads corresponding to repetitive regions are identified as being the same sequence, making it difficult for algorithms of genome assembler softwares to resolve the repetitive region. (Baker, 2012). This results in decreased contiguity in genome assembly. Contiguity can be increased by constructing the assembly using long reads. Long reads, since, can 'read through' these repetitive sequences, or are long enough to be identified as unique reads, they can help resolve these repetitive regions of the genome (Miller et al., 2017).

Long-reads, though they can rescue from the problem of repetitive sequences and the coverage problem caused by GC-bias, tend to be error-prone to a considerable extent (Mikheyev and Tin, 2014). MinION's error rate has been reported up to 38% (Laver et al., 2015). Increasing the coverage can mitigate this drawback to an extent; however, the accuracy of reads is still not on par with the reads from NGS-platforms, such as Illumina's MiSeq or 454 pyrosequencers, that can provide accuracy of up to 99.4% (Dohm et al., 2008) . Resultantly, *de novo* genome assembly constructed from long reads-only would have a high number of infidelities in the sequence of the genome (de Lannoy, 2017). This would, consequently, have an impact on downstream genome analysis. For instance, in genome annotation, incorrect sequence might alter the biological information the sequence contains, leading to incorrect annotation.

A viable alternative to long read- or short reads-only assembly is to construct hybrid genome assembly. In hybrid assembly, genome is constructed from both long and short reads (Antipov et al., 2016). This approach allows to avail the benefits of long reads, as well as short reads (Miller et al., 2017). The respective benefits of using long reads and short reads antagonize each other's flaws, i.e., short reads rectify the inaccuracy of long reads, and long reads rectify the tendency of short reads to be more prone to gaps and uneven coverage.

1.1.4. Genome Annotation

Genome annotation is a process of identifying functional elements in the genome assembly and assigning those regions suitable biological information (Abril and Castellano, 2019). Functional elements can include protein-coding genes, Non-coding RNA genes, such as sequences coding for non-coding RNAs (e.g., transfer RNA (tRNA), small nuclear RNA, long non-coding RNA, ribosomal RNA (rRNA) etc.), and regulatory regions, such as enhancers and promotors (Solovyev et al., 2006; Humann et al., 2019).

The process of genome annotation comprises primarily of two operations: structural annotation and functional annotation. Structural annotation involves identifying probable coding regions in the genome. The procedure of predicting coding regions is, per se, primarily computational. It can be performed independent of any of previous knowledge by identifying possible start and end codons, thus is termed *ab initio* gene prediction (Saraswathy and Ramalingam, 2011). Another approach is to conduct homology-based gene prediction, in which annotation of a closely-related genome are used to annotate the genome under study (Saraswathy and

Ramalingam, 2011). Alternatively, help of experimental data, such as RNA-seq data or expressed sequence tags, can be sought. RNA-seq data can be used in two possible ways for structural annotation: by *de novo* transcriptome construction or genome-guided transcriptome construction (Chen et al., 2017; de Sá et al., 2018). Figure 3 summarizes genome annotation using *de novo* and genome-guided transcriptome assembly construction. In the former, RNA-seq reads are assembled independent of any reference, i.e., *de novo* transcriptome assembly is constructed; the assembled transcripts are aligned to reference genome; those regions of genome which have not been annotated before and are shown to contain information by aligned transcripts are excerpted for further scrutiny; in this way, novel transcript can be discovered. In the genome-guided transcriptome, RNA-seq reads are directly aligned to the genome, and are assembled into transcripts, instead of assembling the reads into transcriptome assembly independent of genome guide as is the case in the *de novo* transcriptome assembly approach. Possible coding regions in the aligned transcripts are then predicted (Yandell and Ence, 2012).

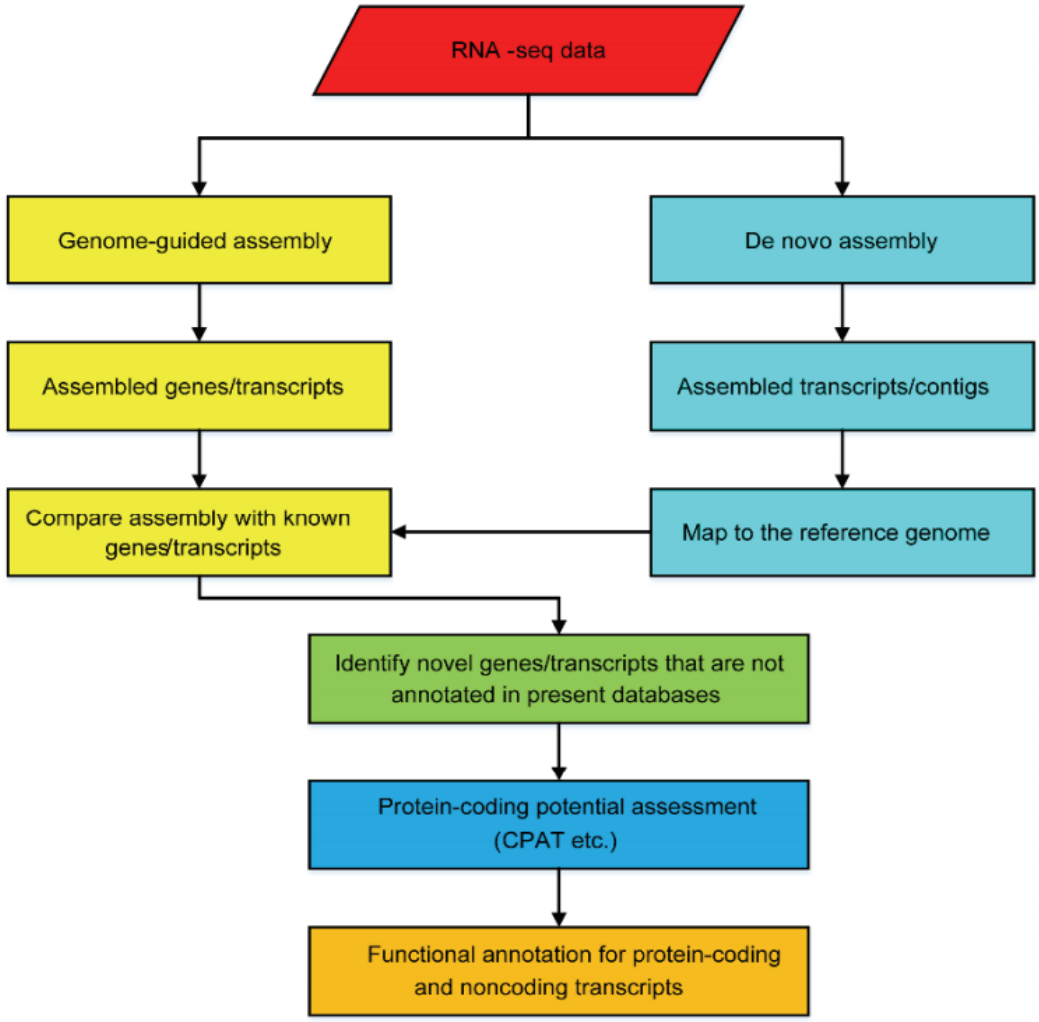


Figure 3. ***Summary of comparison between genome-guided and de novo transcriptome assembly for genome annotation*** *(Chen et al., 2017).*

Genome-guided transcriptome reconstruction approach for genome annotation is relatively more sensitive and accurate as compared to *de novo* transcriptome assembly approach; since the reads are directly aligned to the genome, low abundance reads can also be aligned to genome and can be subsequently assembled into transcript; in this manner, low-abundance transcripts can be discovered; this is difficult to do in *de novo* transcriptome assembly reconstruction as either the low coverage reads are filtered, or even if they are retained, it is difficult to assemble them into transcript because of low coverage**.** Furthermore, low coverage regions within the transcript can be filled using the reference genome (Chen et al., 2017).

Structural annotation is followed by functional annotation. In this process, predicted CDS are BLAST-searched against different biological databases. Biological databases can have nucleotide sequence data, protein sequence data, or both. Different databases differ on the degree of curation (Bhattacharyya, 2009). Therefore, predicted CDS are BLAST-searched against different databases. In this process, predicted CDS are BLAST-searched against different nucleotide and/or protein databases. On the basis of BLAST-results, CDS predicted is annotated with appropriate biological information. (The process to evaluate the results against the searched databases in this study is described in the section 3.4.3)

1.2. PUFA Synthesis in Thraustochytrids

As mentioned earlier, the main attractive attribute of thraustochytrids is their ability to accumulate high lipid content, in particular ω-3 fatty acids. To appreciate the PUFA synthesis and accumulation in thraustochytrids, it is important to illustrate what distinguishes oleaginous microorganisms from non-oleaginous organisms as the basic mechanism for fatty acid synthesis also exists in the latter, and also how fatty acid acids are synthesized in nature generally**.** Oleaginous organisms are those which have the ability to accumulate lipids at more than 20 % of their dry weight (Patel et al., 2020). This would help better comprehend PUFA synthesis in thraustochytrids. Ensuing sub-sections would entail a brief introduction of ω-3 fatty acids and their importance to humans. Then, lipid accumulation in oleaginous organisms would be explained, and at last, PUFA synthesis in thraustochytrids.

1.2.1. Omega (ω)-3 fatty acids and their relevance to humans

ω-3 fatty acids are polyunsaturated fatty acids. To describe their nomenclature briefly, the symbol 'ω' denotes the carbon at the methyl end or farthest from the carboxylic group. They are referred to as '**ω-3**' fatty acids because they have double bond after the carbon number three counting from the methyl end or **ω**-end of the polyunsaturated fatty acid chain. To give another example, ω-6 polyunsaturated fatty acids, another category of PUFAs, would have double bond after the carbon number six counting from the ω end. In literature, names of PUFAs are often written with the number of carbon atoms and double bonds in PUFAs. For instance, ω-3 fatty acid DHA would be referred to as 'DHA (22:6)', where the first numeral denotes the number of carbon atoms and the second denotes the number of double bonds**.** For the illustration of the nomenclature, structural formula of DHA is shown Figure 4.

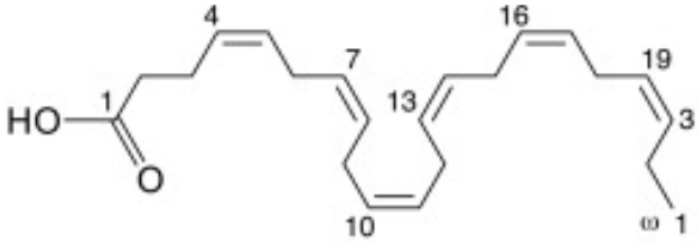


Figure 4. ***Structure of docosahexaenoic acid (DHA)*** *(López-Malo et al., 2020).*

ω-3 fatty acids, are an important dietary supplement. Increasing research is making clear the obvious benefits of ω-3 fatty acids. The most important and relevant ω-3 fatty acids to human health are α-linolenic acid (ALA), eicosapentaenoic acid (EPA), and DHA. ALA (18:3) is precursor to EPA (20:5) and DHA. Humans do have ability to synthesize EPA and DHA from ALA, however, the conversion is very inefficient (Gerster, 1998). Therefore, ω-3 fatty acids' requirement in humans has to be supplemented from dietary sources. ALA is commonly found in plant-based nutritional sources such as rapeseed oil and flax oil. Major sources for EPA (20:5) and DHA are fatty fish (71%), meat (20%), and poultry (6%) (Meyer et al., 2003).

Numerous studies on animal models suggest that DHA is an important factor in neuroprotection of central nervous system. For instance, in one such study, DHA and EPA supplementation was shown to improve performance in cognitive tests and elicit protection against neuroinflammation (Jiang et al., 2009; Labrousse et al., 2012). Studies on rodents have also shown that aged rodents tend to have lower level DHA compared to younger one (Afshordel

et al., 2015). Such findings have prompted scientists to research on whether DHA can be helpful in age-related studies, such as Alzheimer disease, or not. Lower DHA levels negatively impact the level of phosphatidylserine in neural cells**.** Phosphatidylserine is the most abundant phospholipid of inner cell membranes of nerve cells. It has shown to have some role in cognitive function of the central nervous system. For instance, in one study, when 494 elderly people were administrated with phosphatidylserine, it improved their cognitive performance without any side effects (Cenacchi et al., 1993). In another study, it showed memory improvement in non-demented elderly patients (Vakhapova et al., 2011).

1.2.2. Lipid Accumulation in oleaginous Organisms

In citric acid cycle - a sequence of chemical reactions which occurs in mitochondria that oxidizes carbohydrates, proteins and fats to generate energy - one of the intermediates is isocitrate**.** Isocitrate generated is metabolized by an enzyme called isocitrate dehydrogenase (IDH). In oleaginous organisms, IDH is dependent on the nucleotide adenosine monophosphate (AMP) to be functional, whereas IDH in non-oleaginous organisms is AMP-independent (Ratledge, 2004). When an oleaginous microorganism is grown in or encounters nitrogen-limited culture in its natural environment, it begins to utilize AMP as a source of nitrogen**.** The breakdown of AMP to use it as a nitrogen source is catalyzed by AMP deaminase.

$$AMP + H_2O \xrightarrow{AMP\ Deaminase} Inosine\ monophosphate + NH_3$$

This results in depletion of AMP, which makes IDH, which is AMP dependent, inactive. Subsequently, isocitrate accumulates in mitochondria, and equilibrates with its precursor, citrate. Citrate is then transported out into the cytosol in exchange for malate by citrate/malate translocase present in the mitochondrial membrane**.** In cytosol, citrate is broken down by ATP:Citrate lyase, which is found in almost all oleaginous microorganisms but is not found in non-oleaginous microorganisms, into acetyl Coenzyme A (acetyl-CoA) and oxaloacetate. (Ratledge, 2002).

$$ATP + Citrate + Coenzyme\ A \xleftrightarrow{ATP:Citrate\ lyase} ADP + Phosphate + Acetyl\ Coenzyme\ A$$

Acetyl-CoA, which is an essential building block for fatty acid synthesis, is then used for fatty acid synthesis. The accumulated lipids acts as energy reserve, and are utilized for various purposes, such as for the proliferation of the cells (Dellero et al., 2018). So, it is the presence of AMP-dependent isocitrate dehydrogenase in oleaginous organisms, which essentially allows the accumulation of isocitrate, which in turn becomes the source of acetyl-CoA production. Acetyl-CoA is then used as a precursor for fatty acid synthesis, which occurs primarily by two pathways in nature: fatty acid synthase (FAS) pathway and polyketide synthase (PKS) pathway.

1.2.3. FAS Pathway

FAS pathway for fatty acid synthesis exists in all lipid producing organism, independent of whether the organism is oleaginous or not. Basic mechanism for fatty acid synthesis through FAS pathway is shown in Figure 5. Its products include saturated fatty acids, as well as polyunsaturated fatty acids (PUFAs). This pathway involves fatty acid synthase (FAS). There are two types of FAS: type I and type II. FAS type I is a large multidomain protein, where each domain performs a specific function, whereas FAS type II is a system of monofunctional enzymes, where each enzyme performs a specific function (Sul and Smith, 2008; Rock, 2008). Despite the organizational difference of catalytic entities between FAS-I and FAS-II, the primary mechanism for fatty acid synthesis is the same**,** as the enzymes of FAS-II are homologous to the domains of FAS-. 1. At basic level, FAS systems comprises of an acyl carrier protein (ACP), and six catalytic entities (Semenkovich, 1997). The six catalytic entities are acetyl-CoA-ACP transacylase (AT), malonyl-CoA-ACP transacylase (MAT), β-ketoacyl-ACP synthase (KS), β-ketoacyl-ACP reductase (KR), 3-hydroxyacyl-ACP dehydratase (DH), enoyl-ACP reductase (ER), and ACP thioesterase (Semenkovich, 1997). The mechanism to produce fatty acids by FAS system is shown in Figure 5.

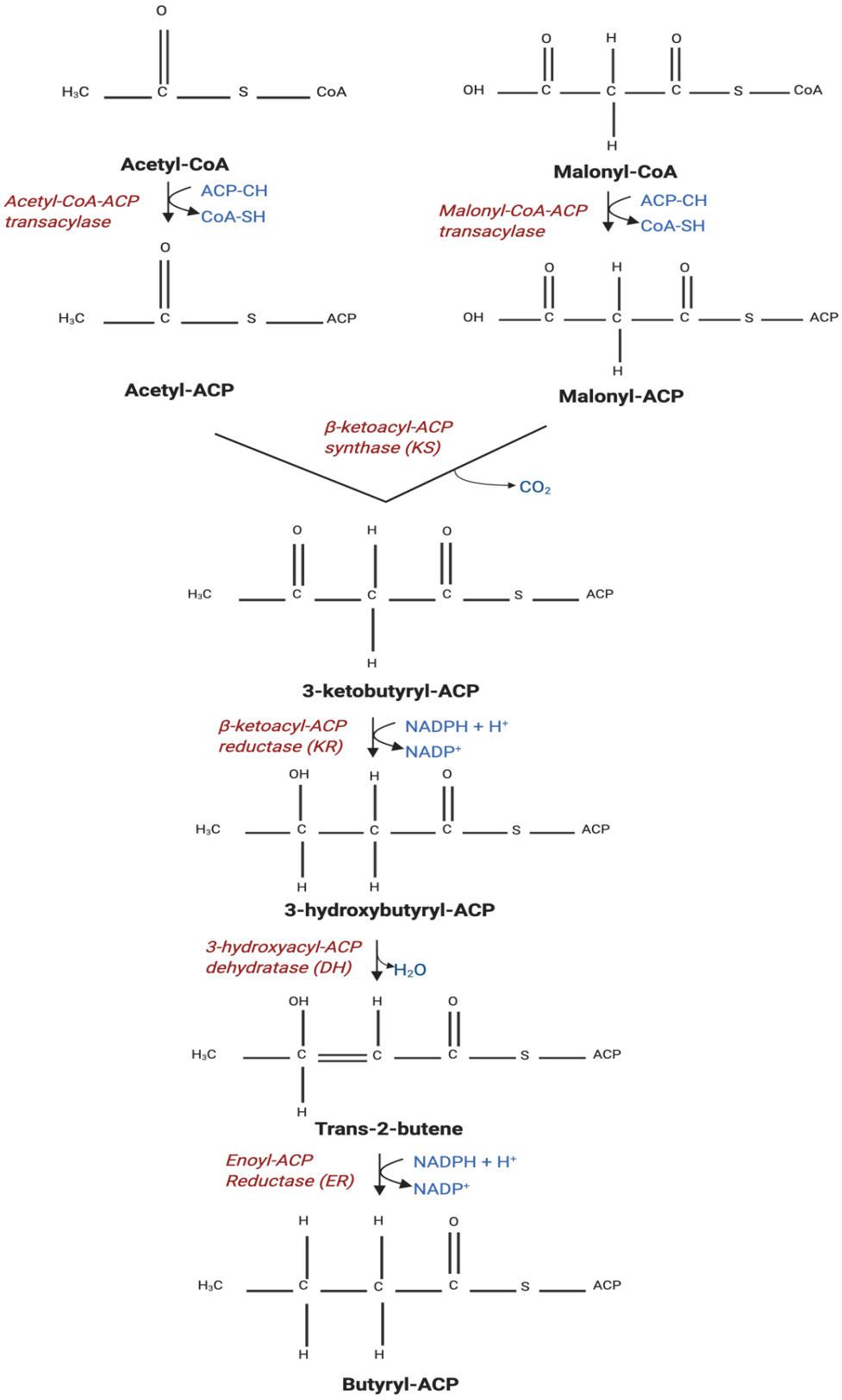


Figure 5. ***General mechanism of fatty acid synthesis by FAS system****. Created with BioRender.com*

With the mechanism shown in Figure 5, the FAS system can produce 14, 16- or 18-carbon long saturated fatty acids (Giordano et al., 2015). In order to synthesize PUFAs through FAS pathway, the saturated fatty acids released from FAS are subject to various elongases and desaturases to make PUFAs (Shanklin and Cahoon, 1998). Conventional sequence of desaturation and elongation to make ω-6 and ω-3 is shown in Figure 6.

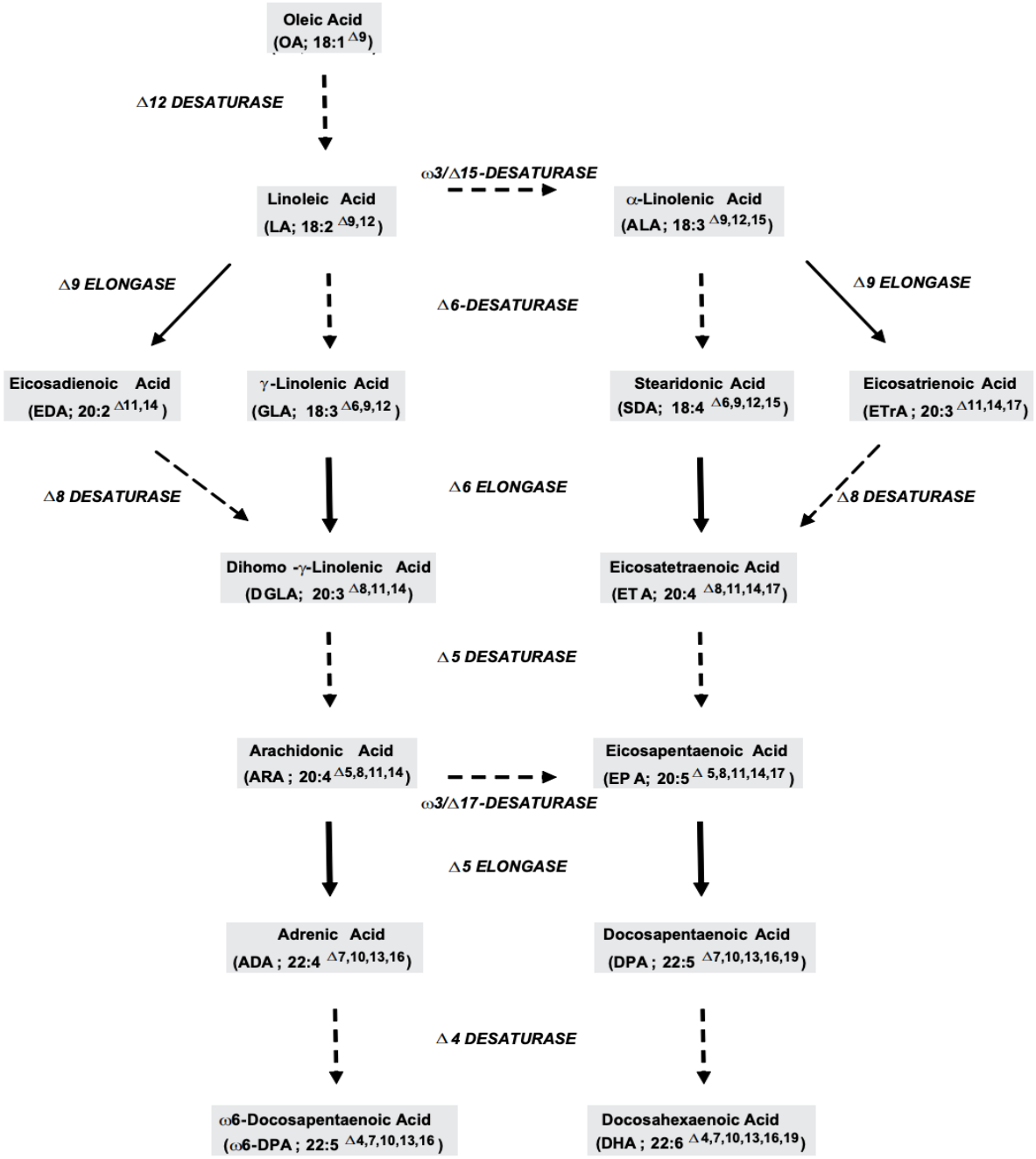


*Figure 6.:**Conventional pathway for PUFA synthesis**.(Vrinten et al., 2007). The nomenclature next to the abbreviations of the compounds shows composition of the compounds. The first numeral represents the number of carbon atoms, the second numeral represents the number of double bonds, and the numerals indicated next to the symbol delta (Δ) indicates the position of the double bonds.*

As is evident from Figure 6, depending on the type of PUFA, the set of desaturases and elongases required also differs. For instance, as shown in Figure 5, to synthesize of ω-6 PUFA, stearic acid (18:0) is acted by upon by Δ9 desaturase to make Oleic acid (18:1), which is followed by conversion into Linoleic acid (18:2) by Δ12 desaturase. Linoleic acid is then subject to Δ6, Δ5, Δ4 and desaturases, with each desaturation reaction flanked by elongase activity, to make DPA, which is an ω-6 fatty acid. Likewise, to make ω-3 fatty acid, Δ12 desaturation is followed by Δ15 desaturase action to produce alpha-linoleic acid. Alpha-linoleic acid is then passed through Δ-6, Δ5, and Δ4 desaturases, with intervening elongase activity, to produce DHA.

1.2.4. PKS Pathway

PKS pathway for PUFA synthesis exists in lower eukaryotes and prokaryotes (Napier, 2002). It comprises of PUFA synthases. PUFA synthase can be a single multi-subunit protein with various domains performing different tasks type I PKS), or it can be a complex of discrete monofunctional enzyme (type II PKS). PUFA synthases have enzymes or domains which are homologues of the catalytic units of the FAS system (Hopwood and Sherman, 1990; Hauvermale et al., 2006). As in the FAS system, the repetitive decarboxylative condensation reaction, which elongates the chain, also happens in the PKS pathway. However, primary difference lies in the last reduction step. In the PKS pathway, the last step is often omitted, resulting in the formation of PUFAs. Unlike the FAS pathway, there is no requirement of desaturases to produce PUFA in the PKS pathway. (Hopwood and Sherman, 1990).
or

1.2.5. PUFA synthesis in Thraustochytrids

It was first thought that thraustochytrids synthesize PUFAs the same way the other oleaginous microorganisms produce, i.e., by standard fatty acid synthase route. This observation probably came from the fact that all the components of standard fatty acid synthesis pathway were also present in thraustochytrids as well. However, in a study conducted by Metz et al., (2001), when *Schizochytrium* was exogenously supplied with ^{14}C-labeled 16:0, 18:1, or 18:3 fatty acids, which are precursor of very-long chain PUFAs if synthesized through FAS pathway, no radioactivity was detected in very-long chain PUFAs such as DHA or DPA. This meant that the organism does not use those putative precursors supplied exogenously in its pathway to synthesis DHA or DPA, and that the strain was using was some different pathway than

previously thought. Then, when the cell-free homogenate derived from *Schizochytrium* cultures was provided with malonyl-CoA labelled with ^{14}C carbon, radioactivity was detected in DHA, DPA, and as well as in saturated fatty acids. This meant that the pathway uses the same precursor as FAS pathway. Later, PUFA subunits were also confirmed to be present in thraustochytrids, suggesting that PUFA synthesis in thraustochytrids occurs through a different route, which has similarities with the PKS system in bacteria (Hauvermale et al., 2006) . Then, in another study, when the *Schizochytrium* PUFA synthase genes were expressed in E. coli, it resulted in the production of both DHA and DPA in E. coli cells. These studies indicated that thraustochytrids likely use a PKS-like pathway to synthesize LC-PUFA (Hauvermale et al., 2006).

Thraustochytrids possess genes for FAS type 1, as well as for PUFA synthase. PUFA synthase is present in the form of multiple subunits, namely subunit A, B, and C. The subunit A has domains for KS, MAT, ACP, KR, and DH. The number of ACP subunits, which are interspaced between MAT and KR domains, vary in number in different species of thraustochytrids (Jiang et al., 2008). The subunit B comprises for KS, chain length factor (CLF), AT, and ER. The CLF domain is considered to be obsolete. The subunit C comprises of two DH domains and one ER domain. To activate the PUFA synthase complex, an enzyme called phosphopantetheinyl transferase is required. (Morabito et al., 2019). It does so by transferring 4′-phosphopantetheine group to the conserved serine residue in ACPs, thereby activating the ACPs (Beld et al., 2014). The exact mechanism of how these subunits work to produce PUFA is not known. The fatty acid chain is elongated by iterative reaction of adding two carbons in each cycle, but how PUFA synthase complex skips the last reduction step by ER, which leads to unsaturated fatty acid production, is not known.

1.3. Aims

As mentioned earlier, genome annotation of T66 has been performed earlier by Heggeset et al., (2019). The assembly was constructed with the reads from Next Generation Sequencing (NGS) technologies, Illumina HiSeq and Roche 454 FLX++. The library consisted of 2 x 100 bp paired-end reads, and 8 kb and 20 kb mate-pair reads. The total data amounted to 13.5 Gbp, giving a total coverage of 300. For the genome annotation, RNA-seq data was also used. In total, 11,683 genes were identified. Generally, if the genome assembly of decent quality is built, it can be subject to annotation since the assembly, despite having gaps, would have

majority of the sequence of the genome (Yandell and Ence, 2012). In the previous annotation of *Aurantiochytrium* sp T66 by Heggeset et al., (2019), majority of the genes have been annotated. However, as mentioned in the section 1.1, the assembly had numerous gaps, despite the use mate-pair reads to increase contiguity. Therefore, there is significant probability that many genes remained unannotated. One of the aims of the study was to sequence the genome of *Aurantiochytrium* sp T66 using MinION sequencer and then build a hybrid assembly with the Illumina short-reads which were used to build the published assembly of *Aurantiochytrium* sp T66, and the long reads generated from MinION in this study, and then, subject the hybrid assembly to annotation using RNA-seq data to see if any genes could be annotated which might have remained unannotated previously.

As mentioned in the section 1.3.6, PUFA synthase in thraustochytrids is present in the form of multiple subunits, namely subunit A, B, and C. Phosphopantetheinyl transferase (*pfaD*) and glutamine synthase are another two enzymes important for PUFA synthesis. Previously, Heggeset et al., (2019) identified PUFA synthase subunit A (*pfaA),* B (*pfaB*), C (*pfaC*) and phosphopantetheinyl transferase (*pfaD*). However, the gene sequences of *pfaA, pfaC,* and *pfaD* were found to be partial. Moreover, glutamine synthetase was found to be split between two contigs. In this study, efforts were made to identify complete sequences of genes *pfaA, pfaC, pfaD,* and glutamine synthase.

As mentioned in the section 1.3.3, ACL is an important enzyme for lipid accumulation, and is found in almost all oleaginous organisms. Though ACL has been identified in other thraustochytrids, however, it could not be detected in *Aurantiochytrium* sp. T66 during the genome annotation of *Aurantiochytrium* sp. T66 by Heggeset et al., (2019). In the expectation that the current hybrid assembly would be of better quality, efforts were made to search for ACL in *Aurantiochytrium* sp. T66.

Mitochondrial genome of *Aurantiochytrium* sp. T66 has not been identified and annotated before. In this study, efforts were made to identify and annotate mitochondrial genome of *Aurantiochytrium* sp. T66.

2. Materials and Methods

2.1. Materials

2.1.1. Growth Medium for *Aurantiochytrium* sp. Strain T66

The growth medium used for growing T66 cells was yeast extract peptone dextrose (YPDS). The YPDS had following constituents: peptone at the concentration of 20g/L, yeast extract at 10g/L, glucose at 2% of the media, and Tropic Marin® sea salt classic to 1.75% of the media, antibiotics ampicillin and streptomycin to the concentration of 200µg/L. The mixture of peptone and yeast extract, and the glucose solution, were autoclaved before mixing them with other constituents of YPDS medium. Solution of Tropic Marin® sea salt classic was also subject to sterile filtration.

2.1.2. Snailase

Enzyme used for cell lysis in this study was Snailase. Snailase is actually a mixture of 20-30 enzymes which primarily contains cellulase, proteolytic enzymes, and pectinases.. Since they contain wide range of enzymes, they can be utilized for several purposes, including cell wall digestion. In this study, it was used for cell wall digestion, and hence, cell lysis.

2.1.3. Cryogenic Grinding

Cryogenic grinding is a process in which cells are exposed to extremely cold temperature, and then the cells are ground using pestle and mortar. In this study, liquid nitrogen was used to flash freeze the cells, which were then subject to grinding.

2.1.4. Genomic DNA extraction kits

Genomic DNA (gDNA) from lysed T66 cells was extracted either with QIAGEN® Genomic DNA for Blood, tissue and cells culture kit according to protocol for yeast, or with NucleoBond® High molecular weight DNA extraction kit's protocol for yeast. Six buffers were used in QIAGEN® Genomic DNA for Blood tissue and cells culture kit. Names and composition of the buffers are as follows:

1. **G2:** 800 mM guanidine HCl; 30 mM Tris·Cl, pH 8.0; 30 mM EDTA, pH 8.0; 5% Tween20; 0.5% Triton X-100
2. **QBT:** 750 mM NaCl; 50 mM MOPS, pH 7.0; 15% isopropanol, 0.15% Triton X-100
3. **QC:** 1.0 M NaCl; 50 mM MOPS, pH 7.0; 15% isopropanol
4. **QF:** 1.25 M NaCl; 50 mM Tris·Cl, pH 8.5; 15% isopropanol
5. **TE:** 10 mM Tris·Cl, pH 8.0; 1 mM EDTA, pH 8.0
6. **Y1:** 1 M sorbitol; 100 mM EDTA; 14 mM **β**-mercaptoethanol

Names of the buffer used from NucleoBond® High molecular weight DNA extraction kit are: **H1**, **H2**, **H3**, **H4** and **H5**.

2.1.5. Instruments used for measuring the concentration of the DNA

Concentration of the gDNA was determined either by Nanodrop™ One spectrophotometer (ThermoFisher Scientific) or Qubit 4.0 fluorometer using Qubit® dsDNA BR Assay Kit (ThermoFisher Scientific). The names of the solutions used from Qubit® dsDNA BR Assay Kit during the measurements were: Qubit® dsDNA HS Reagent, Qubit® dsDNA HS Buffer, Qubit® dsDNA HS Standard 1, and Qubit® dsDNA HS Standard 2.

2.1.6. Preparing the gDNA for MinION Sequencing

Before initiating the MinION sequencing, the gDNA was prepared for sequencing according to the LSK109 protocol provided by ONT. The reagents used were as follows: New England Biolabs (NEB)Next FFPE DNA Repair Buffer, NEBNext FFPE DNA Repair mix, Ultra II End-prep reaction Buffer (NEB), Ultra II End-prep enzyme mix (NEB), AMPure XP beads (Beckman Coulter), Long Fragment Buffer (ONT), 70 % ethanol, Elution Buffer (ONT), Ligation Buffer (ONT), NEBNext Quick T4 DNA Ligase, Adapter Mix (ONT) , Sequencing Buffer (ONT), and Loading Beads (ONT).

2.2. Methods

2.2.1. Growing *Aurantiochytrium* sp. Strain T66

Aurantiochytrium sp. strain T66 was grown in yeast extract peptone dextrose (YPDS) media at 170 rpm at 25°C. T66 cell culture for gDNA extraction was then grown from the pre-culture until the until Optical Density (OD600) was between 7 and 8.

2.2.2. Genomic DNA (gDNA) Isolation

Genomic DNA was isolated either by QIAGEN® Genomic DNA for Blood tissue and cells culture kit with the yeast protocol, or by NucleoBond® High molecular weight DNA extraction kit protocol. In QIAGEN® Genomic DNA for Blood tissue and cells culture kit, the cells were pelleted by centrifuging at 5000xg for 10 minutes. **I)** The cells were then suspended in TE buffer, and again pelleted, and the supernatant was discarded. The intent behind the last step was to wash the cells and purge the remains of YPDS media. **II)** The cells were resuspended in Y1 buffer, and the process of centrifugation and disposition of the supernatant was repeated. **III)** For cell lysis, the cells were either incubated for varying times with Snailase dissolved in either Y1, G2, 1M Sorbitol, or TE, or subject to cryogenic grinding with liquid nitrogen. Afterwards, the process of centrifugation and disposition of the supernatant was repeated for enzymatic lysis. The powdered lysed cells from cryogenic grinding were directly proceeded with the step iv. **IV)** The cells enzymatic lysis or cryogenic grinding were then resuspended in G2 containing Proteinase K**,** and RNase A with the final concentration of 200 µg. Cells were then incubated at 55°C for 90 minutes. The basic aim of this step is to free the DNA from all sorts of proteins, for instance, histone proteins, nucleases etc. G2 denatures proteins, and Proteinase K facilitates the process by digesting the proteins. And the RNase was utilized to digest RNAs, reducing the contamination in the samples. **V)** The mixture was centrifuged, and the supernatant was then subject to column filtration per the instruction in the protocol to extract the gDNA from the supernatant.

Putting the Nucleobond kit extraction kit protocol briefly, the cells were subject to enzymatic treatment as in QIAGEN® Genomic DNA for Blood tissue and cells culture kit protocol. The lysis was further facilitated by incubating the cells in H1 buffer provided in the kit for 30 minutes at 50°C. After the treatment with RNAse, the mixture was subject to column filtration as per instruction of the protocol provided with the kit.

2.2.3. gDNA Quality Control and Quantification

Genomic DNA quality was examined by agarose gel electrophoresis. Agarose gel electrophoresis is a technique used to separate DNA fragments on the basis of their size. Electric field is applied across the agarose gel, and DNA is loaded at the negative end of the electric field. Since DNA possesses negatively charged, they migrate towards positive end of the electric field. DNA fragments then separate on the basis of their size, high molecular weight fragments migrating slower than the low molecular weight fragments. Here, examination of quality means two characteristics of the gDNA: molecular weight and intactness. High molecular weight and no fragmentation is an indication of high quality gDNA.

Purity and concentration were estimated by Nanodrop™ One spectrophotometer (ThermoFisher Scientific) or Qubit 4.0 fluorometer using Qubit® dsDNA BR Assay Kit (ThermoFisher Scientific). To briefly explain the concentration measurement with Qubit 4.0 fluorometer, first the instrument was standardized. To standardize the instrument, two working solutions were prepared by diluting Qubit® dsDNA BR Reagent in Qubit® dsDNA BR Buffer to 1:200 ratio such that the final volume of each solution was 190 µl (it can range from 180-199 µl); Qubit® dsDNA BR Standard 1 was added to one working solution, and Qubit® dsDNA BR Standard 2 to the other working solution, such that total volume of each solution was 200 µl; solutions were vortexed for 2-3 seconds and incubated at room temperature for 2 minutes; readings of these standards were then taken with the instrument to standardize the instrument. To take the readings of the sample DNA, 198 µl working solution was prepared as described above, and then 2 µl sample DNA was added to the working solution such that the total volume was 200 µl; solution was vortexed for 2-3 seconds and incubated at room temperature for 2 minutes; readings of the solution containing the sample DNA were then taken with the instrument.

2.2.4. Preparing the gDNA for MinION Sequencing

Library preparation was performed using LSK109 protocol. To briefly explain the method briefly, DNA was repaired and end-prepared/dA-tailed by mixing 48 µl of the sample DNA (6.4 µg) in 3.5 µl of NEBNext FFPE Repair buffer, 2 µl of NEBNext FFPE Repair mix, 3.5 µl Ultra II End-prep reaction buffer, and 3 µl of Ultra II End-prep enzyme mix, and then incubating this 60 µl of mix first for 5 min at 20 °C, and then for 5 min 65 °C. The DNA was

cleaned up by using 60 µl of AMPure XP beads. At the end of clean up, the DNA was suspended in 61 µl of nuclease free water. The sample DNA was end-ligated with adapters by adding mixing it with 25 µl of Ligation buffer, 10 µl of NEBNext Quick T4 DNA Ligase, and 5 µl of Adapter Mix. After adapter ligation, the sample DNA was again subject to clean up with 40 µl of AMPure XP beads. At the end of clean up, the sample DNA was suspended in 15 µl of Elution Buffer. The prepared DNA library was then made ready for loading into the MinION sequencer by mixing it with 37.5 µl of Sequencing Buffer and 25.5 µl of Loading Beads.

2.2.5. Sequencing

Sequencing was performed using MinION from Oxford Nanopore Technologies. The flow cell R10 (Product code: FLO-MIN110) was used in MinION device. Software released by ONT called MinKNOW (version 19.12.5) was used to initiate the run, which also allows live monitoring of the sequencing. The run was executed on default settings.

2.2.5.1. Base-calling and Data filtration and trimming

Base-calling is a computational procedure to infer the bases from the sequenced reads. In the case of MinION sequencing, electrical signal is generated as a DNA strand passes through a pore, producing a unique pattern of electrical signals. This sensor data is stored in what is called fast5 files by MinKNOW software. Each Fast5 stores information about one DNA strand, or one 'read'. These reads are then subject to base-calling. Base-calling is usually also accompanied by assignation of Phred quality score (Q score), a parameter to evaluate the quality of sequencing (More details in Section Results). In this study, base calling was performed using a software referred to as Guppy, released by ONT. Guppy is able to do base-calling in two modes: fast base-calling and high accuracy base-calling**.** In this case, base calling was carried out with fast mode settings. The filtration of the base-called reads on the bases of Q score was done with a software Nanolift (De Coster et al., 2018), with threshold Q score 7. The base-called reads were further processed to trim the adapters with a software called Qcat, released by ONT. Following the adapter trimming, all the reads were merged in a single file using a software called Cat, also released by ONT.

2.2.5.2. *Quality assessment of the sequencing*

A report generated by MinKNOW at the end of the sequencing run gives an insight to the quality of the sequencing. For further assessment, two quality control tools were put into service: MinIONQC and PycoQC (Lanfear et al., 2019) (Leger and Leonardi, 2019).

The quality of sequencing was assessed on the basis of following parameters:

Phred Quality Score (Q score): Q score is a measure to evaluate the accuracy of the sequenced reads. It is logarithmic representation of the probability of error in the sequenced reads, and is given as $Q = -10log_{10}(P)$, where P is the probability of error in the read or set of reads. To illustrate the parameter further, probability of 1/10, which equals 90% accuracy, would have Q score of 10. Likewise, probability of 1/100 would have Q score of 20.

Total Number of Base pairs (bp) sequenced: As is evident from the term itself, it indicates the total number of base pairs sequenced. This parameter in itself does not offer help in evaluating the quality of sequencing. It is the statistics derived from it which assist in evaluate the sequenced data. Following statistics derived from this parameter were used in this study:

Sequencing Depth: It indicates how much coverage is provided to the genome by total sequenced data. It is given as $coverage = \frac{Total\ Number\ of\ base\ pairs\ sequenced}{Approximate\ size\ of\ the\ genome\ being\ sequenced}$. The equation can be changed to calculate coverage provided by the reads of certain length. In this study, coverage provided by reads greater than 10 kb, 20 kb, 50 kb, and 100 kb also calculated.

Mean read length: As is evident from the term itself, it is mean of read length. It is given as $Mean\ read\ length = \frac{Sum\ of\ lengths\ of\ all\ the\ sequenced\ reads}{Total\ number\ of\ reads}$

2.2.6. . Genome Assembly and Quality Control

With the MinION sequencing data, as well as shotgun data from Illumina Hiseq (2x100bp) sequencing performed by (Liu et al., 2016), a hybrid assembly was generated using an assembly pipeline called Unicycler (Wick et al., 2017). In Unicycler assembly pipeline for hybrid assembly, a Illumina-only assembly is made using a software called SPAdes (Bankevich et al., 2012),; this is followed by hybrid assembly generation collectively from long reads, as well as

contigs of already constructed Illumina-only assembly; after putting assembly through a series of quality enhancing steps, the assembly is 'polished' with Illumina short-reads using a software called Pilon (Walker et al., 2014). In polishing, short reads are mapped to the assembly, and base accuracy is improved by generating consensus sequence between the assembly sequence and mapped short reads. See section 1.1.3 for more detail about using short read data in hybrid assembly.

The assembly's quality was assessed using two tools: GenomeQC (Manchanda et al., 2020) and QUAST (Gurevich et al., 2013). To put the quality assessment into perspective, the current assembly was compared with the *Aurantiochytrium* sp. T66 genome assembly constructed by (Liu et al., 2016), and with the assembly of *Hondaea fermentalgiana* FC1311 (Seddiki et al., 2018).

2.2.7. Genome Annotation

2.2.7.1. Structural Annotation

In structural annotation, genomics features such as protein coding sequences (CDS) and non-coding RNA sequences are predicted. In order to predict CDS on the basis of RNA-seq reads, the reads were first mapped to the genome assembly, which was followed by CDS prediction.

2.2.7.1.1. Read Mapping

RNA-seq reads from were mapped to the assembly using Large Gap Map Reading tool in CLC-genomics Workbench with default settings. The Large Gap mapping tool first maps the reads to the genome; the reads aligned in the first go are referred to as read aligned in the Segment 1, ''seed segment'', or is referred to as uniquely mapped/specific reads in literature; reads that remain unaligned in the first round are tried to be aligned in the subsequent rounds, which are referred to as segment 2, 3, 4…; if reads are aligned in the first round, it indicates high specificity of reads to reference genome.

The quality level of the mapping of the reads to the genome assembly was inferred from the statistics produced by the Large Gap Map Reading tool of CLC-Genomics Workbench after finishing the mapping.

2.2.7.1.2. CDS Prediction

Read mapping was followed by CDS prediction, for which, Transcript Discovery tool of CLC-Genomics Workbench was used. Transcript Discovery predicts CDS on the basis of RNA-seq reads mapped to the genome. It was executed with default settings**.** The CDS predicted by Transcript discovery were extracted. The extracted CDS were then BLAST-searched against the genes of published annotation. From the BLAST result, only those genes for automatic functional annotation were chosen for which no or insignificant similarity to the old annotation was detected.

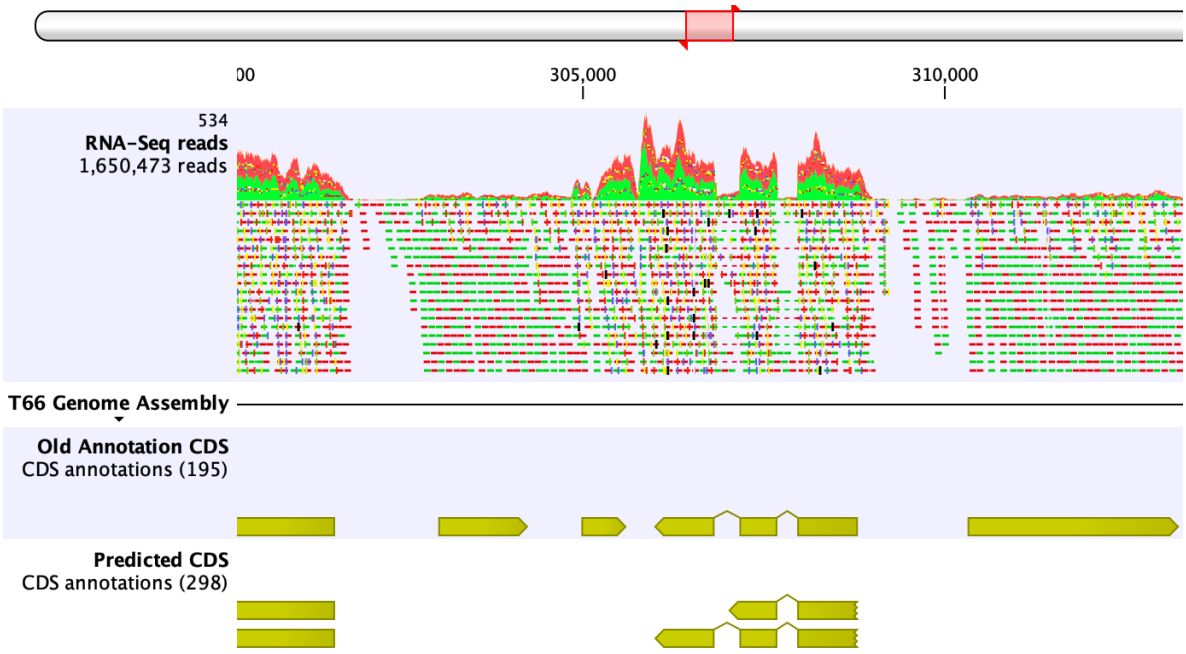


Figure 7. ***Predicted CDS and old annotation CDS in relation tool mapped RNA-seq reads***

For manual functional annotation, CDS predicted in this study and the old annotation were mounted on the genome assembly; the newly predicted CDS and old annotation CDS were analyzed separately in relation to mapped RNA-seq reads, as shown in Figure 7; in this way, only those CDS were fished out which had not been annotated before.

2.2.8. Functional Annotation

Functional Annotation is assignment of biological information to the predicted coding sequences. For automatic functional annotation, the predicted CDS were BLAST-searched against Protein Data bank (PDB) (Berman et al., 2000), Uniprot/Swissprot (Consortium, 2020), and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000)

databases. For manual functional annotation, the predicted CDS were BLAST-searched against Non-redundant, RefSeq (O'Leary et al., 2016), PDB, Uniprot/Swissprot, KEGG, and Pfam databases (Finn et al., 2014).

To further elaborate the annotations, Gene Ontology (GO) (Ashburner et al., 2000; Gene Ontology, 2021) identifiers and terms were assigned using InterProScan (Jones et al., 2014). GO is a platform which provides describes genes in terms of its functions, the location where it performs its function, and the biological process it is part of. Also, the diction for description is very controlled and precise. After assigning the GO identifiers and terms, they were categorized by WEGO (Ye et al., 2018).

3. Results

3.1. Optimization of Chemical Cell lysis Protocol for the Genomic DNA Extraction of *Aurantiochytrium* sp. T66

For long-read sequencing, it is important for the genomic DNA (gDNA) to be of high quality, i.e to have high molecular-weight (MW), minimal fragmentation and contamination of proteins and solvents. The genomic DNA (gDNA) extraction could have been carried out by physical disrupting the cells, such as by cryogenic grinding using liquid nitrogen, which is a common method to lyse rigid cells. However, an effort was made to inspect if the cells could be lysed chemically, and if gDNA extracted by chemically lysing the cell is of quality good enough to perform sequencing with Therefore, different methods were employed to find the optimal chemical cell lysis protocol for the extraction of gDNA, such that the integrity of the gDNA remains intact.

3.1.1. T66 gDNA Extraction Using Snailase as Lytic Enzyme

The aim of this step was to investigate if T66 cells could be lysed using Snailase, and subsequently, if gDNA can be extracted from the lysed cells. T66 cell culture for gDNA extraction was grown from pre-culture until optical density at 600 nm was between 7 and 8. The cells were pelleted and weighed. The weight of the cells was found to be 3.67 grams. The cells were incubated with Snailase. Following incubation, Qiagen Blood and tissue culture Genomic DNA Extraction kit was used to extract gDNA from the cells. Upon adding isopropanol in the supposedly eluted DNA, neither the DNA threads were observed, nor any DNA pelleted upon centrifugation at 5300xg for 15 minutes. Instead of centrifuging at 5300xg for 15 minutes as recommended by the protocol, the elute was centrifuged at 16100xg for 20 minutes. Purpose was to see if it is possible to pellet any DNA, however small amount it was present in. A pellet was obtained, which it was thought, was of DNA. The pellet was dissolved in TE buffer, and the sample was run on agarose gel. The result is shown in Figure 8.

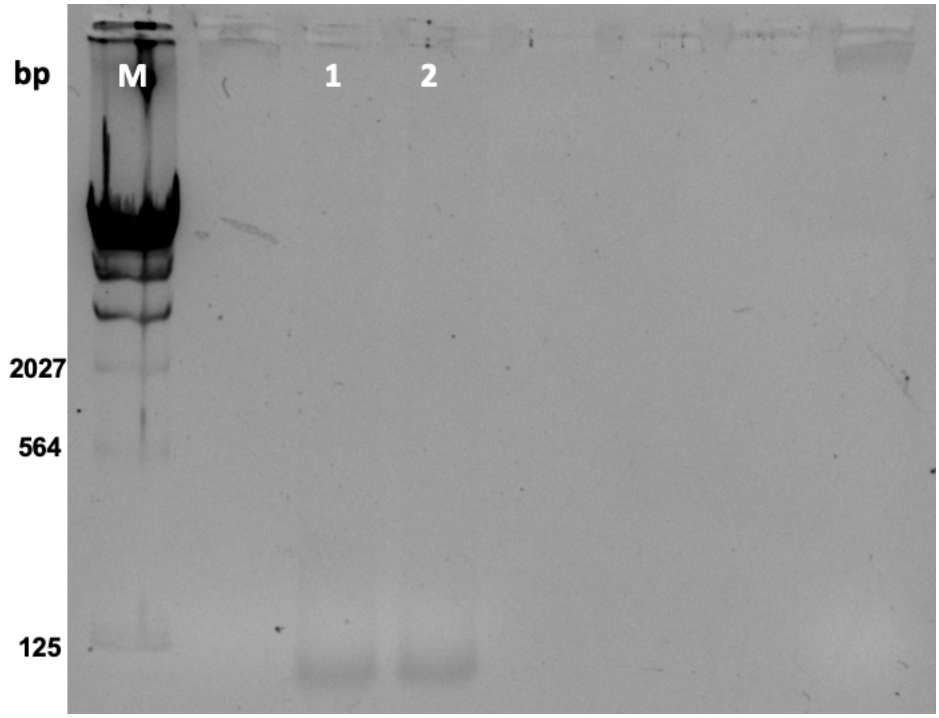


Figure 8. ***Detection of genomic DNA in eluted after high-speed centrifugation by electrophoresis on agarose gel****. Sample was run on 0.5 % agarose gel at 120V. Lambda DNA was used as a ladder, marked as lane M. Lane 1 shows the sample*

In Figure 8, a band can be seen in lane 1 at molecular weight lower than 125 bp, which could be of DNA or RNA considering the low molecular weight. The concentration of the sample was found to be 3ng/ul (measured with Nanodrop™ One spectrophotometer), which is probably inaccurate, as measurements with Nanodrop™ One spectrophotometer at concentrations lower than 10 ng/ul are generally regarded inaccurate (Khetan et al., 2019). But to have such low concentration certainly means that the concentration was lower than 10 ng/ul. To obtain optimum results with MinION sequencing, gDNA of high molecular weight, concentration of 70 ng/ul with total amount of 3.4 ug is recommended (Technologies, 2021). Furthermore, the quality of the sample was not reasonable either; A260/A280 and A260/A230 were measured to be 1.30 and 1.70, which indicates significant contamination of proteins and solvents with the sample, respectively. Thus, it was not possible to proceed with this sample for sequencing.

The procedure was repeated, and the same outcome was observed. This indicated that the cells could not be lysed thoroughly using Snailase, and thus, decent DNA yield could not be attained. There could be two possible reasons for the cells to show resistance to the lysis method being employed: either the buffer being used for Snailase was not effective, or the cells were sensitive

to the incubation times, i.e., they might need longer incubation with Snailase, or Proteinase K + RNase.

3.1.2. Testing the Sensitivity of the Cells to Incubation time, Finding the optimal lysis solution for Snailase, and Testing the Nucleobond Kit for T66 Cells' Lysis

To test the sensitivity of the cells to incubation times and to see which lysis solution Snailase works best with, cells were incubated with Snailase in different buffers with varying incubation times. Cells were then subject to varying incubation times with proteinase K + RNase. Since the aim was to test the cell lysis, not the elution with the kits, the procedure was only performed till cell lysis**,** i.e., till the usage of Snailase. In addition to testing sensitivity of the cells to incubation times and buffer for Snailase, High Molecular Weight (HMW) extraction kit protocol from Nucleobond was also tested for gDNA extraction. As HMW Nucleobond kit had different protocol than Qiagen Blood and tissue culture Genomic DNA Extraction kit, it was used in the anticipation that changing the entire protocol for gDNA extraction might help in obtaining HMW gDNA. Results are shown in Figure 9.

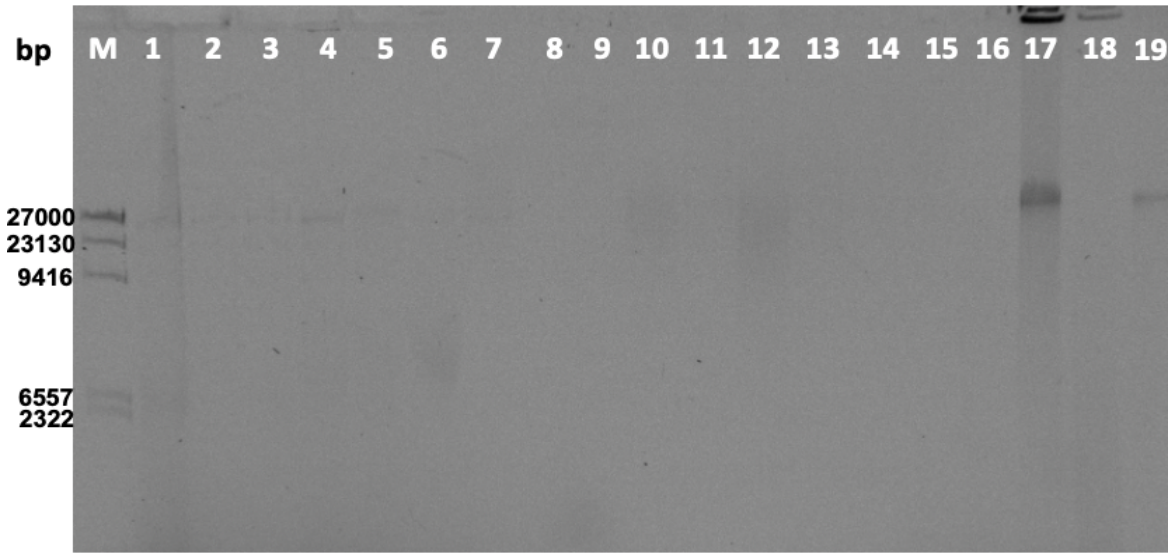


Figure 9. ***Testing the Snailase sensitivity to different buffers and cells sensitivity to different incubation times.*** *Samples were run on 0.5% agarose gel at 120V. Lane 1-3 show Snailase incubation in Y1 for 60 minutes, and G2+RNAse+protease incubation for 60,90 and 120 minutes respectively. Lane 7-9 show Snailase incubation in G2 for 60 minutes, and G2+RNAse+protease incubation for 60,90 and 120 minutes respectively. Lane 10-12 show Snailase incubation in G2 for 60 minutes, and G2+RNAse+protease incubation for 60,90 and 120 minutes respectively. Lane 13-15 show Snailase incubation in Sorbitol for 60 minutes, and G2+RNAse+protease incubation for 60,90 and 120 minutes respectively. Lane 16-18 show Snailase incubation in Sorbitol for 90 minutes, and G2+RNAse+protease incubation for 60,90 and 120 minutes respectively. Lane 19 shows when the samples were subject to Nucleobond HMV protocol.*

As is evident from Figure 9, all the lanes corresponding to Snailase-treated cells, which are represented by lane 1-18, either have very faint bands, or no band at all, except in the lane 17. This means that either the DNA concentration is very low, or no gDNA is present at all. Lane 17 corresponds to the cells which were incubated in Snailase with sorbitol for 90 minutes, and for the same incubation with Proteinase K + RNase. A clear DNA band in lane 17 indicated that the protocol employed to the cells corresponding to lane 17 is effective for cell lysis, and thus, can prove to be useful for gDNA extraction.

Lane 19 corresponds to the cells that were treated with HMW Nucleobond kit. As is clear from the lane 19 in Figure 9**,** the band, though perceptible, is still quite pale, indicating low DNA concentration. Therefore, it was concluded that HMW Nucleobond kit is not effective for gDNA extraction. The only effective method seems to be the one corresponding to the lane 17.

This method was employed for gDNA extraction. However, again, the yield of the DNA was not good enough. It was inferred from the results that T66 cells are too tough to be lysed by the enzymes present in Snailase

3.1.3. Cryogenic Grinding with Liquid Nitrogen

Since the cells could not be lysed with any method tested, it was decided to physically disrupt cells to achieve good cell lysis. Therefore, the cells were subject to cryogenic grinding with liquid nitrogen. After cell lysis with cryogenic grinding, gDNA was then extracted from ground cells using Qiagen Blood and tissue culture Genomic DNA Extraction kit**.** Results are shown in Figure 10.

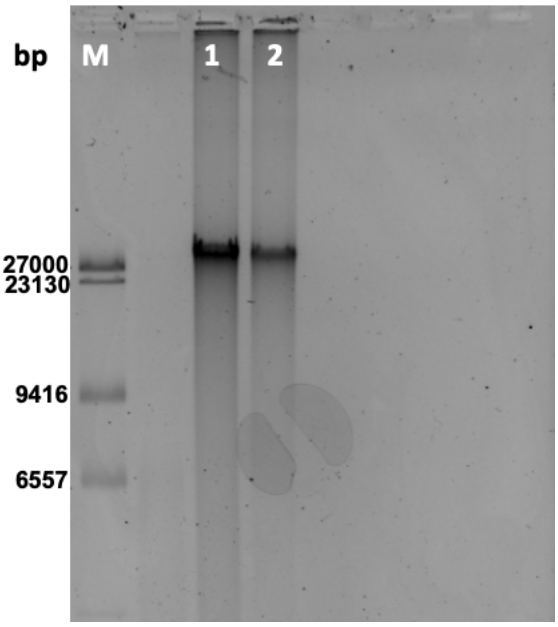


Figure 10. ***Detection of genome DNA in the cells lysed with cryogenic griding.*** *Lambda DNA/HindIII ladder is indicated as the lane M. Lane 1-2 correspond to the cells lysed using liquid nitrogen, with the cells in the lane 1 ground for 5 minutes, and thus have higher concentration, and the cells corresponding to lane 2 ground for 2 minutes, and thus have lower concentration. It was the cells corresponding to lane 1 which were proceeded with for sequencing.*

As is evident from Figure 10, the molecular weight is at least higher than 27 kb. A260/A280 and A260/A230 were measured to be 1.91 and 2.06, respectively, indicating the purity of the sample to be adequate. The concentration of the extracted gDNA was found to be 134 ng/ul.

The parameters employed to check the DNA yield and quality indicated that the extracted gDNA's yield and quality are satisfactory to proceed with sequencing. This suggested that chemical lysis methods employed were not effective enough lyse T66 cells so that quality gDNA could be extracted. Cryogenic grinding with liquid nitrogen proved to be the most effective method out of all the protocols employed to lyse T66 cells. This sample was proceeded with for sequencing.

3.2. Quality Assessment of Sequencing

For sequencing, as it is important to do quality assessment at pre- sequencing stage so that it can be determined whether the DNA sample should be proceeded with for sequencing, or the DNA extraction should be performed again, or any additional measures should be taken in the post-sequencing analysis, likewise, quality assessment of sequence data is of paramount significance. It can suggest if the sequencing output is consistent with the pre-sequencing quality assessment or not. Depending on the purpose of sequencing, it can also help decide if the sequence data requires any additional treatments before proceeding to downstream analysis. In the following sub-sections, the result of quality assessments of the sequencing is described.

3.2.1. Sequencing Run Analysis

The sequencing run was followed in real-time using ONT's published software called MinKNOW, which is designed specifically for ONT's nanopore sequencing devices to follow the sequencing in real-time. By monitoring the start of run, the final output of the run can be forecasted with considerable reliability; and if it does not look to be promising, the run can be stopped, rather than finding the outcome after the run as is the case in next-generation sequencing technologies. The sequence data generated at the end of the run can help analyze the quality of the run**.**

Pore occupancy is a parameter which can help assess the performance of the sequencing run. It is defined as strand pores in fraction to the total number of pores. Low or no pore occupancy either suggests problem with the flow cell or/and the DNA sample, for instance, if the pores of the flow cell have already been utilized, or the DNA amount was not enough, or some problem in the DNA library preparation. Figure 10 shows duty time plot for pore occupancy over the course of sequencing run.

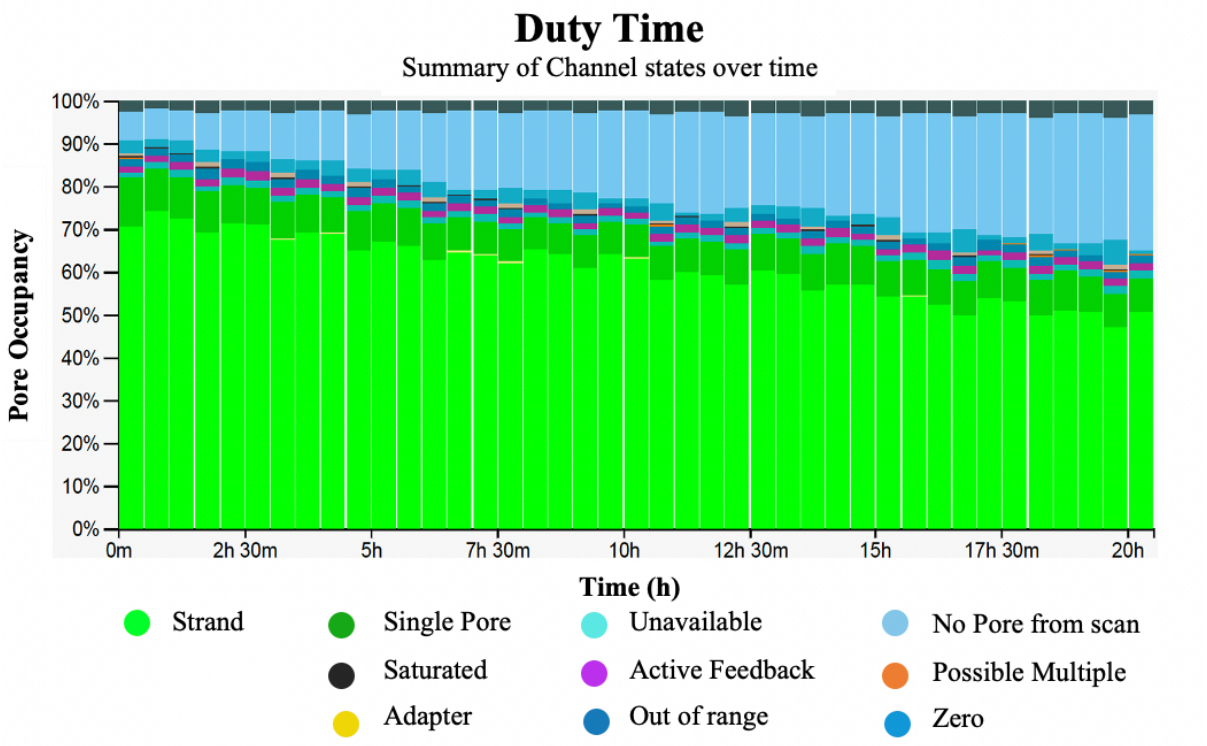


Figure 11. ***Duty time plot for pore occupancy over time.*** *Strand: The channel in which DNA molecule is being translocated through the pore, i.e. the DNA molecule is being sequenced;* ***Adapter:*** *The channel in which the pore is sequencing unligated sequencing adapters.* ***Single Pore:*** *The channel which has single pore available for sequencing, but no sequencing is currently being carried out.* ***Unavailable:*** *The channel in which the pore has been blocked by contaminant.* ***Saturated:*** *The channel in which the pore's membrane is damaged.* ***Active feedback:*** *The channel which is reversing the current flow to remove any blockage in the pore, for instance, to remove the contaminant.* ***Out of Range:*** *Negative current is passing through the channel. It usually happens when the ionic solution in the channel is leaking.* ***Possible multiple:*** *The channel has more than one active pores, and is therefore unavailable for sequencing.* ***No pore from scan:*** *The channel in which no pore has been detected.* ***Zero:*** *The channel which has no current passing through. It could be that the channel was turned off during the sequencing, or that it was not turned on all during the course of run. It should be noted that strand channels are the only channels which are sequencing the DNA in the sample. The rest of the channels, due to their respective problems, were not sequencing the DNA in the sample.*

As is depicted in Figure 11, the experiment starts off with pore occupancy above 70%, and it remains so till almost 5 hours. However, there is a steadily progressive, though gradual, drop in the pore occupancy over time. The plot shows that the drop in the proportion of strand pores, and so the pore occupancy, is in inverse relation to the percentage of inactive pores. This

implies that as the number of inactive pores increase, and thus the pore occupancy drops, the quality of the run also attenuates.

Figure 12 shows the total output of sequencing over the period of total run. As it can be seen from the plot**,** the rate of production of reads attenuates over time; at 7.97 hours, 50% of the data had been generated, and in the 7 next hours, only 25% reads of total output was produced. This pattern is consistent with duty time plot in Figure 11**;** drop in the number of strand pores, and so the overall sequencing, manifests as steadily progressive drop in the rate of read generation.

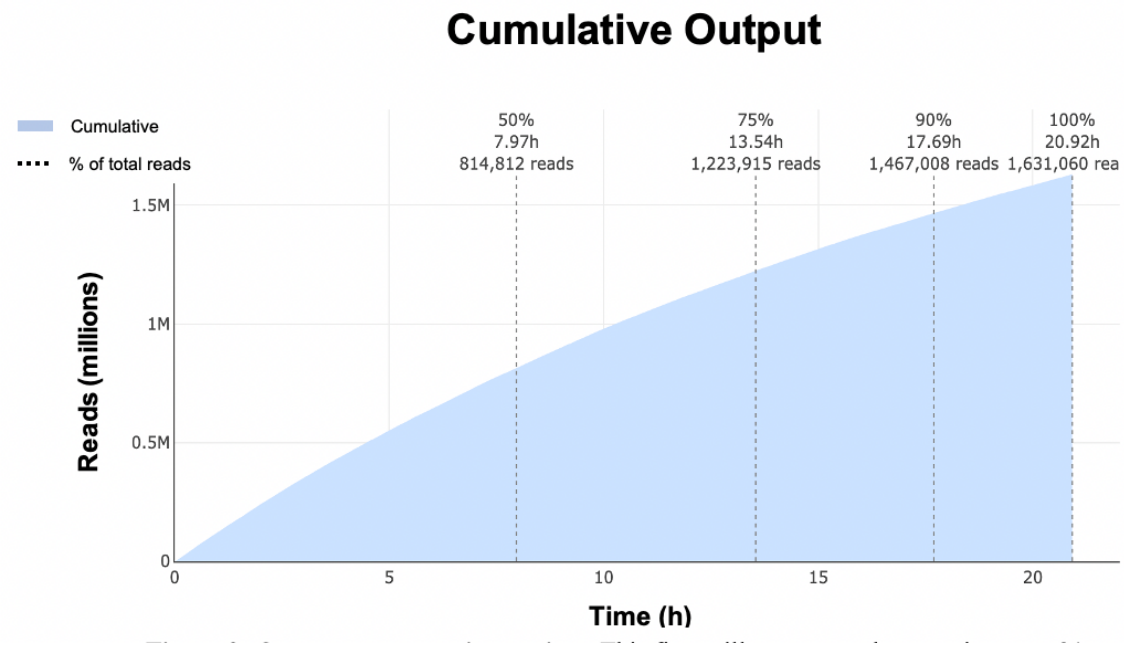


Figure 12. ***Output over experiment time.*** *This figure illustrates read generation over 21-hour sequencing run*

In 21 hours, 1.63 million reads were generated (Figure 12). In terms of base pairs (bp), 5.78 giga base pairs (Gbp) were sequenced. As the T66 genome length is approximately 40 Mbp, this means that 5.78 Gbp of data generated in 21 hours could potentially provide 144-fold coverage (5.78 Gbp/0.04 Gbp). It is written that 5.78 Gbp of data could 'potentially' provide 144-fold coverage. It is written so because after filtering the low-quality data to improve the quality of the sequencing data, the coverage reduces (See section 3.2.2).

In MinION sequencing, it is fairly common to observe the decrease in pore occupancy, and hence the rate of data generation over experiment time. Pore occupancy of 70%, which this

sequencing run maintained for the first 5 hours, is generally regarded high, and is indicative of possibly good final output (Schalamun et al., 2019). For *de novo* genome assembly, 50-60 fold coverage has been shown to be sufficient for genomes under 100 Mbp in size (Desai et al., 2013). The final output of the sequencing run providing potential sequencing depth of 144-fold means that despite the decrease in pore occupancy**,** the run could be considered successful.

3.2.2. Quality Assessment of Sequence Data

As mentioned in the section 1.1.3, for *de novo* genome assembly construction, longer the reads, the better it is for assembly's contiguity and for resolving the repetitive regions in the genome. Long reads in the sequence data are also indication of a good DNA sample and vice versa. Another important aspect to keep in consideration regarding the long reads' generation is that if the long-read data provides adequate coverage or not. Figure 13 displays the distribution of reads on the basis of their length, and Table 1 further dissects Figure 13 to better evaluate the quality of sequencing.

Generally, third-generation sequencing technologies produce reads with average length of 10 kb (Jain et al.; 2016**).** As is evident from Figure 13, the majority of the reads are shorter than 10 kb in length. However, as is exhibited in Table 1, recasting 10 kb reads data in terms of bp makes it 2.11 Gbp, which can solely render coverage of 52.8 (2.11 Gbp/0.04 Mbp).

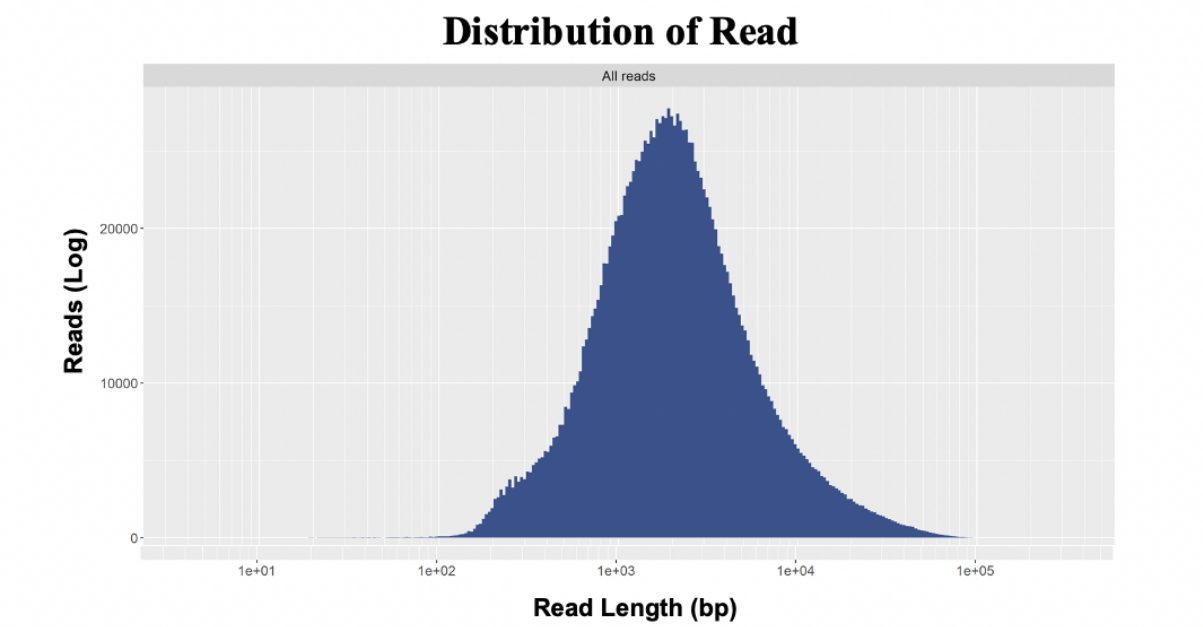


Figure 13. ***Distribution of reads over read length****. Read length is on x-axis is in base pairs (bp), while read count in log scale is on y-axis.*

Table 1. ***Summary of Sequence Data***

Number of reads	Total reads (Gbp)	Mean read length (bp)	N50 Length (bp)	Max length (bp)	10 kb reads	10 kb reads (Gbp)
1631061	5.78	3543	6203	321676	107351	2.11

Number of reads*: total number of reads generated. Total reads in Gbp: size of data in terms of giga bases.* ***Mean read length:*** *mean length of reads.* ***N50:*** *length of shortest read in the set of reads required to cover 50% of genome size.* ***Max Length****: length of the largest read.* ***10 kb reads:*** *total number of reads generated which are at least 10 kb or more in length .* ***10 kb reads in Gbp:*** *Size of 10 kb reads in terms of giga bases.*

Another widely used metric to evaluate the quality of sequencing is to assign Phred Quality Score (Q score) to the reads base called. Assignation of Q score to reads allows to gauge the quality of sequencing in terms of accuracy, and it is primarily dependent on the method of sequencing**,** rather than the DNA sample. Median Q score of sequence data was found to be 7.4. Figure 14A depicts graphical representation of the reads on the bases of Q score.

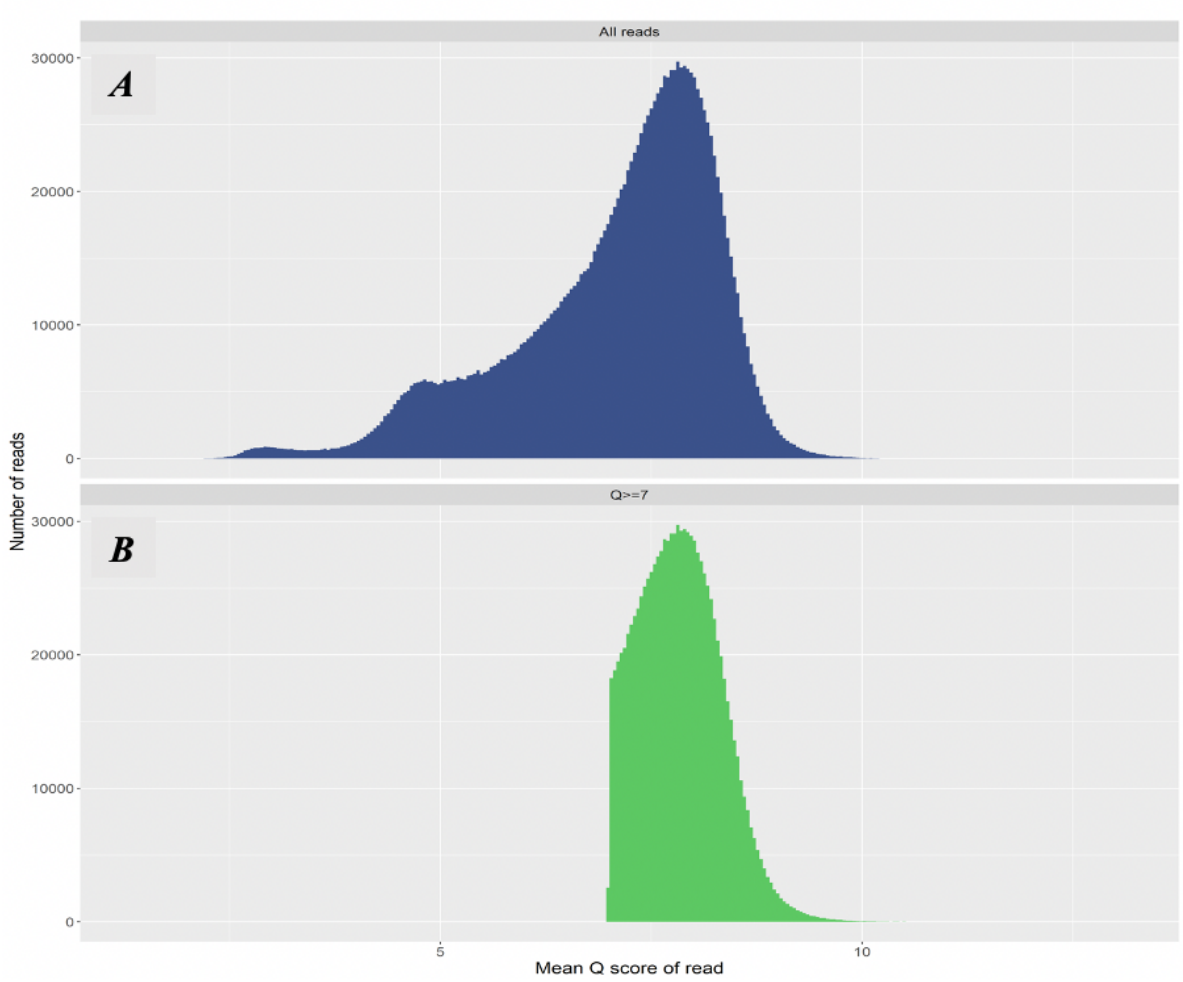


Figure 14. ***Distribution of reads over Q score****. A shows distribution of reads over Q score before filtration, and B shows distribution of reads after filtration. Q score is logarithmic representation of the probability of error in the sequenced reads. It is given as* $Q = -10 \log_{10}(P)$*, where P is the probability of error in the read or set of reads..*

Table 2. ***Summary of comparison between the data before and after Q score filtration****. The data was filtered with threshold Q score 7*

	Pre-filtration		Post-filtration	
	Read Count	No. of Base pairs (Gb)	Read Count	No. of Base pairs (Gb)
Total Reads	1631000	5.78	1013328	3.84
>10 kb Reads	107350	2.1	73888	1.46
>20 kb Reads	35000	1.12	24540	0.78
>50 kb Reads	3000	0.19	2165	0.13
>100 kb Reads	50	0.0062	22	0.0024
N50 (kb)	6.2		6.6	
Mean Read Length (kb)	3.54		3.78	
Median Read Length (kb)	1.94		2.01	
Median Q Score	7.4		7.8	

To concentrate the good quality data, low Q-score reads were filtered out with threshold Q-score of 7. Graphical representation of the filtered data is shown in Fig 14 B. Table 2 summarizes the comparison between filtered and unfiltered data. With the adjustment of threshold Q score to 7, the mean Q-score rose to 7.82. The number of 10-19 kb and 20-49 kb reads decreased by 30 % and 29% respectively; likewise, there was 28% and 56% loss of 50-99kb and >100 kb reads, respectively. However, the estimated genome length of T66 being 40 Mbp, post-filtration data of 3.8 Gbp still provides 96x coverage, which exceeds the recommended 50-60-fold coverage for *de novo* genome assembly (Desai et al., 2013). Despite the loss of significant number of long reads during filtration, reads of length between 10-19 kb still render 36x coverage. Furthermore, filtered 50-99 kb reads provide more than 3x coverage. This filtered data and the short-read data generated previously by (Liu et al., 2016) was used for *de novo* genome assembly.

3.3. Quality Assessment of Genome assembly

Comparison of the current T66 assembly with assembly by Heggeset et al**.,** (2019), and with another thraustochytrid of similar size, called *Hondaea fermentalgiana* FC1311, is summarized in Table 3. As is evident from Table 3, there is stark difference between the N50 of the current assembly, and the other two assemblies; N50 of the current assembly is 25x and 14x greater than the previous T66 assembly and *H.fermentalgiana* FC1311, respectively. Likewise, L50 is 22x and 13x smaller than the previous T66 assembly by (Liu et al., 2016) and *H.fermentalgiana*

FC1311, respectively. From the statistics given in Table 3, it can be concluded that the current assembly has better contiguity than the other two assemblies. Another important difference is the number of unknown bases (Ns); the current assembly has no Ns, while the Ns constitute 11% of and 0.6% of the previous T66 assembly and the *H.fermentalgiana* FC1311 assembly, respectively. This assembly was proceeded with for annotation. (The sequence of the genome assembly of *Aurantiochytrium* sp. T66 constructed in this study can be provided on request by contacting Tonje Marita Bjerken Heggeset from SINTEF.)

Table 3. ***Summary of genome quality comparison of the current T66 assembly, Short-read T66 assembly by (Liu et al., 2016), and thraustochytrid Hondaea fermentalgiana FCC1311***

Parameter	**Illumina shotgun and MinION hybrid T66 assembly**	**Illumina Shotgun and mate-pair T66 assembly**	***Hondaea fermentalgiana* FCC1311**
Total length	40129602	43429441	38716150
Number of Ns	0	5112477	227200
GC content (%)	62.98	62.83	57.13
Total Contigs	980	6833	4504
N50 (bp)	332369	12952	22474
L50 (bp)	40	894	527
L75 (bp)	80	1895	1119
Largest Contig (bp)	1185158	98696	129397
No. of Contigs ≥ 1000 bp	288	4361	2857
No. of Contigs ≥ 10000 bp	198	1281	1293
No. of Contigs ≥ 25000 bp	176	183	439
No. of Contigs ≥ 50000 bp	158	21	75

Parameter's Definition*. a) Total length: total length of the genome****. b) Number of Ns****: number of undetermined bases (Ns) in the assembly.* ***c) GC content (%):*** *the percentage of G and C bases in the assembly****. d) Total Contigs:*** *Total number of contigs; contig refers to a continuous sequence without any Ns.* ***e) N50:*** *the length of shortest contig in the set of contigs required to cover 50% of genome size****. f) L50:*** *the number of contigs required to cover 50% of genome.* ***g) L75:*** *the number of contigs required to cover 75% of genome.* ***h) Largest Contig:*** *length of the largest contig in the assembly.* ***i) No. of Contigs ≥ 1000 bp:*** *Number of contigs with length equal to or greater than 1000bp.* ***j) No. of Contigs ≥ 10000 bp:*** *Similar to (i)****. k) No. of Contigs ≥ 25000 bp:*** *Similar to (i)****. l) No. of Contigs ≥ 50000 bp:*** *Similar to (i)*

3.4. Annotation

The completion of assembling the genome and assessing its quality was followed by genome annotation. As mentioned earlier, the annotation of T66 genome have been performed earlier by Heggeset et al., (2019) . The main objective here was to find protein-coding regions which have not been identified previously. It should be noted here that initially, the identification is of possible coding sequences (CDS), not genes (Yandell and Ence, 2012); the term "gene prediction" is often used synonymously with CDS regions, although the CDS prediction in this case can very much likely be the gene as the intron distribution in the genome of this organism is very sporadic.

3.4.1. Structural Annotation

As mentioned in the section 1.2.4, RNA-seq data can be used in two possible ways for structural annotation: by *de novo* transcriptome construction or genome-guided transcriptome construction. In *de novo* transcriptome reconstruction, since the low-abundance reads are filtered, or even if they are retained, it is difficult to assemble them into transcript because of low coverage, therefore, *de novo* transcriptome reconstruction is not very suitable for discovering novel transcripts and thus, CDS. In genome-guided transcriptome reconstruction, since the reads are directly aligned to the genome, and therefore low abundance reads can also be aligned to the genome, prospects of discovering novel CDS are higher. Since a genome assembly of reasonable quality had already been built, and the purpose of annotation was to find novel transcript rather than annotating the genome from scratch, the genome-guided transcriptome reconstruction method was adopted. First, RNA-seq reads were mapped to the genome, and then the CDS were predicted.

3.4.1.1. Genome-guided Transcriptome Construction (RNA-Seq reads Mapping) and Quality control

For genome-guided transcriptome construction, cleaned RNA-seq reads were acquired from the study conducted by Heggeset et al., (2019). The reads were mapped to the genome. The summary of read mapping is shown in Table 4.

Table 4. ***Summary of RNA-seq reads mapping***

	Number of reads	Percentage (%)
Mapped reads	94.542,133	98
Un-mapped reads	1,702,881	2
- **Invalid match**	1,702,881	2
- **No match**	102,310	0
Total reads	96,245,014	100

As it is evident from the Table 4, 98% of the reads mapped to the assembly; only 2% of reads remained unmapped. The standard RNA-seq read alignment percentage is between 70-90% (Dobin et al., 2013). High alignment percentage means that most of the reads have been aligned to the reference genome, and very few reads were discarded; having 98% of reads aligned indicates that the quality of alignment of the RNA-seq reads to the reference genome was more than adequate.

Table 5 shows the gap summary of mapped reads**.** As indicated by Table 5, most of the reads were un-gapped, i.e., gaps between the matched reads were very low in number, or gaps in the transcriptome assembly were very low in number**.** This means that number of introns are very low, which is consistent with the fact that the genome of T66 has very few introns.

Table 5. ***Gap summary of mapped reads***

	Number of reads	Percentage (%)
Un-gapped reads	93,218,528	99
Gapped reads	1,323,605	1
Total	94,542,133	100

Another important parameter to measure the quality of alignment is distribution of number of segments per read matching the reference; Figure 15 shows the reads distribution against reference number.

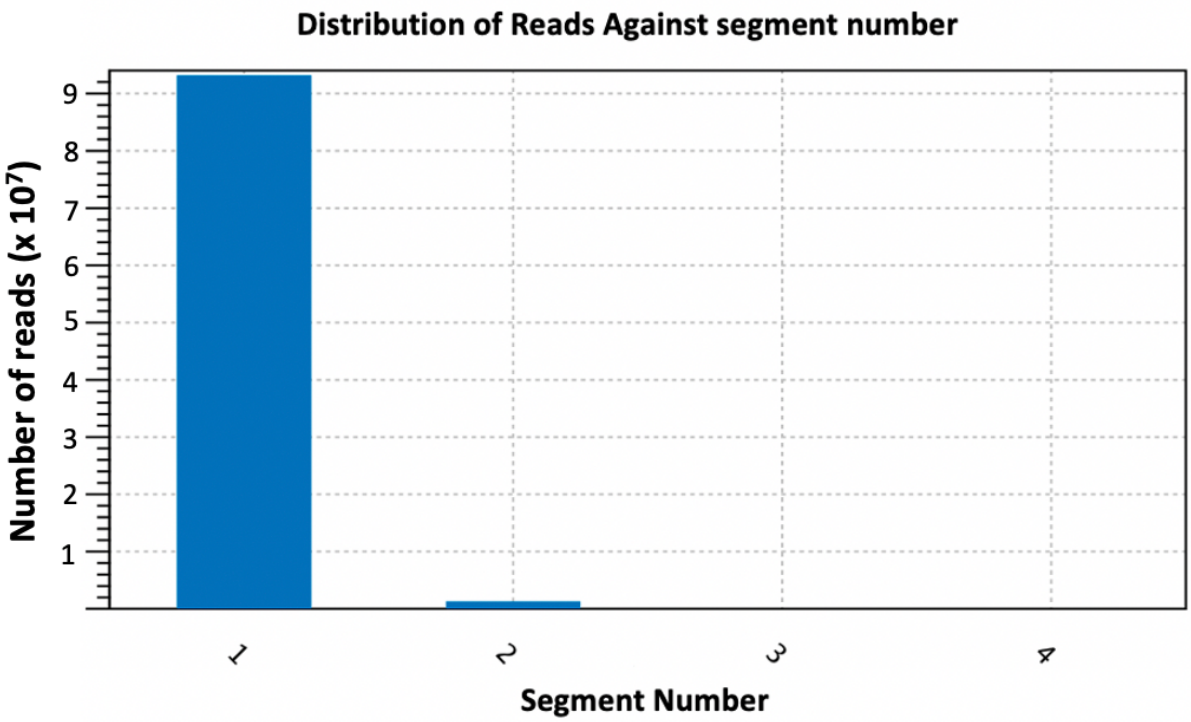


Figure 15

It is evident from the graph that majority of the reads were aligned in Segment 1. Numerically, 90,516,493 reads were aligned in segment 1, which is 96% of the total mapped reads. On the other hand, only 4,025,640, of mapped reads, which makes 4% of total mapped reads, were non-specific i.e., they had multiple regions where they could be aligned equally well, and were therefore aligned in the subsequence segments. Majority of reads being uniquely mapped means that most of the reads were aligned in the 1st round of mapping, indicating high specificity of reads to the reference genome

3.4.1.2. CDS Prediction

CDS were predicted using Transcript Discovery in CLC Genomics Workbench**.** In total, 11,510 possible CDS were predicted. For automatic functional annotation, all the predicted CDS were BLAST-searched against the genes of old annotation. Out of 11,510 predicted CDS, there were 246 CDS for which either no sequence coverage and similarity was found, or it was negligible. These 246 CDs were subject to automatic functional annotation.

For manual functional annotation, since novel CDS had to be identified manually**,** which is considerably time-consuming task, therefore, CDS from only first 15 contigs could be subject to manual functional annotation. Total length of the first 15 contigs is 9,460,284, which equals

almost 24 % of the total genome length. In the first 15 contigs, 453 CDS were identified which were previously unannotated and corresponded with reasonable RNA-seq read count. Number of CDS identified in each contig is shown in Table 6. These genes were then subject to functional annotation.

Table 6. ***Number of CDS identified in each the first 15 contigs***

Contig Number	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Number of Genes	78	44	43	44	35	30	20	26	17	11	22	27	14	17	23

3.4.2. Automatic Functional Annotation

As mention in the section above, there were 246 CDS in total for which either there were no hits against the old annotation, or the hits were insignificant. This means that these predicted CDS had not been annotated previously. These 246 sequences were BLAST-searched against Swissprot/Uniprot and PDB databases.

To make the annotation more meaningful, the sequences were given GO IDs using InterProScan. These identifiers further helped classifying genes on the basis of: **i)** molecular function, which is related to the function performed by the gene that is given GO identifier; **ii**) biological process, which describes what biological process the gene is involved in; **iii)** cellular component, which is related to anatomy of the cell and describes where in the cell the gene performs its functions. Since all these classes represent different aspects of a gene, therefore, the same gene can be found in more than one classification, and so can have one or more than one identifier. Out of 246 genes, 180 genes were assigned 278 GO identifiers. On the basis of GO identifiers, the sequences were categorized as relating to molecular function, biological process, and cellular component. GO terms relating to biological process amounted to 85, 160 to molecular function, and 33 relating to cellular component. Sub-classification of GO terms is shown in Figure 16. **(**All the annotations are given in the excel file 'Annotations'. The sequences of the annotated CDS can be provided on request by contacting Tonje Marita Bjerken Heggeset from SINTEF.)

It should be noted that these 246 CDS are only those for which either no or insignificant sequence coverage and similarity was found in the old annotation. This means that many CDS

were excluded which had high sequence coverage and similarity but were not identical. It's quite possible that many of those excluded CDS were not actually the same gene but another copy of the already annotated genes, or possibly a closely related gene.

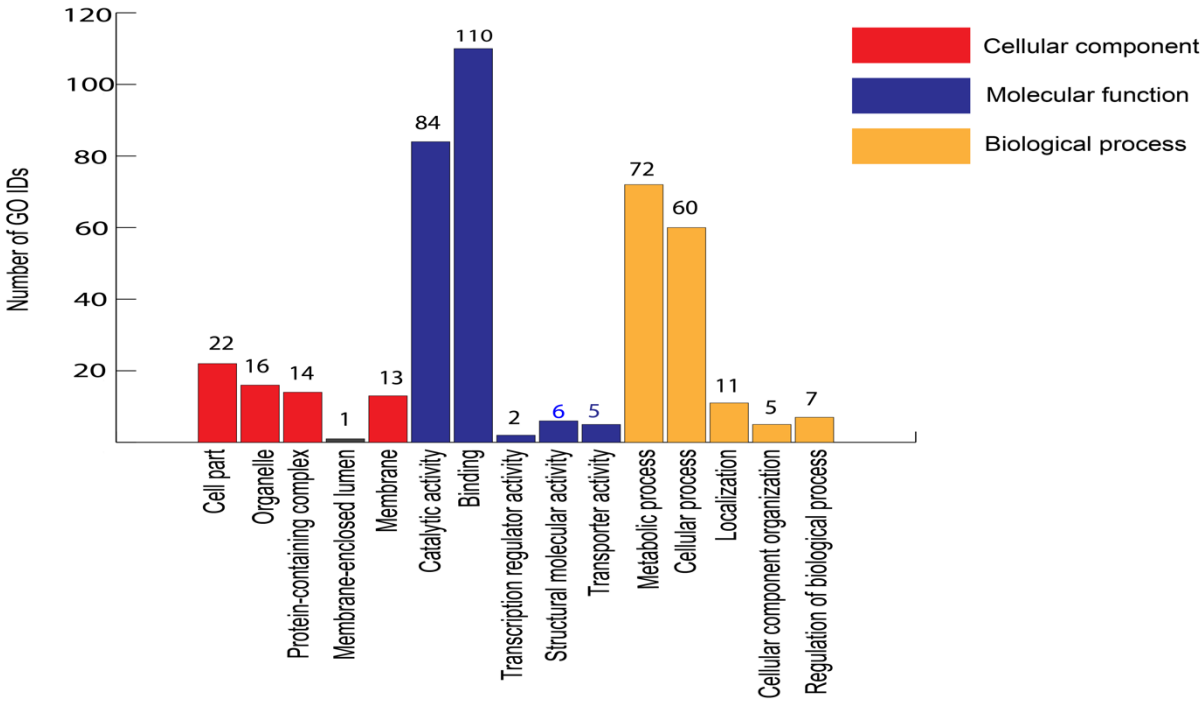


Figure 16. ***Sub-classification of Gene Ontology terms***

3.4.3. Manual Functional Annotation

The purpose of manual functional annotation was to determine what the sequences code for and also examine if the accorded biological information is correct. Candidate 453 CDS were blast-searched against following databases: Non-reductant Protein Sequence (NRDB), Reference Proteins (RefSeq), Protein Data Bank (PDB), and Swiss-Prot. The results of the blasted sequences were further sieved on the following basis: ***a)*** if the hits for a sequence are only against non-curated data base such as Gene Bank; ***b)*** if a sequence aligned against different proteins in different data bases; ***c)*** if the hits for a sequence were insignificant, for instance low sequence similarity or high e-value. No threshold of BLAST statistics was set for the filtration or assigning functions. The reason was that it is quite possible that a CDS aligns against a protein sequence in one database with relatively low similarity but aligns against a homologue of the same protein with higher similarity in another database. Therefore, while filtering and assigning a function to CDS, it was the results against all databases was kept in regard.

Consequently, out of 453 candidate CDS, only 92 passed the filtration. The filtered 92 sequences were further subject to searched against the Pfam database, which is a database of protein families. The purpose was to inspect if the candidate CDSs contain the domains of the proteins against which they aligned in the rest of the databases**.** Additionally, experimental proof for the proteins against which the query sequences aligned was also searched for. All the annotated CDS are given in the excel file 'Annotations'. (The sequences of the annotated CDS can be provided on request by contacting Tonje Marita Bjerken Heggeset from SINTEF.)

To make annotations more meaningful, the genes were assigned Gene Ontology (GO) identifiers. In total, 73 proteins were assigned 118 distinct GO terms. 60 genes were classified as related to molecular function, 43 genes to biological process, and 15 genes to cellular component. Further division of molecular functions, biological processes, and cellular components is shown in Figure 17.

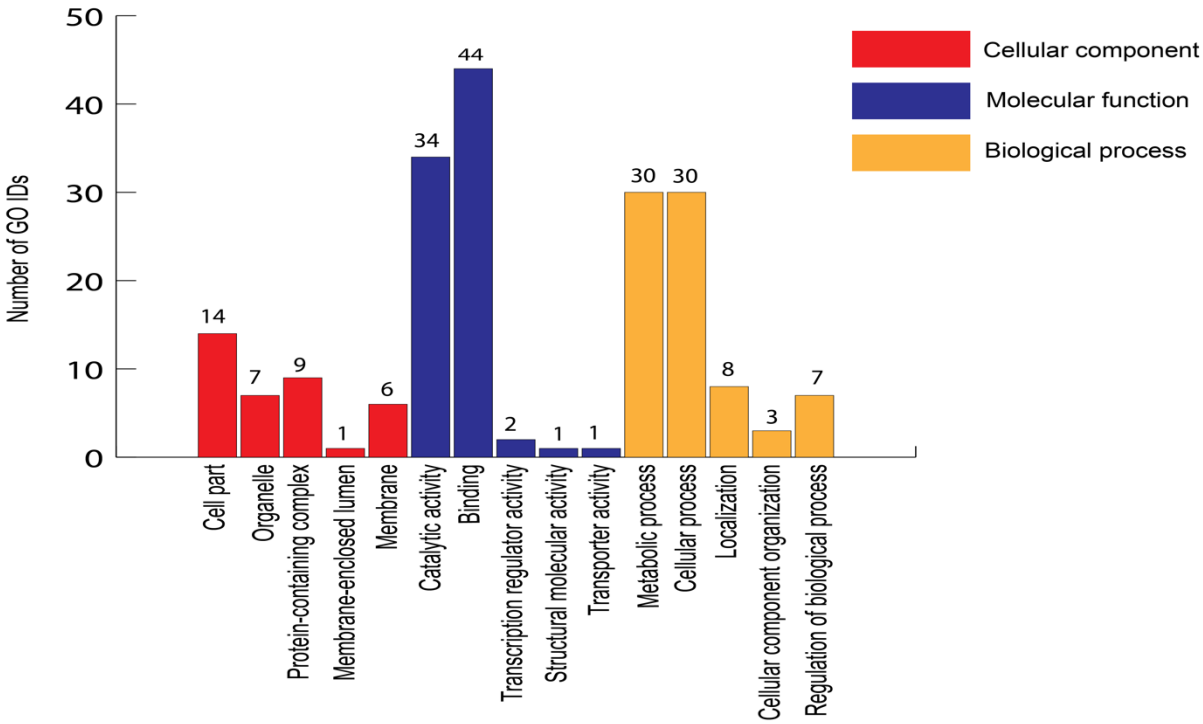


Figure 17. ***Sub-classification of Gene Ontology terms***

In total, 30 genes were annotated as enzymes. No genes were found relating to fatty acid elongation or PUFA synthesis. 21 out 92 genes annotated were also present in the list of genes that were subject to automatic annotation. The rest of the genes are probably those which got

filtered in the automatic annotation process as they aligned against genes of similar, but different sequences in the published annotation of *Aurantiochytrium* sp. T66. The fact that many of the automatically annotated genes of the first 15 contigs were not present in the manually annotated is because many of those genes could not pass the filteration process for manual annotation as described above. Copies of 4 out of 92 genes were in the list of manually annotated genes by Heggeset et al., (2019). The four genes are: Phosphoglycerate mutase (T66001704.1), Dihydrolipoyl dehydrogenase (T66008367.1), Hydroxyacyl-CoA dehydrogenase trifunctional (T66005991.15), and Ornithine transporter, mitochondrial (T66007710.1). Copies of 27 out of 92 genes manually annotated in this study were identified in automatic annotation by Heggeset et al., (2019). Here, it should be noted the word being used is 'copies', instead of claiming that those genes had been annotated previously. This is because while extracting CDS during structural annotation, as described in the section 2.2.6.1**,** only those CDS were extracted which had not been annotated previously.

Instead of manually extracting CDS for manual functional annotation, it could be said that the CDS filtered for automatic functional annotation could also be subject to manual annotation. However, as mentioned earlier, for automatic functional annotation, only those CDS were subject to annotation for which no sequence at all was found in the published annotation; there were many CDS for which highly similar sequences were found, which might have been similar but different genes. In manual extraction, those CDS could also be extracted which got filtered in the process of automatic annotation.

Instead of comparing blast results from different databases, what could be done was to blast different databases, select the best hit for all sequences from the databases, and calculate the percentage of how many sequences were annotated from each database. However, this would have been unavailing in a sense; many a times, the best hit in a database, especially the one not manually curated, is of hypothetical protein, for instance, many of the top hits were of *H.fermentalgiana'*s hypothetical protein; such annotations could not have been very serviceable; therefore, results from different databases were compared to filter any futile results and to append only worthwhile information to the sequences.

An argument could be put forward that since it is the manually curated database which would eventually be the decisive factor in the function assignment, why bother searching in non-

curated database? Manually curated databases do not contain adequate material from the groups doing research on the Thraustochytriaceae family; doing research only in manually curated databases would give sequence alignments organisms of distant classes; on the other hand, non-curated databases, such as GenBank/GenPept, accommodate unreviewed submissions from individual groups. In this study, annotations from non-curated were not made the basis of annotation, as sequences showing hits only against non-curated were filtered; rather, it was used to complement the annotations from manually curated database as the results from the former can be from closely related organism.

3.4.4. Identification of Complete sequences of PUFA synthase subunits, Phosphopantetheinyl transferase *pfaD*, and Glutamine synthase

Previously, Heggeset et al., (2019) identified *pfaA, pfaB, pfaC,* and phosphopantetheinyl transferase (*pfaD*). However, the gene sequences of *pfaA, pfaC,* and *pfaD* were found to be partial. Glutamine synthetase gene identified was found to be split between two contigs. Due to the high similarity between *Thraustochytrium* sp. ATCC 26185 (Zhao et al., 2016) and T66, *pfaA, pfaC*, and *pfaD* were reconstructed by Heggeset et al., (2019) by adapting the sequences from the genes of *Thraustochytrium* sp. ATCC 26185 (Zhao et al., 2016). To reconstitute glutamine synthase, the two split parts were concatenated on the basis of alignment against glutamine synthase gene of *H. fermentalgiana* FCC1311 and *Thraustochytrium* sp. ATCC 26185. *PfaA, pfaC, pfaD*, and glutamine synthase are given as T66011701, T66011702, T66011703, and T66011704 respectively in the annotation submitted by Heggeset et al., (2019). An attempt was made to detect these genes in the newly constructed assembly and check if the complete sequences of these genes can be located. Gene *pfaA*, *pfaC*, *pfaD* and Glutamine synthase from *Thraustochytrium* sp. ATCC 26185 (Genbank: KX651612.1, KX651615.1, KX651614.1, and 1-1047 of MUFY01000151, respectively) were BLAST-searched against the T66 assembly to identify these genes in T66 genome assembly. Results are shown in Table 7**.**

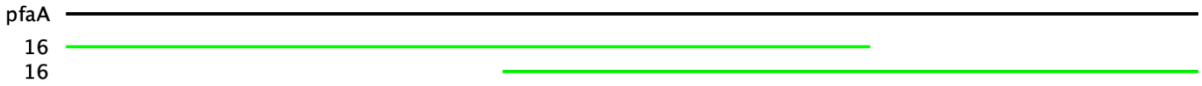
Table 7. **BLAST results of *pfaA, pfaC, pfaD,* and *GS of Thraustochytrium* sp. *ATC26185* against the current T66 genome assembly**

	PfaA	***PfaC***	***PfaD***	***GS***
Strand	–	–	+	+
Contig No	16	108	86	36
Contig region	462165…456172, 458558...453388	85672…81179	102815…103681	250513...249467
E-value	0	0	0	0
Identity (%)	98, 96	100	100	100

Pfa: *Polyunsaturated fatty acid synthase.* ***GS****: Glutamine Synthase. Contig No: Contig number.*

As is evident from Table 7, complete sequences of gene *pfaC, pfaD* and glutamine synthase were identified. The alignment of *pfaA* in the current *Aurantiochytrium sp.* T66 genome assembly was split between in two bits. However, the splits were not flanked by nucleotides, rather they overlapped with each other. The alignment of *pfaA* gene of *Thraustochytrium* sp. ATCC26185 against the T66 assembly is shown in Figure 18

Figure 18. **Alignment of *pfaA of Thraustochytrium* sp. *ATCC26185* against the current T66 genome assembly**



'The number '16' indicates the contig number.

The overlapping of the two fragments gave an indication that there might be more repeat units in the *pfaA* gene of T66 than in the *pfaA* gene of *Thraustochytrium* sp. ATCC26185, as *PfaA* gene tend to have repeating units encoding the ACP domain, and these repeating units vary between different organisms (Jiang et al., 2008). *PfaA* gene of *Aurantiochytrium sp.* T66 and *Thraustochytrium* sp. ATCC26185 were BLAST-searched in the Pfam database to see the domain structure of PUFA synthase subunit A. It was found that the number of ACP domains in the PUFA synthase subunit A of *Thraustochytrium* sp. ATCC26185 are eight, whereas that in *Aurantiochytrium sp.* T66 are nine. The domain structure *pfaA* of *Thraustochytrium* sp. ATCC26185 and T66 are shown in Figure 19. This means that the annotation of *pfaA* gene of T66 by Heggeset et al., (2019) was not correct as it was based on *pfaA* gene of *Thraustochytrium* sp. ATCC26185. (The complete sequences of *pfaA, pfaC, pfaD,* and

glutamine synthase of T66 can be provided on request by contacting Tonje Marita Bjerken Heggeset from SINTEF)

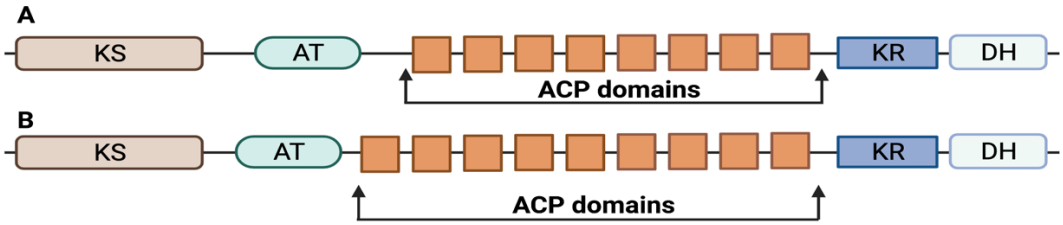


Figure 19. ***Domain structure of PUFA synthase subunit A*** *(Created with BioRender.com).* **A** shows the domain structure of *Thraustochytrium* sp. ATCC26185, and **B** shows the domain structure of *Aurantiochytrium* sp. T66. KS: β-ketoacyl-ACP synthase (KS). AT: Acetyl-CoA-ACP transacylase (AT). ACP: Acyl carrier protein. KR: β-ketoacyl-ACP reductase.

3.4.5. ATP-citrate synthase (ACL)

As mentioned in the section 1.1.3.4, ACL is an enzyme which catalyzes the conversion of citrate into oxaloacetate and acetyl-CoA, and is important for fatty acid biosynthesis. ACL has been shown to be present in thraustochytrids (Meesapyodsuk and Qiu, 2016; Nazir et al., 2020; Ren et al., 2009; Chang et al., 2013). Since the gene is present in other thraustochytrids, it is not unreasonable to assume the high probability of its presence in strain T66 as well. However, during the genome annotation of *Aurantiochytrium.* sp T66 by Heggeset et al., (2019), ACL could not be detected. One of the explanations that the ACL gene remained undetected could be that the assembly constructed was not good enough. As shown in Table Table 3 in section 3.3, the assembly constructed by Heggeset et al., (2019) had 5112477 unknown bases (Number of Ns). Therefore, it is quite possible that the *ACL* is present was present in those undetermined bases, which have been resolved in the current assembly. Thus, the *ACL* gene from *Schizochytrium aggregatum* ATCC 28209 was BLAST-searched against the current T66 assembly. Table 8 shows the results. The only significant hit was against the Succinyl-CoA ligase Subunit A (Contig: 57, Region: 248588…249030), which *ACL* belongs belong to the same protein family as *ACL* and shares a similar mechanism to phosphorylate a histidine residue in their active sites. Succinyl-CoA ligase Subunit A has already been annotated before as T66006817.1 by Heggeset et al., (2019).

Table 8. **BLAST results of the gene ACL of *Schizochytrium aggregatum ATC 28209* against the current T66 genome assembly**

Strand	Contig No	Contig region	Locus	E-value	% identity
+	157	248588.. 249030	T66006817.1	1 x 10^{-19}	58

ACL: *ATP-dependent citrate lyase.* ***T66006817.1:*** *Succinyl-CoA ligase subunit A.*

3.4.3. Mitochondrial Genome Identification

Mitochondrial genome of T66 had not been identified before. Here, the purpose is to identify and annotate mitochondrial genome (MG) of the organism. To identify the mitochondrial genome, the two already submitted MG of *Aurantiochytrium acetophilum* sp HS-399 (GenBank accession number MH259702) and *Schizochytrium* sp. TIO1101 (GenBank accession number KU183024) were BLAST against the entire assembly. Since they are from the same family, it was expected that these strains would "fish out" the MG of T66 from the entire genome assembly. Both MGs used as BLAST queries aligned against the contig 157, which was a circular contig, of the T66 genome assembly. The sequence was further chosen for annotation. The size of the MG of *Aurantiochytrium* sp. T66 was 55,160 bp

In total 56 genes were detected: 33 protein-coding genes (PCGs), 21 transfer RNAs (tRNAs), and 2 ribosomal RNAs (rRNAs). The 33 PCGs comprised 15 ribosomal protein-coding genes (*rpl* and *rps*), 10 NADH dehydrogenase genes (*nad*), three ATP synthase genes (*atp*), three cytochrome c oxidase genes (*cox*), one cytochrome b gene (*cob*) and one sec-independent protein translocase component (*tatC*). (The sequence and annotations of the MG genome can be provided on request by contacting Tonje Marita Bjerken Heggeset from SINTEF.)

..

4. Discussion

The genome annotation of *Aurantiochytrium* sp. T66 has been performed previously by Heggeset et al., (2019), which was based on the assembly that was constructed using shortgun and mate-pair reads from NGS platforms. Short reads, though have high accuracy, the Q-score usually being above 30 (Ma et al., 2019; Fox et al., 2014), however, since the reads are short, it is difficult to resolve repetitive sequences in the genome, resulting in gaps in the assembly. Though mate-pair reads can lessen the problem to an extent, however, the problem of gaps persists due to other drawbacks of NGS platforms. NGS platforms are prone to GC-biasedness due to the protocols involved in preparing the sample for sequencing. This results in low-converge of GC-rich region, contributing negatively to the contiguity of the assembly. Furthermore, the problem of 'dephasing' in NGS platforms contributes to the errors in the assembly. Owing to these limitations, there is high probability that in the annotation of the assembly constructed previously, some genes might have missed the annotation process. In the anticipation that resolving the above-mentioned issues would help annotate those genes, the genome of *Aurantiochytrium* sp. T66 was resequenced in this study. The genome was sequenced using ONT's MinION sequencer, which generates long reads. Long reads, though far less accurate than the short reads from NGS platforms, can help resolve repetitive sequences in the genome and improve contiguity. Since the protocols of MinION sequencing used in this study for preparing the sample does not involve amplifying PCR, it could also reduce the effect of GC-biasedness. To attain the benefits of shotgun from NGS platforms and MinION's long reads, a hybrid assembly was built, which was then subject to genome annotation.

4.1 The Quality of Sequencing and Genome Assembly

The most cited drawback of TGS platforms, including MinION, is the high error rate. The accuracy of MinION has ranged from median Q-score of 6 to12 in different studies (Plesivkova et al., 2019). The difference in the accuracy is mainly because of the improvement in the technical details of MinION sequencing over time. In the initial period of MinION release, the accuracy was low. With the refinement of MinION sequencing technology over time, the accuracy of the sequencing improved. In this study, bases were called with the median Q-score of 7.8 (after filtering the reads with the threshold Q-score at 8), which equals the error rate of 16.6%. The accuracy of the MinION reads in this study, therefore, can be considered to be typical. However, since the intention was to build hybrid assembly using Illumina's short reads,

which have accuracy of 99.99%, and the MinION reads, the issue of low accuracy of the MinION reads was not a problem as such. It is the feature of MinION to generate long reads that was benefitted from in this study.

The median read length varies between different studies because of the same reason the accuracy varies. There have been studies with the mean read length of 20 kb with MinION sequencing (Tyson et al., 2018), and also with the mean read length of 2.2 kb (Brancaccio et al., 2021). The mean read length in this study was around 3.78 (after filtering the reads with the threshold Q-score at 8). So, the quality of the sequencing in this study in terms of average read length cannot be considered exceptional. However, as mentioned earlier, the total number of 10 kb reads generated provide 96x coverage. Therefore, not having exceptionally good mean length was not a problem. The highest read length with MinION over 100 kb is routinely reported. Also, in this study, the highest read length generated was 134 kb (highest read length before the filtration was 321 kb). Even though the highest read length in this study is not exceptional if it were to be compared with the literature, it is the highest read length attained in this lab with MinION.

In comparison to the contig N50 of 12 kb and L50 of 849 of the published genome assembly of *Aurantiochytrium* sp. T66, N50 and L50 of the current assembly is around 332 kb and 40, respectively (see Table 3). Evidently, the contiguity of the current assembly is higher than the published assembly of T66. If the quality of the assembly were to be gauged on the basis of contiguity, the approach of constructing hybrid assembly using Illumina short-reads and MinION reads has proven to be quite effective. Comparing the current assembly to the published assemblies of other thraustochytrids, out of the 11 assemblies of thraustochytrids submitted at NCBI, only 1 assembly had better contiguity, which was of *Schizochytrium* sp. TIO01 that had contig N50 of 2.8 Mbp, and contig L50 of 9 (Hu et al., 2020). As mentioned earlier the aim was to annotate the genes which could have been possibly missed in the previous annotation because of the gaps in the assembly**.** The current assembly certainly has improved contiguity, and thus, has fewer number of gaps. Whether the assembly proved to be helpful in finding putatively unannotated genes, or if there were unannotated genes at all, would depend on the outcome of the annotation process in this study.

4.2. Genome Annotation

Putative genes in the current assembly were BLAST-searched against the list of genes annotated by Heggeset et al., (2019) to find the genes which might have missed the annotation process previously. In total, 246 putative genes in the current assembly were identified which were not present in the annotated list of genes. As mentioned in the section 3.4.2, these 246 genes are those sequences for which either no or insignificant similarity was found in the previously annotated list of genes. The rest of the genes were not all 100 % identical, though all the genes had identity at least above 65 % identity. Although it is quite possible that the difference in identity of the aligned sequences is because of the correction of the assembly and they were actually the same genes, however, it is also quite probable that many of putative genes in the current assembly which aligned and showed high percentage identity to the genes in the old annotation were closely related but different genes. The manual annotation did address the issue, identifying 71 genes which could not be identified in automatic annotation; however, it was only for the first 15 contigs which make 24% of the genome. What could be done was to annotate all the putative genes which showed significant similarity to the genes in the old annotation and compare their annotations. In this way, it could be found whether the putative genes in the current assembly are the same as genes against which they aligned or are they closely related but different genes.

In this study, complete sequences of *pfaA, pfaC, pfaD,* and *GS* were also identified. The sequences of *pfaA, pfaC* and *pfaD* identified in the previous assembly were incomplete, highlighting the errors in the assembly. *GS,* though complete, was split between two contigs, highlighting the mis-assemblies. As mentioned in the section 1.3.5, *pfaA* of different organisms tend to have varying number of ACP domains. The incorrect annotation of *pfaA* gene in the previous assembly was probably due to the presence of varying number of repeats in the gene. The ability to correctly identify the number of repeats in the *pfaA* is one of the examples how hybrid assemblies using the data from NGS platforms and TGS platforms can help.

Another important gene for lipid synthesis is ACL. It provides acetyl-CoA, which is a precursor for lipid synthesis. *ACL*, which could not be detected in *Aurantiochytrium* sp. T66 previously, was not found in this study either. Though the gene has been detected in some thraustochytrids, however, apparently, not all thraustochytrids possess it (Heggeset et al., 2019). So, it is not only the genome of *Aurantiochytrium* sp. T66 which does not seem to contain *ACL* gene. This

suggests that *Aurantiochytrium* sp. T66, and those thraustochytrids in which *ACL* could not be detected, might have some other gene(s) that provide(s) acetyl-CoA for lipid synthesis.

Regarding the MG of *Aurantiochytrium* sp. T66, the identified sequence of MG was 55,160 bp in length. A large part of this 55,160 bp sequence is unannotated, where no genes were found. This is region between 9,500 and 35,000. When Illumina's shotgun reads were mapped to this region, it was found that the reads belonged to Segment 2 (see section 2.2.7.1.1 for more detail about Segments), or in other words, the reads were non-specific. The size of the MG of *Aurantiochytrium acetophilum* sp HS-399 and *Schizochytrium* sp. TIO1101 being 30,886 and 31,494, respectively, also indicates of the oddity of the sequence. Considering that the MG of other closely related species mentioned above is shorter by almost the same length as the length of this odd sequence, it very likely that this sequence is either contamination from some other species, or is mis-assembled. But it is certain that the sequence is not part of the MG of *Aurantiochytrium sp.*T66. What could be done, and could not be done in this study, was to BLAST-search this 25,000 bp sequence against different databases to find any genes, and in the case of the presence of the genes which are not likely to be present in the MG or not finding any genes at all, the sequence should be removed.

4.3. Conclusions

By constructing a hybrid assembly with MinION's and Illumina's shotgun reads, this study aimed at improving the genome assembly and annotation of *Aurantiochytrium* sp. T66. As explained earlier, the quality of the assembly has certainly improved. Regarding the annotation, the current annotation process added 317 genes, 246 through automatic annotation and 71 through manual annual annotation, to the published annotation of *Aurantiochytrium* sp. T66, in which 11,683 genes have been identified. In terms of annotation of the entire genome, it could be said that the current annotation improvement on top of the existing annotation has been mild, and that the published annotation of *Aurantiochytrium* sp. T66 was quite complete. However, the current annotation process also helped in identifying complete sequences of genes relating to PUFA synthase complex, which are core of the most studied pathway in thraustochytrids, i.e. lipid synthesis. Even though it is not novel discovery in that sense, as PUFA synthase subunits have been identified previously in thraustochytrids, it does build upon or improve the genome annotation of the organism under study in this lab, i.e.

Aurantiochytrium sp. T66. Furthermore, mitochondrial genome was also identified. If the overall annotation of the organism was to be looked at, this study has definitely been helpful.

Even though the current study has been worthwhile, there is still room for improvement. For instance, even though the current T66 assembly is significantly of better quality than the previous assembly and is also better than the most thraustochytrids assemblies built so far, its quality can be further enhanced. The contiguity of the assembly can be further improved by closing the remaining gaps in the assembly, which could not be done in this study due to time restraints. For this purpose, there are different softwares available, such CLC finishing module of CLC Genomics Workbench. Resolving the existing gaps in the assembly might help find more genes, as it did in this study. By doing so, the genome annotation of *Aurantiochytrium* sp. T66 can be further improved.

5. References

Aasen, I. M., Ertesvåg, H., Heggeset, T. M. B., Liu, B., Brautaset, T., Vadstein, O. & Ellingsen, T. E. 2016. Thraustochytrids as production organisms for docosahexaenoic acid (DHA), squalene, and carotenoids. *Applied Microbiology and Biotechnology,* 100(10)**,** pp 4309-4321.

Abril, J. F. & Castellano, S. 2019. Genome Annotation. *In:* Ranganathan, S., Gribskov, M., Nakai, K. & Schönbach, C. (eds.) *Encyclopedia of Bioinformatics and Computational Biology.* Oxford: Academic Press.

Afshordel, S., Hagl, S., Werner, D., Röhner, N., Kögel, D., Bazan, N. G. & Eckert, G. P. 2015. Omega-3 polyunsaturated fatty acids improve mitochondrial dysfunction in brain aging – Impact of Bcl-2 and NPD-1 like metabolites. *Prostaglandins, Leukotrienes and Essential Fatty Acids,* 92(23-31.

Aird, D., Ross, M. G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D. B., Nusbaum, C. & Gnirke, A. 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biology,* 12(2)**,** pp R18.

Aki, T., Hachida, K., Yoshinaga, M., Katai, Y., Yamasaki, T., Kawamoto, S., Kakizono, T., Maoka, T., Shigeta, S., Suzuki, O. & Ono, K. 2003. Thraustochytrid as a potential source of carotenoids. *Journal of the American Oil Chemists' Society,* 80(8)**,** pp 789.

Antipov, D., Korobeynikov, A., McLean, J. S. & Pevzner, P. A. 2016. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics (Oxford, England),* 32(7)**,** pp 1009-1015.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. & Sherlock, G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature genetics,* 25(1)**,** pp 25-29.

Bajpai, P. K., Bajpai, P. & Ward, O. P. 1991. Optimization of production of docosahexaenoic acid (DHA) byThraustochytrium aureum ATCC 34304. *Journal of the American Oil Chemists' Society,* 68(7)**,** pp 509-514.

Baker, M. 2012. De novo genome assembly: what every biologist should know. *Nature Methods,* 9(4)**,** pp 333-337.

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., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A. & Pevzner, P. A. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of computational biology : a journal of computational molecular cell biology,* 19(5)**,** pp 455-477.

Barclay, W., Weaver, C., Metz, J. & Hansen, J. 2010. Development of a Docosahexaenoic Acid Production Technology Using Schizochytrium: Historical Perspective and Update. *Single Cell Oils,* 75-96.

Barnhart, M. M. & Chapman, M. R. 2006. Curli biogenesis and function. *Annual review of microbiology,* 60(131-147.

Beld, J., Sonnenschein, E. C., Vickery, C. R., Noel, J. P. & Burkart, M. D. 2014. The phosphopantetheinyl transferases: catalysis of a post-translational modification crucial for life. *Natural Product Reports,* 31(1)**,** pp 61-108.

Berglund, E. C., Kiialainen, A. & Syvänen, A.-C. 2011. Next-generation sequencing technologies and applications for human genetic history and forensics. *Investigative genetics,* 2(23-23.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. 2000. The Protein Data Bank. *Nucleic Acids Research,* 28(1)**,** pp 235-242.

Bhattacharyya, A. 2009. Genome Sequence Databases: Annotation. *In:* Schaechter, M. (ed.) *Encyclopedia of Microbiology (Third Edition).* Oxford: Academic Press.

Bleidorn, C. 2016. Third generation sequencing: technology and its potential impact on evolutionary biodiversity research. *Systematics and Biodiversity,* 14(1)**,** pp 1-8.

Bolisetty, M. T., Rajadinakaran, G. & Graveley, B. R. 2015. Determining exon connectivity in complex mRNAs by nanopore sequencing. *Genome Biology,* 16(1)**,** pp 204.

Brancaccio, R. N., Robitaille, A., Dutta, S., Rollison, D. E., Tommasino, M. & Gheit, T. 2021. MinION nanopore sequencing and assembly of a complete human papillomavirus genome. *Journal of Virological Methods,* 294(114180.

Branton, D., Deamer, D. W., Marziali, A., Bayley, H., Benner, S. A., Butler, T., Di Ventra, M., Garaj, S., Hibbs, A., Huang, X., Jovanovich, S. B., Krstic, P. S., Lindsay, S., Ling, X. S., Mastrangelo, C. H., Meller, A., Oliver, J. S., Pershin, Y. V., Ramsey, J. M., Riehn, R., Soni, G. V., Tabard-Cossa, V., Wanunu, M., Wiggin, M. & Schloss, J. A. 2008. The potential and challenges of nanopore sequencing. *Nature Biotechnology,* 26(10)**,** pp 1146-1153.

Carter, J.-M. & Hussain, S. 2017. Robust long-read native DNA sequencing using the ONT CsgG Nanopore system.

Cenacchi, T., Bertoldin, T., Farina, C., Fiori, M. G., Crepaldi, G., Azzini, C. F., Girardello, R., Bagozzi, B., Garuti, R., Vivaldi, P., Belloni, G., Bordin, A., Durando, M., Lo Storto, M., Bertoni, L., Battistoni, A., Cacace, C., Arduini, P., Bonini, A., Caramia, M. P., Vaglieri, G., Brusomini, A., Donà, G., March, A., Campi, N., Cannas, P., Casson, F., Cavallarin, G., Delia, M., Cristianini, G., Louvier, O., Mello, F., Fameli, R., Urbani de Gheltoff, N., De Candia, O., Nante, G., Cattoni, C., Forte, P. L., Loreggian, M., Targa, A., Mansoldo, G., Noro, G., Meggio, A., Pedrazzi, F., Bonmartini, F., Ruggiano, C., Peruzza, M., Olivari, G., Recaldin, E., Bellunato, C., Rigo, G., Marin, M., Marinangeli,

L., Saracino, A., Miceli, O., Lovo, G., Scarpa, R., Battistello, L., Tomat, E., Bernava, B., Olivo, P., Verga, G., Merli, G., Zerman, A. M., Crivellaro, R., Vozza, A., Ziliotto, G. R., Favaretto, V. & Allegro, L. 1993. Cognitive decline in the elderly: A double-blind, placebo- controlled multicenter study on efficacy of phosphatidylserine administration. *Aging Clinical and Experimental Research,* 5(2)**,** pp 123-133.

Chang, G., Luo, Z., Gu, S., Wu, Q., Chang, M. & Wang, X. 2013. Fatty acid shifts and metabolic activity changes of Schizochytrium sp. S31 cultured on glycerol. *Bioresource Technology,* 142(255-260.

Chen, G., Shi, T. & Shi, L. 2017. Characterizing and annotating the genome using RNA-seq data. *Science China Life Sciences,* 60(2)**,** pp 116-125.

Choudhuri, S. 2014. Additional Bioinformatic Analyses Involving Nucleic-Acid Sequences.

Churko, J. M., Mantalas, G. L., Snyder, M. P. & Wu, J. C. 2013. Overview of high throughput sequencing technologies to elucidate molecular pathways in cardiovascular diseases. *Circulation research,* 112(12)**,** pp 1613-1623.

Consortium, T. U. 2020. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research,* 49(D1)**,** pp D480-D489.

De Coster, W., D'Hert, S., Schultz, D. T., Cruts, M. & Van Broeckhoven, C. 2018. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics,* 34(15)**,** pp 2666-2669.

de Lannoy, C. a. d. R., D and Risse, J 2017. The long reads ahead: de novo genome assembly using the MinION [version 2; peer review: 2 approved]. *F1000Research,* 6(1083)**,** pp.

de Sá, P. H. C. G., Guimarães, L. C., das Graças, D. A., de Oliveira Veras, A. A., Barh, D., Azevedo, V., da Costa da Silva, A. L. & Ramos, R. T. J. 2018. Chapter 11 - Next-Generation Sequencing and Data Analysis: Strategies, Tools, Pipelines and Protocols. *In:* Barh, D. & Azevedo, V. (eds.) *Omics Technologies and Bio-Engineering.* Academic Press.

Deamer, D., Akeson, M. & Branton, D. 2016. Three decades of nanopore sequencing. *Nature Biotechnology,* 34(5)**,** pp 518-524.

Dellero, Y., Rose, S., Metton, C., Morabito, C., Lupette, J., Jouhet, J., Maréchal, E., Rébeillé, F. & Amato, A. 2018. Ecophysiology and lipid dynamics of a eukaryotic mangrove decomposer. *Environmental Microbiology,* 20(8)**,** pp 3057-3068.

Desai, A., Marwah, V. S., Yadav, A., Jha, V., Dhaygude, K., Bangar, U., Kulkarni, V. & Jere, A. 2013. Identification of optimum sequencing depth especially for de novo genome assembly of small genomes using next generation sequencing data. *PloS one,* 8(4)**,** pp e60204-e60204.

Dijk, E. L., van, Jaszczyszyn, Y., Naquin, D. & Thermes, C. 2018. The Third Revolution in Sequencing Technology. *Trends in Genetics,* 34(9)**,** pp 666--681.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M. & Gingeras, T. R. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England),* 29(1)**,** pp 15-21.

Dohm, J. C., Lottaz, C., Borodina, T. & Himmelbauer, H. 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic acids research,* 36(16)**,** pp e105-e105.

Erlich, Y., Mitra, P. P., delaBastide, M., McCombie, W. R. & Hannon, G. J. 2008. Alta-Cyclic: a self-optimizing base caller for next-generation sequencing. *Nat Methods,* 5(8)**,** pp 679-82.

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J. & Punta, M. 2014. Pfam: the protein families database. *Nucleic acids research,* 42(Database issue)**,** pp D222-D230.

Fox, E. J., Reid-Bayliss, K. S., Emond, M. J. & Loeb, L. A. 2014. Accuracy of Next Generation Sequencing Platforms. *Next generation, sequencing & applications,* 1(1000106.

Foxman, B. 2012. Chapter 5 - A Primer of Molecular Biology. *In:* Foxman, B. (ed.) *Molecular Tools and Infectious Disease Epidemiology.* San Diego: Academic Press.

Gao, G. & Smith, D. I. 2015. Mate-Pair Sequencing as a Powerful Clinical Tool for the Characterization of Cancers with a DNA Viral Etiology. *Viruses,* 7(8)**,** pp 4507-4528.

Gene Ontology, C. 2021. The Gene Ontology resource: enriching a GOld mine. *Nucleic acids research,* 49(D1)**,** pp D325-D334.

Gerster, H. 1998. Can adults adequately convert ??-linolenic acid (18:3n-3) to eicosapentaenoic acid (20:5n-3) and docosahexaenoic acid (22:6n-3)? *International journal for vitamin and nutrition research. Internationale Zeitschrift für Vitamin- und Ernährungsforschung. Journal international de vitaminologie et de nutrition,* 68(159-73.

Giani, A. M., Gallo, G. R., Gianfranceschi, L. & Formenti, G. 2019. Long walk to genomics: History and current approaches to genome sequencing and assembly. *Computational and structural biotechnology journal,* 18(9-19.

Giordano, D., Coppola, D., Russo, R., Denaro, R., Giuliano, L., Lauro, F. M., di Prisco, G. & Verde, C. 2015. Chapter Four - Marine Microbial Secondary Metabolites: Pathways, Evolution and Physiological Roles. *In:* Poole, R. K. (ed.) *Advances in Microbial Physiology.* Academic Press.

Goodwin, S., McPherson, J. D. & McCombie, W. R. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics,* 17(6)**,** pp 333-351.

Goyal, P., Krasteva, P. V., Van Gerven, N., Gubellini, F., Van den Broeck, I., Troupiotis-Tsaïlaki, A., Jonckheere, W., Péhau-Arnaudet, G., Pinkner, J. S., Chapman, M. R., Hultgren, S. J., Howorka, S., Fronzes, R. & Remaut, H. 2014. Structural and

mechanistic insights into the bacterial amyloid secretion channel CsgG. *Nature,* 516(7530)**,** pp 250-253.

Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics (Oxford, England),* 29(8)**,** pp 1072-1075.

Hampton, O. A., English, A. C., Wang, M., Salerno, W. J., Liu, Y., Muzny, D. M., Han, Y., Wheeler, D. A., Worley, K. C., Lupski, J. R. & Gibbs, R. A. 2017. SVachra: a tool to identify genomic structural variation in mate pair sequencing data containing inward and outward facing reads. *BMC Genomics,* 18(6)**,** pp 691.

Hauvermale, A., Kuner, J., Rosenzweig, B., Guerra, D., Diltz, S. & Metz, J. G. 2006. Fatty acid production in Schizochytrium sp.: Involvement of a polyunsaturated fatty acid synthase and a type I fatty acid synthase. *Lipids,* 41(8)**,** pp 739-747.

Heggeset, T. M. B., Ertesvåg, H., Liu, B., Ellingsen, T. E., Vadstein, O. & Aasen, I. M. 2019. Lipid and DHA-production in Aurantiochytrium sp. – Responses to nitrogen starvation and oxygen limitation revealed by analyses of production kinetics and global transcriptomes. *Scientific Reports,* 9(1)**,** pp 19470.

Hopwood, D. A. & Sherman, D. H. 1990. MOLECULAR GENETICS OF POLYKETIDES AND ITS COMPARISON TO FATTY ACID BIOSYNTHESIS. *Annual Review of Genetics,* 24(1)**,** pp 37-62.

Hu, F., Clevenger, A. L., Zheng, P., Huang, Q. & Wang, Z. 2020. Low-temperature effects on docosahexaenoic acid biosynthesis in Schizochytrium sp. TIO01 and its proposed underlying mechanism. *Biotechnology for Biofuels,* 13(1)**,** pp 172.

Humann, J. L., Lee, T., Ficklin, S. & Main, D. 2019. Structural and Functional Annotation of Eukaryotic Genomes with GenSAS. *In:* Kollmar, M. (ed.) *Gene Prediction: Methods and Protocols.* New York, NY: Springer New York.

Jain, M., Fiddes, I. T., Miga, K. H., Olsen, H. E., Paten, B. & Akeson, M. 2015. Improved data analysis for the MinION nanopore sequencer. *Nature Methods,* 12(4)**,** pp 351-356.

Jain, M., Olsen, H. E., Paten, B. & Akeson, M. 2016. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology,* 17(1)**,** pp 239.

Jiang, H., Zirkle, R., Metz, J. G., Braun, L., Richter, L., Van Lanen, S. G. & Shen, B. 2008. The Role of Tandem Acyl Carrier Protein Domains in Polyunsaturated Fatty Acid Biosynthesis. *Journal of the American Chemical Society,* 130(20)**,** pp 6336-6337.

Jiang, L.-h., Shi, Y., Wang, L.-s. & Yang, Z.-r. 2009. The influence of orally administered docosahexaenoic acid on cognitive ability in aged mice. *The Journal of Nutritional Biochemistry,* 20(9)**,** pp 735-741.

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R. & Hunter, S. 2014. InterProScan 5: genome-

scale protein function classification. *Bioinformatics (Oxford, England),* 30(9)**,** pp 1236-1240.

Juber, B. A., Jackson, K. H., Johnson, K. B., Harris, W. S. & Baack, M. L. 2017. Breast milk DHA levels may increase after informing women: a community-based cohort study from South Dakota USA. *International breastfeeding journal,* 12(7-7.

Kalyanaraman, A. 2011. Genome Assembly. *In:* Padua, D. (ed.) *Encyclopedia of Parallel Computing.* Boston, MA: Springer US.

Kanehisa, M. & Goto, S. 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research,* 28(1)**,** pp 27-30.

Khetan, D., Gupta, N., Chaudhary, R. & Shukla, J. S. 2019. Comparison of UV spectrometry and fluorometry-based methods for quantification of cell-free DNA in red cell components. *Asian journal of transfusion science,* 13(2)**,** pp 95-99.

Kircher, M. & Kelso, J. 2010. High-throughput DNA sequencing – concepts and limitations. *BioEssays,* 32(6)**,** pp 524-536.

Labrousse, V. F., Nadjar, A., Joffre, C., Costes, L., Aubert, A., Grégoire, S., Bretillon, L. & Layé, S. 2012. Short-term long chain omega3 diet protects from neuroinflammatory processes and memory impairment in aged mice. *PloS one,* 7(5)**,** pp e36861-e36861.

Lanfear, R., Schalamun, M., Kainer, D., Wang, W. & Schwessinger, B. 2019. MinIONQC: fast and simple quality control for MinION sequencing data. *Bioinformatics,* 35(3)**,** pp 523-525.

Laszlo, A. H., Derrington, I. M., Brinkerhoff, H., Langford, K. W., Nova, I. C., Samson, J. M., Bartlett, J. J., Pavlenok, M. & Gundlach, J. H. 2013. Detection and mapping of 5-methylcytosine and 5-hydroxymethylcytosine with nanopore MspA. *Proc Natl Acad Sci U S A,* 110(47)**,** pp 18904-9.

Laver, T., Harrison, J., O'Neill, P. A., Moore, K., Farbos, A., Paszkiewicz, K. & Studholme, D. J. 2015. Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomolecular Detection and Quantification,* 3(1-8.

Leger, A. & Leonardi, T. 2019. pycoQC, interactive quality control for Oxford Nanopore Sequencing. *Journal of Open Source Software,* 4(1236.

Liu, B., Ertesvåg, H., Aasen, I. M., Vadstein, O., Brautaset, T. & Heggeset, T. M. B. 2016. Draft genome sequence of the docosahexaenoic acid producing thraustochytrid Aurantiochytrium sp. T66. 8(115-116.

Loferer, H., Hammar, M. & Normark, S. 1997. Availability of the fibre subunit CsgA and the nucleator protein CsgB during assembly of fibronectin-binding curli is limited by the intracellular concentration of the novel lipoprotein CsgG. *Molecular Microbiology,* 26(1)**,** pp 11-23.

Lu, H., Giordano, F. & Ning, Z. 2016. Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics, Proteomics & Bioinformatics,* 14(5)**,** pp 265-279.

López-Malo, D., Arnal, E., Miranda, M., Johnsen-Soriano, S. & Romero, F. J. 2020. Chapter 16 - Antioxidative component of docosahexaenoic acid in the brain in diabetes. *In:* Preedy, V. R. (ed.) *Diabetes (Second Edition).* Academic Press.

Ma, X., Shao, Y., Tian, L., Flasch, D. A., Mulder, H. L., Edmonson, M. N., Liu, Y., Chen, X., Newman, S., Nakitandwe, J., Li, Y., Li, B., Shen, S., Wang, Z., Shurtleff, S., Robison, L. L., Levy, S., Easton, J. & Zhang, J. 2019. Analysis of error profiles in deep next-generation sequencing data. *Genome Biology,* 20(1)**,** pp 50.

Manchanda, N., Portwood, J. L., Woodhouse, M. R., Seetharam, A. S., Lawrence-Dill, C. J., Andorf, C. M. & Hufford, M. B. 2020. GenomeQC: a quality assessment tool for genome assemblies and gene structure annotations. *BMC Genomics,* 21(1)**,** pp 193.

Manrao, E. A., Derrington, I. M., Laszlo, A. H., Langford, K. W., Hopper, M. K., Gillgren, N., Pavlenok, M., Niederweis, M. & Gundlach, J. H. 2012. Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. *Nature Biotechnology,* 30(4)**,** pp 349-353.

Meesapyodsuk, D. & Qiu, X. 2016. Biosynthetic mechanism of very long chain polyunsaturated fatty acids in Thraustochytrium sp. 26185. *Journal of lipid research,* 57(10)**,** pp 1854-1864.

Metz, J. G., Roessler, P., Facciotti, D., Levering, C., Dittrich, F., Lassner, M., Valentine, R., Lardizabal, K., Domergue, F., Yamada, A., Yazawa, K., Knauf, V. & Browse, J. 2001. Production of Polyunsaturated Fatty Acids by Polyketide Synthases in Both Prokaryotes and Eukaryotes. *Science,* 293(5528)**,** pp 290.

Metzker, M. L. 2010. Sequencing technologies — the next generation. *Nature Reviews Genetics,* 11(1)**,** pp 31-46.

Meyer, B. J., Mann, N. J., Lewis, J. L., Milligan, G. C., Sinclair, A. J. & Howe, P. R. C. 2003. Dietary intakes and food sources of omega-6 and omega-3 polyunsaturated fatty acids. *Lipids,* 38(4)**,** pp 391-398.

Mikheyev, A. S. & Tin, M. M. Y. 2014. A first look at the Oxford Nanopore MinION sequencer. *Molecular Ecology Resources,* 14(6)**,** pp 1097-1102.

Miller, J. R., Zhou, P., Mudge, J., Gurtowski, J., Lee, H., Ramaraj, T., Walenz, B. P., Liu, J., Stupar, R. M., Denny, R., Song, L., Singh, N., Maron, L. G., McCouch, S. R., McCombie, W. R., Schatz, M. C., Tiffin, P., Young, N. D. & Silverstein, K. A. T. 2017. Hybrid assembly with long and short reads improves discovery of gene family expansions. *BMC Genomics,* 18(1)**,** pp 541.

Miyamoto, M., Motooka, D., Gotoh, K., Imai, T., Yoshitake, K., Goto, N., Iida, T., Yasunaga, T., Horii, T., Arakawa, K., Kasahara, M. & Nakamura, S. 2014. Performance comparison of second- and third-generation sequencers using a bacterial genome with two chromosomes. *BMC Genomics,* 15(1)**,** pp 699.

Morabito, C., Bournaud, C., Maës, C., Schuler, M., Aiese Cigliano, R., Dellero, Y., Maréchal, E., Amato, A. & Rébeillé, F. 2019. The lipid metabolism in thraustochytrids. *Progress in Lipid Research,* 76(101007.

Napier, J. A. 2002. Plumbing the depths of PUFA biosynthesis: a novel polyketide synthase-like pathway from marine organisms. *Trends in Plant Science,* 7(2)**,** pp 51-54.

Nazir, Y., Halim, H., Prabhakaran, P., Ren, X., Naz, T., Mohamed, H., Nosheen, S., Mustafa, K., Yang, W., Abdul Hamid, A. & Song, Y. 2020. Different Classes of Phytohormones Act Synergistically to Enhance the Growth, Lipid and DHA Biosynthetic Capacity of Aurantiochytrium sp. SW1. *Biomolecules,* 10(5)**,** pp 755.

Niederweis, M., Ehrt, S., Heinz, C., Klöcker, U., Karosi, S., Swiderek, K. M., Riley, L. W. & Benz, R. 1999. Cloning of the mspA gene encoding a porin from Mycobacterium smegmatis. *Molecular Microbiology,* 33(5)**,** pp 933-945.

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C. M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V. S., Kodali, V. K., Li, W., Maglott, D., Masterson, P., McGarvey, K. M., Murphy, M. R., O'Neill, K., Pujar, S., Rangwala, S. H., Rausch, D., Riddick, L. D., Schoch, C., Shkeda, A., Storz, S. S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R. E., Vatsan, A. R., Wallin, C., Webb, D., Wu, W., Landrum, M. J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T. D. & Pruitt, K. D. 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research,* 44(D1)**,** pp D733-D745.

Park, H., Kwak, M., Seo, J., Ju, J., Heo, S., Park, S. & Hong, W. 2018. Enhanced production of carotenoids using a Thraustochytrid microalgal strain containing high levels of docosahexaenoic acid-rich oil. *Bioprocess and Biosystems Engineering,* 41(9)**,** pp 1355-1370.

Patel, A., Karageorgou, D., Rova, E., Katapodis, P., Rova, U., Christakopoulos, P. & Matsakas, L. 2020. An Overview of Potential Oleaginous Microorganisms and Their Role in Biodiesel and Omega-3 Fatty Acid-Based Industries. *Microorganisms,* 8(3)**,** pp 434.

Patel, A., Rova, U., Christakopoulos, P. & Matsakas, L. 2019. Simultaneous production of DHA and squalene from Aurantiochytrium sp. grown on forest biomass hydrolysates. *Biotechnology for Biofuels,* 12(1)**,** pp 255.

Plesivkova, D., Richards, R. & Harbison, S. 2019. A review of the potential of the MinION™ single-molecule sequencing system for forensic applications. *WIREs Forensic Science,* 1(1)**,** pp e1323.

Quail, M. A., Smith, M., Coupland, P., Otto, T. D., Harris, S. R., Connor, T. R., Bertoni, A., Swerdlow, H. P. & Gu, Y. 2012. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics,* 13(1)**,** pp 341.

Raghukumar, S. 2002. Ecology of the marine protists, the Labyrinthulomycetes (Thraustochytrids and Labyrinthulids). *European Journal of Protistology,* 38(2)**,** pp 127-145.

Raghukumar, S. & Raghukumar, C. 1999. Thraustochytrid fungoid protists in faecal pellets of the tunicate Pegea confoederata, their tolerance to deep-sea conditions and implication in degradation processes. *Marine Ecology Progress Series,* 190(133-140.

Ratledge, C. 1993. Single cell oils — have they a biotechnological future? *Trends in Biotechnology,* 11(7)**,** pp 278-284.

Ratledge, C. 2002. Regulation of lipid accumulation in oleaginous micro-organisms. *Biochemical Society Transactions,* 30(6)**,** pp 1047-1050.

Ratledge, C. 2004. Fatty acid biosynthesis in microorganisms being used for Single Cell Oil production. *Biochimie,* 86(11)**,** pp 807-815.

Ratledge, C. 2012. Omega-3 biotechnology: Errors and omissions. *Biotechnology Advances,* 30(6)**,** pp 1746-1747.

Ren, L.-J., Huang, H., Xiao, A.-H., Lian, M., Jin, L.-J. & Ji, X.-J. 2009. Enhanced docosahexaenoic acid production by reinforcing acetyl-CoA and NADPH supply in Schizochytrium sp. HX-308. *Bioprocess and Biosystems Engineering,* 32(6)**,** pp 837.

Rock, C. O. 2008. CHAPTER 3 - Fatty acid and phospholipid metabolism in prokaryotes. *In:* Vance, D. E. & Vance, J. E. (eds.) *Biochemistry of Lipids, Lipoproteins and Membranes (Fifth Edition).* San Diego: Elsevier.

Saraswathy, N. & Ramalingam, P. 2011. 8 - Genome sequence assembly and annotation. *In:* Saraswathy, N. & Ramalingam, P. (eds.) *Concepts and Techniques in Genomics and Proteomics.* Woodhead Publishing.

Schadt, E. E., Turner, S. & Kasarskis, A. 2010. A window into third-generation sequencing. *Human Molecular Genetics,* 19(R2)**,** pp R227-R240.

Schalamun, M., Nagar, R., Kainer, D., Beavan, E., Eccles, D., Rathjen, J. P., Lanfear, R. & Schwessinger, B. 2019. Harnessing the MinION: An example of how to establish long-read sequencing in a laboratory using challenging plant tissue from Eucalyptus pauciflora. *Molecular ecology resources,* 19(1)**,** pp 77-89.

Schirmer, M., D'Amore, R., Ijaz, U. Z., Hall, N. & Quince, C. 2016. Illumina error profiles: resolving fine-scale variation in metagenomic sequencing data. *BMC Bioinformatics,* 17(1)**,** pp 125.

Seddiki, K., Godart, F., Aiese Cigliano, R., Sanseverino, W., Barakat, M., Ortet, P., Rébeillé, F., Maréchal, E., Cagnac, O. & Amato, A. 2018. Sequencing, De Novo Assembly, and Annotation of the Complete Genome of a New Thraustochytrid Species, Strain CCAP_4062/3. *Genome announcements,* 6(11)**,** pp e01335-17.

Semenkovich, C. F. 1997. Regulation of fatty acid synthase (FAS). *Progress in Lipid Research,* 36(1)**,** pp 43-53.

Shanklin, J. & Cahoon, E. B. 1998. DESATURATION AND RELATED MODIFICATIONS OF FATTY ACIDS. *Annual Review of Plant Physiology and Plant Molecular Biology,* 49(1)**,** pp 611-641.

Shendure, J. & Ji, H. 2008. Next-generation DNA sequencing. *Nature Biotechnology,* 26(10)**,** pp 1135-1145.

Solovyev, V., Kosarev, P., Seledsov, I. & Vorobyev, D. 2006. Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biology,* 7(1)**,** pp S10.

Sul, H. S. & Smith, S. 2008. CHAPTER 6 - Fatty acid synthesis in eukaryotes. *In:* Vance, D. E. & Vance, J. E. (eds.) *Biochemistry of Lipids, Lipoproteins and Membranes (Fifth Edition).* San Diego: Elsevier.

Sutton, J. R. M. S. K. G. 2010. Assembly algorithms for next-generation sequencing data. *Genomics,* 95(6)**,** pp 315 - 327.

Technologies, O. N. 2021. *Nanopore Store* [Online]. Available: https://store.nanoporetech.com/sample-prep/ligation-sequencing-kit.html [Accessed 11 June; 2021.

Teytelman, L., Özaydın, B., Zill, O., Lefrançois, P., Snyder, M., Rine, J. & Eisen, M. B. 2009. Impact of Chromatin Structures on DNA Processing for Genomic Analyses. *PLOS ONE,* 4(8)**,** pp e6700.

Tyler, A. D., Mataseje, L., Urfano, C. J., Schmidt, L., Antonation, K. S., Mulvey, M. R. & Corbett, C. R. 2018. Evaluation of Oxford Nanopore's MinION Sequencing Device for Microbial Whole Genome Sequencing Applications. *Scientific Reports,* 8(1)**,** pp 10931.

Tyson, J. R., O'Neil, N. J., Jain, M., Olsen, H. E., Hieter, P. & Snutch, T. P. 2018. MinION-based long-read sequencing and assembly extends the Caenorhabditis elegans reference genome. *Genome research,* 28(2)**,** pp 266-274.

Vakhapova, V., Richter, Y., Cohen, T., Herzog, Y. & Korczyn, A. D. 2011. Safety of phosphatidylserine containing omega-3 fatty acids in non-demented elderly: a double-blind placebo-controlled trial followed by an open-label extension. *BMC Neurology,* 11(1)**,** pp 79.

Van der Verren, S. E., Van Gerven, N., Jonckheere, W., Hambley, R., Singh, P., Kilgour, J., Jordan, M., Wallace, E. J., Jayasinghe, L. & Remaut, H. 2020. A dual-constriction biological nanopore resolves homonucleotide sequences with high fidelity. *Nature Biotechnology,* 38(12)**,** pp 1415-1420.

Van Nieuwerburgh, F., Thompson, R. C., Ledesma, J., Deforce, D., Gaasterland, T., Ordoukhanian, P. & Head, S. R. 2012. Illumina mate-paired DNA sequencing-library preparation using Cre-Lox recombination. *Nucleic acids research,* 40(3)**,** pp e24-e24.

Vrinten, P., Wu, G., Truksa, M. & Qiu, X. 2007. Production of Polyunsaturated Fatty Acids in Transgenic Plants. *Biotechnology and Genetic Engineering Reviews,* 24(1)**,** pp 263-280.

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K. & Earl, A. M. 2014. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLOS ONE,* 9(11)**,** pp e112963.

Wescoe, Z. L., Schreiber, J. & Akeson, M. 2014. Nanopores Discriminate among Five C5-Cytosine Variants in DNA. *Journal of the American Chemical Society,* 136(47)**,** pp 16582-16587.

Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. 2017. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Computational Biology,* 13(6)**,** pp e1005595.

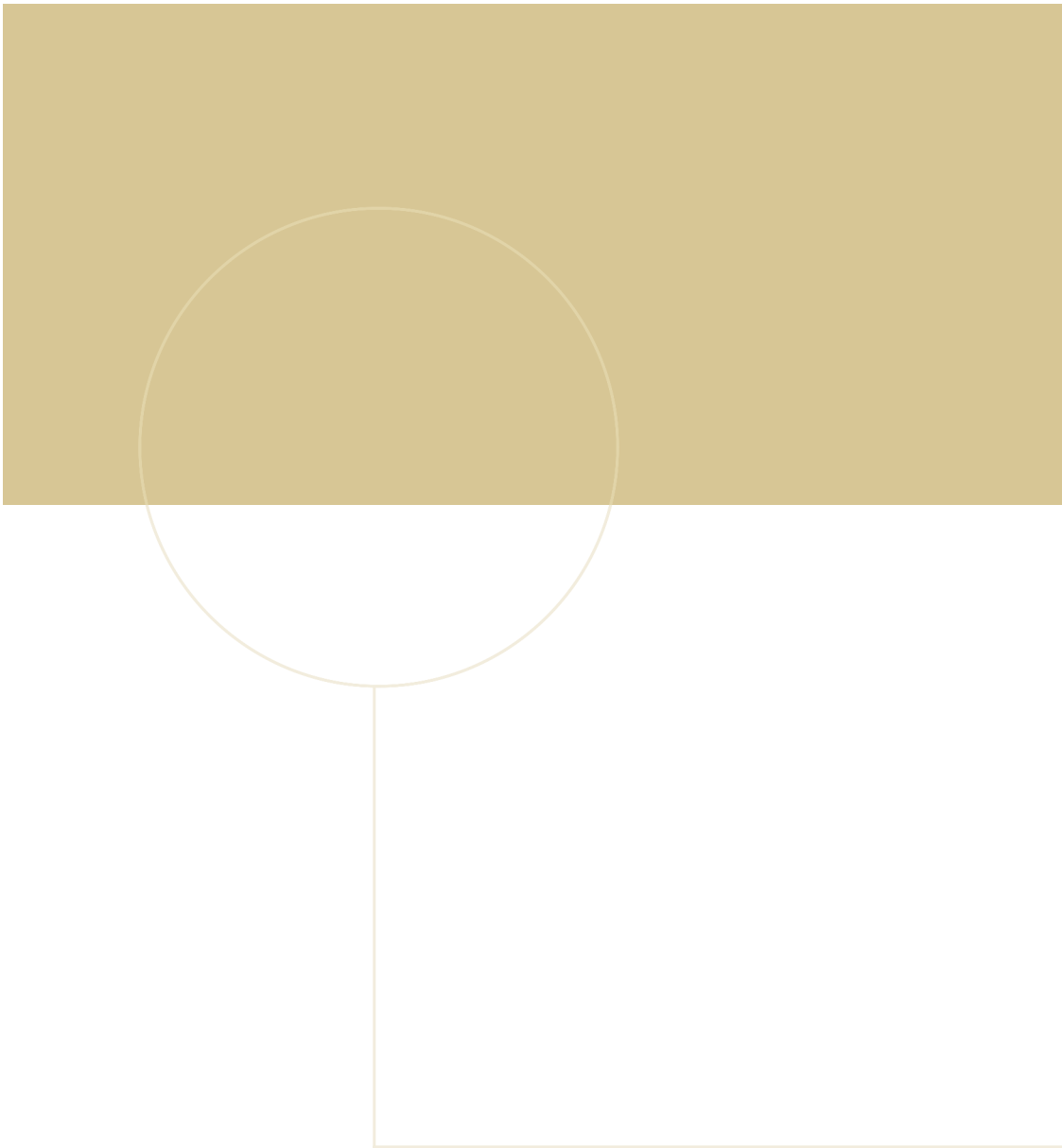
Wu, C.-C., Ye, R., Jasinovica, S., Wagner, M., Godiska, R., Tong, A. H.-Y., Lok, S., Krerowicz, A., Knox, C., Mead, D. & Lodes, M. 2012. Long-span, mate-pair scaffolding and other methods for faster next-generation sequencing library creation. *Nature Methods,* 9(9)**,** pp i-ii.

Yandell, M. & Ence, D. 2012. A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics,* 13(5)**,** pp 329-342.

Ye, J., Zhang, Y., Cui, H., Liu, J., Wu, Y., Cheng, Y., Xu, H., Huang, X., Li, S., Zhou, A., Zhang, X., Bolund, L., Chen, Q., Wang, J., Yang, H., Fang, L. & Shi, C. 2018. WEGO 2.0: a web tool for analyzing and plotting GO annotations, 2018 update. *Nucleic Acids Research,* 46(W1)**,** pp W71-W75.

Zhang, W., Chen, J., Yang, Y., Tang, Y., Shang, J. & Shen, B. 2011. A Practical Comparison of De Novo Genome Assembly Software Tools for Next-Generation Sequencing Technologies. *PLOS ONE,* 6(3)**,** pp e17915.

Zhao, X., Dauenpen, M., Qu, C. & Qiu, X. 2016. Genomic Analysis of Genes Involved in the Biosynthesis of Very Long Chain Polyunsaturated Fatty Acids in Thraustochytrium sp. 26185. *Lipids,* 51(9)**,** pp 1065-1075.



NTNU

Norwegian University of
Science and Technology