ABANT IZZET BAYSAL UNIVERSITY

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

ANNOTATION OF *CAULERPA CYLINDRACEA* GENOME BY HIGH-THROUGHPUT SEQUENCING DATA ANALYSIS

MASTER OF SCIENCE

MERYEM AYDIN

BOLU, MARCH 2018

ABANT IZZET BAYSAL UNIVERSITY

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

DEPARTMENT OF CHEMISTRY

ANNOTATION OF *CAULERPA CYLINDRACEA* GENOME BY HIGH-THROUGHPUT SEQUENCING DATA ANALYSIS

MASTER OF SCIENCE

MERYEM AYDIN

BOLU, MARCH 2018

APPROVAL OF THE THESIS

ANNOTATION OF *CAULERPA CYLINDRACEA* GENOME BY HIGH-THROUGHPUT SEQUENCING DATA ANALYSIS submitted by **Meryem AYDIN** in partial fulfillment of the requirements for the degree of **Master of Science** in **Department of Chemistry, The Graduate School of Natural and Applied Sciences of ABANT IZZET BAYSAL UNIVERSITY in 26/03/2018** by

Examining Committee Members	**Signature**
Supervisor Dr. Ercan Selçuk ÜNLÜ Abant İzzet Baysal University	.....
Member Prof. Dr. Azra BOZCAARMUTLU Abant İzzet Baysal University	.....
Member Prof. Dr. Necmi AKSOY Düzce University	.....

Graduation Date :

Assoc. Prof. Dr. Ömer ÖZYURT .....

Director of Graduate School of Natural and Applied Sciences

To the beauty of every imperfection

DECLARATION

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Meryem AYDIN

_____

ABSTRACT

ANNOTATION OF THE CAULERPA CYLINDRACEA GENOME BY HIGH-THROUGHPUT SEQUENCING DATA ANALYSIS
MSC THESIS
MERYEM AYDIN
ABANT IZZET BAYSAL UNIVERSITY GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
DEPARTMENT OF CHEMISTRY
(SUPERVISOR: DR ERCAN SELÇUK ÜNLÜ)

BOLU, MARCH 2018

Algae are a diverse photosynthetic species with lack of true roots, stems and leaves. They can classify based on their thallus pigmentation such as red (*Rhodophyta*), brown (*Phaeophyta*) and green algae (*Chlorophyta*). The green algae *Caulerpa cylindracea* is an invasive member of Bryopsidales that spread around tropical to subtropical territories. It has been reported in the 100 worst invaders of alien marine species in the Mediterranean Sea. It can rapidly spread and change ecological biodiversity dramatically. Despite this potential threat on coastal ecosytem there has been reported very limited genetic information about *C. cylindracea*. An extensive genomic research is essential to reveal biological, molecular and biochemical mechanisms of this species. Due to these reasons, this study represents a novel research for *C. cylindracea* genome. We aimed to annotate transcriptome data of *C. cylindracea* which obtained from previous study by using next-generation sequencing. 83,071,493.00 of raw reads were *de novo* assembled in 47400 contigs representing by 33340 unigenes using Trinity software. Swiss-Prot, Pfam and EggNOG databases were used for functional annotation of assembled *C. cylindracea* transcriptome. Total of 14147 unigenes belonging to 21001 transcripts were functionally annotated. Highest coverage with 13012 unigenes was obtained from Swiss-Prot database. Gene ontology annotations resulted that 40.94% of the GO terms matches (23842 hits) with biological process, followed by 32.05% cellular component (18666 hits) and 27.01% molecular function (15734 hits). In addition, *C. cylindracea* genome contributed to 88.85% of eukarya, 9.86% bacteria and 0.67% of viruses domains and 69.76% contribution to *Arabidopsis thaliana* at species level. Simple sequence repeat (SSR) markers were identified using MISA tool. A total of 16,156 and 10,207 SSRs were identified for (>500 bp) and (>1000 bp) nucleotide sequences respectively with 1 sequences containing more than one SSR. It is expected that the annotated transcriptome data and found results will lead to future studies and enlighten to molecular pathways of *Caulerpa cylindracea*.

KEYWORDS: *Caulerpa cylindracea*, Next-generation sequencing, *De novo* assembly, Functional annotation, Bioinformatics

ÖZET

CAULERPA CYLINDRACEA GENOMUNUN YÜKSEK İŞLEM HACİMLİ DİZİLEME VERİ ANALİZİ İLE BELİRLENMESİ

YÜKSEK LİSANS TEZİ

MERYEM AYDIN

ABANT İZZET BAYSAL ÜNİVERSİTESİ FEN BİLİMLERİ ENSTİTÜSÜ KİMYA ANABİLİM DALI

(TEZ DANIŞMANI: DR ERCAN SELÇUK ÜNLÜ)

BOLU, MART - 2018

Algler kök, gövde ve yaprakları farklılaşmamış; geniş fotosentetik bir türdür. Tallus pigmentasyoları baz alınarak kırmızı (*Rhodophyta*), kahverengi (*Phaeophyta*) ve yeşil algler (*Chlorophyta*) olarak sınıflandırılabilirler. Yeşil bir alg olan *Caulerpa cylindracea* tropikal ve yarı tropikal bölgelere yayılan istilacı bir su yosunu türüdür. Akdeniz'de en kötü 100 istilacı yabancı deniz türünden biri olarak bildirilmiştir. Çok kolaylıkla yayılabilir ve ekolojik biyoçeşitliliği dramatik olarak değiştirebilir. Kıyı ekosistemi üzerine teşkil ettiği tüm potansiyel tehdide rağmen, hakkında bulunan genetik bilgi çok kısıtlıdır. Biyolojik, moleküler ve biyokimyasal mekanizmalarının ortaya çıkarılması için türe ait kapsamlı genomik bir araştırmanın yapılması zorunludur. Bu sebeplerden dolayı, bu çalışma *C. cylindracea* genomuna özgün bir yaklaşım sunmaktadır. Bir önceki çalışmada yeni nesil DNA dizileme yöntemiyle elde edilen *C. cylindracea* transkriptomunun analizi hedeflenmiştir. Elde edilen 83.071.493,00 okuma Trinity programı ile referans genom kullanmaksızın *de novo* birleştirilmiş ve 33340 adet belirli fonksiyonlara sahip gen kümesi tarafından temsil edilen 47400 kontig belirlenmiştir. İşlevsel analiz için Swiss-Prot, Pfam, ve EggNOG veri tabanları kullanılmıştır. 21001 transkripte ait 14147 adet gen kümesi belirlenmiş ve en yüksek kapsama 13012 gen kümesi ile Swiss-Prot veri tabanından elde edilmiştir. Gen ontolojisi analizleri %40,94 biyolojik proses, ardından %32,05 hücresel bileşen ve %27,01 moleküler fonksiyon şeklinde sonuçlanmıştır. Ek olarak *C. cylindracea* genomu %88,85 ökaryot, %9,86 bakteri ve %0,67 virüs alemine katılmış olup, tür seviyesinde *Arabidopsis thaliana*'ya %69,76 katılım göstermiştir. Basit dizilim tekrarları MISA programıyla tanımlanmıştır. Sırasıyla 16.156 ve 10.207 basit dizilim tekrarı (>500 baz çifti) ve (>1000 baz çifti) nükleotid sekansları için tanımlanmıştır. Analizi yapılan transkriptom bilgisinin ve bulunan sonuçların gelecekteki çalışmalara zemin hazırlaması ve *Caulerpa cylindracea* türüne ait moleküler yolakları aydınlatması beklenmektedir.

ANAHTAR KELİMELER: *Caulerpa cylindracea*, Yeni nesil DNA dizileme, *De novo* birleştirme, İşlevsel analiz, Biyoinformatik

TABLE OF CONTENTS

Page

ABSTRACT ..... **v**

ÖZET ..... **vi**

TABLE OF CONTENTS ..... **vii**

LIST OF FIGURES ..... **viii**

LIST OF TABLES ..... **ix**

LIST OF ABBREVIATIONS AND SYMBOLS ..... **x**

1. INTRODUCTION **1**

1.1 Algae 1
1.2 *Caulerpa* Species 3
1.3 *Caulerpa racemosa* var. *cylindracea* 4
1.4 Sequencing Technology 5
1.5 Analyses Techniques 6
1.5.1 Next-Generation Sequencing Technology 8
1.5.2 RNA-seq 10
1.5.2.1 Trinity 11

2. AIM AND SCOPE OF THE STUDY **13**

3. MATERIALS AND METHODS **14**

3.1 Sampling, Total RNA and Messenger RNA Isolation 14
3.2 Transcriptome Analysis 18
3.3 *De Novo* Assembly 19
3.4 Bioinformatics Analyses 20

4. RESULTS AND DISCUSSION **22**

4.1 Sequencing and *de novo* assembly of *C. cylindracea* transcriptome 22
4.2 Functional Annotation of *C. cylindracea* transcriptome assembly 23
4.3 Identification of the Simple Sequence Repeat (SSR) 29

5. CONCLUSIONS AND RECOMMENDATIONS **32**

REFERENCES **34**

6. CURRICULUM VITAE **37**

LIST OF FIGURES

Page

Figure 1.1. Phylogenetic tree analysis of Ulvophyceae class (Modified from Lewis and McCourt (2004)) 3
Figure 1.2. Photograph of *C. cylindracea* (Taken by Ercan Selçuk Ünlü) 5
Figure 1.3. Complete scheme of the first generation DNA sequencing (Modified from Heather and Chain (2016)) 11
Figure 1.4. Outline of Trinity assembly (Modified from Haas et al. (2013)) .. 12
Figure 3.1. Sampling of *C. cylindracea* (Dr. Ercan Selçuk Ünlü (CMAS 2* diver)) 14
Figure 3.2. Collected samples and parts of *C. cylindracea* 15
Figure 3.3. Workflow of total RNA isolation 16
Figure 3.4. Workflow of mRNA isolation 17
Figure 3.5. Schematic representation of the overall sequencing and annotation workflow for the *C. cylindracea* transcriptome 21
Figure 4.1. Length distribution of *de novo* assembled contigs 23
Figure 4.2. Venn diagram for the annotated transcripts based on databases ... 24
Figure 4.3. Histogram representation of the gene ontology of C. cylindracea unigenes 25
Figure 4.4. GO terms distribution of unigenes 26
Figure 4.5. Comparison of database match values for *C. cylindracea* unigenes between domains, species and Viridiplantae 28
Figure 4.6. Frequency distribution of simple sequence repeats (SSRs) based on type of motifs 31

LIST OF TABLES

Page

Table 1.1. Commercial applications of algae (Modified from Vassilev and Vassileva (2016)) 2
Table 1.2. Advantage, mechanism and applications of significant sequencers (Modified from Liu et al. (2012)) 9
Table 4.1. Summary for the *C. cylindracea* transcriptome assembly 22
Table 4.2. Summary of *C. cylindracea* transcriptome functional annotation analysis 24
Table 4.3. Results of simple sequence repeat (SSR) analyses using MISA 29
Table 4.4. Summary of simple sequence repeat (SSR) types (>500 bp) based on the number of repeats 30

LIST OF ABBREVIATIONS AND SYMBOLS

A : Adenine

Bp : Base pair

C : Cytosine

cDNA : Complementary deoxyribonucleic acid

crRNA : CRISPR ribonucleic acid

DNA : Deoxyribonucleic acid

g : Gram

G : Guanine

Gb : Gigabase

GO : Gene ontology

Kb : Kilobase

lncRNA : Long non-coding ribonucleic acid

mRNA : Messenger ribonucleic acid

miRNA : Micro ribonucleic acid

MISA : MIcroSAtellite software

ng : Nanogram

NGS : Next-generation sequencing

PCR : Polymerase chain reaction

PE : Paired end

piRNA : Piwi-interacting ribonucleic acid

RNA : Ribonucleic acid

RNA-seq : Next-generation sequencing of ribonucleic acid

rRNA : Ribosomal ribonucleic acid

SA-PMP : Streptavidin-paramagnatic particles

SBS : Sequence-by-synthesis

SE : Single end

siRNA : Small interfering ribonucleic acid

SMS : Single molecule sequencing

SMRT : Single molecule real-time sequencing

snRNA : Small nuclear ribonucleic acid

snoRNA : Small nucleolar ribonucleic acid

SRP RNA : Signal recognition particle ribonucleic acid

SSC **:** Saline-sodium citrate buffer

T **:** Thymine

tRNA **:** Transfer ribonucleic acid

tmRNA **:** Transfer-messenger ribonucleic acid

µl **:** Microliter

°C **:** Centigrade degree

ACKNOWLEDGEMENTS

I would like to express my gratitude to my supervisor Dr. Ercan Selçuk Ünlü for his guidance, remarks and encouragement through the learning process of this master thesis.

Furthermore I would like to thank Prof. Dr. Azra Bozcaarmutlu and Prof. Dr. Necmi Aksoy for their advice, criticism and support on the way.

I like to thank the members of my research lab especially Dr. Gülgez Gökçe Yıldız for her sincerity and unlimited support, also Özge Kaya and Ömer Can Ünüvar for their help and collaboration.

I would like to express my most sincere gratitude to my family who have supported me throughout my entire life and loved unconditionally.

Finally I would like to thank my beloved friends Erva Özkan, Erhan Çimen, Gözde Şahin, Mehmet Örgeç and Elif Tecen for helping me putting pieces together. I will be grateful forever for your friendship.

This study was supported by the TÜBİTAK, Grant No: 113Z785

1. INTRODUCTION

1.1 Algae

Algae are specified as thalloid plants. They differentiate from plants with the absence of specialized generative organs. They have no true roots, stems and leaves. Their size-range differentiates depending on the types of algae. They can be microscopic and unicellular to macroscopic marine algae or seaweeds that can grow up to 100 m height. They can live in the oceans, sea shores, freshwaters, and they are also found terrestrial lands such as soils, rocks and trees. Both prokaryotic and eukaryotic taxa are classified in the algae family. They are classified based on the nuclear organization, cell wall components, pigmentation, carbon source, motility and reproduction. Algae are autotrophic organisms that they synthesize their own food and energy. In this respect, the specie plays an outstanding role in food chains and oxygen circulation. Chlorophyll A is common among algae due to its primary role in photosynthesis. In addition, they can synthesize various forms of chlorophylls and pigments such as other chlorophylls (B, C, D, E, and F), carotenoids, phycobilins, phycoerythrin and phycocyanin.

Macroalgae are classified into three main groups according to their photosynthetic pigmentation alterations: red (Rhodophyta), brown (Phaeophyta) and green (Chlorophyta) (Chen et al., 2015). The green algae also called Chlorophyta are the earliest members of the kingdom Plantae which evolved from land plants nearly 500 million years ago (Qin et al., 2012). A connection between green algae and land plants has been solved based on morphology, gene analyses and genome studies and green algae found as primitive plants. Due to this close relation between terrestrial plants and green algae, they share common specialities such as chlorophyll B as the major accessory pigment and starch linkage between microtubule structures. They are the most diverse group of algae including eleven classes and more than 10.000 members belonged to two different phyla that generally live in the freshwater and marine ecosystem.

Marine algae, also named as macroalgae or seaweed, can live in harsh environmental conditions such as extreme temperature, salinity, pH and sunlight. Their physiology can change to adapt environmental conditions, and as a result of this reason, they produce various secondary metabolites (H. D. Wang et al., 2017). These secondary metabolites are maintained as bioactive compounds such as polysaccharides, polyunsaturated fatty acids, and tannins. In many cases, the seaweeds are applied in human or animal nutrients for their distinguished mineral ingredient and supplementary effects of their polysaccharides and protein content. Since macroalgae become one of the most widely studied marine organisms, they are maintained as remarkable sources for health researches. Several studies are carried out for potential usage of their bioactive compounds such as antioxidant, anti-tumor and anti-inflammatory effects. In addition to potential usage of bioactive compounds, marine algae are commonly used as feedstock for the renewable energy production. Due to their importance of both complexity and evolutionary perspectives, algae have been involved with several scientific studies and commercial applications. They are often demanded by health, cosmetic, food and energy industry.

Table 1.1. Commercial applications of algae (Modified from Vassilev and Vassileva (2016))

USAGE	PRODUCT
Food Industry (as an additive, ingredient, preservative, supplement)	Agar, alginate, astaxanthin, carotenoids, carrageenan, phytosterols
Energy Industry	Alcohols, aviation gas, biodiesel, biofuel, biooil, gasoline
Pharmaceutical Industry	Antioxidants, supplement and precursor for specific vitamins, therapeutic materials
Industrial Usage	Cosmetic, dyeing, glass, paper and textile applications

1.2 *Caulerpa* Species

Caulerpa is defined as an alga with a coenocytic thallus organization under Ulvophyceae class Bryopsidales order (Famà et al., 2002). The thallus of *Caulerpa* species is made of a creeping stolon part, it is joined to the substrata by root-like structures called rhizoids, and leaves-like structures named as fronds. The shapes of fronds may have different forms depending on the species (Infantes et al., 2011). The genus *Caulerpa* includes nearly 85 species which is spread around tropical to semitropical areas. It usually grows sandy or muddy parts of the marine environments and aquariums. The fronds differentiate among 15-30 cm (6-12 inches) in length. Some members of *Caulerpa* species are the most known invasive species. The complex combination of secondary metabolites is thought to be one of the biggest advantages of the *Caulerpa* species over other surrounding species (Maric et al., 2017). Most of them show toxic effects. Caulerpales are powerful competitors in the context of eliminate to other algae and marine species. Moreover the genus *Caulerpa* is also utilized as nutritient source in some Asian countries due to low in lipid high in protein capacity.

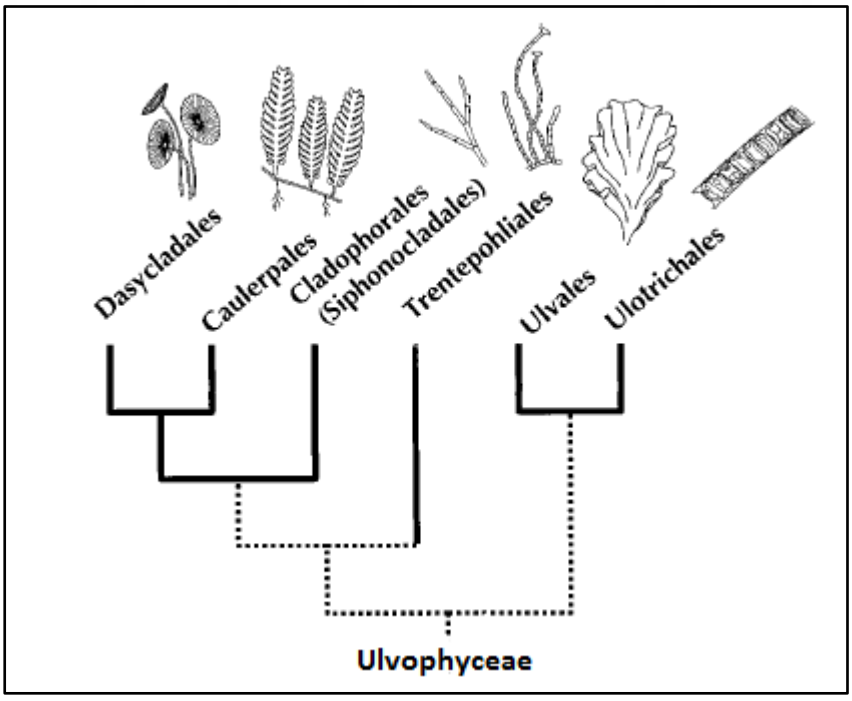

Figure 1.1. Phylogenetic tree analysis of Ulvophyceae class (Modified from Lewis and McCourt (2004))

1.3 *Caulerpa racemosa* var. *cylindracea*

The seaweed *Caulerpa cylindracea* (Sonder), formerly known as *Caulerpa racemosa var. cylindracea* (Sonder) Verlaque, Huisman and Boudouresque, is an invasive algal species in the Atlantic-Mediterranean Sea (Rizzo et al., 2016). The latest molecular studies have shown the genetic background differences of *Caulerpa racemosa var. cylindracea* at species-level existence. Thus, it has been suggested to rename the original strain *Caulerpa cylindracea* Sonder (Belton et al., 2014). It has been specified as an Australian strain Verlaque et al. (2003) that has been invading all types of substrata between 0 and 60 m of depth (Casu et al., 2009). *Caulerpa cylindracea* is formed of thallus, stolon, rhizoids and fronds parts like other *Caulerpa* species. They also have branches named as trabeculae that strength the stolon part. The species is able to reproduce both sexually and vegetatively. When compared to reproduction frequency, vegetative or also called asexual reproduction is widely observed due to effects of human or sea organisms and simplicity of fragmentation.

Non-indigenous species are major concern for countries because of the difficulties on obliteration or control. In addition, they can interact with native species dramatically and strongly affect ecological biodiversity. Taking into consideration, its effective and rapid spreading capacity on local basins, *C. cylindracea* has been introduced in the 100 worst invasive species of the Mediterranean Sea (Stabili et al., 2017). The mechanisms behind the invasive behaviour of this species are still unclear. The biological properties of *C. cylindracea* (vegetative and rapid propagation etc.) determine its extraordinary colonisation potency and its great talent to outcompete. After the first identity in Mediterranean Sea at the beginning of the 1990s, *Caulerpa cylindracea* has been experienced a powerful rate of spreading during the last three decade. It can easily changed the native benthic areas, so this species is a specific potential danger for the Mediterranean coastal ecosystem (Ruiz et al., 2011). In addition to biological properties, caulerpenyne synthesized as a secondary metabolite is another explanation for invasion. Caulerpenyne is a sesquiterpene excreted by *Caulerpa* species which has cytotoxic effects on different marine organisms especially fishes and sea urchins (Cavas and Yurdakoc, 2005). Due to these cytotoxic properties that may be involved in invasive characterization and chemical defence mechanisms.

Figure 1.2. Photograph of *C. cylindracea* (Taken by Ercan Selçuk Ünlü)

1.4 Sequencing Technology

Deoxyribonucleic acid (DNA) is the hereditary material in all living organisms. The genetic information in DNA is stored as a code represented by four nitrogenous bases: adenine (A), guanine (G), cytosine (C) and thymine (T). DNA is first transcribed to mRNA and translated into functional gene products within the cell. The information flow from DNA to RNA to functional gene products is named gene expression. These gene products are frequently proteins such as hormones, enzymes or receptors, but in non-protein coding genes for example ribosomal or transfer RNA genes, final product is functional RNA. Despite the genome that represents the whole DNA of a cell, transcriptome refers to RNA transcripts. Three major types of RNA which are messenger RNA (mRNA), ribosomal RNA (rRNA) and transfer RNA (tRNA) found in all organisms and showed function on protein

synthesis. Apart from that, there is many other RNA types are found within the different organisms. Some of them are CRISPR RNA (crRNA), guide RNA, long non-coding RNA (lncRNA), micro RNA (miRNA), piwi-interacting RNA (piRNA), short interfering RNA (siRNA), signal recognition particle RNA (SRP RNA), small nuclear RNA (snRNA), small nucleolar RNA (snoRNA), telomerase RNA and transfer-messenger RNA (tmRNA).

Since the first development of DNA sequencing technology in 1977 by Frederick Sanger, sequencing approach has a critical role on genetic studies. In the upcoming years of the first sequencing, this technology has been made great progress such as automation, sequencing large nucleotides at a time and more precise detection. Over the last decade, various alternative sequencing techniques have become accessible under the title of high-throughput or next-generation sequencing. Sequencing of whole genome as well as of individual parts and genes has been main concern of molecular biology and genetics. The sequencing of genomes is crucial to understand gene function, gene structure and evolutionary perspectives for organisms. In this context, several sequence-based technologies have been improved and showed useful strategies for gene discovery and annotation.

1.5 Analyses Techniques

Genomic analyses are based on the identification, evaluation and comparison of DNA sequence and corresponding structural properties. It covers information related with gene expression, function and regulation mechanisms at genomic levels. Along with several wet-lab techniques including conventional microarray and modern sequencing technologies, diverse number of bioinformatics approaches should be incorporated into data processing steps. Nowadays to accomplish sequence data, different methods are used. If sequencing technology is compared the earlier methods such as microarray analyses, this technology has been offered more precise results and high resolution (Roberts et al., 2011). It is becoming widely used technology because of answering important questions such as what is all made of or basis. Sequencing analyses are arisen from generational differences in among them.

First-generation high-throughput sequencing is standard Sanger sequencing also known as chain-termination sequencing or dideoxy sequencing. The classical dideoxy Sanger sequencing developed by Frederick Sanger in the middle of 1970s uses an enzymatic reaction. This technique is based on the detection of chain-end nucleotides probed with different coloring fluorescence materials where DNA polymerase is used for the extention of complementary strands. It results in a noticeable fluorescent labelled chain termination and consequently allows the determination of the unknown DNA sequence (Mutz et al., 2013). Despite its breakthrough impact on genetic studies, the Sanger sequencing has huge limitations such as large sequences separation by using gels or total automation of overall sequencing.

Second-generation sequencing refers to next-generation sequencing (NGS) technology. This sequencing is based on DNA molecules or clonally amplified DNA molecules arrayed on a two-dimensional layer with bases and high-throughput resolution in time. Commercial sequencing platforms use different sequencing techniques such as pyrosequencing, sequencing-by-synthesis (SBS) of every single nucleotide and sequencing by ligation. Despite the differences between sequencing techniques, all available second generation sequencing platforms have certain features with regard to cDNA library preparation, library amplification and the sequencing. Next-generation DNA sequencing offers that accurate analysis of whole or partial RNA transcripts, analysis of DNA parts for indentification of functional regions and novel applications such as mRNAs, small RNAs detection and transcription factors profiling (Ansorge, 2009).

Third-generation high-throughput sequencing maintains individual molecules of DNA or RNA as templates. This means that single molecule sequencing (SMS) or single molecule real-time sequencing (SMRT) at a time is also utilized by this technique. The visualization system of this technique enables of individual fluorescently labelled nucleotides detection that are attached to a DNA polymerase molecule during the synthesis reaction. It performs less sequencing reactions per cycle but the length of sequence can be larger. Third-generation sequencing is promising technology in terms of detection accuracy, low error rate and potential sequencing of double stranded DNA.

1.5.1 Next-Generation Sequencing Technology

In principle DNA sequencing technologies should be fast, precise, easy handled and inexpensive. The next-generation technologies differ from the Sanger sequencing in terms of simultaneous analysis, high-throughput resolution, and reduced cost and time. The impact of next-generation sequencing (NGS) technologies on genomic studies is humongous since the first announced to the sale in 2005. These technologies has been utilized for typical sequencing analyses, for example genome sequencing and resequencing or different practices for former Sanger sequencing analyses (Morozova and Marra, 2008). The main approach proposed by NGS is the ability to generate a large amount of data inexpensively. A standard next-generation sequencing run is capable of 6000 millions sequencing reactions of 100 nucleotides concluded 600 billion bases of raw sequence data. This ability permits large-scale evolutionary and comparative investigations to be accomplished. Different major NGS companies have in prominent that are the 454 instruments from Roche, the Illumina sequencing platforms and the Sequencing by Oligonucleotide Ligation and Detection (SOLiD) systems.

The whole set of expressed or transcribed genes from genomic DNA is defined as transcriptome. The tanscriptome is extremely dynamic and flexible compared to genome information stored in chromosomal sequences. While stored sequence information is supplied by genome, transcriptome is more related with presence and proportional abundance of RNA transcripts. Therefore, it offers a better perspective for active components in the cell rather than the genome (Jamers et al., 2009). The transcriptome analysis studies are frequently used since the invention of microarray technology but this technology is low sensitive, relatively expensive and cannot allow to process large amount of data for gene expression studies. In contrast with the microarray technology the next-generation sequencing offers rapid, cheap and high-througput analysis for various genome studies. NGS-based methods allow identification and quantification of transcripts without any previous data or information for a specific gene. Besides, NGS technologies can be alternatively used for *de novo* analyses of large unknown genomes, genomic differences among the same or different species, characterization of DNA-binding protein spectrums and genome based epigenetic alterations profiles (Tarazona et al., 2011).

Table 1.2. Advantage, mechanism and applications of significant sequencers (Modified from Liu et al. (2012))

Sequencer	**Roche 454 System**	**Illumina GA/Hiseq System**	**AB Solid System**	**Sanger Sequencing**
Sequencing mechanism	Pyrosequencing	Sequencing by synthesis	Ligation and two-base coding	Dideoxy chain termination
Read length	700 bp	50SE, 50PE, 101PE	50 + 35 bp or 50 + 50 bp	400~900 bp
Accuracy	99.9%	98% (100PE)	99.94% ∗raw data	99.999%
Reads	1M	3G	1200~1400M	-
Output data/run	0.7 Gb	600 Gb	120 Gb	1.9~84 Kb
Advantage	Read length, fast	High throughput	Accuracy	High quality, long read length
Disadvantage	Error rate with polybase more than 6, high cost, low throughput	Short read assembly	Short read assembly	High cost low throughput
Resequencing	-	Yes	Yes	-
De novo	Yes	Yes	-	Yes
Array	Yes	Yes	Yes	Yes
High GC sample	Yes	Yes	Yes	-
Bacterial	Yes	Yes	Yes	
Large genome	Yes	Yes	-	-

1.5.2 RNA-seq

RNA-seq is a set of experimental and computational sequencing technique for the detection of RNA sequences in biological specimens. It is a high-throughput and the most comprehensive next-generation sequencing application (Xiong et al., 2010). During the last decade, RNA-seq has been evolved as one of the most multi-purpose NGS-based application and led to tremendous studies on transcriptome. The high-throughput RNA-seq analyses are the most reliable alternative to provide unique vision on transcriptome annotation. Next-generation sequencing of RNA (RNA-seq) does not only supply the whole transcriptome of an organism, but also it provides data on alternative splicing, allele spesific expression and details on sequence specific variations (Young et al., 2010). Different from the microarray technologies, RNA-seq can determine accurate transcript levels of both sequenced and not sequenced organisms, identify novel isoforms, detect formerly annotated cDNA ends, exon/intron based mapping and show different sequence variations such as single nucleotide polymorphisms. As much as connected to wet-lab techniques RNA-seq is closely related with computational analyses by bioinformatics tools and softwares. These tools have been coded by researchers. Some of them are open-accessed for research applications through online sources. Researchers can carry out different analyses such as sequence mapping, assembly, transcriptome annotation or expression level comparison (Pareek, 2015).

The experimental part of RNA-seq consists of three main steps: i- isolation of total RNA or mRNA; ii- cDNA library preparation; iii- high-throughput analysis of whole transcriptome. Library preparation is conversion of the all RNAs in specimen to complementary DNA. The sequencing of cDNA is more preferable than genomic DNA since the cDNA is directly synthesized from transcribed RNA molecules rather than whole genome. This specification decreases the required number of targeted sequence reads. For RNA-seq analyses, there are different procedures to prepare the library depending on the sequencer of choice. In general, high molecular weight RNA molecules are chopped into smaller sizes, followed by producing blunt ended DNA fragments which are ligated with specific adapters. After ligation each of these sequencers use a different strategy such as sequencing directly amplificated library or pre-amplification step before sequencing (Buermans and den Dunnen, 2014).

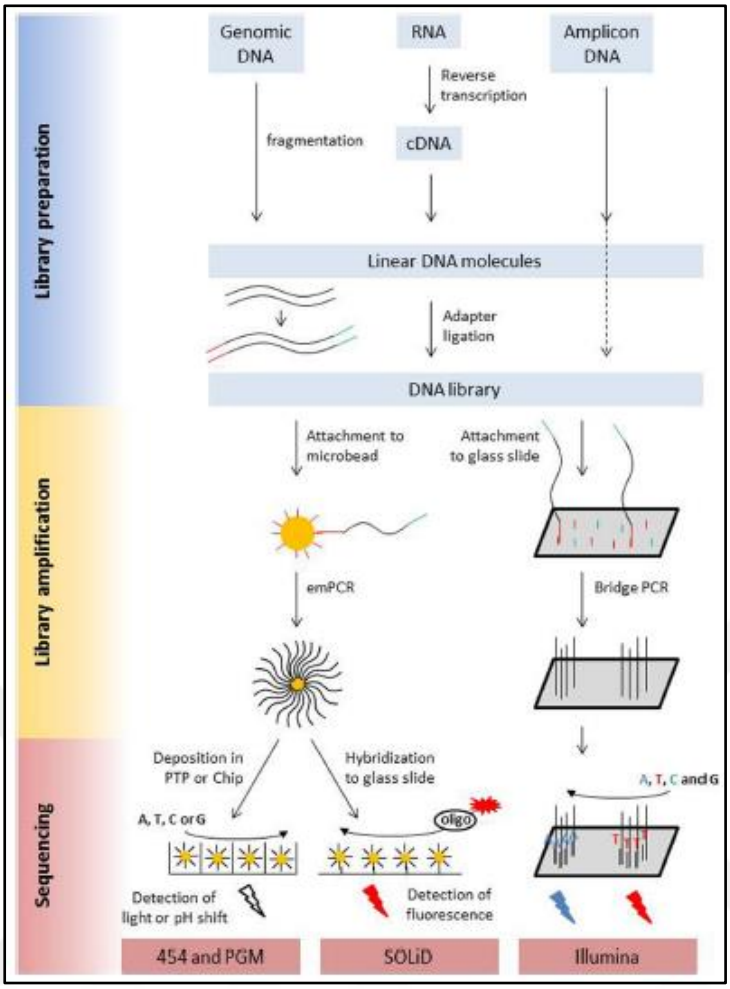

Figure 1.3. The library preparation and sequencing process of the major next generation sequencing platforms (Knief, 2014)

1.5.2.1 Trinity

RNA-seq assembly is based on alignment of short overlapped fragments which are obtained from the next-generation sequencing. Alignment is defined as lining up reads to discover where the sequences are similar and how high the similarity level is. Main goal of RNA-seq assembly is to get full-length transcriptome of studied organism. Depending on the availability of reference genomic sequence information, there are two available strategies. First assembly method requires a reference genome where short sequence fragment reads are primarily mapped on the genome and then assembled by joining overlapping sequences to obtain whole corresponding transcripts. This mapping-based assembly deals with some challengings such as comparing large genomes to short reads, mismatches due to

pseudogenes or nonunique repeats. Second assembly method, *de novo* assembly, is preferred if there is no available reference genome. This alternative approach joins the short reads by comparing the sequences to each other for overlapping regions.

Various programs are accessible for *de novo* assembly of RNA-seq analysis. Trans-ABySS, Velvet-Oases, and SOAPdenovo-trans are previously developed and used to assemble studied genomes. Besides all these programs, Trinity is an alternative and novel approach for assembly of transcriptome (Haas et al., 2013). Trinity uses Illumina RNA-seq sequences data for assembly process. It consists of three separate modules which are Inchworm, Chryalis and Butterfly. This combination allows processing large volumes of RNA-seq data with different perspectives. Inchworm initiates the assembly and identifies overlapping short reads called as contigs. Chrysalis clusters these contigs generated by Inchworm and produces de Bruijin graph for each of them. De Bruijin graph is a mathematical method to find out repetitive sequences of transcriptome. Each produced cluster represents the full set of transcriptional characteristics for a specific gene. Butterfly processes each de Bruijin graphs simultaneously and extracts alternative splicing generated isoforms.

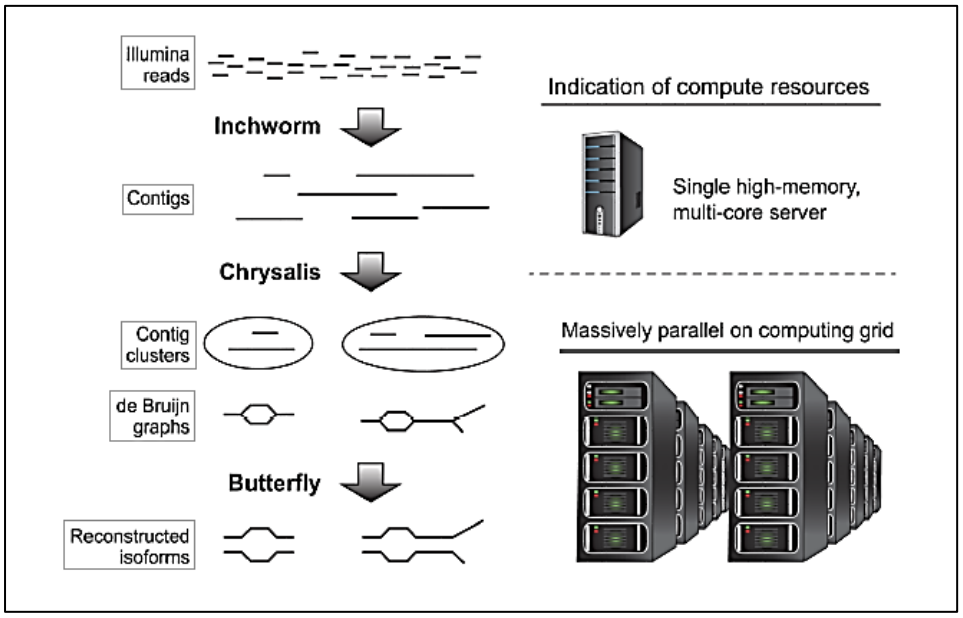

Figure 1.4. Outline of Trinity assembly (Modified from Haas et al. (2013))

2. AIM AND SCOPE OF THE STUDY

The fundamental aim of this study is annotation of *Caulerpa cylindracea* transcriptome by high-throughput sequencing data analysis. To achieve this aim, analysis of whole trancriptome sequencing data belonged to *C. cylindracea* was carried out under *de novo* assembly, bioinformatics analyses and identification of simple sequence repeat (SSR) titles. This is the most comprehensive study for *C. cylindracea* genome. Therefore, it is strongly expected that annotation of *C. cylindracea* transcriptome will guide to researchers for new studies of this species.

3. MATERIALS AND METHODS

3.1 Sampling, Total RNA and Messenger RNA Isolation

Collecting of *Caulerpa cylindracea* samples, isolation of total and messenger RNA (mRNA) steps were carried out by Ömer Can Ünüvar which are previously described (Unuvar, 2016). Briefly, *Caulerpa cylindracea* was collected by Dr. Ercan Selçuk Ünlü (CMAS 2* diver) from the Sığacık coast (Seferihisar, İzmir, Turkey) in September 9th, 2015 at shallow water (12 m) on rocky bottoms of the Aegean Sea. The exact location of the collecting area is 38°10'38.80"N - 26°45'57.24"E.

Figure 3.1. Sampling of *C. cylindracea* (Dr. Ercan Selçuk Ünlü (CMAS 2* diver))

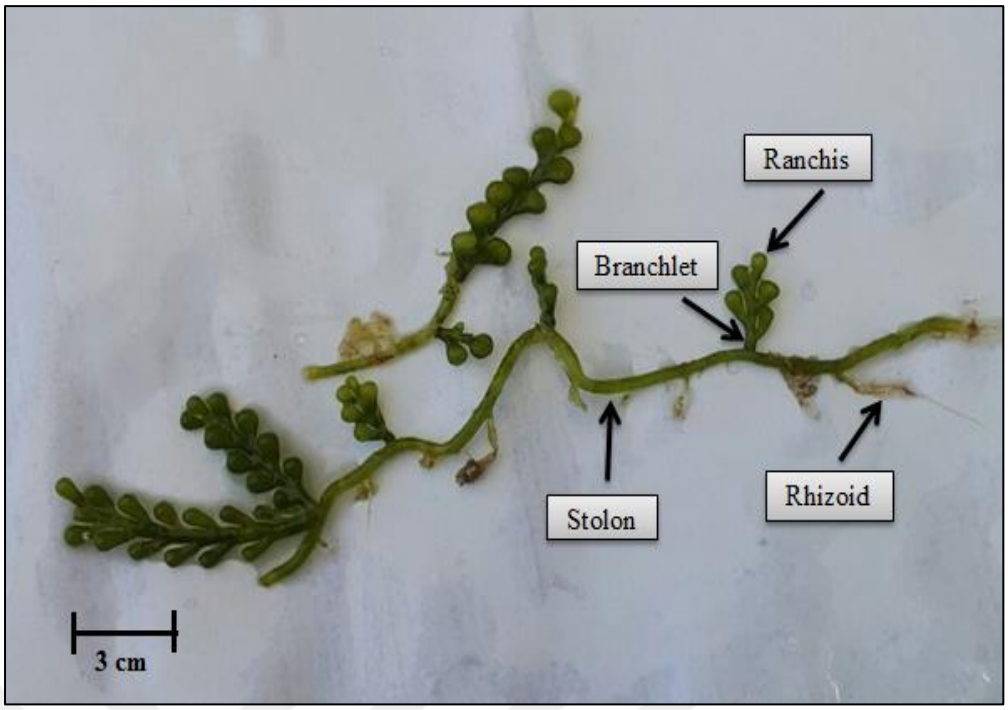

Figure 3.2. Collected samples and parts of *C. cylindracea*

After sampling, collected algae washed with sterile water without any harm to tissues and cytoplasm of the algae. The samples were divided into 5 grams portions and packaged. A total of 1 kilogram (wet weight) of algae sample was transported to the research laboratory in the liquid nitrogen transfer tank. Total RNA and mRNA isolation of collected *Caulerpa cylindracea* samples are given in Figure 3.3 and Figure 3.4 respectively.

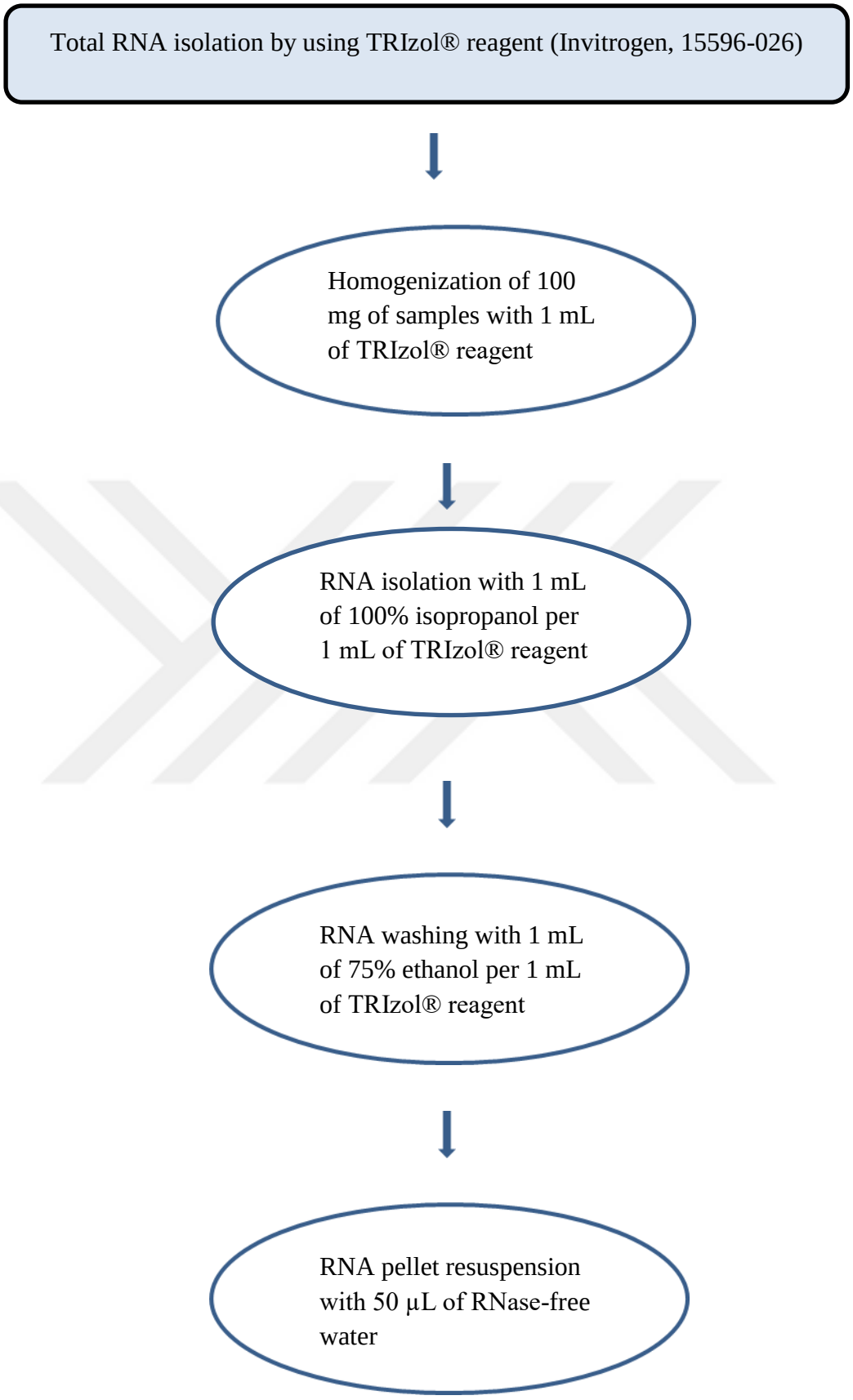

Figure 3.3. Workflow of total RNA isolation (Unuvar, 2016)

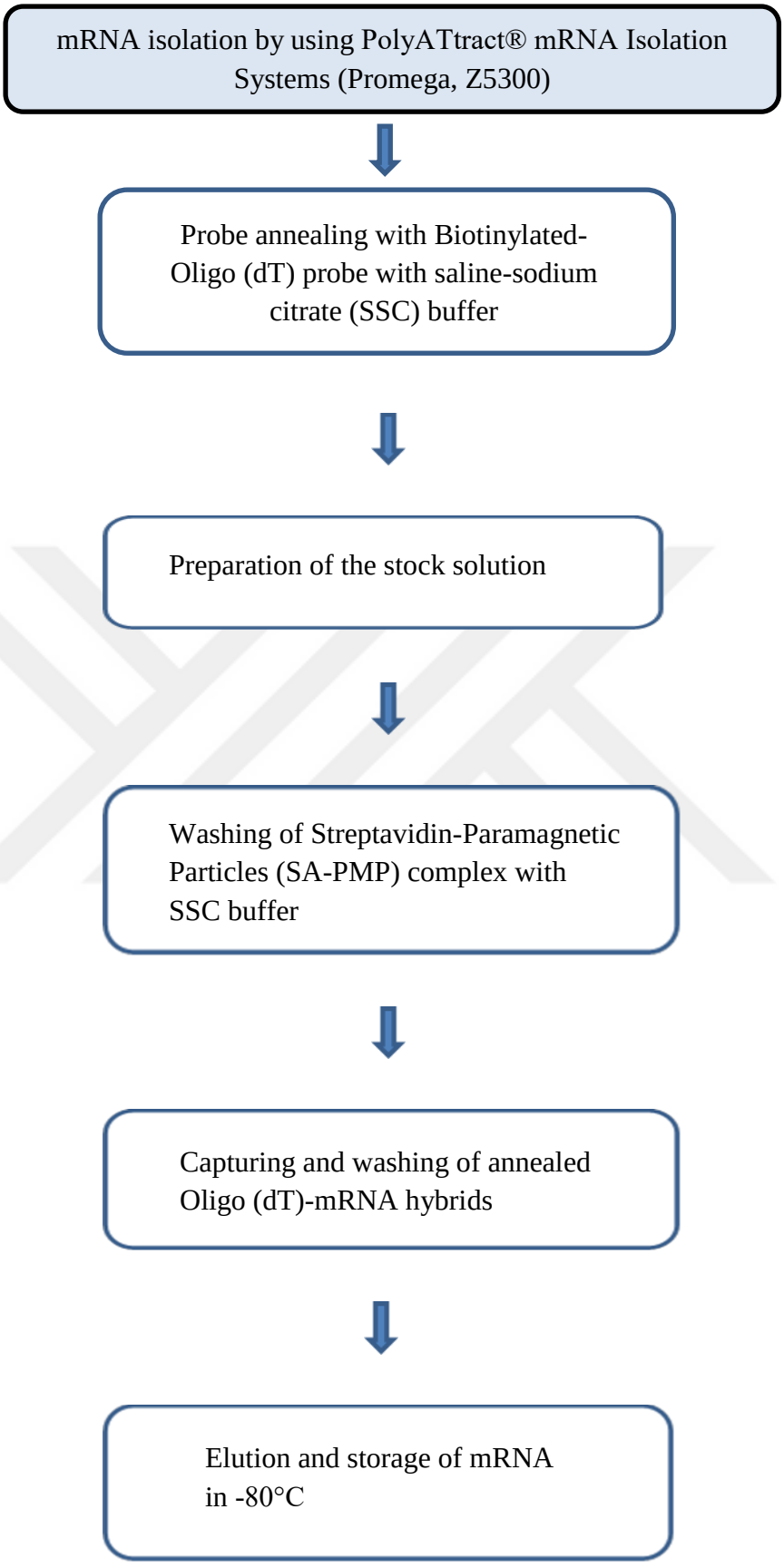

Figure 3.4. Workflow of mRNA isolation (Unuvar, 2016)

3.2 Transcriptome Analysis

The transcriptome is the entire set of transcripts in a cell of an organism under specific conditions. Analyzing the transcriptome is necessary for revealing the essential components of the genome and interpreting the tissues and cells elements at the molecular level (Z. Wang et al., 2009). The transcriptome provides inclusive sequence information for genetic studies and contributes as resources for genetic databases of the species. RNA-seq is a revolutionizing method for the purpose of transcriptome analysis. Various recent studies show that it is highly sensitive and accurate sequencing method. In the main, this method consists of to convert total or fractionated RNA samples with poly(A)+ tail to cDNA library. Every molecule in the cDNA library is then sequenced in a high-throughput way to get short reads from one or both ends. The produced short reads can be useful for different transcriptome analyses such as transcript quantification, expression level, genome-guided annotation or *de novo* assembly (Li and Dewey, 2011). Despite its invasive behaviour and secondary metabolites, limited genetic information about *C. cylindracea* is not enough for extensive molecular studies on the species. In this context, RNA-seq is a valuable method to build whole transcriptome of this organism in cases where no genetic information.

Isolated mRNA samples from four separate tissues which are Branchlet, Ranchis, Stolon and Rhizoid were collected as 10 ng/µL individually with the aim of verification. TruSeq Stranded mRNA Library Prep Kit (Illumina, RS-122-2101) for cDNA library construction was used for sequencing. Collected mRNA samples were divided into fragments in divalent cation solution. After the fragmentation, complementary strand synthesis was carried out helped by random primers and reverse transcriptase enzyme. The samples were incubated combination of both DNA polymerase I and RNase H to remove mRNA chain and synthesize new strand of DNA. Adenine molecules were joined to 3' tails of the fragmentated DNA's against the possible nonspecific ligation between cDNA fragments. After that adapters were joined to the tails of the fragments by ligation reaction. The samples were amplified by using PCR including 15 replication reactions and concentrations were measured as 2.41 ng/µL.

RNA isolation and sequencing of cDNA steps were carried out by Ligand Biotechnology Company which is the laboratory placed in İzmir. Transcriptome sequencing was accomplished by sequencing-by-synthesis based Illumina HiSeq™ 2000 from cDNA library with 100 nucleotide reads. This sequencing-by-synthesis based method also allows that measuring gene expression levels. 83,071,493.00 raw sequences were get at the end of the sequencing. After removing adapter from sequence fragments and assembling process 27737 individual gene sequences were identified.

3.3 *De Novo* Assembly

RNA-seq analyses generate millions of short reads from the one or both ends of cDNAs. The primary aim of RNA-seq assembly is to rebuild full-length transcripts by using these short-read sequences. Principally there are two assembly methods to reconstruct transcripts based on availability of a reference genome. If there is a reference genome to compare, reconstruction can be performed with its guide. This method is named as mapping-based assembly. On the other hand, second assembly method can be utilized absence of a reference genome which is named *de novo* assembly. *De novo* is a Latin word to express literally anew or afresh, over again but in a different perspective. In terms of *de novo* assembly, it refers a computational analysis which is useful for lacking a reference genome sequence (Paszkiewicz and Studholme, 2010). It is an effective practise to allow construction a transcriptome where no assembled genome exists against to compare short-read sequences of an organism. Current studies reveal that it is a cheap and high-throughput assembly method.

De novo assembly mainly uses two ways to construct a transcriptome. First approach is based on the calculation of pairwise overlaps between the short-reads to draw topological assembly graph. Second approach uses a de Bruijn graph to construct a transcriptome which is a mathematical method invented before the sequencing era. Common *de novo* assemblers such as Trinity, Oases and trans-ABySS depend on the de Bruijn graph data system McGettigan (2013), which reads large amount of data in shorter time. The assembly graphs stand for novel segments

of the genome or the transcriptome. In this regard, the main aim of *de novo* assembly is to obtain as long as potential contiguous segments (contigs) from the assembly graph. *De novo* assembly protocol was performed by using Trinity software package (version 2014-07-17) Haas et al. (2013), downloaded from Trinity software website (https://github.com/trinityrnaseq/trinityrnaseq/wiki). This software firstly assembles the short-reads of a certain length to produce longer contigs without any gaps. This process repeats until all overlaps are handled to obtain the longest sequences. As a result of this process, total number of contigs, mean length of contigs represented by N50 value and %GC content can be identified. Obtained contigs are clustered to get no longer elongated sequences that specified as unigenes. They describe expressed assembled sequences instead of functional genes (Patnaik et al., 2016). Various protocols and tutorials for how to set up and run Trinity are maintained at the website. *De novo* RNA-seq assembly using Trinity was carried out by servers supplied from Ligand Biotechnology Company.

3.4 Bioinformatics Analyses

Main goal of the bioinformatics analyses is the annotation of vast amount of data to get true ontology of the gene. The Gene Ontology (GO) consists of three categories which are molecular function, biological process and cellular component. Molecular function describes examination of activities at the molecular level such as catalytic or binding activities. Biological process is related with biological properties gathered with molecular function such as cell division, ageing and cell death. Cellular component part gives information about localization of subcellular components and macromolecules of the cell. Bioinformatics analyses were performed to annotate assembled contigs which were obtained by using Trinity software. The most studied and common databases such as Swiss-Prot, Pfam, and EggNOG were used for annotation. Swiss-Prot is a database with annotated protein sequences which was created in 1987. It includes various sequence entries, each with their own arrangement. Swiss-Prot allows to annotate the sequence in terms of function, post-translational modifications, domains and sites, secondary and possible quarternary structures, similarities and variations between other proteins (Bairoch and Apweiler, 2000). After the assembly and annotation, obtained *C. cylindracea*

unigenes were scanned for identification of simple sequence repeat (SSR) markers also known as microsatellites using the MIcroSAtellite (MISA,http://pgrc.ipk-gatersleben.de/misa/) software. Microsatellites are DNA parts consisting of short tandem repetitive units (1 to 6 bp in length) located within the both coding and non-coding parts of all eukaryotes (Orquera-Tornakian et al., 2017). SSR markers can be used for detection of genetic diversity, breeding via marker assisted-selection and mapping studies. Results were collected and ordered into an inclusive annotation report by using Trinotate software.

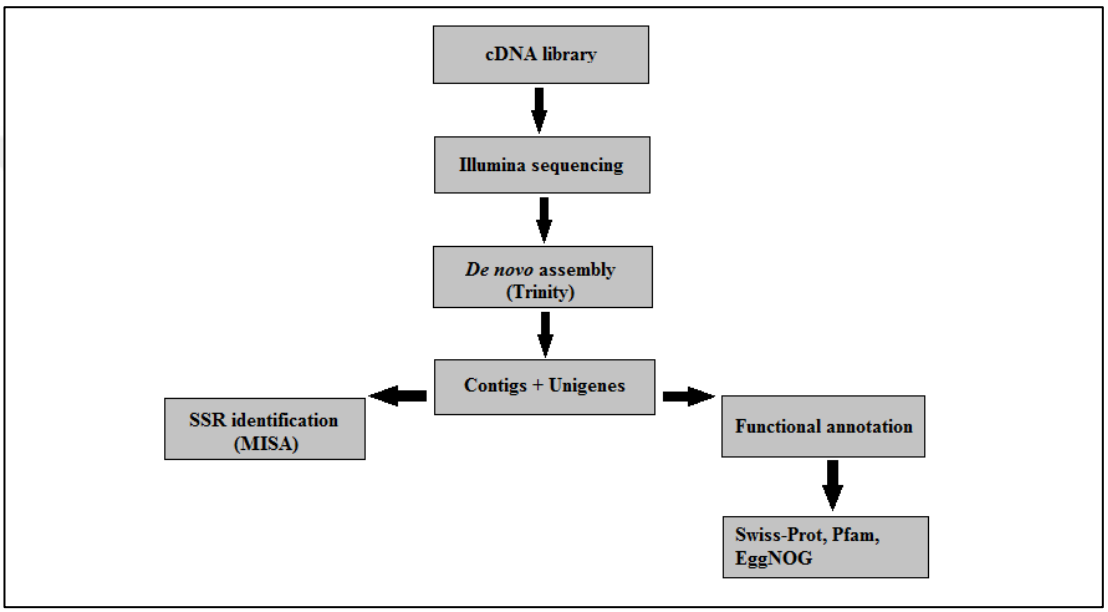

Figure 3.5. Schematic representation of the overall sequencing and annotation workflow for the *C. cylindracea* transcriptome

4. RESULTS AND DISCUSSION

4.1 Sequencing and *de novo* assembly of *C. cylindracea* transcriptome

Next-generation sequencing analysis was carried out from pooled RNA samples from different tissues (Rhizoid, Branchlet, Ranchis, Stolon) to represent entire algae. After the sequencing analysis 83,071,493.00 of raw reads were obtained. The raw reads are cleaned from adaptors and low-quality reads before assembly. Adaptor and quality trimming were performed by using Sickle tool to remain only high-quality reads. The reads clustered in 47400 assembled contigs representing by 33340 unigenes using Trinity software followed to the sequence trimming. The total number of contigs and unigenes, the percentage of GC content, N50 value and the average contig length were found identified under the *de novo* assembly. N50 is a measure to define the quality of assembled sequences that are fragmented in contigs with different length. The sequence N50 value, which is larger than the average contig length, indicates that at least 50% of the *C. cylindracea* transcripts were significantly assembled. The summary for sequence analysis is given in Table 4.1.

Table 4.1. Summary for the *C. cylindracea* transcriptome assembly

Number of Unigenes	33340
Number of Contigs	47400
Percent GC	42.18
Contig N50	1197
Average Contig Length	759.13

De novo assembly resulted in different length among obtained contiguous consensus nucleotide sequences which defined as contigs. Highest contig length found as 24994 nucleotide sequences in totaly 500 contigs. A total of 1000 contigs resulted in 10586 nucleotides, followed by approximately 1500 contigs with 5647 nucleotides with average length. More than 5000 contigs showed at least length with 122 nucleotide sequences.

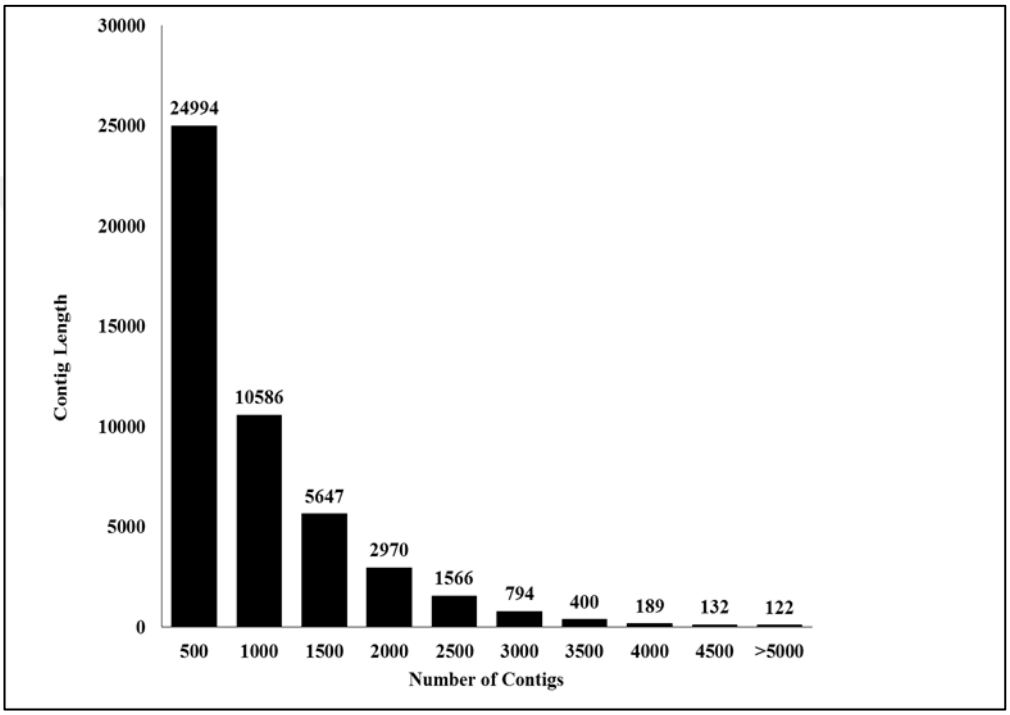

Figure 4.1. Length distribution of *de novo* assembled contigs

4.2 Functional Annotation of *C. cylindracea* transcriptome assembly

Using the Trinotate (v3.0.0) program (http://trinotate.github.io), well-known databases such as Swiss-Prot, Pfam, and EggNOG were scanned to annotate assembled unigenes putatively to known protein sequences based on similarity between their sequences. A local BlastX tool was used to compare assembled contigs against the Swiss-Prot database for searching protein sequences using translated nucleotides. The Pfam, the EggNOG and BlastP databases were searched for protein families, multiple sequence alignments and orthologous groups to to carry out complete functional annotation. The analysis results showed that 21001 transcripts

were matched at least in one protein annotation database. Total of 14147 unigenes belonging to 21001 transcripts were annotated. The highest coverage was attained from Swiss-Prot database analysis with 13012 unigenes. Distribution of annotated transcripts based on database type is given in Table 4.2 and Figure 4.2 respectively.

Table 3.2. Summary of *C. cylindracea* transcriptome functional annotation analysis

	Number of Annotated Sequences	
Database	***Contigs***	***Unigenes***
Swiss-Prot (BlastX)	19267	13012
(Blast P)	14613	9543
Pfam	13586	8837
EggNOG	11320	7829
TOTAL	21001	14147

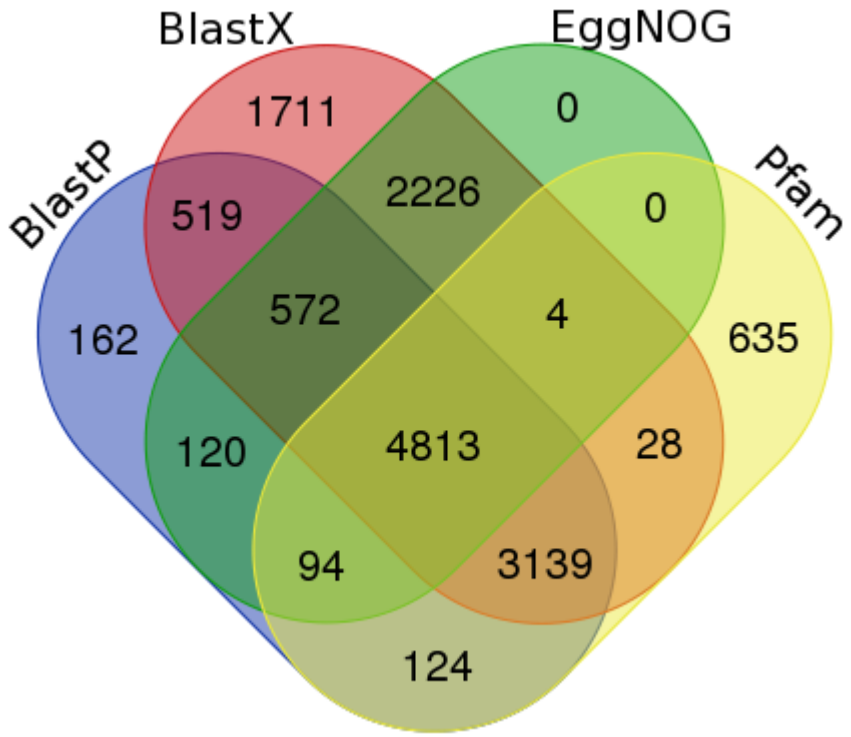

Figure 4.2. Venn diagram for the annotated transcripts based on databases

Gene ontology annotations clustering included “biological process”, “molecular function” and “cellular component” were performed by retrieving GO terms with the highest first ten hits from Swiss-Prot BlastX/BlastP match results for 13512 unigenes. Following cluster analysis a total of 58242 GO term matches were retrieved for 8622 unigenes. The highest distribution of unigenes were obtained under cellular component clusters with 8605 unigenes, followed by molecular function with 4806 unigenes and biological process with 2043 unigenes. According to the results the unigenes belonged to the cellular components were found mainly related with nucleus and cytoplasm. Obtained unigenes under the biological process resulted in mostly binding function. In the main, the unigenes associated with the molecular function resulted in transcriptional activities. In terms of gen expression, the transcription occurs during G1 and S phases that are the beginning of cell division. According to this, when the samples were collected *C. cylindracea* could initiate the cell division. GO terms distrubition of the unigenes based on function is given in Figure 4.3.

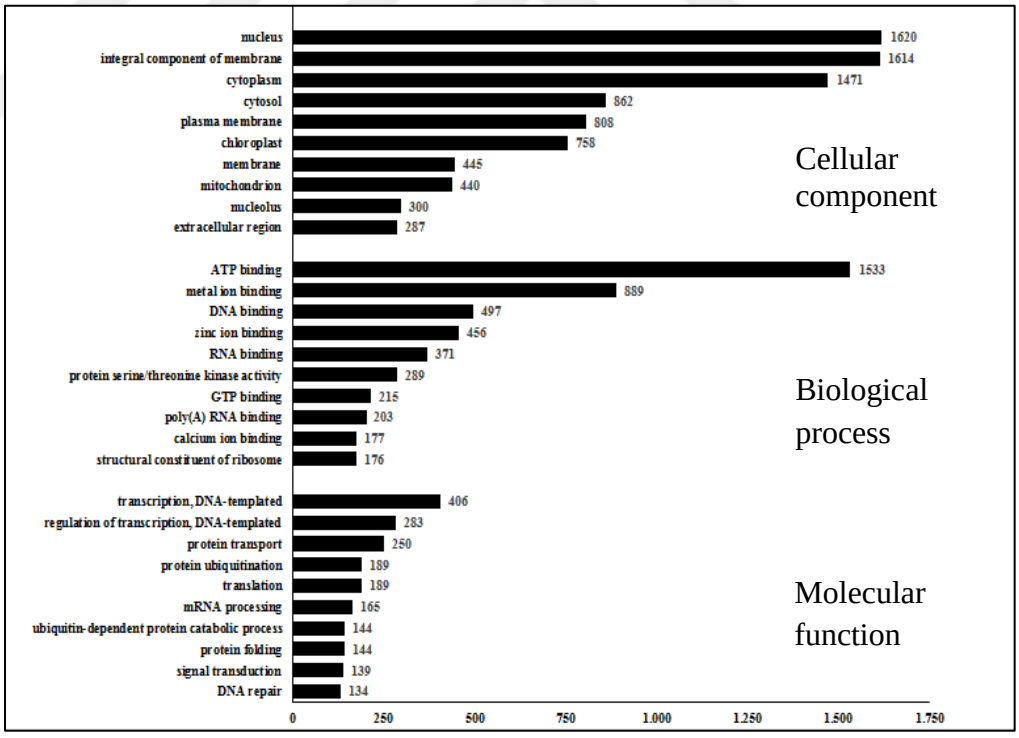

Figure 4.3. GO terms distribution of the unigenes

40.94% of the GO terms matches (23842 hits) with the highest distribution rate were obtained for biological process, followed by 32.05% cellular component (18666 hits) and 27.01% molecular function (15734 hits). Histogram representation of the gene ontology classification of *C. cylindracea* unigenes is given in Figure 4.4.

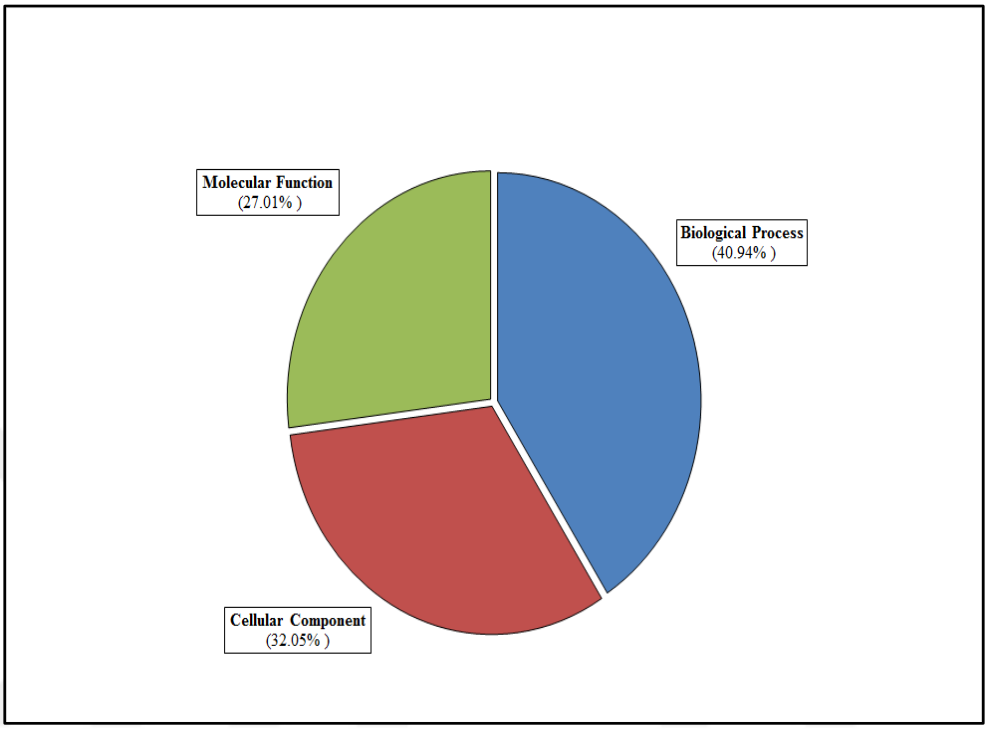

Figure 4.4. Histogram representation of the gene ontology of *C. cylindracea* unigenes

According to broadest database analysis results, 88.85% of eukarya, 9.86% bacteria and 0.67% of virus genes were used to obtain GO term annotations for 8622 unigenes analyzed. If specified, highest contribution level (69.76%) obtained from *Arabidopsis thaliana* records under Viridiplantae clade. It is an expected result since *A. thaliana* is one of the most studied and well-known model organism in terms of plant genetics. In addition, considering the contribution of algae genome data, highest contribution was found 3.97% for *Chlamydomonas reinhardtii* genes. It is obvious that more genomics data is requiered for better comparison of transcriptome data among algal species. Detailed contribution comparison for species in terms of gene similarity matches against *C. cylindracea* unigenes is given in Figure 4.5.

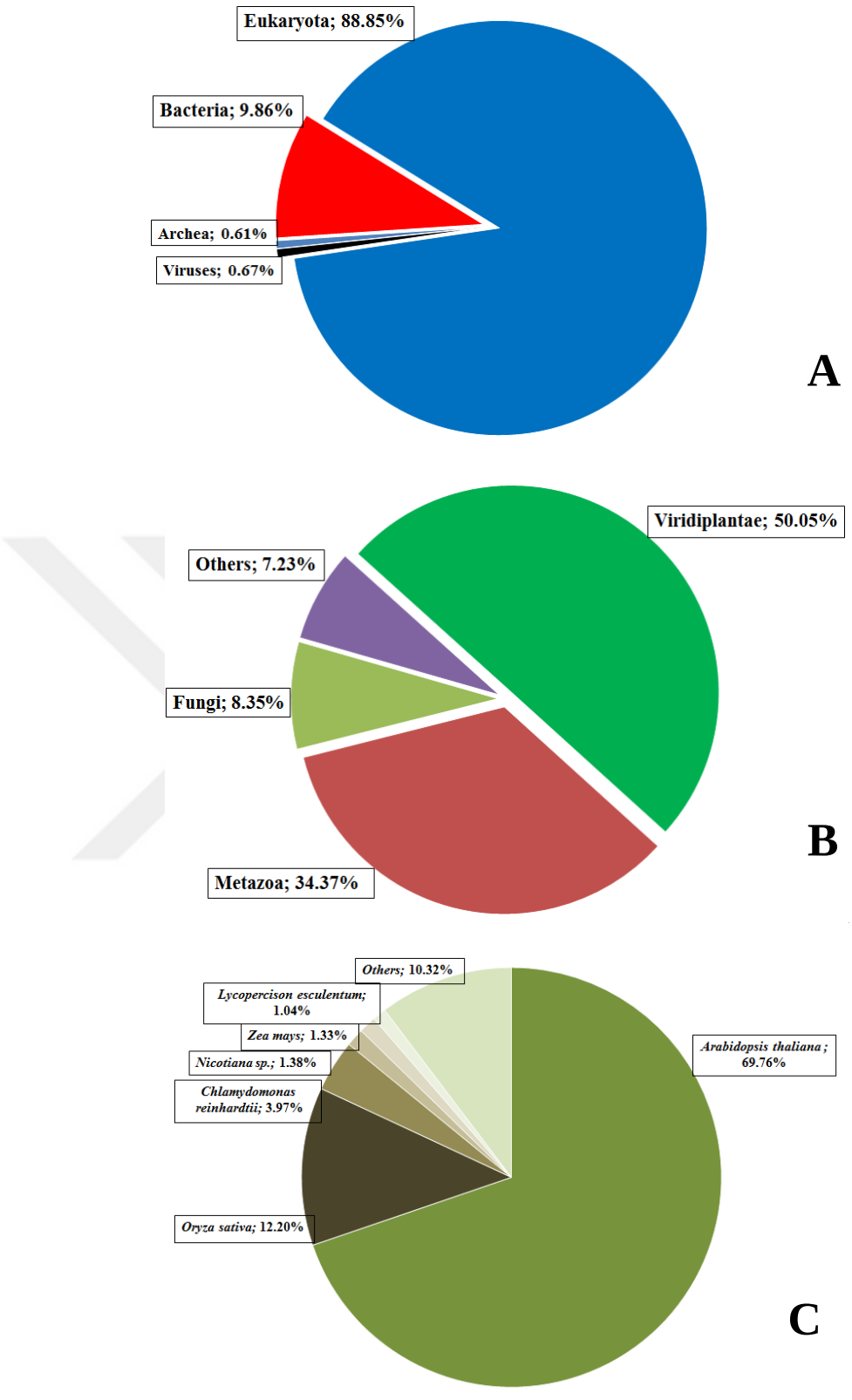

Figure 4.5. Comparison of database match values for *C. cylindracea* unigenes. Species based distribution of matching genes for *C. cylindracea* unigenes among broadest groups (A) under Eukaryota domain (B) and under Viridiplantae clade (C) were provided.

4.3 Identification of the Simple Sequence Repeat (SSR)

Microsatellite identification tool (MISA, http://pgrc.ipk-gatersleben.de/misa/) was applied to define simple sequence repeat (SSR) markers. The minimum number of repeats for mono-, di-, tri-, tetra-, penta- and hexa- nucleotide units were set as 10, 6, 5, 5, 5 and 5, respectively. The maximal number of bases interrupting two SSR markers in a compound microsatellite was set as 100. For SSR identification >500 bp and >1000 bp sequences of unigenes were took in consideration. All types of SSRs, from dimers to hexamers were examined. Results of SSR analyses using MISA is given in Table 4.3.

Table 4.3. Results of simple sequence repeat (SSR) analyses using MISA

Unit size	**Number of SSRs**		
	>1 bp	**>500 bp**	**>1000 bp**
2	273	217	177
3	1560	1172	963
4	50	45	42
5	6	5	0
6	5	2	0

Totaly 1441 unigenes (>500 bp) containing 16,156 SSRs were identified, with 1 sequences containing more than one SSR. From maximum to minimum presence of SSRs were trinucleotide repeats (1172), followed by di- (217), tetra- (45), penta- (5), and hexanucleotide repeats (2) in order. A total of 1182 unigenes (>1000 bp) including 10,207 SSRs were identified, within 1 sequences containing more than one SSR. The highest presence of SSRs were trinucleotide repeats (963), followed by di- (177), and tetranucleotide repeats (42) respectively. There was no obtained SSR data for penta- and hexanucleotide repeats. The summary of simple sequence repeat (SSR) types (>500 bp) depending on the number of repeats is given Table 4.4.

Table 4.4. Summary of simple sequence repeat (SSR) types (>500 bp) based on the number of repeats

Repeat Numbers		**5**	**6**	**7**	**8**	**9**	**10**	**11**
Motif Length	**di**	N/A	142	52	41	25	5	8
		N/A	115	43	30	19	2	8
		N/A	96	35	22	17	0	7
	tri	841	433	252	28	3	3	0
		636	332	171	27	3	3	0
		529	276	136	19	3	0	0
	tetra	47	3	0	0	0	0	0
		42	3	0	0	0	0	0
		39	3	0	0	0	0	0
	penta	6	0	0	0	0	0	0
		5	0	0	0	0	0	0
		0	0	0	0	0	0	0
	hexa	3	1	1	0	0	0	0
		2	0	0	0	0	0	0
		0	0	0	0	0	0	0

The frequency results are shown that both >500 bp and >1000 bp unigene sequences respectively. The most abundant SSR type in C. cylindracea transcriptome was identified as trinucleotide repeats included AAC/GTT (389; 26.94% - 329; 27.81%), AGG/CCT (315; 21.81% - 263; 22.23%), AAG/CTT (179; 12.40% - 148. 12,51%), ACC/GGT (108; 7.48% - 98; 8.28%), ATC/ATG (105; 7.27% - 64; 5.41%), AGC/CTG (43; 2.98%; - 35. 2,96%) and ACT/AGT (18; 1.25% - 17; 1.44%). After the outstanding trinucleotides, dinucleotide repeats were common within AG/CT (125; 8.66% - 100; 8.45%) and AC/GT (85; 5.89% - 72; 6.09%). In general, distribution of simple sequence repeat (SSR) graph revealed that SSRs found widely within less than 500 bp (>1 bp) unigene sequences. Tetranucleotide repeats were observed within >1000 bp according to >500 bp sequences. Penta- and hexanucleotide repeats found between 500 and 1000 bp unigene sequences. An analysis of the frequency distribution of SSRs depending on types of motifs is given in Figure 4.6.

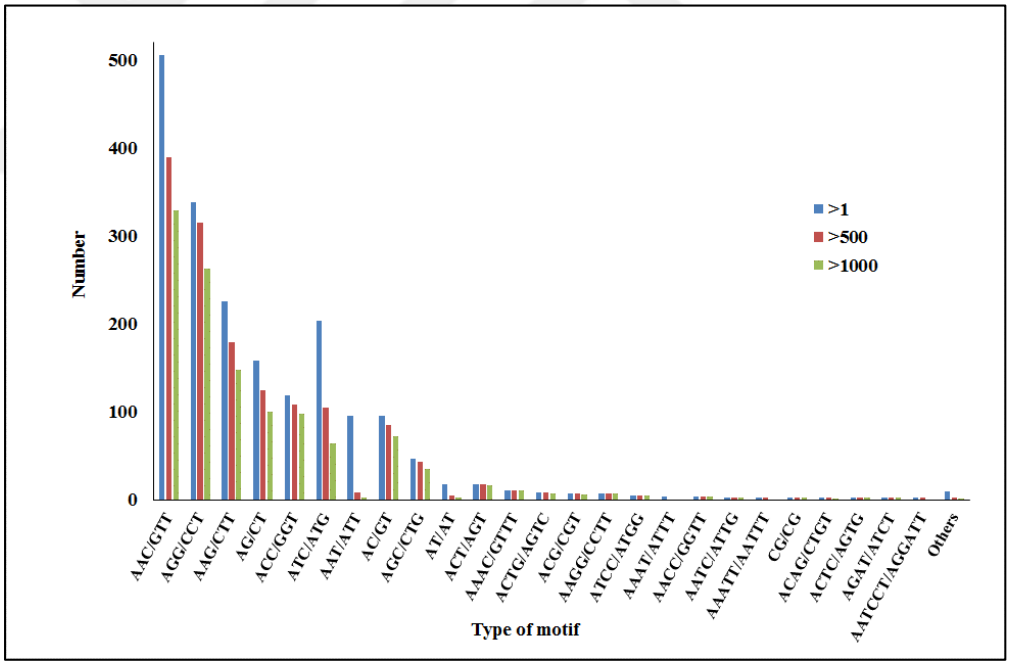

Figure 4.6. Frequency distribution of simple sequence repeats (SSRs) based on type of motifs

5. CONCLUSIONS AND RECOMMENDATIONS

In this study, it is aimed to annotate *Caulerpa cylindracea* transcriptome by high-throughput sequencing data analysis. So, raw RNA-Seq data produced by using next-generation sequencing technology for *Caulerpa cylindracea* samples was used for further bioinformatics analysis. Bioinformatics workflow can be summarized under *de novo* assembly, functional annotation of transcriptome data and identification of simple sequnce repeats (SSRs) titles. We used robust bioinformatics tools and databases such as Trinity, Swiss-Prot, Pfam, EggNOG and MISA during the study. This study represents the first genomic resource for the *Caulerpa cylindracea* including an extensive report of *de novo* assembled and functionally annotated cDNAs. Extended genetic studies have been carried out for the significant members of the Caulerpaceae family (*C. sertularioides*, *C. taxifolia* and *C. racemosa*) compared to the *C. cylindracea*, but there were only a couple of partial gene sequences belonged to the species in public databases. In this regard, this is the most comprehensive research of *C. cylindracea* genome.

After de novo assembly protocol total number of 47400 contigs were clustered and 33340 assembled unigenes were assigned for a particular function. Calculated N50 value represents that at least 50% of the *C. cylindracea* transcripts were significantly assembled. Gene ontology (GO) terms classification was carried out under three categories: cellular component, biological process and molecular function. Identified unigenes under GO terms can be adapted to many forthcoming research approaches such as gene regulation studies, enzymatic studies, or localization and expression based studies of proteins. Identified simple sequence repeats (SSRs) will provide potential molecular markers for population studies among *Caulerpa* species. They can be preferable for some practical applications such as genetic diversity, evolutinary analysis and genetic mapping studies.

Assembled and annotated transcriptome will serve as an milestone for gene discovery in *C. cylindracea*. The findings of this study will provide a remarkable source for future functional studies focusing on genetic background lying under invasive behaviour of the species, developing industrial applications for identified or

novel secondary metabolites. In summary this study will aid to a better explanation for all competencies assigned to *Caulerpa cylindracea*.

REFERENCES

Ansorge WJ (2009) “Next-generation DNA sequencing techniques”, N Biotechnol, 25(4): 195-203.

Bairoch A and Apweiler R (2000) “The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000”, Nucleic Acids Research, 28(1): 45-48.

Belton GS, van Reine WF, Huisman JM, Draisma SG and Gurgel CFD (2014) “Resolving phenotypic plasticity and species designation in the morphologically challenging *Caulerpa racemosa-peltata* complex (Chlorophyta, Caulerpaceae)”, J Phycol, 50(1): 32-54.

Buermans HP and den Dunnen JT (2014) “Next generation sequencing technology: Advances and applications”, Biochim Biophys Acta, 1842(10): 1932-1941.

Casu D, Ceccherelli G, Sechi N, Rumolo P and Sarà G (2009) “*Caulerpa racemosa* var. *cylindracea* as a potential source of organic matter for benthic consumers: evidences from a stable isotope analysis”, Aquat Ecol, 43: 1023-1029.

Cavas L and Yurdakoc K (2005) “An investigation on the antioxidant status of the invasive alga *Caulerpa racemosa* var. *cylindracea* (Sonder) Verlaque, Huisman, et Boudouresque (Caulerpales, Chlorophyta)”, Journal of Experimental Marine Biology and Ecology, 325: 189-200.

Famà P, Wysor B, Kooistra WHCF and Zuccarello GC (2002) “Molecular Phylogeny of the Genus *Caulerpa* (Caulerpales, Chlorophyta) Inferred from Chloroplast *tuf*A Gene”, Journal of Phycology, 38: 1040-1050.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, MacManes MD, Ott M, Orvis J, Pochet N, Strozzi F, Weeks N, Westerman R, William T, Dewey CN, Henschel R, LeDuc RD, Friedman N and Regev A (2013) “*De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis”, Nat Protoc, 8(8): 1494-1512.

Infantes E, Terrados J and Orfila A (2011) “Assessment of substratum effect on the distribution of two invasive *Caulerpa* (Chlorophyta) species”, Estuarine, Coastal and Shelf Science, 91: 434-441.

Jamers A, Blust R and De Coen W (2009) “Omics in algae: paving the way for a systems biological understanding of algal stress phenomena?”, Aquat Toxicol, 92(3): 114-121.

Knief C (2014) “Analysis of plant microbe interactions in the era of next generation sequencing technologies”, Front Plant Sci, 5: 216.

Lewis LA and McCourt RM (2004) “Green algae and the origin of land plants”, Am J Bot, 91(10): 1535-1556.

Li B and Dewey CN (2011) “RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome”, BMC Bioinformatics, 12: 323.

Liu L, Li Y, Li S, Hu N, He Y, Pong R, Lin D, Lu L and Law M (2012) “Comparison of next-generation sequencing systems”, J Biomed Biotechnol, 2012: 251364.

Maric P, Ahel M, Senta I, Terzic S, Mikac I, Zuljevic A and Smital T (2017) “Effect-directed analysis reveals inhibition of zebrafish uptake transporter Oatp1d1 by caulerpenyne, a major secondary metabolite from the invasive marine alga *Caulerpa taxifolia*”, Chemosphere, 174: 643-654.

McGettigan PA (2013) “Transcriptomics in the RNA-seq era”, Curr Opin Chem Biol, 17(1): 4-11.

Morozova O and Marra MA (2008) “Applications of next-generation sequencing technologies in functional genomics”, Genomics, 92(5): 255-264.

Mutz KO, Heilkenbrinker A, Lonne M, Walter JG and Stahl F (2013) “Transcriptome analysis using next-generation sequencing”, Curr Opin Biotechnol, 24(1): 22-30.

Orquera-Tornakian GK, Garrido P, Kronmiller B, Hunger R, Tyler BM, Garzon CD and Marek SM (2017) “Identification and characterization of simple sequence repeats (SSRs) for population studies of *Puccinia novopanici*”, J Microbiol Methods, 139: 113-122.

Pareek CS (2015) “Transcriptome Analysis on RNA-seq Data”, Journal of Next Generation Sequencing & Applications, S1.

Paszkiewicz K and Studholme DJ (2010) “*De novo* assembly of short sequence reads”, Brief Bioinform, 11(5): 457-472.

Patnaik BB, Wang TH, Kang SW, Hwang HJ, Park SY, Park EB, Chung JM, Song DK, Kim C, Kim S, Lee JS, Han YS, Park HS and Lee YS (2016) “Sequencing, De Novo Assembly, and Annotation of the Transcriptome of the Endangered Freshwater Pearl Bivalve, Cristaria plicata, Provides Novel Insights into Functional Genes and Marker Discovery”, PLoS One, 11(2): e0148622.

Rizzo L, Fraschetti S, Alifano P, Tredici MS and Stabili L (2016) “Association of Vibrio community with the Atlantic Mediterranean invasive alga *Caulerpa cylindracea*”, Journal of Experimental Marine Biology and Ecology, 475: 129–136.

Roberts A, Pimentel H, Trapnell C and Pachter L (2011) “Identification of novel transcripts in annotated genomes using RNA-Seq”, Bioinformatics, 27(17): 2325-2329.

Ruiz JM, Marín–Guirao L, Bernardeau–Esteller J, A. Ramos–Segura, García–Muñoz R and Sandoval–Gil JM (2011) “Spread of the invasive alga *Caulerpa racemosa* var. *cylindracea* (Caulerpales, Chlorophyta) along the Mediterranean coast of the Murcia region (SE Spain)”, Animal Biodiversity and Conservation, 34(1): 73-82.

Stabili L, Rizzo L, Pizzolante G, Alifano P and Fraschetti S (2017) “Spatial distribution of the culturable bacterial community associated with the invasive alga *Caulerpa cylindracea* in the Mediterranean Sea”, Marine Environmental Research, 125: 90-98.

Tarazona S, Garcia-Alcalde F, Dopazo J, Ferrer A and Conesa A (2011) “Differential expression in RNA-seq: a matter of depth”, Genome Res, 21(12): 2213-2223.

Unuvar OC (2016) Identification of Alternative Oxidase Encoding Genes in *Caulerpa Cylindracea*, Abant Izzet Baysal University The Graduate School Of Natural And Applied Sciences, Bolu.

Vassilev SV and Vassileva CG (2016) “Composition, properties and challenges of algae biomass for biofuel application: An overview”, Fuel, 181: 1-33.

Verlaque M, Durand C, Huisman JM, Boudouresque C-F and Parco YL (2003) “On the identity and origin of the Mediterranean invasive *Caulerpa racemosa* (Caulerpales, Chlorophyta)”, European Journal of Phycology, 38: 325-339.

Wang HD, Li XC, Lee DJ and Chang JS (2017) “Potential biomedical applications of marine algae”, Bioresour Technol, 244(Pt 2): 1407-1415.

Wang Z, Gerstein M and Snyder M (2009) “RNA-Seq: a revolutionary tool for transcriptomics”, Nat Rev Genet, 10(1): 57-63.

Young MD, Wakefield MJ, Smyth GK and Oshlack A (2010) “Gene ontology analysis for RNA-seq: accounting for selection bias”, Genome Biol, 11(2): R14.

6. CURRICULUM VITAE

Name SURNAME **:** Meryem AYDIN

Place and Date of Birth **:** Silivri, 21/08/1992

Universities

Bachelor's Degree **:** Ege University

e-mail **:** mrymaydn118@gmail.com

Address **:** Karaçayır Mah. Özlem Sitesi G1 Blok Kat:3 No:6 Merkez/Bolu