PRODUCTION OF TAXOL PRECURSORS

WITH YEAST CELL FACTORIES

Master's Thesis

Ahmad Mohammad Abdo ALKOTB

Eskişehir 2019

PRODUCTION OF TAXOL PRECURSORS WITH YEAST CELL FACTORIES

Ahmad Mohammad Abdo ALKOTB

Master of Sciences Thesis

Department of Advanced Technologies

Programme in Biotechnology

Supervisor: Asst. Prof. Dr. Hülya Karaca GENÇER

Eskişehir

Anadolu University

Graduate School of Sciences

May 2019

FINAL APPROVAL FOR THESIS

This thesis titled "Production of Taxol Precursors with Yeast Cell Factories" has been prepared and submitted by Ahmad Mohammad Abdo ALKOTB in partial fullfillment of the requirements in "Anadolu University Directive on Graduate Education and Examination" for the Degree of Master of Science in Biotechnology. Department has been examined and approved on 30 / 05 / 2019

Committee Members	**Title, Name and Surname**	**Signature**
Member (Supervisor):	Asst. Prof. Dr. Hülya KARACA GENÇER	.....
Member:	Assoc. Prof. Dr. MEHMET BURÇİN MUTLU	.....
Member:	Prof. Dr. Ahmet ÇABUK	.....

Prof. Dr. Murat TANIŞLI

Director of Graduate School of Science

ÖZET

OLUŞTURULACAK MAYA HÜCRE FABRİKLARI İLE TAKSOL ÖNCÜLERİNİN ELDESİ

Ahmad Mohammad Abdo ALKOTB

YÜKSEK LİSANS TEZİ

İleri Tekenolojiler

Biyoteknoloji (Tezli)

Anadolu Üniversitesi, Fen Bilimleri Enstitüsü, Mayıs 2019

Danışman: Dr. Öğr. Ü. Hülya Karaca GENÇER

Taksol *Taxus brevifolia* tarafından üretilen önemli bir anti-kanser ajandır. Bitkilerde istenilen maddenin üretiminde değişimler görülmesi ve üretimin istikrarsız olmasından dolayı üretim işleminin devamlı ve daha ekonomik olması için, farklı bir kaynağa gereksinim vardır. Bir çok isoprenoid bileşiğin *Saccharomyces cerevisiae'* da üretimi başarılı bir şekilde gerçekleştirilmiş ve metabolizma mühendisliği çalışmaları kullanılarak, mevalonat yolağında düzenlemeler yapılıp, verimli sonuçlar alınmıştır. Taksolde, yine bu yolla *S. cerevisiae'* da üretilebilmiştir. Mayada, mevalonat yolağında metabolizma mühendisliği çalışılarak Taksol'ün öncü molekülü olan Taksadien'in üretimini arttırmak için pek çok girişimde bulunulmuştur. Bu çalışmada ise; *S.cerevisiae*'da taksadien miktarını arttırmak için farnesil difosfat sentaz (*ERG20*), taksadien sentaz (*TS*) ve geranyl geranyl diphosphate sentaz (*GGPPS*) genleri 5 epizomal plazmit ile *S. cerevisiae*'ya transfer edilmiştir. *S.cerevisiae*'da taksadien üretimi için *GGPPS* genleri *Taxus canadensis*, S*ulfolobus acidophilus*, *Taxus baccata*, ve *Taxus cuspidata* olmak üzere dört farklı suştan elde edilmiş ve taksadien üretimi için en etkili gen belirlenmeye çalışılmıştır. Referans suş ile karşılaştırıldığında *Sulfolobus acidophilus*'tan alınan GGPPS ile, *ERG20*, ve *TS* genleri içeren suşun ile çalkalamalı kültürde taksadien üretiminin kontrol suşa oranla 31 kat arttığı belirlenmişti.

Anahtar kelimeler: Metabolizma mühendisliği; Taksadien; *GGPPS*; *S. cerevisiae*

ABSTRACT

PRODUCTION OF TAXOL PRECURSES WITH YEAST CELL FACTORIES

Ahmad Mohammad Abdo ALKOTB

Department of Advanced Technologies

Programme in Biotechnology

Anadolu University, Graduate School of Science, May 2019

Supervisor: Asst. Prof. Dr. Hülya Karaca GENÇER

Taxol is an important anticancer agent that is naturally produced by plant. As the plants yield is not consistent and is subjected to fluctuations, the need for another source became a must to make the production process more economical and consistent. Many isoprenoid compounds have been successfully produced in *S. cerevesiae* and reached good yields through modifying the mevalonate pathway by metabolic engineering tools. Taxol, in the same way, can be produced in *Saccharomyces cerevesiae*. Taxadiene is a precursor to Taxol and there have been many attempts to increase its yield in yeast through engineering the mevalonate pathway. In this study, in order to increase taxadiene yield in *S. cerevesiae*, farnesyl diphosphate synthase (*ERG20*), taxadeine synthase (*TS*) and geranyl geranyl diphosphate synthase (*GGPPS*) genes were introduced into *S. cerevesiae* via episomal plasmids. Effect of *GGPSS* genes from four different strains; *Taxus canadensis*, *Sulfolobus acidophilus*, *Taxus baccata*, and *Taxus cuspidata* was investigated to select best *GGPPS* gene for producion of taxadiene in *S. cerevesiae.* In comparison to the to the reference strain Taxadiene production increased by 31 folds by strain contained overexpressed *GGPPS* gene from *Sulfolobus acidophilus*, *ERG20* and *TS* genes.

Key Words: Metabolic engineering; Taxadiene; *GGPPS*; *S. cerevesiae.*

STATEMENT OF COMPLIANCE WITH ETHICAL PRINCIPLES AND RULES

I hereby truthfully declare that this thesis is an original work prepared by me; that I have behaved in accordance with the scientific ethical principles and rules throughout the stages of preparation, data collection, analysis and presentation of my work; that I have cited the sources of all the data and information that could be obtained within the scope of this study, and included these sources in the references section; and that this study has been scanned for plagiarism with "scientific plagiarism detection program" used by Eskişehir Technical University, and that "it does not have any plagiarism" whatsoever. I also declare that, if a case contrary to my declaration is detected in my work at any time, I hereby express my consent to all the ethical and legal consequences that are involved.

.....

Ahmad Mohammad Abdo ALKOTB

Tables List

Figures List

Abbreviations

DNA	Deoxyribonucleic acid
URA	uracil requiring
Yeps	Yeast episomal plasmids
Yips	Yeast Integrative plasmids
Yrps	Yeast Replicating plasmids
RNA	Ribonucleic acid
mRNA	Messenger ribonucleic acid
CPSF	Cleavage and polyadenylation specificity factor
CstF	Cleavage stimulation factor
IPP	Isopentenyl diphosphate
DMAPP	Dimethylallyl diphosphate
MVA	Mevalonate
MEP	2-C-methyl-D- erythritol 4-phosphate
HMGR	Hydroxy-3-methylglutaryl-CoA reductase
HMG1	Hydroxy-3-methylglutaryl-CoA reductase
ATP	Adenosine triphosphate
ERG20	Farnesyl diphosphate synthase.
GGPPS	Geranylgeranyl diphosphate synthase
TS	Taxadiene synthase
GGOH	Geranylgeraniol
MSA	Multiple Sequence Alignment
BLAST	Basic Local Alignment Search Tool
NCBI	National center of biotechnology information
CDS	CoDing Sequence
SRA	Sequence Read Archive
PCR	Polymerase chain reaction

Contents

1. INTRODUCTION

Synthetic biology is an interdisciplinary field of science that emerges from combination of biology and engineering science fields. It aims at modelling and constructing the biological systems to make them understandable and easier to solve biological problems and challenges.

There are many application fields that emerged in this science. Invention of **biosensors** that depend on existent biological processes to detect various elements such as toxins and heavy metals. Another breakthrough of synthetic biology is the production of **synthetic bacteria** where the whole bacterial chromosome was completely synthesized. This was achieved by Craig Venter institute in 2010 where *Mycoplasma mycoides* genome was completely synthesized and introduced into empty *M. capricolum* cell and a new *M. mycoides* cell has been produced [1]. One of the most applications of synthetic biology is **metabolic engineering** and cell transformation which led to a revolution in production of heterogeneous compounds of economic value in microorganisms. From the industrial prospective, metabolic engineering is now gaining a positive momentum in being a convenient and cost-effective way to produce biofuels and many other economical chemicals [8].

The aim of this study is to investigate the effect of overexpression of *GGPPS* genes from four different microorganisms; *Taxus canadensis*, S*ulfolobus acidophilus*, *Taxus baccata*, and *Taxus cuspidate* on taxadiene production in *S. cerevesiae.*

1.1. Metabolic Engineering

Genetic engineering refers to the modification of an organism's DNA either by deletion or introduction of genes into organism's genome.

First successful attempt to introduce foreign genes into cell was done by Cohen & Boyer [2], they managed to introduce foreign DNA into *E. coli* cells using restriction enzymes and plasmids.

Insulin was successfully produced through biotechnological engineering of *E. coli* through introducing insulin peptide chains and producing genes into *E. coli*. Then there

was a need to introduce, delete more than one gene, or modifying whole metabolic pathways to produce certain products, thus metabolic engineering evolved.

Metabolic engineering field was defined as "directed improvement of product formation or cellular properties through the modification of specific biochemical reactions or the introduction of new genes with the use of recombinant DNA technology." [4].

Metabolic pathways include many enzymatic steps and thus the metabolic engineering focuses on overexpressing and/or introducing the genes responsible for the production of these enzymes.

Metabolic engineering is the field that modifies metabolic pathways inside microorganisms with using recombinant DNA technology to either increase the production of native metabolites [5,6] or entirely new chemical products that are not naturally found in the host organism [7].

1.2 Metabolic Engineering Tools

Metabolic engineering aims at increasing the production of key products through many different ways such as (1) overexpressing genes that catalyzes the rate limiting steps in a metabolic pathway (2) deleting the enzymes that catalyzes the competing metabolic pathways inside the microorganism, and (3) introduction of heterologous genes if the organism does not naturally produce the key product [9].

Metabolic engineering depends on DNA recombinant technology to carry out these modifications and thus there are many tools that must be adjusted to perform effective cloning and gene expression process.

Plasmids are small circular double stranded DNA molecules that are present separately from chromosomal DNA and replicate in an independent way. They are found in bacterial cells and may be found in some eukaryote organisms. Plasmids carry useful genes for the survival of microorganism in drastic and lethal conditions such as antibiotics resistance genes. Plasmids are amongst the most essential tools of molecular biology. It is useful in molecular cloning for transferring heterologous genes into cells. For a plasmid

to be able to independently replicate inside a cell, it must have an origin of replication [11].

Plasmids are classified into five types:

- Fertility plasmids: They are responsible for sexual conjugation between cells.
- Resistance plasmids: They are responsible for resisting lethal conditions such as antibiotics and heavy toxic metals.
- Col plasmids: They are responsible for the production of toxic proteins that kill other competent cells in the environment.
- Degradative plasmids: They aid in digestion of unwanted materials in the environment such as salicylic acid and xylene.
- Virulence plasmids: They are responsible for making bacteria pathogenic.

Plasmids are used to amplify or express a certain gene [12]. Artificial vectors are synthesized to do that job. They must contain an origin of replication and a selection gene as antibiotic resistance (in case of bacteria) or URA3 (in case of yeast) and a multiple cloning site that will allow for the cloning of genes. Normal cloning plasmids are used to transfer DNA fragments up to 15 kilo base pair. Cosmids, yeast artificial chromosomes or bacterial artificial chromosomes are used for larger DNA fragments [13]. Episomal vectors are plasmids that are able to integrate themselves into the chromosomal DNA and also can replicate independently inside the cell [14]. There are many advantages associated with using episomal vectors; there is no risk of gene interruption that usually happens during gene integration into chromosomal DNA. Moreover, the risk of cell's chromosomal DNA disruption after integration is diminished. Also, episomal vectors amplify the inserted gene as they can stay in multiple copies inside the cells [20].

Most of the vectors used in *S. cerevisiae* are shuttle vectors [17]. Shuttle plasmid vectors are the vectors that can replicate in 2 different types of organisms, usually *E. coli* and yeasts [18]. They allow for genes modification and amplification in *E. coli* and then to be used in yeast that are slower in growth than *E. coli*. Shuttle vectors contain origin of replication and antibiotic resistance gene (for *E. coli)*. and autonomously replicating sequence, centromere and a selective gene such as URA3 (for yeasts).

Heterogenous DNA can be transferred to yeast through two ways; homologous integration and yeast plasmids. Chromosomal integration is a process that modifies genes inside chromosomes using homologous recombination to insert new genes, delete or disrupt existing genes inside the genome using gene targeting [15]. DNA double strand breaks repair mechanism in *S. cerevisiae* depends mainly on homologous recombination. So, yeast cells will deal with introduced linear DNA as a broken DNA and will integrate it inside its genome into the target site that contains identical sequence as the inserted DNA fragment [16].

Yeasts contain endogenous small circular 2 μm plasmids. There are also different types of plasmids used in yeast cloning and transformation processes [17]. Yeast Integrative plasmids (Yips) are synthesized from 2 μm yeast plasmids and do not contain an origin of replication and only depend on integration into cell's chromosomal DNA. They normally contain a selective gene such as URA3. Yeast episomal plasmids (Yeps) contain a replication origin based on the endogenous 2 μm yeast plasmid's origin of replication. Yeast Replicating plasmids (YRps) contain an autonomous replicating sequence which is different from endogenous 2 μm yeast plasmid's origin of replication. They have higher copy number but are less stable than YEps.

Restriction enzymes are naturally produced enzymes by bacteria as a defensive mechanism against invading viruses as it cleaves any introduced foreign DNA molecules and thus making it inactive. Restriction enzymes were discovered during the studies of bacteriophage in 1950s. Where some λ phages' yields were not as much as expected in some of *E. coli* strains and the reason discovered to be the cleavage of phage's DNA by restriction enzymes [19].

During an invasion of foreign organism or introduction of a foreign DNA molecule, the restriction enzymes inside the bacteria digests only the foreign DNA while the endogenous bacterial DNA is protected by methyltransferase enzymes that methylate adenine or cytosine bases within the recognition sites to protect it from the effect of its own restriction endonucleases. Restriction endonuclease and its corresponding methyltransferase form what is called the restriction modification system [21].

Restriction enzymes are a subtype of endonucleases that make incisions within (endo) the sugar backbone of the DNA molecule by breaking the phosphodiester bonds. Thus, they are called restriction endonucleases. While the exonuclease only cut at the recognition sites' endings. Restriction endonuclease cuts sequences of a length of 4-8 nucleotide called restriction sites. Most of these sites are palindromic where their nucleotide sequence can be read the same backwards and forwards within the same DNA strand (mirror palindromic) or in the complementary DNA strands (inverted repeat palindromic) [22]. Around 3500 restriction enzymes have been discovered so far [23].

There are 4 types of restriction enzymes. Type 1 and Type 3 are similar, and their restriction and methylations activity are done by one enzyme complex. The restriction occurs randomly, as they cleave the DNA in a random manner, sometimes hundreds of bases away from the recognition site. Type 2 has separate restriction enzyme and methylase. It cuts at specific site within the recognition site not in a random manner. This type of restriction enzymes is very important for cloning processes as they can cut the DNA forming blunt or steaky ends that helps in easing the ligation process (**Figure 1.1**) [24]. Type 4 only carries out restriction on modified DNA such as methylated DNA.

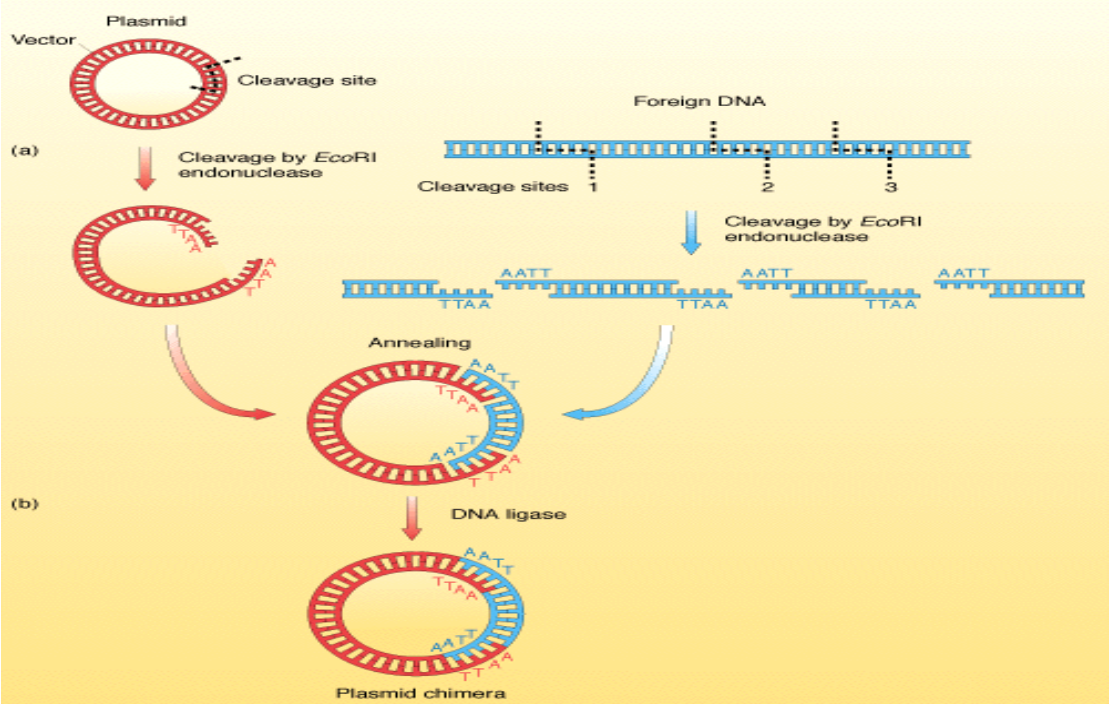


Figure 1.1. *Restriction enzymes and cloning process [24].*

Promoter is the part of DNA that is responsible for the initiation of DNA transcription process. Length of promoters varies between 100-1000 bp and they are located near transcription start regions of genes. Promoters provide a binding site for the polymerase enzymes. Upon binding conformational changes occur in the shape of polymerase enzyme and DNA and this initiates the transcription. Promoters in eukaryotes contain a core promoter, proximal, distal regions upstream of the core promoter. Core promoter is approximately 40 nucleotides upstream of the transcription start site and it contains general transcription initiation factor binding sites such as TATA box [61], and cis-elements that bind to the transcription factors and RNA polymerase [62]. Proximal and distal regions of the promoter contain sequences that are responsible for transcription regulation too such as enhancers, silencers, insulators. [63]. Distal regions have lower effect on the transcription level than proximal region. Promoters are either constitutive or regulatable. Constitutive promoters are active in all cell conditions without any additional stimulating substances while regulatable promoters are induced or repressed by additional substances. One of the important factors for achieving the desired trancription levels of cloned genes is the choice of promoters. *pTEF1*, *pADH1* are strong constitutive promoters commonly used for high-level gene expression [17].

Terminator is a sequence of nucleotides that marks the end of a gene signaling the transcription. Polymerase II is associated with some protein factors that are responsible for recognition of terminal signals during transcription of mRNA.
In eukaryotes, termination occurs when the poly-A tail is formed during transcription, cleavage and polyadenylation specificity factor (CPSF) and cleavage stimulation factor (CstF) transfer from RNA polymerase II to the poly-A signal and this signals the cleavage of mRNA from transcription complex [64]. In prokaryotes, there are two types of terminators. Rho-dependent terminators, where rho factor protein is responsible for disruption of transcriptional complex. [65] In case of rho-independent terminators, the disruption of transcriptional complex is triggered by a self-annealing hairpin structure that forms on the transcript upon transcripting the terminator sequence [66]. Terminators commonly used for gene expression in *S. cerevisiae* [17] are derived from the *PGK1*, *ADH1*, and *CYC1* genes.

Bioinformatics is an interdisciplinary science that emerged from combination of biological science, statistics, mathematics, and computational sciences. This branch of science enables us to save a huge amount of data that can be processed to extract useful data to help in developing solutions to biological problems.
Areas of interest in bioinformatics:

- Bioinformatics and computational biology.
- Genetics, genomics, metabolomics, trancriptomics and proteomics,
- Systemic biology

Genomics is the study of genome which is the complete genes profile of an organism. It studies the sequence analysis of genomes. It differs from genetics in such that it focuses on the genes as a whole and studies their structure, mapping, and functions [52]. Metabolomics is the study of metabolites, that are resulting from certain cellular processes (such as hormones, metabolic intermediates, and secondary metabolites) and their chemical reactions that are found within certain organism [53]. Transcriptome is the collection of all the RNA molecules inside a certain organism. Sometimes it refers to only mRNA molecules. It depends on the environment as it reflects the active expressed genes in the cells at a certain time and environmental conditions. Transcriptomic studies focus on examining mRNA expression levels in the cells. Proteomics is the study of all proteins produced by a certain organism [54]. Proteome complexity increases due to post translational modifications. Proteomics helps in identification of increasing and different types of proteins.

The ultimate goal of bioinformatics is to combine all the biological data to form a comprehensive image of all the biological activities. Consequently, this increases our understanding of all biological processes. Thus, bioinformatics involves analysis of many data such as nucleotide and amino acids sequences, protein domains and structures. Computational biology is the science of data analysis process [55]. Bioinformatics allows the storage of genomes and proteomes sequences and their mutations; thus, we can retrieve the sequences of genes of relative microorganisms that are not restored yet on the databases. In this study, we have been able to utilize this feature to get the geranyl geranyl diphosphate synthase (*GGPPS*) sequences of *Taxus baccata, and Taxus cuspidata*.

Sequence analysis is an important branch of bioinformatics. There is more than one type of sequence analysis [56].

- Pairwise sequence alignment: in this type of alignment the level of similarity between two genes or amino acid sequences is measured through a specific scoring process (algorithm).
- Local sequence alignment: In this type, the query is being matched with a portion of the reference. The algorithm used for this type of alignment is Smith–Waterman [57]. It is usually performed to detect homologous domains in non-homologous gene sequences.
- Global sequence alignment: The matching is being carried out with the whole structure of the reference from end to end. Needleman–Wunsch is the algorithm used for this sequence alignment [58]. It is usually performed on homologous genes.
- Multiple Sequence Alignment (MSA): Aligning more than two sequencing is called multiple sequence alignment, this is one of the most used algorithm in molecular biology and it is used to find conserved regions which are very important in predicting the structure of proteins from members of shared biological ancestry and assessing the conservation of protein structures [56].

Some important algorithmic programs used in bioinformatic:

- BLAST: Basic Local Alignment Search Tool. It is a program developed by the national center of biotechnology information (NCBI). This program carries out research task in the sequences database. There are many different BLAST tools such as BLASTP that is related to protein database, BLASTN that is related to nucleotides database, BLASTX that searches for nucleotide sequence after translating a protein query, and TBLASTN that searches for protein sequence after translating a nucleotide query.
- Crustal: is a program developed by Thompson et. al. in 1994 [59]. For multiple sequence alignment of amino acids or nucleotides.

- Open Reading Frame Finder (ORF finder): It is a tool developed by NCBI and it looks up query DNA sequence for open reading frames. It is mainly used to check for potential protein coding segments in DNA sequences
- FASTA: It an algorithm for protein and DNA sequence alignment that has been developed by Lipman and Pearson in 1988 [60].

1.3. Terpenes

There are more than 50,000 terpenoid compounds of important function in the world [25]. They are used as flavors, perfume agents, coloring agents, and pharmaceuticals. Terpenoids are composed of five-carbon isoprene units. [26]. Isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP) are precursors of isoprene, they are synthesized in eukaryotes, archaea by mevalonate (MVA) pathway and in prokaryotes and plants by 2-C-methyl-D- erythritol 4-phosphate (MEP) pathway [27].

Terpenoids are subdivided into hemiterpenoids (C 5), monoterpenoids (C 10), sesquiterpenoids (C 15), diterpenoids (C 20), triterpenoids (C 30), tetraterpenoids (C 40), and long chain isoprenyl products (**Figure 1.2**) [26].

There are many terpenoid products that produced in *S. cerevisiae* such as 11β-hydroxymanoyl oxide which is aprecurser of forskolin, [28].

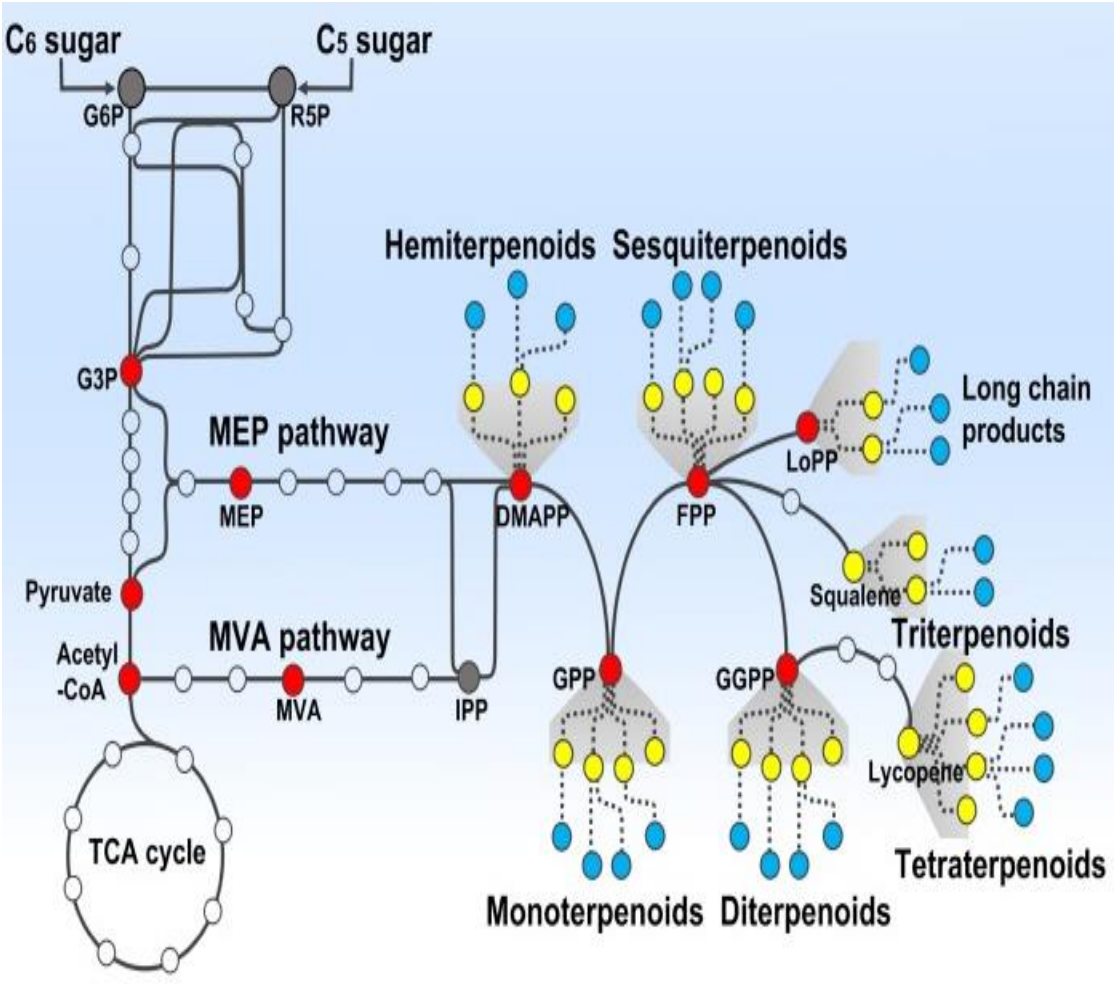


Figure 1.2. *Different types of terepnoid compounds and their synthesis from MVA and MEP pathways [26].*

Plants are the source from which most terpenoids have been discovered. Terpenoids are mainly produced through extraction from these plants, but the major problem is the low yield as there are many challenges in depending on plants as the only source for obtaining these compounds. Plants are slowly growing, and sometimes weather conditions add limitations to the growth. In most cases, the amount of compound is very low inside plants. It is difficult to engineer plants for increased isoprenoids production in plants. On the other hand, microorganisms grow fast and are not dependent on the weather conditions [26].

Microbials host for industrial production should be a fast-growing robust organism that can withstand harsh conditions of osmotic pressure and pH. It also should have a

well-defined and studied metabolic pathways and well-developed genetic tools. Moreover, it must be economical in terms of consuming cheap carbon sources for growth.

Both *S. cerevisiae* and *E. coli* are widely used as metabolic hosts for production of terpenoids. *S. cerevisiae* is a unicellular eukaryote that combine the advantages of prokaryotic organisms as fast growth and genetic manipulation; and the advantages of eukaryotes of post transitional modifications of proteins [28]. Yeast is a robust microorganism that can withstand drastic industrial conditions such as low pH and high osmotic pressure and has an immune system for phage infections [29]. *S. cerevisiae* has an advantage over *E. coli*, namely; the cytochrome P450 system that allows for post translational modification and as a result this leads to easier extraction process and more diversive terpenoids production [26], The native MEP pathway could be less efficient in IPP and DMAPP supply for terpenoid production. *E. coli* has been engineered to improve IPP and DMAPP synthesis by augmenting enzymes of MEP pathway. Yeast is considered as a safe organism in comparison to other organisms that can be used in metabolic engineering processes without safety dangers. Moreover, genetic engineering tools for *S. cerevisiae* are well developed [17].

Yeast and other eukaryotes synthesize isoprenoids through mevalonate pathway (**Figure 1.3**) [61, 62]. The pathway starts with acetyl-CoA that is converted into acetoacetyl-CoA (AA-CoA) in the cytosol [62], and proceeds to 3-hydroxy-3-methylglutaryl-CoA (HMG)-CoA, HMG-CoA is converted to mevalonic acid by 3-hydroxy-3-methylglutaryl-CoA reductase (HMGR), the rate-limiting enzyme of mevalonate pathway [63]. Mevalonic acid is phosphorylated with mevalonate kinase to mevalonate-5-phosphate. Mevalonate-5-phosphate is converted to mevalonate 5-diphosphate by Phosphomevalonate kinase consuming one ATP molecule. Mevalonate 5-diphosphate is decarboxylated by Mevalonate diphosphate decarboxylase into isopentyl diphosphate (IPP). IPP is isomerized to dimethylallyl diphosphate DMAPP by IPP isomerase. Isoprenoids are derived from the combination of IPP and DMAPP the two molecules.

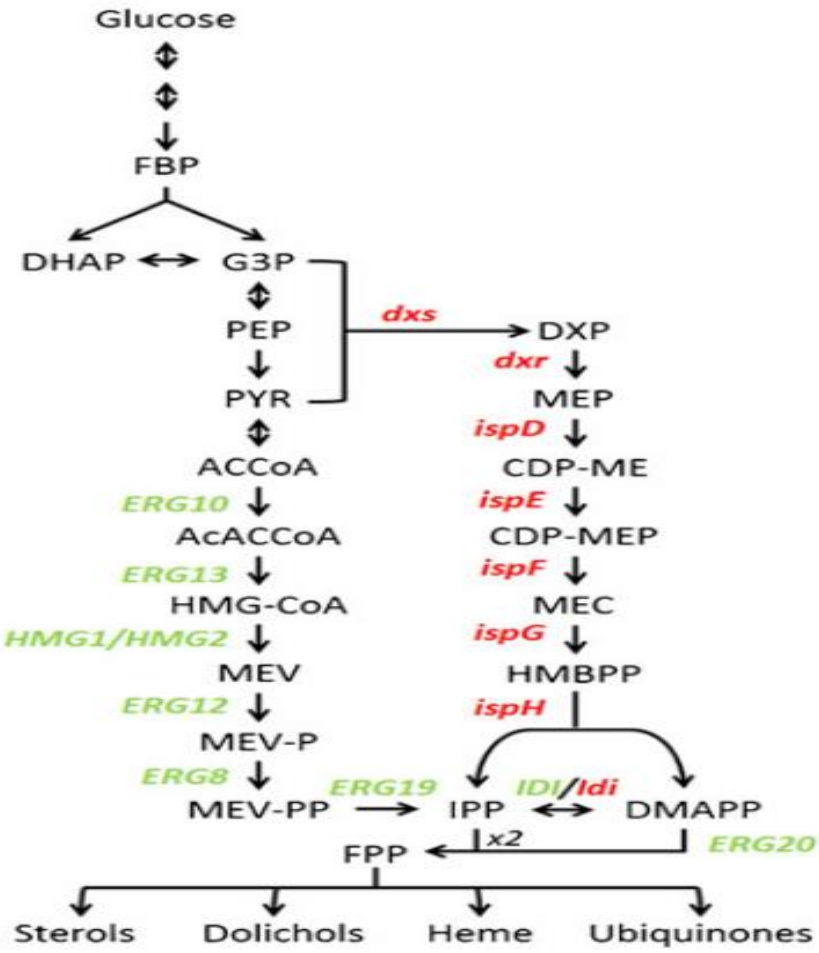


Figure 1.3. *Isoprenoid pathways in Yeast and E. coli [62]*.

1.3.1. Paclitaxel (taxol):

Paclitaxel (Taxol) is a terpenoid compound produced by *Taxus brevifolia* [29]. Pharmaceutical importance of Paclitaxel was not discovered until late seventies because of the extraction difficulties [30]. Paclitaxel is used as anticancer agent for numerous types of cancers as breast cancer, non-small cell lung cancer, ovarian cancer, and AIDS-related Kaposi's sarcoma [31]. Sales of Paclitaxel reaches over $1 billion per year [32]. Paclitaxel is derived from geranylgeranyl diphosphate which is derived from mevalonate. Biosynthesis pathway of Taxol includes 19 enzymatic steps from geranylgeranyl diphosphate (GGPP) precursor [33]. So far 15 enzymes have been well characterized. [34]. These enzymes are Geranylgeranyl diphosphate synthase (*GGPPS*), Taxadiene synthase (TS), Taxadiene-5α-hydroxylase (T5αOH), Taxadiene-13α-hydroxylase

(T13αOH), Taxadiene-5α-ol-O-acetyl transferase (TAT), Taxane-10β-hydroxylase (T10βOH), Taxane-10β-hydroxylase a (T10βOH), Taxane-9α-hydroxylase b (T9αOH), Taxane-7β-hydroxylase b (T7βOH), Taxadiene-2α-hydroxylase (T2αOH), Taxane-1β-hydroxylase (T1βOH), Taxane-2α-O-benzoyl transferase (TBT), C4β,C20-epoxidase (EPOX), Oxomutase (OXM), Taxane-9α-dehydrogenase (T9αDH), 10-deacetylbaccatin III-10-O-acetyl transferase (DBAT), Phenylalanine aminomutase (PAM), β-phenylalanoyl-CoA ligase c (PCL), Baccatin III: 3-amino, 13-phenylpropanoyltransferase (BAPT), Taxane-2'α-hydroxylase (T2'αOH), N-benzoyl transferase (DBTNBT), NADPH:cytochrome P450 reductase (NADPH:P450-Red), and β-phenylalanine-CoA ligase (**Figure 1.4**) [34].

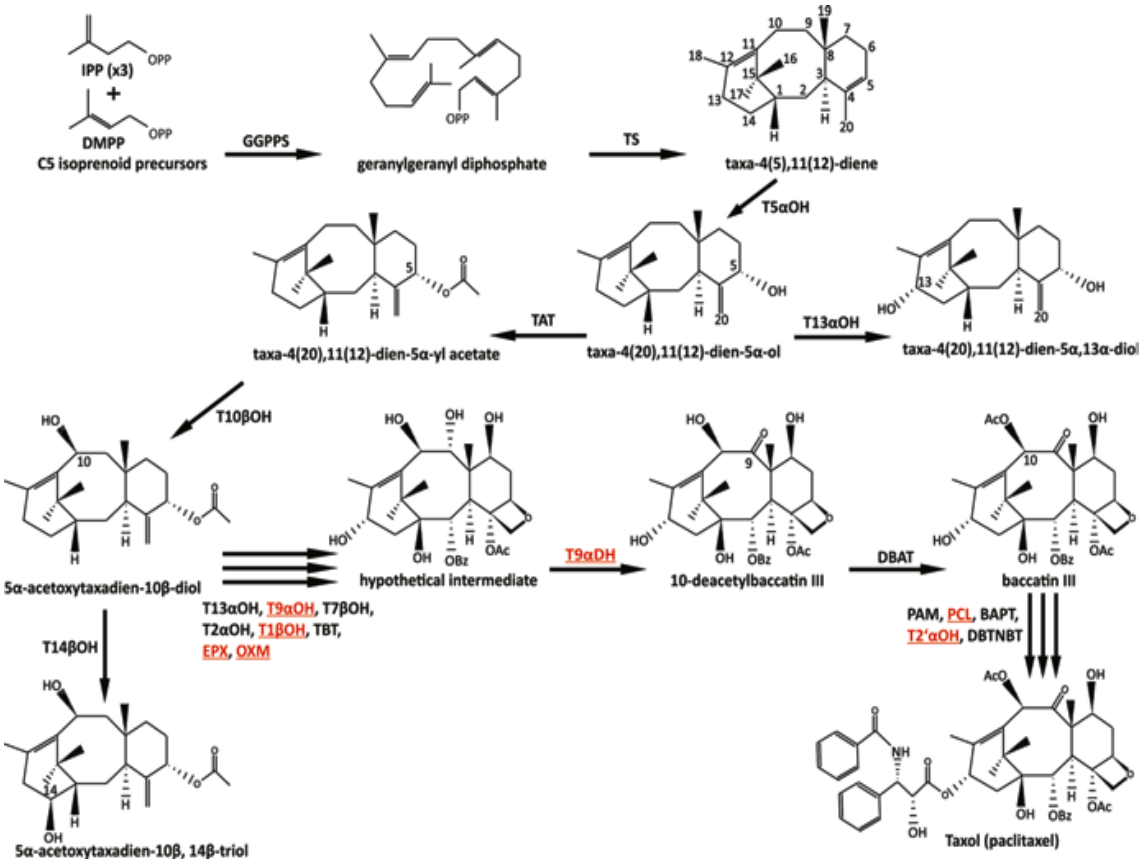


Figure 1.4. *An overview of the production pathway of Taxol starting from geranylgeranyl diphosphate through Taxol [34].*

Paclitaxel is mainly extracted from barks of *Taxus brevifolia* (yew), but the plant's slow growth is a challenging problem, the paclitaxel content is only 0.01% of the dry

weight of the bark. One tree can only produce up to 300 mg [35, 36]. This leads to the high cost of production which reaches $600,000 per kilogram [37]. This led to searching for alternatives to produce paclitaxel such as natural plant extraction, plant cell culture, chemical synthesis, and semi-synthesis using biosynthetic intermediates.

The total chemical production is being developed but it has still a complex multistep synthesis, also there is the stereoisomerism problem. For all of this, it is very costly and not a practical economical way to produce taxol through chemical synthesis on a large scale. Currently paclitaxel costs every patient of Breast cancer around $5000. [38] And thus it is very costly and hard for all patients to get access to treatment. And that's what makes it worth finding a lower costly production method to make it available as artemisinin (antimalarial drug) has been available at low cost using the means of metabolic engineering [39].

In the meantime, Taxol is produced by semisynthesis as taxus cell cultures are used to produce taxol precursors like baccatin III and 10-deacetylbaccatin III. Production of paclitaxel through metabolic engineering will provide a stable source of paclitaxel in comparison to extraction from the natural plant. In addition, it is more economical than chemical synthesis [40].

Taxadiene is a key precursor in paclitaxel synthesis. Through metabolic engineering tools terpenes production can be achieved at good yields in S. *cerevisiae* and *E. coli* [41]. Taxadiene yield in *E. coli* reached 1020 mg/L. [42] However, *S. cerevisiae* is more suitable than E. coli for downstream processing and industrial production. [43].

Many attempts have been done to produce taxadiene through modifying mevalonate pathway, IPP and DMAPP are combined by geranyl diphosphate to produce granyl diphosphate, which is converted to farnesyl diphosphate by farnesyl diphosphate synthase (*ERG20*). Then, geranylgeranyl diphosphate synthase (*GGPPS*) produces geranylgeranyl diphosphate (GGPP) from farnesyl diphosphate. Geranylgeranyl diphosphate is then converted to taxadiene by taxadiene synthase (TS).

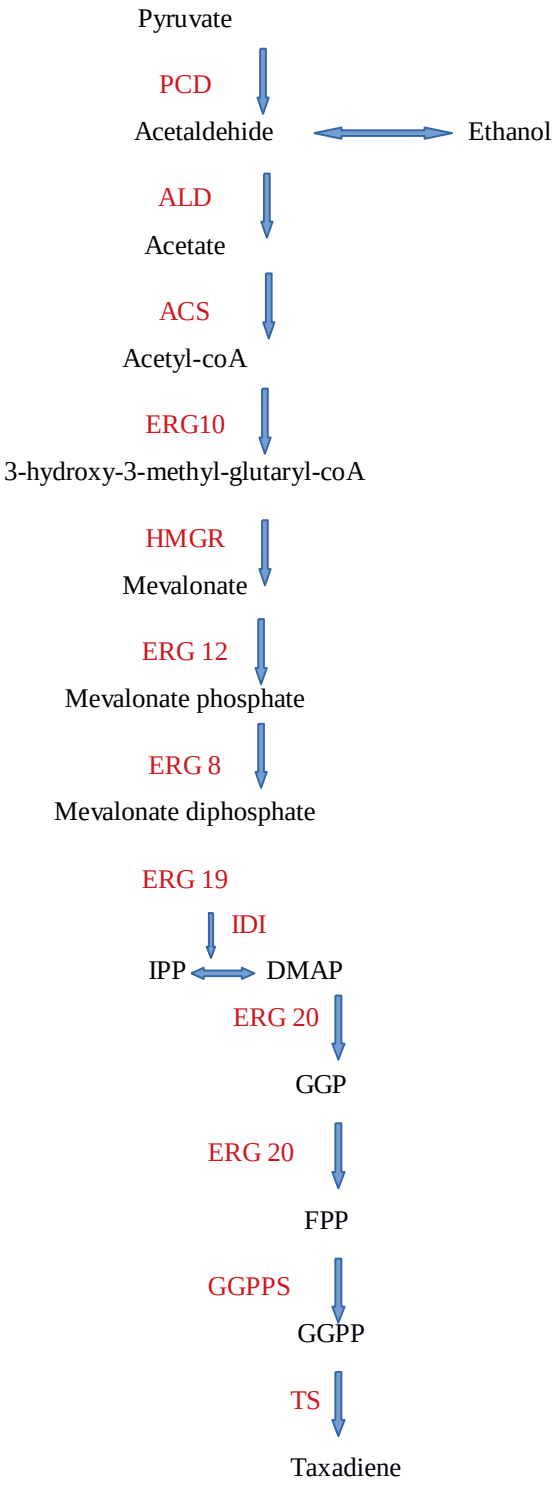


Figure 1.5. *Mevalonate pathway in yeast and relationship with taxadiene synthesis*.

1.3.1.1 *Geranylgeranyl diphosphate synthase GGPPS as an intermediated enzyme in taxol production*

In order to increase terpene production, the supply of precursor metabolites in mevalonate pathways has to be increased [43]. Geranylgeranyl diphosphate (GGPP) is a precursor of many terpenoid compounds. The natural produced content of geranylgeranyl diphosphate (GGPP) in *S. cerevisiae* is very low [44]. Thus many studies have been done to enhance the terpenoids production through improving and enhancing GGPP production in order to increase the production of its derivative terpenoid compounds. Geranylgeranyl diphosphate synthase (*GGPPS*) exists in a variety of plants, animals and microbes with the similar functions of catalyzing DMAPP or FPP to GGPP. However, the catalytic capabilities of GGPP synthase may be different.

The following studies will present the effect of enhanced *GGPPS* gene expression from different microorganisms on the overall yield of different terpenes. The taxol level has been increased in *Paraconithyrium* fungus after integration of *T. canadensis* geranylgeranyl diphosphate synthase (*GGPPS*) gene [45].

In another study, when taxadiene synthase gene was only integrated into episomal vector and introduced into yeasts, taxadiene was not detected. Geranylgeranyl diphosphate synthases from *Erwinia herbicola*, *S. cerevisiae* and *T. baccata x T. cuspidate* strains were selected to be constructed into episomal plasmids, the highest taxadiene amount was produced by *T. baccata* x *T. cuspidate* strain, it reached a level of 72.8 mg/L [46].

In another study *GGPPS* from *Sulfolobus acidocaldarius* was integrated into *S.* cerevisiae to increase the production of Geranylgeraniol (GGOH) [47]. Geranylgeraniol is produced through the dephosphorylation of geranylgeranyl pyrophosphate. It is an important ingredient in perfume industry and can be used in production of many important vitamins such as vitamin A and E [48, 49]. The Geranylgeraniol yield reached 437.52 mg/L. Lycopene yield was increased by 770 folds after the integration of *GGPPS* from *S. acidocaldirius* [50]. In this study four key enzymes' genes were overexpressed. These enzymes were truncated 3-hydroxy-3-methylglutaryl-CoA reductase gene (tHMG1), a fusion gene of FPP synthase (*ERG20*) and endogenous GGPP synthase and

GGPP synthase gene from *S. acidocaldirius*. Also, the lycopene synthetic genes from *Pantoea agglomerans* were integrated. The lycopene level reached 135.21 mg/L.

The production of miltiradiene was enhanced by introduction of *GGPPS* from *S. acidocaldarius* [51]. Miltiradiene is a precursor to tanshinones which are important terpenoid compounds used in the traditional Chinese medicine to treat various cardiovascular diseases. The geranylgeranyl diphosphate pool was increased by overexpression of FPP synthase (*ERG20*), endogenous *GGPPS* and together a heterologous GGPP synthase from *S. acidocaldarius*. This increased miltiradiene production up to 8.8 mg/L.

In our study, four different *GGPPS* genes from *S. acidocaldarius, T. canadensis, T. baccata,* and *T. cuspidata* activity will be screened to detect which gene has the ability to enhance the yield of taxadiene. The amounts of produced taxadiene will be measured and accordingly the best strains will used for further studies. The amounts of produced taxadiene will be measured and accordingly the best strains will used for further studies. There were some challenges in finding the gene sequences of geranylgeranyl diphosphate synthase (*GGPPS*) gene of *T. baccata* and *T. cuspidata* so available bioinformatics tools have been used to retrieve their sequences.

2. MATERIALS AND METHODS

2.1 *GGPPS* Gene Sequences Searching In NCBI

GGPPS genes sequences of *S. acidocaldirus, T. baccata* and *T. cuspidata* were not found in National Center for Biotechnology Information (NCBI) database so bioinformatic screeninig tools were used to get the sequences. These tools included BLAST, ORF finder, SRA, Clustal and genome DeNovo assembly.

2.1.1. *Sulfolobus acidocaldirus GGPPS* gene

Amino acid sequence of *S. acidocaldirus GGPPS* protein was retrived from Unibrot (accession number: ID UPI000012B3FE), then it was reverse translated into nucleotide sequence and its sequence synthesis was orderd after making codon optimization according to *S. cerevisae.*

2.1.2. *Taxus baccata* and *Taxus cuspidata GGPPS* gene

Neither the nucleotide sequence nor the protein sequence has been found on NCBI database so the following approach was taken.

- *GGPPS* reference sequence finding

To get the sequence of *GGPPS* gene of both *Taxus baccata* and *Taxus cuspidata*, refenece *GGPPS* sequence should be found. Complete CDS (CoDing Sequence) of available *GGPPS* genes' mRNA in *Taxus spp.* was searched. As the complete CDS of the found sequences is longer than *GGPPS* sequence, NCBI ORF finder tool was used to find the sequences that contain open reading frames (ORF) i.e. the sequences coding for *GGPPS* gene. After getting the ORF maps of the analyzed mRNA sequences, *GGPPS*' protein and nucleotide sequences of 3 members in Taxus spp. (T. *canadensis, Taxus x media, and T. chinensis)* were retrieved and their aminoacid similarities were analyzed to get a refenece sequence (that includes the common aminoacids in all the *GGPPS* genes of *Taxus spp.).* The *GGPPS* of *T. canadensis* was chosen to be the reference sequence as it conserves most of the similarities of *GGPPS* genes in *Taxus spp*. (**Figure 2.1)**

```
x_media          MAYTAMAAGTQSMQLRTVASYQECNSMRSCFKLTPFKSFHGVNFNFPSLGAANCGIMGHL
canadensis       MAYTAMAAGTQSLQLRTVASYQECNSMRSCFKLTPFKSFHGVNFNVPSLGAANCEIMGHL
chinenses        MAYTAMAAGTQSMQLRTVASYQECNSMRSCFKLTPFKSFHGVNFNFPSLGAANCEIMGHL

x_media          KLGSLPYKQCSVSSRSTKTMAQLVDLAETEKAKGKDIEFDFNEHMKSKAVAVDAALDKAI
canadensis       KLGSLPYKQCSVSSKSTKTMAQLVDLAETEKAEGKDIEFDFNEYMKSKAVAVDAALDKAI
chinenses        KLGSLPYKQCSVSSRSTKTMAQLVDLAETEKAEGKDIEFDFNEYMKSKAVAVDAALDKAI

x_media          PLEYPEKIHEPMRYSLLAGGKRVRPALCIAACELVGGSQDLAMPTACAMEMIHTMSLIHD
canadensis       PLEYPEKIHESMRYSLLAGGKRVRPALCIAACELVGGSQDLAMPTACAMEMIHTMSLIHD
chinenses        PLEYPEKIHESMRYSLLAGGKRVRPALCIAACELVGGSQDLAMPTACAMEMIHTMSLIHD

x_media          DLPCMDNDDFRRGKPTNHKVFGEDTAVLAGDALLSFAFEHIAVATSKTVPSDRTLRVISE
canadensis       DLPCMDNDDFRRGKPTNHKVFGEDTAVLAGDALLSFAFEHIAVATSKTVPSDRTLRVISE
chinenses        DLPCMDNDDFRRGKPTNHKVFGEDTAVLAGDALLSFAFEHIAVATSKTVPSDRTLRVISE

x_media          LGKTIGSQGLVGGQVVDIASEGDANVDLKTLEWIHIHKTAVLLECSVVSGGILGGATEDE
canadensis       LGKTIGSQGLVGGQVVDITSEGDANVDLKTLEWIHIHKTAVLLECSVVSGGILGGATEDE
chinenses        LGKTIGSQGLVGGQVVDITSEGDANVDLKTLEWIHIHKTAVLLECSVVSGGILGGATEDE

x_media          IARIRRYARCVGLLFQVVDDILDVTKSSEELGKTAGKDLLTDKATYPKLMGLEKAKEFAA
canadensis       IARIRRYARCVGLLFQVVDDILDVTKSSEELGKTAGKDLLTDKATYPKLMGLEKAKEFAA
chinenses        IARIRRYARCVGLLLQVVDDILDVTKSSEELGKTAGKDLLTDKATYPKLMGLEKAKEFAA

x_media          ELATRAKEELSSFDQIKAAPLLGLADYIAFRQN
canadensis       ELATRAKEELSSFDQIKAAPLLGLADYIAFRQN
chinenses        ELATRAKEELSSFDQIKAAPLLGLADYIAFRQN
```

Figure 2.1. *CLUSTAL O (1.2.4) multiple sequence alignment of GGPPS aminoacids sequences of Taxus canadensis, Taxus x media, and Taxus chinensis*

GGPPS of *T. canadensis* was used to find the *GGPPS* gene sequences of both *T. baccata* and *T. cuspidata* as follows.

- Finding and analysing transcriptome projects of both *T. baccata and T. cuspidata*

The transcriptomes of both orgsanisms were searched on the Sequence Read Archive (SRA) on NCBI, we found seveal transcriptome projects and the following projects were selected.

- ✓ *For T. baccata*, 454 sequencing of *T. baccata* EST project (accn : SRX026383)
- ✓ *for T. cuspidata*, 454 sequencing of *T. cuspidata* transcriptome (accn: SRX015126) and Transcriptome Analysis of Japanese yew (accn: ERX008014)

To find similar sequences to *GGPPS* gene in these transcriptome data we made BLAST analysis to make a sequence alignment of available *GGPPS CDS* sequence of *T. canadensis* against all the sequences found in these transcriptomic data. All the results of sequence alignments were downloaded.

- Finding the consensus sequence of both genes.

The downloaded sequences alignments were assembled using Genome De-Novo Assembly to get the consensus sequence of both genes. Consensus sequences were aligned with *GGPPS* sequences of other three *Taxus spp*. family members. Then after making reverse translation of consensus amino acid sequences and codon optimization according to *S. cerivisae, GGPPS* genes sequences of both *T. baccata and T. cuspidata* were achieved.

T. baccata GGPPS gene was orderd from Oligomer ® company after codon optimization according to *S. cerevisiae*. There are only 3 amino acids differences in protein structure of *GGPPS* genes of *T. baccata* and *T. cuspidata* (**Figure 2.2**) So *GGPPS* gene of *T. cuspidata* was synthesized by site directed mutation (point mutation) of *T. baccata GGPPS* gene using PCR reactions. Four different DNA parts (1st part 70 bp, 2nd part 117 bp, 3rd part 166 bp, 4th part 930 bp) of the gene were amplified and fuesd through PCR reactions using Fcusp_comp, FcusPM1, RcusPM1, FcusPM2, RcusPM2, FcusPM3, RcusPM3, and Rcus_Comp primers bearing the mutations intented to be done. PCR reactions were done using PrimeSTAR® HS DNA Polymerase.

```
T.baccata     MAYTAMAAGTQSLQLRTVASYQECNSMRSCFKLTPFKSFHGVNFNVPSLGAANCEIMGHL
T.cuspıdata   MAYTAMAAGTQSMQLRTVASYQECNSMRSCFKLTPFKSFHGVNFNFPSLGAANCEIMGHL

T.baccata     KLGSLPYKQCSVSSRSTKTMAQLVDLAETEKAEGKDIEFDFNEYMKSKAVAVDAALDKAI
T.cuspıdata   KLGSLPYKQCSVSSRSTKTMAQLVDLAETEKAKGKDIEFDFNEYMKSKAVAVDAALDKAI

T.baccata     PLEYPEKIHESMRYSLLAGGKRVRPALCIAACELVGGSQDLAMPTACAMEMIHTMSLIHD
T.cuspıdata   PLEYPEKIHESMRYSLLAGGKRVRPALCIAACELVGGSQDLAMPTACAMEMIHTMSLIHD

T.baccata     DLPCMDNDDFRRGKPTNHKVFGEDTAVLAGDALLSFAFEHIAVATSKTVPSDRTLRVISE
T.cuspıdata   DLPCMDNDDFRRGKPTNHKVFGEDTAVLAGDALLSFAFEHIAVATSKTVPSDRTLRVISE

T.baccata     LGKTIGSQGLVGGQVVDITSEGDANVDLKTLEWIHIHKTAVLLECSVVSGGILGGATEDE
T.cuspıdata   LGKTIGSQGLVGGQVVDITSEGDANVDLKTLEWIHIHKTAVLLECSVVSGGILGGATEDE

T.baccata     IARIRRYARCVGLLFQVVDDILDVTKSSEELGKTAGKDLLTDKATYPKLMGLEKAKEFAA
T.cuspıdata   IARIRRYARCVGLLFQVVDDILDVTKSSEELGKTAGKDLLTDKATYPKLMGLEKAKEFAA

T.baccata     ELATRAKEELSSFDQIKAAPLLGLADYIAFRQN
T.cuspıdata   ELATRAKEELSSFDQIKAAPLLGLADYIAFRQN
```

Figure 2.2. BLAST *multiple sequence alignment of GGPPS aminoacids sequences of T. baccata, cuspidata*

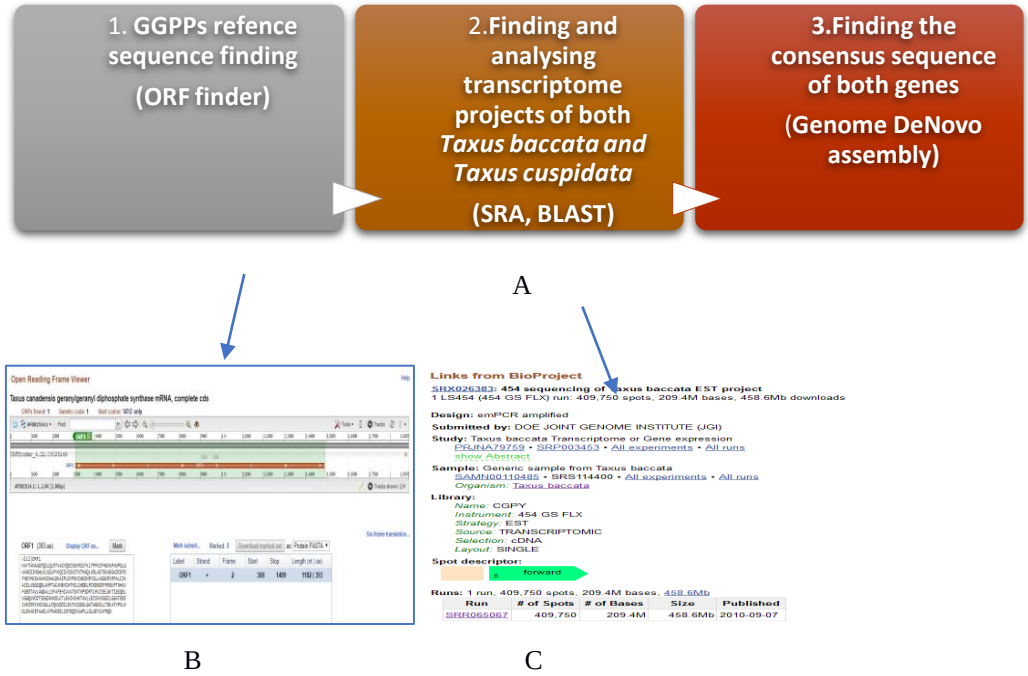


Figure 2.3. *A) Flow chart of steps included in finding GGPPS sequences of T. baccata and T. cuspidata. B) Open reading frame finder tool in NCBI. C) example of a transcriptome project of T. baccata in SRA in NCBI.*

2.2. Plasmids Construction

In this study, five episomal plasmids were designed for each strain, pSP-GM1 (7404 bp) was cloned with *GGPPS*, *TS* genes to design plasmid 1 (B); *TS-GGPPS* fusion gene to design plasmid 2 (D); *GGPPS-TS* fusion gene to design plasmid 3 (E); *TS*, *ERG20-GGPPS* genes to design plasmid 4 (F); and *TS*, *GGPPS-ERG20* genes to design Plasmid 5 (G). (**Figure 2.4**). *S. cerevisae* SCIGS22a strain (*MATa MAL2-8c SUC2 ura3-52 lpp1Δ::loxP dpp1Δ::loxP*P_{ERG9}*Δ::loxP-*P_{HXT1}*gdhΔ::loxP* P_{TEF1}*-ERG20P;*P_{PGK1}*-GDH2;*P_{TEF1}*-tHMG1)* [75] was used as background strain.

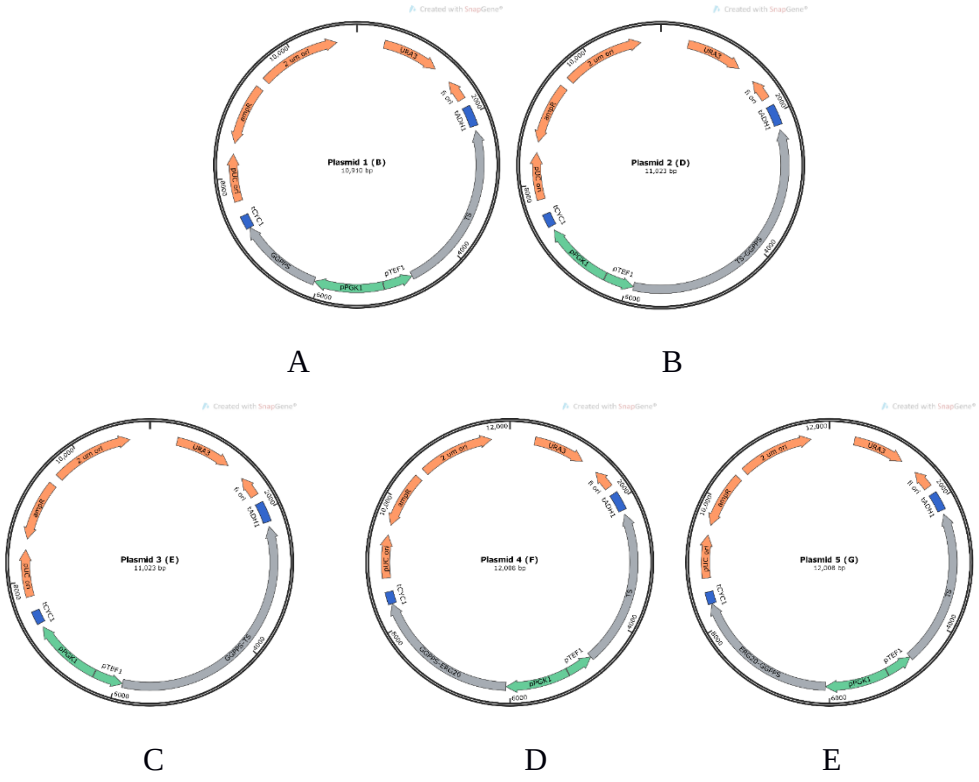


Figure 2.4. *Episomal plasmids*

A) Plasmid 1 (B): *GGPPS*, *TS*. B) Plasmid 2 (D): *TS-GGPPS C)* Plasmid 3 (E), *GGPPS-TS;* D)Plasmid 4 (F): *TS*, *ERG20-GGPPS;* F) Plasmid 5 (G): *TS*, *GGPPS-ERG20*

2.2.1. Plasmids containing *GGPPS* gene of *Taxus canadensis*

2.2.1.1. *Backbone plasmid pSP-GM1, P_{TEF1}-TS-T_{ADH1}*

Plasmid ligation was carried out by Thermo Scientific™ T4 DNA Ligase (catalog number: EL0011). Gene was transformed into *E. coli* according to heat shock transformation protocol. *E. coli* DH5α cells were made competent using calcium chloride protocol by making an overnight culture of the cells in LB broth and incubating them at 37 ℃ and 200 rpm for 12-16 hours. In the second day, 0.3 mL of the culture was added to 100 mL LB broth and incubated at 37 ℃ till OD600 0.5-0.6. Cells were centrifuged for 5 minutes at 4 ℃ and 5000 g and resuspended in 50 mL mM calcium chloride and left for 15 minutes inside ice box. Cells were centrifuged again for 10 minutes at 4 °C and 5000 G and resuspended in 4 mL calcium chloride + 80% sterile glycerol and divided into 200 µL aliquotes into precooled tubes. The tubes were placed inside ice for 5 minutes and then stored at -80 °C. To start transformation, *E. coli* DH5α chemically competent cells stored at -80 ℃ were placed into an ice box for 5 minutes to thaw. At the same time, a tube containing vector was added to the ice box for 5 minutes. 2 µL vector plasmid was added to 100 µLcomponents cells and place inside ice box for 30 minutes. The tubes were placed inside 42 ℃ water bath for 2 minutes and then in ice for 3 minutes. 500 µLof LB broth media was added to the tube and incubated in a shaking incubator for 60 minutes at 37 ℃ and 150 rpm. The cells were then centrifuged at 8000 rpm for 3 minutes and then inoculated to ampicillin LB agar for 14-16 hours [67].

2.2.1.2. *Plasmid 1 (B) pSP-GM1-TS-GGGPPS P_{TEF1}-TS-T_{ADH1}; P_{PGK1}-GGPPS-T_{CYC1}*

E.coli was transformed with pUC57-*GGPPS* plasmid. The plasmid was isolated and *GGPPS* gene was cut from plasmid and then cloned to base plasmid pSP-GM1-*TS* by using BamHI/ NheI enzymes. Then *E. coli* was transformed with the resultant plasmid.

2.2.1.3 *Plasmid 2 (D) pSP-GM1-TSGGGPPS P_{TEF} -GGPPS-TS-T_{ADH}*

TS gene was amplified from pUC57-*TS* plasmid using *TS*-SpeI-F1 and *TS*-R1 primers. *GGPPS* gene was amplified from pUC57-*GGPPS* using *GGPPS*-F1 and *GGPPS*-SacI-R1 primers. Both genes were fused and amplified using *TS*-SpeI-F1and *GGPPS*-SacI-R1

primers and then cloned to pSP-GM1 plasmid using SpeI/SacI enzymes. Then *E. coli* was transformed with the resultant plasmid.

2.2.1.4 *Plasmid 3 (E) pSP-GM1-TSGGGPPS P_{TEF} -TS-GGPPS-T_{ADH1}*

GGPPS gene was amplified form pUC57-*GGPPS* using *GGPPS*-SpeI-F2 and *GGPPS*-R2 primers. *TS* gene was amplified from pUC57-*TS* plasmid using *TS*-F2 and *TS*-SacI-R2 primers. Both genes were fused and amplified using *GGPPS*-SpeI-F2 and *TS*-SacI-R2 then cloned to pSP-GM1 plasmid using SpeI/SacI enzymes. Then *E. coli* was transformed with the resultant plasmid.

2.2.1.5 *Plasmid 4 (F) pSP-GM1-TS-ERG20GGGPPS P_{TEF} -TS- T_{ADH1}, P_{PGK1} - ERG20-GGPPS-T_{CYC1}*

ERG20 gene was amplified from pUC57-*TS* plasmid using *Erg20*-BamHI-F1 and *Erg20*-R1 primers *GGPPS* gene was amplified form pUC57-*GGPPS* using *GGPPS*-F3 and *GGPPS*-F3 primers. Both genes were fused and amplified using *Erg20*-BamHI-F1 and *GGPPS*-NheI-R3. Then it was cloned to base plasmid pSP-GM1-*TS* using BamHI/NheI enzymes. Then *E. coli* was transformed with the resultant plasmid.

2.2.1.6 *Plasmid 5 (G) pSP-GM1-TS-GGGPPSERG20 P_{TEF}-TS- T_{ADH1}, P_{PGK1} - GGPPS-ERG20-T_{CYC1}*

GGPPS gene was amplified form pUC57-GGPPS using TB-*GGPPS*-BamH1-F and TB- *GGPPS*-R2 primers. ERG20 gene was amplified from pUC57-*TS* plasmid using TB-*ERG20*-F2 and *Erg20*-NheI-R2 primers. Both genes were fused and amplified using TB-*GGPPS*-BamH1-F and *Erg20*-NheI-R2 primer and then cloned to base plasmid pSP-GM1-*TS* using BamHI/NheI enzymes. Then *E. coli* was transformed with the resultant plasmid.

2.2.2. Plasmids containing GGPPS gene of Sulfolobus acidocaldirus and Taxus baccata

For conustrcting the plasmids containing *S. acidocaldirus and T. baccata GGPPS* gene; *TS*-G*GGPPS*, *GGPPS-TS*, *GGPPS-ERG20*, and *ERG20-GGPPS* gene fusions were

prepared the same way as the plasmids containing *T. canadensis GGPPS* gene but with using pTZ57R_T-*GGPPS* and pBluescipt ll SK (+)Baccata plasmids respectively for amplification of *GGPPS* gene, then they were cloned into their corresponding plasmids previously prepared. *TS*-G*GGPPS* fusion of plasmid 2 (D) was colned using Spe1/Pacl enzymes. Plasmid 3 (E) was cancelled. Primers used are shown in (**Table 2.2**. in supporting files)

2.2.3. *Plasmids containing GGPPS gene of Taxus cuspidata*

The *GGPPS* gene of *T. cuspidata* that was synthesized as shown in section 2.2. was cloned to plasmid number 1 (B plasmid) that contains *T. baccata GGPPS* gene after treating both the gene and plasmid with Spe1/Sac1 restriction enzymes, Then *E. coli* was transformed with the resultant gene.

After all the plasmids were transformed into E. coli, they were isolated from *E. coli* using Thermo Scientific™ GeneJET Plasmid Miniprep Kit # K0502 and then they were transformed into SCIGS22a S. cerevisiae strain [75] using two either chemical transformation or elctroporation transformation. Resulting in 15 strains (Table 2.2.).

For transformation using lithium acetate/single-stranded carrier DNA/PEG protocol [68], yeast was inoculated into 5 mL YPD mL overnight culture. In the second day 50 mL fresh culture YPD media was prepared from the overnight culture with initial OD600 of 0.05 till it reached OD600 0.8. The cells were washed with 25 mL distilled water by centrifugation at 3000 g for 5 minutes, resuspended in 1 mL 100 mM lithium acetate and centrifugated at max g speed for 8 seconds. The cells were resuspended slowly in 400 µL 100 mM lithium acetate and left in room temperature for 5 minutes. Then, they were centrifuged and the cells were resuspended in 500 µL 10-15% glycerol, divided into 50 µL aliquotes and stored at -80 °C. To start the transformation protocol, yeast chemical competent cells were brought from - 80 °C and left in 37 °C water bath for 15-30 seconds to thaw. Then centrifuged at 13000 g for 2 minutes. Supernatant was removed and 260 µL PEG 3350 (50%), 36 µL LiAc 1 M, 2 µLsingle stranded carrier DNA, and 14 µL sterile water containing plasmid were added by order and cells pellet was mixed by vortexing after each addition. The cells were incubated in 42 °C water bath for 20-60-

minute, centrifuge at 13000 g for 30 seconds. Supernatant was removed and resuspended in 200 µL sterile water and inoculated to the SD-URA selective agar plates.

For electroporation transformation protocol [69], a yeast colony was inoculated into 5 mL YPD media and incubated overnight at 30 °C and 200 rpm. In the next day, overnight culture was re-cultured into 50 mL YPD broth at 30 °C and 200 with starting OD600 of 0.05-0.1 till it reached OD600 of 0.8-1.0. Cells were then centrifuged for 5 minutes at 1000 g at 4 °C. Supernatant was discarded and cells were resuspended in 16 mL sorbitol (1M), 2 mL 10X SE buffer (pH 7.5) and 2 mL LioAc (1M) and incubated at 30 °C and 150 rpm for 30 min. 200 µLDDT (1M) were added to the mixture and incubated at 30 °C and 150 rpm for 15 min. Cells were centrifuged at 4 °C, 1000 g for 5 min, then washed twice using 20 mL 1M ice cooled sorbitol. Cells were resuspended in 300 µL1 M ice cooled sorbitol and divided in 50 µL aliquotes and placed on ice and DNA (total volume <5 µl) was added to the cells and placed on ice for 15 min. Cells were then pulsed in Bio-rad micropulse Electroporation device at voltage of 1500 V. 1 mL cold sorbitol was immediately added to the cells and then incubated at 30 °C and 200 rpm for about 2 hours. Cells were centrifuged for 3 minutes at 3000 g and inoculated onto the SD-URA selective agar plates and incubated at 30 °C until the clones appear.

Table 2.1. *Yeast strains used and made in this study.*

Strain Code	Description	Source
SCIGS22a	*MATa MAL2-8c SUC2 ura3-52 lpp1Δ::loxP dpp1Δ::loxP* PP_{ERG9}*Δ::loxP-*P_{HXT1}*gdhΔ::loxP* P_{TEF1}*-ERG20P;*P_{PGK1}*-GDH2;*P_{TEF1}*-tHMG1*	Lopez et. al., 2015 [75]
SCGI22a-0	SCIGS22a + P_{tHXT7}*-tHMG1-*T_{HMG1}*;* P_{TEF1}*-TS-*T_{ADH1}*;,* P_{PGK1}*-ERG20-GGPPStc-*T_{CYC1}	Dr. Hülya Karaca Gençer
SCGI22a-1	SCIGS22a + P_{TEF1}*-TS-*T_{ADH1}*;* P_{PGK1}*-GGPPStc-*T_{CYC1}	This study
SCGI22a-2	SCIGS22a + P_{TEF1}*-GGPPStc-TS-*T_{ADH1}	This study
SCGI22a-3	SCIGS22a + P_{TEF1}*-TS-GGPPStc-*T_{ADH1}	This study
SCGI22a-4	SCIGS22a + P_{TEF1}*-TS-*$T_{ADH1,}$ P_{PGK1}*-GGPPStc-ERG20-*T_{CYC1}	This study
SCGI22a-5	SCIGS22a + P_{TEF1}*-TS-*$T_{ADH1,}$ P_{PGK1}*-ERG20-GGPPStc-*T_{CYC1}	This study
SCGI22a-6	SCIGS22a + P_{TEF1}*-TS-*T_{ADH1}*;* P_{PGK1}*-GGPPSsa-*T_{CYC1}	This study
SCGI22a-7	SCIGS22a + P_{TEF1}*-GGPPSsa-TS-*T_{ADH1}	This study
SCGI22a-8	SCIGS22a + P_{TEF1}*-TS-GGPPSsa-*T_{ADH1}	This study
SCGI22a-9	SCIGS22a + P_{TEF1}*-TS-*$T_{ADH1,}$ P_{PGK1}*-GGPPSsa-ERG20-*T_{CYC1}	This study
SCGI22a-10	SCIGS22a + P_{TEF1}*-TS-*$T_{ADH1,}$ P_{PGK1}*-ERG20-GGPPSsa-*T_{CYC1}	This study
SCGI22a-11	SCIGS22a + P_{TEF1}*-TS-*T_{ADH1}*;* P_{PGK1}*-GGPPStb-*T_{CYC1}	This study
SCGI22a-12	SCIGS22a + P_{TEF1}*-GGPPStb-TS-*T_{ADH1}	This study
SCGI22a-13	SCIGS22a + P_{TEF1}*-TS-*$T_{ADH1,}$ P_{PGK1}*-GGPPStb-ERG20-*T_{CYC1}	This study
SCGI22a-14	SCIGS22a + P_{TEF1}*-TS-*$T_{ADH1,}$ P_{PGK1}*-ERG20-GGPPStb-*T_{CYC1}	This study
SCGI22a-15	SCIGS22a + P_{TEF1}*-TS-*T_{ADH1}*;* P_{PGK1}*-GGPPStcusp-*T_{CYC1}	This study

2.3. Taxadiene Extraction And GC Analysis

After transformation, 5 colonies containig each plasmid were cultured into 10 mL DELFT media in shake flasks at 30 °C and 200 rpm. Then a seed was cultured into a flask containing 18 mL DELFT media and 2 mL dodecene with intial OD600 of 0.1 and left for 72 hours in 30 °C and 200 rpm. The taxadiene layer was collected after 72 hours and

sent for GC analysis according to Tippmann et. al. protocol [78] as 2 µLof the sample solution was analyzed by GC-MS (Shimadzu GCMS-QP2010 ultra) equipped with the capillary column DB-5MS (30 m x 0.25 mm x 0.25 µm) under the following condition: Injection port temperature, 250 °C; Carrier gas, He (99.999%); Injection mode, Splitless; Pressure, 49 kPa; Total flow, 13.4 mL/min; Oven Temperature, 50 °C to 300 °C at a speed of 10 °C/min. MS conditions are Ion Source Temperature, 200 °C; Interface Temp., 320 °C; Acquisition Mode, SIM; Selected m/z, 272, 122, 107."

3. RESULTS

After genes amplification and fusions, PCR purification was carried out. Gel electrophoresis run (agarose 1%) was done and the gel results are shown below [**Figure 3.1:3.6**]. After cloning of cassettes into pSP-GM1 plasmid and transformation into *E.coli,* plasmids were isolated from positive colonies and treated with the same restriction enzymes used for cloning as plasmid 1 (B) was treated with BamHI/NheI to check for *GGPPS* gene, plasmid 2 (D) was treated with SpeI/SacI to check for *TS-GGPPS* fusion, plasmid 3 (E) was treated with SpeI/SacI to check for *GGPPS-TS* fusion, plasmid 4 (F) was treated with BamHI/NheI to check for *ERG20-GGPPS* fusion, and plasmid 5 (G) was treated with BamHI/NheI to check for *GGPPS-ERG20* fusion. Reaction mixture was run into gel electrophoresis (agarose 1%) and the gel results are shown below.

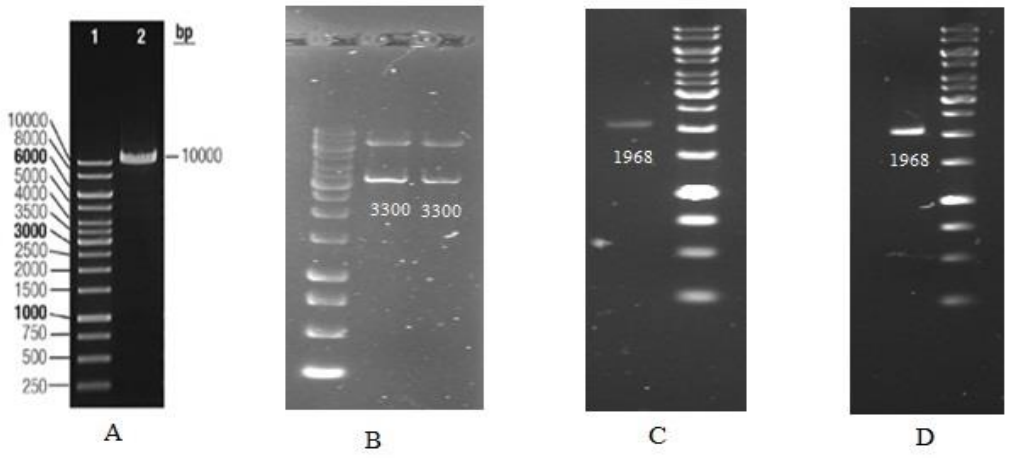


Figure 3.1. *T. canadensis GGPPS containing cassettes.*
A: GeneRuler TM 1kb DNA Ladder, B: Left, D plasmid (TS-GGPPS fusion "3300 bp"), Right: E plasmid (GGPPS-TS fusion "3300 bp"), C: F plasmid (ERG20-GGPPS fusion "1968 bp"), D: G plasmid (GGPPS-ERG20 fusion "1968 bp").

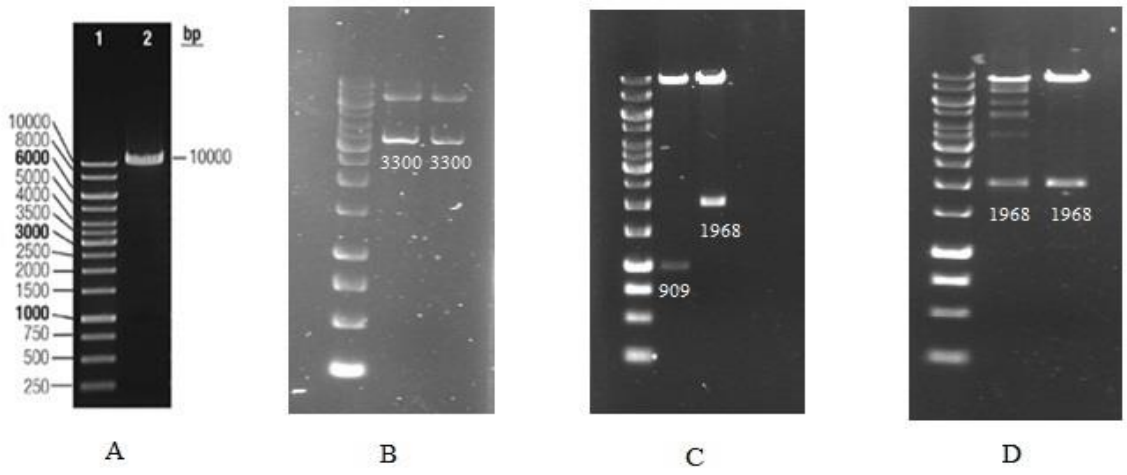


Figure 3.2. *T. canadensis GGPPS containing plasmids.*

A: GeneRuler TM 1kb DNA Ladder, B: Left, D plasmid after treatment with SpeI/SacI (TS-GPPS insert "3300 bp"), Right: E plasmid after treatment with SpeI/SacI (GGPPS-TS fusion "3300 bp"), C: Left: B plasmid after treatment with BamHI/NheI (GGPPS insert "909 bp"), Right: F plasmid after treatment with BamHI/NheI (ERG20-GGPPS insert "1968 bp"), D: G plasmid after treatment with BamHI/NheI (GGPPS-ERG20 insert "1968 bp").

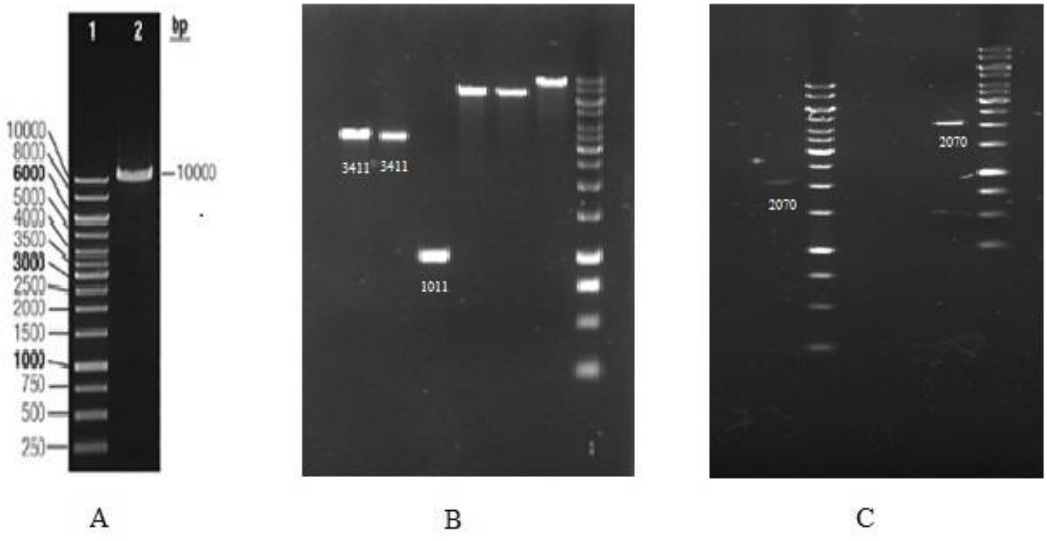


Figure 3.3. *S. acidocaldirus GGPPS* containing cassettes.

A: GeneRuler TM 1kb DNA Ladder, B: Left first, D plasmid (TS-GGPPS fusion "2411 bp"), Left second: E plasmid (GGPPS-TS fusion "3411 bp"). Left third: B plasmid (GGPPS gene "1011 bp"), C: Left: F plasmid (ERG20-GGPPS fusion "2070 bp"). Right: G plasmid (GGPPS-ERG20 fusion "2070 bp").

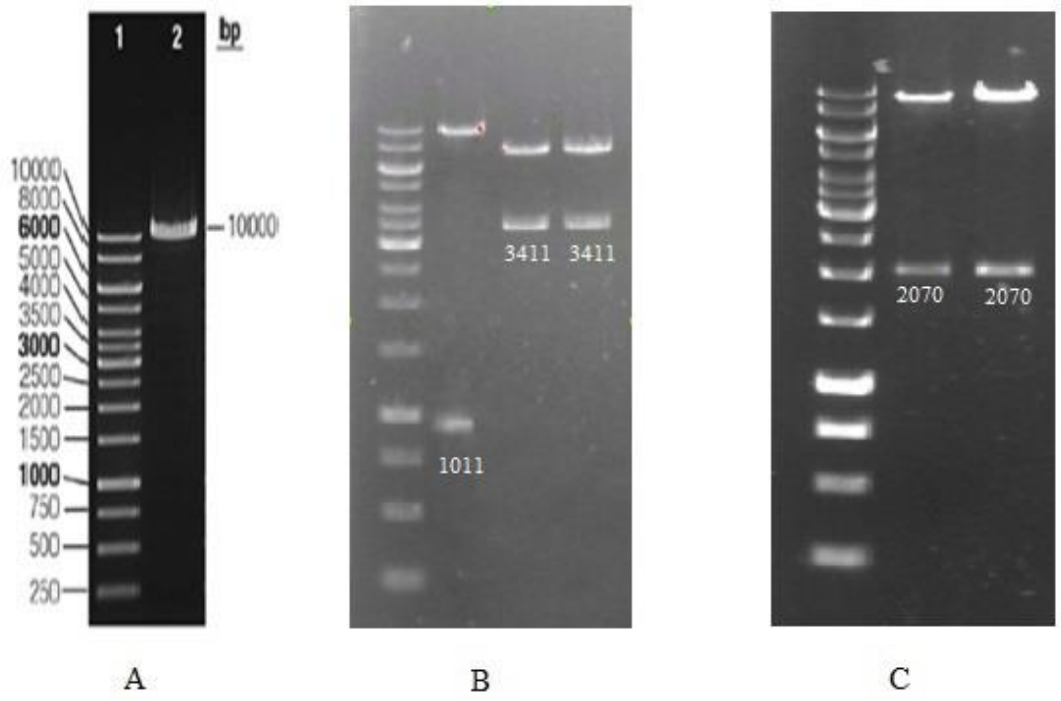


Figure 3.4. *S. acidocaldirus GGPPS containing plasmids*

A: GeneRuler™ 1kb DNA Ladder, B: Left: B plasmid after treatment with BamHI/NheI (GGPPS insert "1011 bp") Middle: D plasmid after treatment with SpeI/SacI (TS-GPPS insert "3411 bp"), Right: E plasmid after treatment with SpeI/SacI (GGPPS-TS fusion "2411 bp"), C: Left: F plasmid after treatment with BamHI/NheI (ERG20-GGPPS insert "2070 bp"), Right: G plasmid after treatment with BamHI/NheI (GGPPS-ERG20 insert "2070 bp").

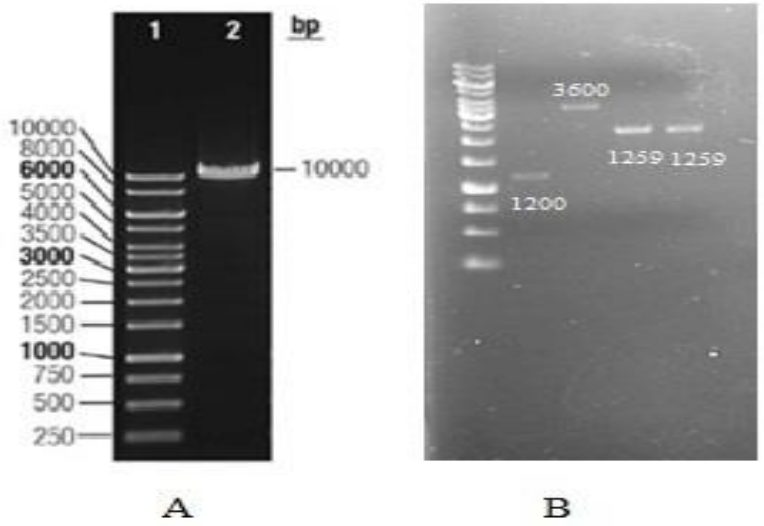


Figure 3.5. *T. baccata GGPPS containing cassettes.*

A: GeneRuler™ 1kb DNA Ladder, B: Order from left, 1) B plasmid (GGPPS gene "1200 bp"), 2) D plasmid (TS-GGPPS fusion "3600 bp"). 3) F plasmid (ERG20-GGPPS fusion "2259 bp") 4) G plasmid (GGPPS-ERG20 fusion "2259 bp").

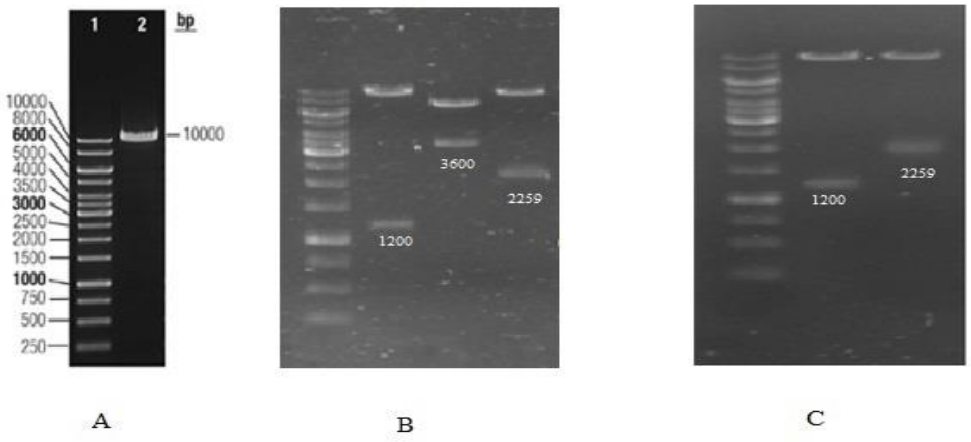


Figure 3.6. *T. baccata GGPPS containing plasmids.*
A: GeneRuler TM 1kb DNA Ladder, B: Left: B plasmid after treatment with BamHI/NheI (GGPPS insert "1200 bp"), Middle: D plasmid after treatment with SpeI/PacI (TS-GPPSS insert "3600 bp"), Right: F plasmid after treatment with BamHI/NheI (ERG20-GGPPS insert "2259 bp"), C: B plasmid after treatment with BamHI/NheI (GGPPS insert "1200 bp"), Right: G plasmid after treatment with BamHI/NheI (GGPPS-ERG20 insert "2259 bp").

For plasmid containing *T. cuspidata GGPPS* gene explained in section 2.2., four parts of the gene were amplified from pBluescipt ll SK (+)Baccata plasmid that contains the *GGPPS* gene of *T. cuspidata* and fused. Gel electrophoresis results are shown in **Figure 3.7** After cloning to plasmid 1 (B) of that contains *GGPPS* of T. *baccata*, the plsmid was isolated from positive colonies and treated with BamHI/NheI enzymes to check for the presence of *GGPPS* gene and it was run into gel electrophoresis and the results are shown in **Figure 3.7**.

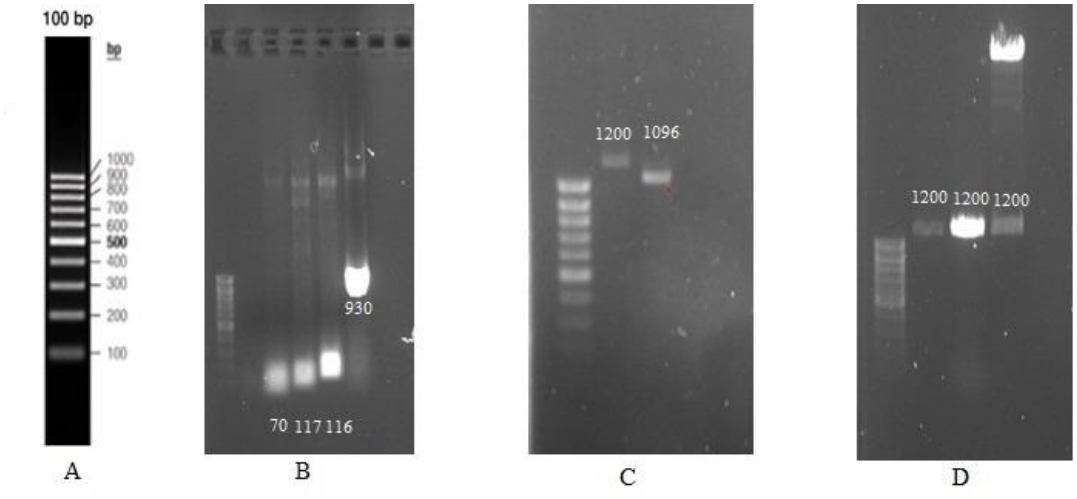


Figure 3.7. *T. cuspidata GGPPS gene.*

A: GeneRuler 100 bp DNA Ladder, B: 1st part (70 bp), 2nd part (117 bp), 3rd part (166 bp), 4th part (930 bp), C: T. cuspidata GGPPS gene, fusion of third and fouth parts (1096 bp), D: PCR purified T. cuspidata GGPPS, plasmid B treated with BamH1/Nhe1.

SCGI22a *S. cerevesiae* cells were transformed by plasmids and the postitive colonies were checked for the following inserts *GGPPS* gene for plasmid 1 (B), *TS-GGPPS* for plasmid 2 (D), *GGPPS-TS* for plasmid 3 (E), *ERG20-GGPPS* for plasmid 4 (F), and *GGPPS-ERG20* for plasmid 5 (G). Colony PCR products were run on agarose gel and electrophoresis run results are shown in **Figure 3.8.**

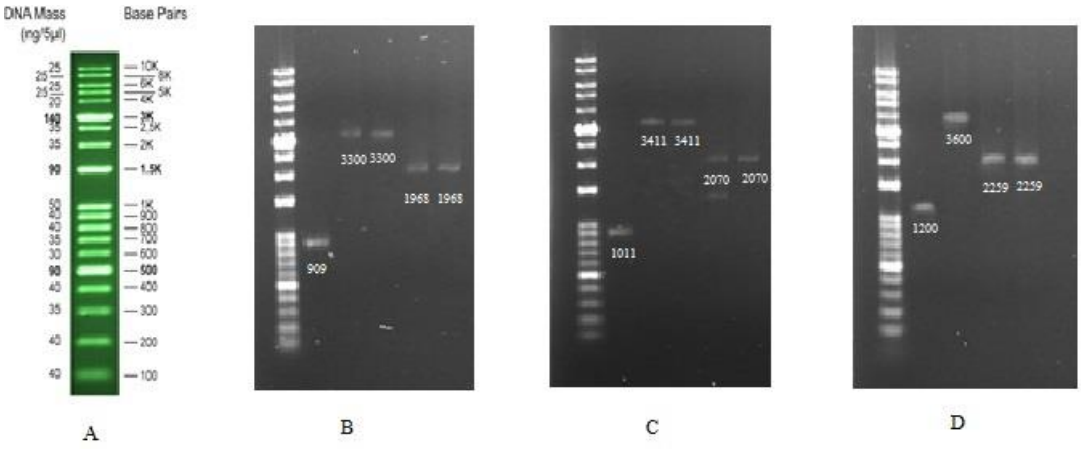


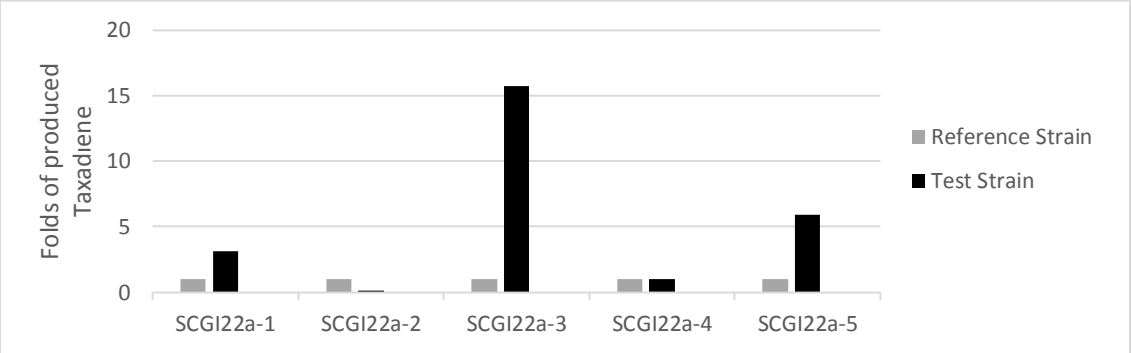
Figure 3.8. *Colony PCR results.*

A) Safe-Green™ 1kb Opti-DNA Marker, B) Plasmids of *T.canadensis* (plasmids 1, 2, 3, 4, 5), C) Plasmids for *Sulfolobus acidocaldirus* (plasmids 1, 2, 3, 4, 5), D) Plasmids of *T.baccata* (plasmids 1, 2, 3, 5)

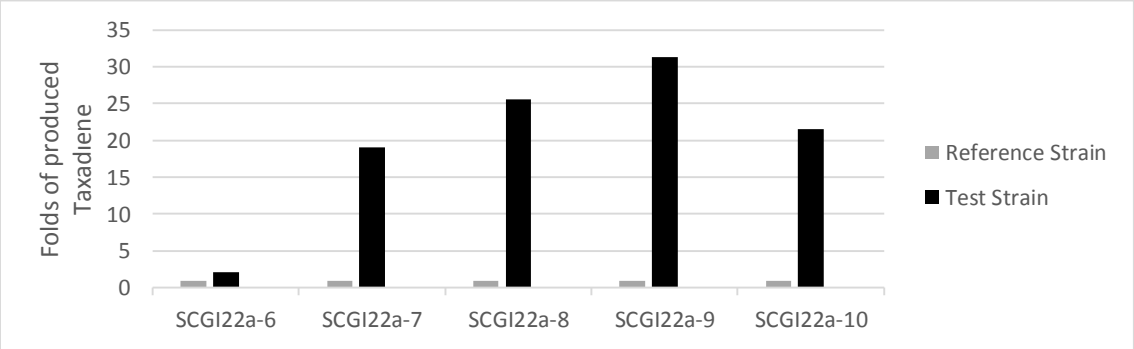
For analysis of taxadiene produced by different strain under investigation in this study, a culture of five colonies of each strain was carried out and the produced taxadiene was collected by dodecene and analyzed by GC-MS. As taxadiene standard was not available, the reference was another strain (SCGI22a-0) developed by Dr. Hülya Karaca Gençer. Taxadiene yield of SCGI22a-0 strain was 1.83 mg/L The results were then calculated in comparison to this strain and are shown in (**Figure:3.9, Table 3.1**) SGGI22a was used as a negative control as it lacks *TS* gene and so it lacks the ability to produce taxadiene.

Table 3.1. *Results of GC analysis*

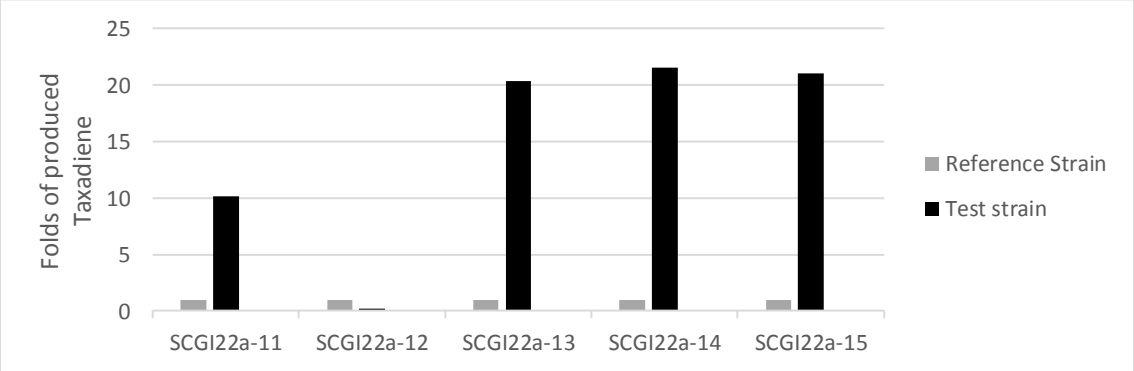
Strains	Area from GC analysis	Folds of refeence	Taxadiene produced (mg/L)
SCGI22a-0	621311.2	1	1.83
SCGI22a-1	1963936.3	3.161	5.784
SCGI22a-2	69318.2	0.112	0.205
SCGI22a-3	9779879	15.741	28.806
SCGI22a-4	638979	1.028	1.881
SCGI22a-5	3671424	5.909	10.813
SCGI22a-6	1279016	2.059	3.768
SCGI22a-7	11860371	19.089	34.932
SCGI22a-8	15849531	25.51	46.683
SCGI22a-9	19454547	31.312	57.3
SCGI22a-10	13400632	21.568	39.469
SCGI22a-11	6317009	10.167	18.506
SCGI22a-12	49395.2	0.08	0.146
SCGI22a-13	12661469	20.379	37.294
SCGI22a-14	184053.7	21.568	39.469
SCGI22a-15	7149168	21.058	11.507



A



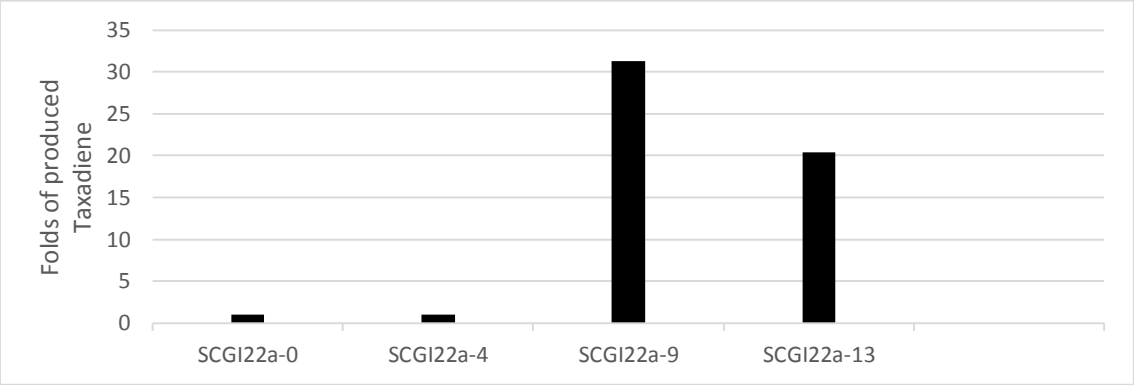
B



C

Figure 3.9. *Taxadiene production results*.

A) Taxadiene produced by strains containig *T. canadensis GGPPS* gene. B) Taxadiene produced by strains containig *S. acidocaldirus GGPPS gene. C) Taxadiene produced by strains containig T. baccata and T. cuspidata GGPPS genes. D) Taxadiene produced by stains containing plasmid F.*



D

Figure 3.9. *Taxadiene production results. (Continued)*

A) Taxadiene produced by strains containig *T. canadensis GGPPS* gene. B) Taxadiene produced by strains containi*g S. acidocaldirus GGPPS gene. C) Taxadiene produced by strains containig T. baccata and T. cuspidata GGPPS genes. D) Taxadiene produced by stains containing plasmid F.*

4. DISCUSSION

So far taxadiene yield achieved 72.8 mg/L in engineered S. *cerevisiae* [25]. For increasing the taxadiene production in *S. cerevisiae*, its precursor geranylgeranyl diphosphate supply has to be increased. Thus, it is crucial to increase the GGPP pool through expression of heterologous *GGPPS* in *S. cerevisiae.* In this study, the effect of four different *GGPPS* genes from *T. canadensis, S. acidocaldirus, T. baccata,* and *T. cuspidata* on the taxadiene yield was investigated.

According to the taxadiene levels produced, strains containing GGPPS from *S. acidocaldirus* (SCGI22a-5:10) produced more taxadiene in comparison to their counterpart strains. SCGI22a-9 strain produced up to 57.3 mg/L taxadiene**. Figure 3.9. section D** shows a comparison made between taxadiene amounts produced by SCGI22a-0. SCGI22a-4, SCGI22a-9, and SCGI22a-13 strains. These strains contain the same overespressed genes of TS, and *ERG20-GGPPS*. SCGI22a-9 showed increased taxadiene as GGPPS source was from *S. acidocaldarius*. This suggests that *S. acidocaldarius GGPPSsa* gene is more efficient in production of geranylgeranyl diphosphate and consequently other terpenoid products. This finding can be augmented by other studies

that included S. acidocaldarius *GGPPSsa* gene. A combinatorial study has been done for geranylgeraniol (GGOH) overproduction in *S. cerevisiae* [47]. Three different constructs of endogenous gene *tHMGR*, fusion of endogenous *GGPPS-ERG20*, and a heterologous gene *GGPPS* from *S. acidocaldarius* were integrated separately into *S. cerevisiae*. It was found that geranylgeraniol was produced more from strains with integrated *GGPPS* from *S. acidocaldarius* than strains with endogenous *GGPPS* from *S. cerevisiae*, geranylgeraniol yield reached 437.52 mg/L. As (GOH) is formed by the dephosphorylation of geranylgeranyl pyrophosphate, it can be used as a direct reporter for GGPP overproduction. Also, in another study, taxadiene production increased upon expression of *S. acidocaldarius* geranylgeranyl diphosphate synthase in *S. cerevisiae* [70]. Lycopene (another terpenoid product derived from GGPP) production was increased by 770 folds when *GGPPS* gene from *S. acidocaldarius*, (tHMG1), and a fusion gene of ERG20 and endogenous *GGPPS (BTS1)* were overexpressed in *S. cerevisiae* strain BY4742 [50].

Strains containing *ERG20, TS, and GGPPS* genes produced more taxadiene than those contained only *TS* and GGPPS genes, this indicates the importance of *ERG20* overexpression in the production of taxadiene [25]. *ERG20* plays an important role in increasing the *GGPPS* pool inside the cells. Acyclic monoterpene alcohol geraniol was synthesized in *S. cerevisiae* by expression of endogenous *ERG20* variants *ERG20*(G), *ERG20* (K197G) and *ERG20* (WW), *ERG20* (F96W-N127W) and overexpression of *IDI1*, *tHMG1*, and *UPC2-1* genes. Overexpressing *ERG20* (WW) improved geraniol production. Geraniol level reached 293 mg/L [73]. *ERG20* has been identified to be a limiting factor in the production of monoterpenes. After engineering modification of *ERG20* and fusing it to *GGPPS* followed by integration into *S. cereveisae*, the yield of sabinene monoterpene has increased significantly by 340 folds. [74]

Taxadiene levels produced by SCGI22a-2, SCGI22a-7 strains that contained *TS-GGPPS* construct was relatively higher than the taxadiene yield of SCGI22a-3, SCGI22a-8 strains that contained *GGPPS-TS* constructs. This indicated that changing ther order of both genes affected their expression and consequently the level of taxadiene produced. When the order of both gene changed from *GGPPS-TS* to *TS-GGPPS* during isoprenoid

pathway optimization in *E. col*i, it resulted in 2-3 folds ıncrease ın taxadiene production [42].

SCGI22a-5 strain contains one copy of *HMG1* enzyme integrated into yeast's chromosome and it produced more taxadiene than SCGI22a-0 strain contains an additional copy expressed in the episomal plasmid. This suggests that the presence of several copies of HMG1 enzymes led to metabolic stress. HMG1 enzyme is a rate limiting enzymes in mevalonate pathway and its overexpression leads to increasing of sterol production. Cubeol's, a sesquiterpene compound, yield increased upon overexpression of HMG1 enzyme's active domain [76]. For production of geranylgeraniol (GGOH), HMG1 gene, and *BTS1-ERG20* fusion gene were expressed in *S. cerevisae*. It was found that overexpression of HMG1 only led to more squalene accumlation (191.9 mg/L) rather than GGOH (0.2 mg/L). Upon coexpression of endogenous *BTS1-ERG20*, the metabolic flux was shifted completely towards GGOH (228.8 mg/L) rather than squalene (6.5 mg/L) [77].

In our study, *TEF1* and *PGK1* constitutive promoters were used in the design of the episomal plasmids used in this study. Both promoters exhibited the most constant activity independent of the level of glucose used as a carbon source when the activities of *pTEF1, pTPI1, pTDH3, pADH1, pPGK1, pHXT7, pPYK1, pGAL1,* and *pGAL10* promoters were compared using integrative plasmid pSF011 in *S. cerevesiae* [72].

5. CONCLUSION

Taxol is an important anticancer agent that is produced by *Taxus brevifolia*. Its production in plant is prone to fluctuations due to weather conditions and plant slow growth. Many attempts have been done to produce taxol in microorganisms such as *E. coli* and *S. cerevesiae* through metabolic engineering tool to provide a constant source of taxol. Its production in *E. coli* reached 1g/L. However, yeast is better than *E. coli* as it can withstand low pH and high osmotic pressure in industrial production. The highest taxadiene yield in *S. cerevisiae* reached 72.8 mg/L [25]. In this study, modification of the mevalonate pathway in *S. cerevesiae* has been attempted to increase taxadiene yield. Five episomal plasmids have been designed to contain farensyl diphosphate synthase (*ERG20*), taxadeine synthase (*TS*) and geranyl geranyl diphosphate synthase (*GGPPS*) genes in different combinatorial arranges to select for the best set and arrangement of genes. Moreover, efficiency of *GGPPS* genes from four different strains; *Taxus canadensis*, s*ulfolobus acidophilus*, *Taxus baccata*, and *Taxus cuspidata* was investigated to select the best *GGPPS* gene for production of taxadiene in *S. cerevesiae*. Taxadiene yield in shaking flasks increased by 31 folds yield upon using a fusion gene of *Sulfolobus acidophilus GGPPS* and *ERG20*, and *TS* gene from *Taxus brevifolia*.

References:

1. Daniel G. Gibson et al. (2010). "Creation of a bacterial cell controlled by a chemically synthesized genome". Science. 329 (5987): 52–6.
2. Cohen SN, Chang ACY, Boyer HW, Helling RB. (1973). Construction of biologically functional bacterial plasmids in vitro. Proc. Natl. Acad. Sci. USA 70(11):3240–44
3. Kampranis SC and Makris AM (2012) Developing a yeast cell factory for the production of terpenoids. Comput Struct Biotechnol J. 3: e201210006.
4. Stephanopoulos GN, Aristidou AA, Nielsen J. (1998.) Metabolic Engineering: Principles and Methodologies. San Diego, CA: Acad. Press
5. Bailey, J. E. (1991). Toward a science of metabolic engineering. Science 252, 1668–1675.
6. Woolston, B. M., Edgar, S. & Stephanopoulos, G. (2013). Metabolic engineering: past and future. Annu. Rev. Chem. Biomol. Eng. 4, 259–288.
7. Nakamura, C. E. & Whited, G. M. Metabolic engineering for the microbial production of 1,3-propanediol. Curr. Opin. Biotechnol. 14, 454–459 (2003).
8. Keasling D.,Jay (2010). *Advanced Biofuel production in microbes.* Biotechnol.J.,5,147-162
9. Kulkarni R, (2016). Metabolic Engineering: Biological Art of Producing Useful Chemicals. Resonance, 21 (3), 233-237.
10. Genetics Home Reference, Cells and DNA (2019) Lister Hill National Center for Biomedical Communications U.S. National Library of Medicine, National Institutes of Health.
11. T. A. Brown (2010). “Chapter 2: Vectors for Gene Cloning: Plasmids and Bacteriophages.” Gene Cloning and DNA Analysis: An Introduction (6th ed.). Wiley-Blackwell.
12. Russell et al. (2001). Molecular cloning: a laboratory manual. Cold Spring Harbor, N.Y: Cold Spring Harbor Laboratory.
13. Norman Stone et al. (1996). "Construction of a 750-kb bacterial clone contig and restriction map in the region of human chromosome 21 containing the progressive myoclonus epilepsy gene". Genome Research. 6 (3): 218–225.

14. Terry Brown. (2011) Introduction to Genetics: A Molecular Approach. Garland Science. p. 238.
15. Rothstein, R. J. (1983) One-step gene disruption in yeast. Methods Enzymol. 101: 202–211.
16. Orr-Weaver, T. L., J. W. Szostak, and R. J. Rothstein. (1981) Yeast transformation: a model system for the study of recombination. Proc. Natl. Acad. Sci. USA 78: 6354–6358.
17. Siewers V, Mortensen U, Nielsen J. 2010. Genetic Engineering Tools for Saccharomyces cerevisiae, p 287-301. In Baltz R, Demain A, Davies J, Bull A, Junker B, Katz L, Lynd L, Masurekar P, Reeves C, Zhao H (ed), Manual of Industrial Microbiology and Biotechnology, Third Edition. ASM Press, Washington, DC.
18. Anthony J.F. Griffiths. (2000) "An Introduction to Genetic Analysis" W.H. Freeman.
19. Winnacker E.L. (1987). "Chapter 2: Isolation, Identification, and Characterisation of DNA fragments". From Genes to Clones. VCH.
20. Kathleen V. C., Peter V. and Guy H. (2000) "Episomal vectors for gene expression in mammalian cells" Europen journal of Biochemistry 267, 5665-5678
21. Kobayashi I. (2001). "Behavior of Restriction-modification systems as selfish mobile elements and their impact on genome evolution." Nucleic Acids Research. 29 (18): 3742–56.
22. Clark D.P. (2005). Molecular biology. Amsterdam: Elsevier Academic Press. ISBN 0-12-175551-7.
23. Pingoud A. (2004). Restriction Endonucleases (Nucleic Acids and Molecular Biology) Springer. p.3.
24. College of arts and science, Miami universioty website, http://www.bio.miami.edu/dana/dox/restrictionenzymes.html Retrieved: 17.04.2019
25. Ding M-z, Yan H-f, Li L-f, Zhai F, Shang L-q, et al. (2014) Biosynthesis of Taxadiene in Saccharomyces cerevisiae : Selection of Geranylgeranyl

Diphosphate Synthase Directed by a Computer-Aided Docking Strategy. PLoS ONE 9(10)

26. Wang C, Liwei M, Park J-B, Jeong S-H, Wei G, Wang Y and Kim S-W (2018) Microbial Platform for Terpenoid Production: Escherichia coli and Yeast. Front. Microbiol.
27. Liao, P., Hemmerlin, A., Bach, T. J., and Chye, M. L. (2016). The potential of the mevalonate pathway for enhanced isoprenoid production. Biotechnol. Adv. 34,697–713.
28. Shang-Tian,Yang Xiaoguang, LiuYaliZhang. (2007) Bioprocessing for Value-Added Products from Renewable Resources. New Technologies and Applications 2007, Pages 73-118.
29. Wani MC, Taylor HL, Wall ME, Coggon P, McPhail AT (1971) Plant Antitumor Agents VI. The Isolation and Structure of Taxol, a Novel Antileukemic and Antitumor Agent from Taxus brevifolia. Journal of the American Chemical Society, 93, 2325-2327.
30. Kampranis SC and Makris AM (2012) Developing a yeast cell factory for the production of terpenoids. Comput Struct Biotechnol J. 3: e201210006.
31. Rowinsky EK, Cazenave LA, Donehower RC. Taxol: a novel investigational antimicrotubule agent. Natl Cancer Inst. J. 1990 Aug 1;82(15):1247–1259
32. Malik S, Cusido RM, Mirjalili MH, Moyano E, Palazon J, Bonfill M. Production of the anticancer drug taxol in *Taxus baccata* suspension cultures: a review. Process. Biochem 2011;46:23–34
33. Croteau R, Ketchum RE, Long RM, Kaspera R, Wildung MR. (2006) Taxol biosynthesis and molecular genetics. Phytochem .5(1):75–97.
34. Christopher McElroy and Stefan Jennewein, (2018) "Biosynthesis and Production: From Forests to Fermenters" Biotechnology of Natural Products pp 145-185|
35. Whiterup KM, Look SA, Stasko MW, Ghiorzi TJ, Muschik GM, Cragg GM (1990) Taxus spp. needles contain amounts of taxol comparable to the bark of Taxus brevifolia: analysis and isolation. Journal of Natural Products . 53:1249
36. Vidensek N, Lim P, Campbell A, Carlson C (1990) Taxol content in bark, wood,

root, leaf, twig, and seedling from several Taxus species. Journal of Natural Products. 53:1609

37. Smetanska I. (2008) Production of secondary metabolites using plant cell cultures. Food Biotechnology. 111:187–228.
38. Havrilesky, L. (| June 27, 2012). Paclitaxel Shortage May Add Millions to Cost of OvarianCancer Care. Retrieved on 13.05.2019 from https://www.oncnursingnews.com/publications/oncology-nurse/2011/august-2011/paclitaxel-shortage-may-add-millions-to-cost-of-ovarian-cancer-care.
39. Chris J. Paddon and Jay D. Keasling. (2014) "Semi-synthetic artemisinin: a model for the use of synthetic biology in pharmaceutical development" Nature Reviews Microbiology volume 12, pages 355–367
40. Denis JN, Greene AE, Guenard D, Gueritte-Voegelein F, Mangatal L, Potier P. (1988) Highly efficient, practical approach to natural taxol. Journal of American Chem istry Soceity. 10(17):5917–9.
41. Martin VJJ, Pitera DJ, Withers ST, Newman JD, Keasling JD (2003) Engineering a mevalonate pathway in Escherichia coli for production of terpenoids. Nature Biotechnology 21: 796–802.
42. Ajikumar PK, Xiao WH, Tyo KEJ, Wang Y, Simeon F, et al. (2010) Isoprenoid pathway optimization for taxol precursor overproduction in Escherichia coli. Science 330: 70–74.
43. Kirby J, Keasling JD (2009) Biosynthesis of plant isoprenoids: perspectives for microbial engineering. Annu Rev Plant Biol 60: 335–355.
44. Jiang Y, Proteau P, Poulter D and Ferro-Novick S (1995) BTS1 encodes a geranylgeranyl diphosphate synthase in Saccharomyces cerevisiae. Journal of Biological Chemitry 270: 21793–21799.
45. Soliman SSM, et al. (2017) Exogenous and endogenous increase in fungal GGPP increased fungal Taxol production. Applied Microbial Biotechnology. Volume 101, Issue 20, pp 7523-7533
46. Ding M-z, Yan H-f, Li L-f, Zhai F, Shang L-q, Yin Z, et al. (2014) Biosynthesis of Taxadiene in Saccharomyces cerevisiae: Selection of Geranylgeranyl

Diphosphate Synthase Directed by a Computer-Aided Docking Strategy. PLoS ONE 9(10): e109348.

47. Tian-Qing Song et al. (2017) Engineering Saccharomyces cerevisiae for geranylgeraniol overproduction by combinatorial design. Scientific Reports volume 7, Article number: 14991

48. Gershenzon, J. & Dudareva, N. (2007) The function of terpene natural products in the natural world. Jouenal of Nature Chemical Biology 3, 408–414,

49. You, L. F., Guo, L. Q., Lin, J. F., Ren, T. & Wang, J. R. (2014) Overproduction of geranylgeraniol in Coprinopsis cinerea by the expression of geranylgeranyl diphosphate synthase gene. Journal of Basic Microbiology 54, 1387–1394,

50. Shi MY, et al. (2014) Construction of Saccharomyces cerevisiae cell factories for lycopene production. Zhongguo Zhong Yao Za Zhi 39(20):3978-85
51. Dai Z, et al. (2012) Production of miltiradiene by metabolically engineered Saccharomyces cerevisiae. Biotechnol Bioeng 109(11):2845-53).
52. National Human Genome Research Institute. "A Brief Guide to Genomics". https://www.genome.gov/19016904/ Retrieved 20.04.2019.
53. Daviss, Bennett (April 2005). “Growing pains for metabolomics”. The Scientist. 19 (8): 25–28.
54. Anderson NL, Anderson NG; Anderson (1998). "Proteome and proteomics: new technologies, new concepts, and new words". Electrophoresis. 19 (11): 1853–61.
55. Attwood TK, Gisel A, Eriksson NE, Bongcam-Rudloff E (2011). "Concepts, Historical Milestones and the Central Place of Bioinformatics in Modern Biology: A European Perspective". Bioinformatics – Trends and Methodologies. November 2011
56. Tolga Can, “Introduction to Bioinformatics” Methods Molelcular Biology 2014;1107:51-71.
57. Smith TF, Waterman MS (1981) Identification of common molecularubsequences. Journal of Molecular Biology 147:195–197
58. Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48(3):443–453

59. Thompson JD, Higgins DG, Gibson TJ. (1994) “CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.” Nucleic Acids Res. 1994 Nov 11;22(22):4673-80.
60. Pearson, WR; Lipman, DJ (1988). “Improved tools for biological sequence comparison”. Proceedings of the National Academy of Sciences of the United States of America. 85 (8): 2444–8.
61. C. Molina, E. Grotewold Genome wide analysis of *Arabidopsis* core promoters BMC Genomics, 6 (2005), p. 25
62. S.K. Burley, R.G. Roeder. Biochemistry and structural biology of transcription factor IID (TFIID) Annual Review of Biochemistry, 65 (1996), pp. 769-799
63. T.I. Lee, R.A. Young Transcription of eukaryotic protein-coding genes. Annual Review of Genetics, 34 (2000), pp. 77-137
64. Watson, J. (2008). Molecular Biology of the Gene. Cold Spring Harbor Laboratory Press. pp. 410–411.
65. Richardson, J. P. (1996). "Rho-dependent Termination of Transcription Is Governed Primarily by the Upstream Rho Utilization (rut) Sequences of a Terminator". Journal of Biological Chemistry. 271 (35): 21597–21603.
66. Von Hippel, P. H. (1998). "An Integrated Model of the Transcription Complex in Elongation, Termination, and Editing". Science. 281 (5377): 660–665.
67. Chang, Angela Y., Chau, Vivian WY., Landas, Julius A., Pang, Yvonneecoli (2017) preparation of calcium competent Escherichia coli and heatshock transformation. Undergraduate Methods Paper vol 1:22-25
68. R Daniel Gietz 1 & Robert H Schiestl. (2007) Frozen competent yeast cells that can be transformed with high efficiency using the LiAc/SS carrier DNA/PEG method. Nature Protocols Vol 2(1).
69. Suga M, Hatakeyama T. High-efficiency electroporation by freezing intact yeast cells with addition of calcium. Curr Genet. 2003;43:206–211.
70. Benedikt Engels, Pia Dahm, Stefan Jennewein. (2008) Metabolic engineering of taxadiene biosynthesis in yeast as a first step towards Taxol (Paclitaxel) production. Metabolic Engineering 10 (2008) 201– 206

71. Koepp, A.E., Hezari, M., Zajicek, J., Vogel, B.S., LaFever, R.E., Lewis, N.G., Croteau, R., 1995. Cyclization of geranylgeranyl diphosphate to taxa-4(5),11(12)-diene is the committed step of Taxol biosynthesis in Pacific yew. J. Biol. Chem. 270, 8686.
72. Siavash Partow, Verena Siewers, Sara Bjørn, Jens Nielsen and Jérôme Maury. (2010). Characterization of different promoters for designing. Yeast 2010; 27: 955–96
73. Reference: Zhao J, et al. (2016) Improving monoterpene geraniol production through geranyl diphosphate synthesis regulation in Saccharomyces cerevisiae. Appl Microbiol Biotechnol 100(10):4561-71
74. Ignea C, et al. (2015) Efficient diterpene production in yeast by engineering Erg20p into a geranylgeranyl diphosphate synthase. Metab Eng 27:65-75
75. Lopez J, Essus K, Kim IK, Pereira R, Herzog J, Siewers V, et al. Production of ß-ionone by combined expression of carotenogenic and plant CCD1 genes in *Saccharomyces cerevisiae*, Microb Cell Fact. 2015;14: 84.
76. Asadollahi MA, et al. (2010) Enhancement of farnesyl diphosphate pool as direct precursor of sesquiterpenes through metabolic engineering of the mevalonate pathway in Saccharomyces cerevisiae. *Biotechnol Bioeng* 106(1):86-96
77. Tokuhiro K. et al. (2009) Overproduction of geranylgeraniol by metabolically engineered Saccharomyces cerevisiae. Applied environ Microbiology 75(17):5536-43
78. Tippmann, S., Scalcinati, G., Siewers, V., & Nielsen, J. (2016) Production of farneseneand santalene by Saccharomyces cerevisiae using fed-batch cultivations with RQcontrolled feed. Biotechnology and Bioengineering, 113(1), 72–81.

Supporting files

Table 2.2. *Primers used in the amplification of genes and gene fusions.*

Primer Code	Primer Seqeuence
Plasmids containing GGPPS **of Taxus canadensis:**	
For D Plasmid	
TS-SpeI-F1:	GTTGTTACTAGTAAAACAATGTCATCATCCAC
TS-R1:	TACTTGGATTGGGTCGATGTAA
GGPPS-F1:	AGAAAGGTTTACATCGACCCAATCCAAGTAGGTTCTGGTATGTTC GATTTCAA- CGAATACATG
GGPPS-SacI-R1:	GTTGTTGAGCTC TTAATTTTGTCTGAATGCGATGTA
For E Plasmid	
GGPPS-SpeI-F2:	GTTGTTACTAGTAAAACAATGTTCGATTTCAACGAATACATG
GGPPS-R2:	ATTTTGTCTGAATGCGATGTA
TS-F2:	TTAGCAGACTACATCGCATTCAGACAAAAT GGT TCT GGT ATGTCATCATCCACAGGTACTTC
TS-SacI-R2:	GTTGTTGAGCTC TTATACTTGGATTGGG
For F plasmid	
Erg20-BamHI-F1:	GTTGTT GGATCC AAAACA ATGGCTTCAGAAAAAGAAATTAGG
Erg20-R1:	TTTGCTTCTCTTGTAAACTTTG
GGPPS-F3:	TTCTTGAACAAAGTTTACAAGAGAAGCAAAGGTTCT- GGTATGTTCGATTTCAACGAATACATG
GGPPS-NheI-R3:	GTTGTT GCTAGC TTAATTTTGTCTGAATGCGATGTA
For G plasmid	
GGPPS-BamHI-F:	GTTGTT GGATCC AAAACA ATGTTCGATTTCAACGAATACATG
Erg20-F2:	TTAGCAGACTACATCGCATTCAGACAAAAT GGTTCTGGTATGGCTTCAGAAAAAGAAATTAGG
Erg20-NheI-R2:	GTTGTT GCTAGC CTATTTGCTTCTCTTGTAAACTTTG

For plasmids containing *GGPPS* of *Sulfolobus acidocaldirus*

D plazmid :

D2-TS-SpeI-F1: GTTGTTACTAGTAAAACAATGTCATCATCCAC

D2-TS-R1: TACTTGGATTGGGTCGATGTAA

D2-GGPPS F1: AGAAAGGTTTACATCGACCCAATCCAAGTAGGTTCTGGTATGAG TTATTTCGACAACTACTTCAACG

D2-GGPPS-rev: GTTGTTGAGCTCTTATTTTCTCCTCCTGATTGTAAATTCAGC

E plazmid:

E2-GGPPS-spe Fw: GTTGTTACTAGTAAAACAATGAGTTATTTCGACAACTACTTCAACG

E2 GGPPS rw: TTTTCTCCTCCTGATTGTAAATTCAGCC

E2-Ts fw: GCTGAATTTACAATCAGGAGGAGAAAAGGTTCTGGTATGTCAT CATCCACA

GGTACTTC

E2- TS-SacI-R2: GTTGTTGAGCTCTTATACTTGGATTGGG

F plasmid:

F2 Erg20-BamHI-F1:GTTGTTGGATCCAAAACAATGGCTTCAGAAAAAGAAATTAGG

F2 Erg20-R1: TTTGCTTCTCTTGTAAACTTTG

F2-GGPPS-Fw: TTCTTGAACAAAGTTTACAAGAGAAGCAAAGGTTCTATGA GTTATTTCGACAA CTACTTCAACG

F2- GGPPS-Nhe1-rw:GTTGTTGCTAGCTTTTCTCCTCCTGATTGTAAATTCAGC

G plasmid

G2- GGPPS-BamH1–F: GTTGTTGGATCCAAAACAATGAGTTATTTCGACAAC TACTTCAACG

E2.GGPPS rw: TTTTCTCCTCCTGATTGTAAATTCAGCC

G2-ErG20 Frward: GCTGAATTTACAATCAGGAGGAGAAAAGGTTCTGGTATGGC TTCAGAAAAAGAAATTAGG

G2- Erg20-NheI-R2: GTTGTTGCTAGCCTATTTGCTTCTCTTGTAAACTTTG

For plasmids containing *GGPPS* of *Taxus baccata*:

For Plasmid D

TB-TS-SpeI-F1: GTTGTT ACTAGTAAAACA ATGTCATCATCCAC

TB-TS-R1: TACTTGGATTGGGTCGATGTAA

TB-GGPPS-F1: AGAAAGGTTTACATCGACCCAATCCAAGTAGGTTCTGGT-ATGGCGTATACGGCCATGGCAGC

TB-GGPPS-PacI-R1: GTTGTTTTAATTAATTAATTTTGACGGAATGCAATATA

For Plasmid E

TB- GGPPS-SpeI-F2: GTTGTT ACTAGTAAAACA ATGGCGTATACGGCCATGGCAGC

TB-GGPPS-R2 : ATTTTGACGGAATGCAATATA

TB-TS-F2: CTTGCAGACTATATTGCATTCCGTCAAAATGGTTCTGGTA TGTCATCATCCAC AGGTACTTC

TB-TS-PacI-R2: GTTGTT TTAATTAA TTATACTTGGATTGGG

For Plasmid F

TB- ERG20-BamH1-F1: GTTGTTGGATCCAAAACAATGGCTTCAGAAAAAGAAATTAGG

ERG20 R1: TTTGCTTCTCTTGTAAACTTTG

GGPPS-F3 : TTCTTGAACAAAGTTTACAAGAGAAGCAAA GGT TCT GGT-ATGGCGTATACGGCCATGGCAGCTG

GGPPS-NheI-R3: GTTGTT GCTAGC TTAATTTTGACGGAATGCAATAT

For Plasmid G

TB-GGPPS-BamH1-F: GTTGTTGGATCC AAAACA ATGGCGTATACGGCCATGGCAGCTG

TB- GGPPS-R2: ATTTTGTCTGAATGCGATGTA

TB- ERG20-F2: TTGCAGACTATATTGCATTCCGTCAAAAT GGT TCT-GGTATGGCTTCAGAAAAAGAAATTAGG

Erg20-NheI-R2: GTTGTTGCTAGCCTATTTGCTTCTCTTGTAAACTTTG

***Taxus cuspidata* GGPPS gene primers**

Fcusp_comp:	GTTGTTGGATCCAAAACAATGGCGTATACGGCCATG
FcusPM1:	CGCAGAGCATGCAATTAAGGACGGTCGCC
RcusPM1:	CCGTCCTTAATTGCATGCTCTGCGTACCAGCT
FcusPM2:	GGAGTAAACTTTAATTTTCCTTCTCTGGGAGCAG
RcusPM2:	GAGAAGGAAAATTAAAGTTTACTCCGTGAAATGAC
FcusPM3:	CGAGACCGAGAAAGCCAAAGGGAAAGATATTG
RcusPM3:	CTTTCCCTTTGGCTTTCTCGGTCTCGGCTAGATCG
Rcus_Comp:	GCTAGCTTAATTTTGACGGAATGCAATATAGTC

List of the devices used

Stuart SBH1300 Block heater

STAR LAB AirClean 600 PCR Workstation

DYMO LabelWriter 450

Biorad Electrophoresis device

Biorad ChemiDoc XRS+ Molecular Imager

Biorad GenePulser Xcell Electroporation device

IMK-20 Hotplate Stirrer (magnetic)

MAESTROGEN Illuminator for cutting DNA from agraose gel

ZEISS AXIO scope a1 Microscope

INOVIA technology MINO-10K MINI CENTRIFUGE

Thermo Scientific maxQ 8000 thermo Shaking Incubator

Thermo Scientific Nanodrop one C

LightCycler Nano Roche Realtime PCR

Sartorius cp225d Analytical Balance

Eppendorf minispin plus small centrifuge

Eppendorf 5424 R Table top centrifuge

Applied Biosystems Standard PCR – Veriti 96 well Thermal Cycler

DAIHAN Scientific Vortex

JELO TECH Vortex

Media preparation.

YPD broth was prepared by adding yeast extract (10 g/L), peptone (10 g/L), and dextrose (20 g/L). Then the solution was autoclaved and stored at room temperature.

YPD agar plates were prepared by adding yeast extract (10 g/L), peptone (10 g/L), dextrose (20 g/L), and agar (15 g/L) and then the solution was autoclaved and poured into plates inside a laminar flow cabinet and stored at 4 °C

SDA-URA gar plates were prepared by adding yeast nitrogen base (1.7 g/L), ammonium sulfate (5 g/L), dextrose (20 g/L), CSM-URA (0.7 g/L), and agar (20 g/L) and then the solution was autoclaved and poured into plates inside a laminar flow cabinet and stored at 4 °C.

Delft media was prepared by preparing four solutions adding ammonium sulphate (7.5 g/L), potassium phosphate (14.4 g/L), magnesium sulphate (0.5 g/L), Glucose (20 g/L) trace metal solution (2 mL/L), and vitamin solution (1mL/L).

Trace metal solution contained FeSO4.7H2O (3.0 g/L), ZnSO4. 7 H2O (4.5 g/L), CaCl2.2H2O (4.5 g/L), MnCl2.4H2o (1 g/L), CoCl2.6H2O (300 mg/L), CuSO4.5H2O (300 mg/L), N2mOo4.2h2o (400 mg/L), H3BO3 (1 g/L), KI (100 mg/L), and Na2EDTA.2H2O (19 g/L) and its pH was adjusted to 6. and stored at 4 °C *in amber glass bottle after sterilization.*

Vitamin Solution contained Biotin (50 mg/L), D-Pantothenic acid hemicalcium salt (1g/L), Thiamin-HCl (1g/L), Pyridoxin-Hcl (1g/L), Nicotinic acid (1 g/L), 4-aminobenzoic acid (0.2 g/L), and m-Inositol (25 g/L). The pH of the final solution was adjusted to 6.5 and stored at 4 °C *after sterilization.*

***Sulfolobus acidocaldirus* GGPPS protein sequence from uniprot**

MSYFDNYFNEIVNSVNDIIKSYISGDVPKLYEASYHLFTSGGKRLRPLILTISSDL
FGGQRERAYYAGAAIEVLHTFTLVHDDIMDQDNIRRGLPTVHVKYGLPLAILA
GDLLHAKAFQLLTQALRGLPSETIIKAFDIFTRSIIIISEGQAVDMEFEDRIDIKEQE
YLDMISRKTAALFSASSSIGALIAGANDNDVRLMSDFGTNLGIAFQIVDDILGLT
ADEKELGKPVFSDIREGKKTILVIKTLELCKEDEKKIVLKALGNKSASKEELMSS
ADIIKKYSLDYAYNLAEKYYKNAIDSLNQVSSKSDIPGKALKYLAEFTIRRRK

***Sulfolobus acidocaldirus GGPPS* gene nucleotide sequence after reverese translation**

ATGAGCTATTTTGATAACTATTTTAACGAAATTGTGAACAGCGTGAACGATA
TTATTAAAAGCTATATTAGCGGCGATGTGCCGAAACTGTATGAAGCGAGCT
ATCATCTGTTTACCAGCGGCGGCAAACGCCTGCGCCCGCTGATTCTGACCAT
TAGCAGCGATCTGTTTGGCGGCCAGCGCGAACGCGCGTATTATGCGGGCGC
GGCGATTGAAGTGCTGCATACCTTTACCCTGGTGCATGATGATATTATGGAT
CAGGATAACATTCGCCGCGGCCTGCCGACCGTGCATGTGAAATATGGCCTG
CCGCTGGCGATTCTGGCGGGCGATCTGCTGCATGCGAAAGCGTTTCAGCTGC
TGACCCAGGCGCTGCGCGGCCTGCCGAGCGAAACCATTATTAAAGCGTTTG
ATATTTTTACCCGCAGCATTATTATTATTAGCGAAGGCCAGGCGGTGGATAT
GGAATTTGAAGATCGCATTGATATTAAAGAACAGGAATATCTGGATATGAT
TAGCCGCAAAACCGCGGCGCTGTTTAGCGCGAGCAGCAGCATTGGCGCGCT
GATTGCGGGCGCGAACGATAACGATGTGCGCCTGATGAGCGATTTTGGCAC
CAACCTGGGCATTGCGTTTCAGATTGTGGATGATATTCTGGGCCTGACCGCG
GATGAAAAAGAACTGGGCAAACCGGTGTTTAGCGATATTCGCGAAGGCAAA
AAAACCATTCTGGTGATTAAAACCCTGGAACTGTGCAAAGAAGATGAAAAA
AAAATTGTGCTGAAAGCGCTGGGCAACAAAAGCGCGAGCAAAGAAGAACT

GATGAGCAGCGCGGATATTATTAAAAAATATAGCCTGGATTATGCGTATAA
CCTGGCGGAAAAATATTATAAAAACGCGATTGATAGCCTGAACCAGGTGAG
CAGCAAAAGCGATATTCCGGGCAAAGCGCTGAAATATCTGGCGGAATTTAC
CATTCGCCGCCGCAAA

Bioinformatics data of *Taxus baccata* and *Taxus cuspidata GGPPS* genes finding

>baccata_consensus (*GGPPS* gene)

ATGGCTTACACGGCAATGGCAGCAGGGACCCAGAGCTTGCAACTCCGCACT
GTTGCTTCCTATCAAGAATGCAATAGTATGAGGAGTTGTTTTAAATTGACAC
CTTTTAAAAGTTTTCATGGAGTGAATTTCAATGTTCCCTCACTGGGTGCTGCT
AATTGTGAGATTATGGGTCACCTGAAACTTGGGTCATTGCCATATAAACAAT
GTTCGGTGTCATCTAGATCCACAAAAACAATGGCCCAGTTGGTTGATTTGGC
TGAAACAGAGAAGGCGGAGGGAAAGGATATTGAATTTGATTTCAACGAGTA
TATGAAGTCCAAGGCTGTGGCAGTGGATGCGGCACTGGATAAGGCAATCCC
ACTTGAATATCCTGAAAAAATACATGAATCAATGAGGTATTCACTTCTAGCA
GGAGGTAAGCGCGTCAGGCCTGCTCTGTGCATTGCAGCATGTGAGCTTGTA
GGAGGGAGTCAGGACCTTGCCATGCCAACTGCCTGTGCAATGGAGATGATT
CATACCATGTCTCTGATTCATGATGACTTGCCGTGCATGGATAATGATGATT
TCAGAAGAGGGAAGCCCACAAATCACAAGGTCTTTGGAGAGGACACTGCTG
TTCTTGCAGGGGACGCCCTGCTTTCATTTGCATTTGAGCATATTGCTGTGGCT
ACAAGCAAGACTGTGCCTAGTGATAGGACTTTAAGGGTGATATCTGAATTG
GGTAAGACAATAGGCTCTCAAGGGCTTGTAGGGGGGCAGGTGGTTGATATT
ACATCCGAGGGGGATGCTAATGTGGACCTGAAAACCCTGGAATGGATTCAT
ATACACAAGACTGCTGTGCTCTTGGAATGTTCAGTTGTGAGTGGAGGGATCC
TTGGTGGTGCTACAGAGGACGAGATTGCGAGAATTCGGCGGTACGCCCGGT
GTGTGGGGCTTCTGTTTCAGGTTGTGGATGACATACTTGATGTCACTAAATC
TTCTGAAGAATTGGGAAAGACTGCAGGAAAGGAT

>baccata_amioacid (**GGPPS protein)**

MAYTAMAAGTQSLQLRTVASYQECNSMRSCFKLTPFKSFHGVNFNVPSLGAA
NCEIMGHLKLGSLPYKQCSVSSRSTKTMAQLVDLAETEKAEGKDIEFDFNEYM
KSKAVAVDAALDKAIPLEYPEKIHESMRYSLLAGGKRVRPALCIAACELVGGS
QDLAMPTACAMEMIHTMSLIHDDLPCMDNDDFRRGKPTNHKVFGEDTAVLAG
DALLSFAFEHIAVATSKTVPSDRTLRVISELGKTIGSQGLVGGQVVDITSEGDAN
VDLKTLEWIHIHKTAVLLECSVVSGGILGGATEDEIARIRRYARCVGLLFQVVD
DILDVTKSSEELGKTAGKDLLTDKATYPKLMGLEKAKEFAAELATRAKEELSSF
DQIKAAPLLGLADYIAFRQN-

>cuspidate_consensus (*GGPPS* gene)

ATGGCTTACACGGCAATGGCAGCGGGTACCCAGAGCATGCAACTCCGCACT
GTTGCTTCCTATCAAGAATGCAATAGTATGAGGAGTTGTTTTAAATTGACAC
CTTTTAAAAGTTTTCATGGAGTGAATTTCAATTTTCCCTCACTGGGTGCTGCT
AATTGTGAGATTATGGGTCACCTGAAACTTGGGTCATTGCCATATAAACAAT
GTTCGGTGTCATCTAGATCCACAAAAACAATGGCCCAGTTGGTTGATTTGGC
TGAAACAGAGAAGGCGAAGGGAAAGGATATTGAATTTGATTTCAACGAGTA
TATGAAGTCCAAGGCTGTGGCAGTGGATGCGGCACTGGATAAGGCAATCCC
ACTTGAATATCCTGAAAAAATACATGAATCAATGAGGTATTCACTTCTAGCA
GGAGGTAAGCGCGTTAGGCCTGCTCTGTGCATTGCAGCATGTGAGCTTGTA
GGAGGGAGTCAGGACCTTGCCATGCCAACTGCCTGTGCAATGGAGATGATT
CATACCATGTCTCTGATTCATGATGACTTGCCCTGCATGGATAATGATGATT
TCAGAAGAGGGAAGCCCACAAATCACAAGGTCTTTGGAGAGGACACTGCTG
TTCTTGCAGGGGATGCCCTGCTTTCATTTGCATTTGAGCATATTGCTGTGGCT
ACAAGCAAGACTGTGCCTAGTGATAGGACTTTAAGGGTGATATCTGAATTG
GGTAAGACAATAGGCTCTCAAGGGCTTGTAGGGGGACAGGTGGTTGATATT
ACATCCGAGGGGGATGCTAATGTGGACCTGAAAACCCTGGAATGGATTCAT
ATACACAAGACTGCTGTGCTCTTGGAATGTTCAGTTGTAAGTGGAGGGATCC
TTGGTGGTGCTACAGAGGATGAGATTGCGAGAATTCGGCGGTACGCCCGGT
GTGTGGGGCTTCTGTTTCAGGTTGTGGATGACATACTTGATGTCACTAAATC

TTCTGAAGAATTGGGAAAGACTGCAGGGAAGGATTTGCTTACTGATAAGGC
CACTTA

TCCCAAGTTGATGGGCCTGGAGAAAGCAAAAGAATTTGCCGCTGAATTGGC
AACGAGAGCCAAGGAAGAGCTGTCATCCTTTGATCAGATTAAGGCTGCACC
TTTGTTGGGTCTTGCAGATTACATTGCATTCAGGCAAAACTGA

>cuspidata_aminoacid (**GGPPS protein)**

MAYTAMAAGTQSMQLRTVASYQECNSMRSCFKLTPFKSFHGVNFNFPSLGAA
NCEIMGHLKLGSLPYKQCSVSSRSTKTMAQLVDLAETEKAKGKDIEFDFNEYM
KSKAVAVDAALDKAIPLEYPEKIHESMRYSLLAGGKRVRPALCIAACELVGGS
QDLAMPTACAMEMIHTMSLIHDDLPCMDNDDFRRGKPTNHKVFGEDTAVLAG
DALLSFAFEHIAVATSKTVPSDRTLRVISELGKTIGSQGLVGGQVVDITSEGDAN
VDLKTLEWIHIHKTAVLLECSVVSGGILGGATEDEIARIRRYARCVGLLFQVVD
DILDVTKSSEELGKTAGKDLLTDKATYPKLMGLEKAKEFAAELATRAKEELSSF
DQIKAAPLLGLADYIAFRQN-

Gene sequence of TS of *Taxus brevifolia:*

ATGTCATCATCCACAGGTACTTCCAAGGTTGTATCCGAAACTTCCTCCACCA
TCGTAGACGACATTCCAAGATTATCCGCCAATTACCATGGTGACTTGTGGCA
TCACAATGTTATTCAAACATTGGAAACCCCTTTTAGAGAATCTTCAACTTAC
CAAGAAAGAGCTGATGAATTAGTTGTCAAAATTAAGGACATGTTCAACGCC
TTGGGTGACGGTGACATATCTCCTTCAGCTTATGATACAGCATGGGTTGCCA
GATTAGCTACCATTTCCAGTGACGGTTCAGAAAAGCCAAGATTCCCTCAAG
CATTGAACTGGGTCTTCAATAACCAATTGCAAGATGGTTCTTGGGGTATAGA
AAGTCATTTCTCTTTGTGTGATAGATTGTTGAACACTACAAACTCAGTCATC
GCTTTGTCCGTATGGAAGACTGGTCACTCTCAAGTTCAACAAGGTGCCGAAT
TCATTGCTGAAAATTTGAGATTGTTGAACGAAGAAGATGAATTGTCACCAG
ACTTCCAAATCATCTTCCCTGCATTGTTGCAAAAGGCTAAGGCATTGGGTAT
TAATTTGCCATACGATTTGCCTTTTATAAAGTATTTGAGTACCACTAGAGAA
GCTAGATTGACAGATGTATCTGCTGCAGCCGACAATATTCCAGCCAATATGT
TGAACGCTTTAGAAGGTTTGGAAGAAGTTATAGATTGGAACAAAATCATGA

GATTCCAATCAAAGGACGGTTCCTTCTTATCTTCACCAGCATCAACCGCCTG
TGTTTTGATGAATACTGGTGACGAAAAGTGCTTCACATTCTTGAACAACTTG
TTGGACAAGTTCGGTGGTTGTGTCCCTTGCATGTACTCAATCGATTTGTTGG
AAAGATTGTCCTTAGTAGACAACATTGAACATTTGGGTATCGGTAGACACTT
CAAGCAAGAAATCAAGGGTGCATTGGATTATGTTTACAGACATTGGTCAGA
AAGAGGTATTGGTTGGGGTAGAGATTCCTTGGTCCCAGACTTAAATACAAC
CGCTTTGGGTTTAAGAACCTTAAGAATGCATGGTTACAACGTATCCAGTGAT
GTTTTGAACAACTTCAAGGACGAAAACGGTAGATTTTTCTCTTCAGCTGGTC
AAACTCACGTTGAATTGAGAAGTGTAGTTAATTTGTTTAGAGCATCTGATTT
GGCCTTCCCAGACGAAAGAGCTATGGATGACGCAAGAAAATTTGCCGAACC
TTACTTGAGAGAAGCCTTAGCTACAAAGATCTCAACTAACACAAAGTTATTC
AAGGAAATCGAATACGTCGTAGAATACCCATGGCATATGTCCATCCCTAGA
TTGGAAGCTAGAAGTTACATCGATTCTTACGATGACAACTACGTCTGGCAA
AGAAAGACTTTGTACAGAATGCCATCCTTGAGTAACTCTAAGTGTTTGGAAT
TAGCAAAGTTGGATTTCAACATAGTTCAAAGTTTGCATCAAGAAGAATTAA
AGTTGTTGACAAGATGGTGGAAGGAATCTGGTATGGCTGATATCAACTTCA
CTAGACACAGAGTTGCTGAAGTCTACTTTTCCAGTGCAACATTCGAACCAGA
ATATTCTGCTACTAGAATCGCATTCACAAAGATAGGTTGTTTGCAAGTTTTG
TTTGATGACATGGCCGATATCTTCGCTACTTTGGACGAATTGAAGTCTTTTA
CTGAAGGTGTCAAGAGATGGGATACCTCCTTGTTACATGAAATACCTGAAT
GTATGCAAACTTGCTTTAAAGTATGGTTCAAGTTGATGGAAGAAGTTAATA
ACGATGTTGTCAAAGTCCAAGGTAGAGACATGTTAGCTCACATTAGAAAGC
CATGGGAATTGTACTTCAACTGTTACGTACAAGAAAGAGAATGGTTAGAAG
CAGGTTACATCCCTACCTTCGAAGAATACTTGAAAACTTATGCCATTTCTGT
TGGTTTGGGTCCATGCACATTGCAACCTATCTTGTTGATGGGTGAATTAGTA
AAGGATGACGTAGTTGAAAAGGTTCATTACCCATCTAACATGTTCGAATTA
GTTTCATTGTCCTGGAGATTGACAAACGATACCAAAACTTACCAAGCTGAA
AAGGCAAGAGGTCAACAAGCATCAGGTATTGCCTGTTATATGAAAGATAAT
CCTGGTGCAACCGAAGAAGACGCCATCAAGCACATTTGCAGAGTCGTAGAT
AGAGCTTTAAAGGAAGCAAGTTTCGAATACTTCAAGCCATCTAACGATATC
CCTATGGGTTGTAAGTCTTTTATCTTCAACTTGAGATTGTGCGTTCAAATATT

TTACAAGTTCATCGATGGTTATGGTATTGCCAACGAAGAAATCAAGGACTA
CATCAGAAAGGTTTACATCGACCCAATCCAAGTATAA

Gene sequence of *ERG20*:

ATGGCTTCAGAAAAAGAAATTAGGAGAGAGAGATTCTTGAACGTTTTCCCT
AAATTAGTAGAGGAATTGAACGCATCGCTTTTGGCTTACGGTATGCCTAAG
GAAGCATGTGACTGGTATGCCCACTCATTGAACTACAACACTCCAGGCGGT
AAGCTAAATAGAGGTTTGTCCGTTGTGGACACGTATGCTATTCTCTCCAACA
AGACCGTTGAACAATTGGGGCAAGAAGAATACGAAAAGGTTGCCATTCTAG
GTTGGTGCATTGAGTTGTTGCAGGCTTACTTCTTGGTCGCCGATGATATGAT
GGACAAGTCCATTACCAGAAGAGGCCAACCATGTTGGTACAAGGTTCCTGA
AGTTGGGGAAATTGCCATCAATGACGCATTCATGTTAGAGGCTGCTATCTAC
AAGCTTTTGAAATCTCACTTCAGAAACGAAAAATACTACATAGATATCACC
GAATTGTTCCATGAGGTCACCTTCCAAACCGAATTGGGCCAATTGATGGACT
TAATCACTGCACCTGAAGACAAAGTCGACTTGAGTAAGTTCTCCCTAAAGA
AGCACTCCTTCATAGTTACTTTCAAGACTGCTTACTATTCTTTCTACTTGCCT
GTCGCATTGGCCATGTACGTTGCCGGTATCACGGATGAAAAGGATTTGAAA
CAAGCCAGAGATGTCTTGATTCCATTGGGTGAATACTTCCAAATTCAAGATG
ACTACTTAGACTGCTTCGGTACCCCAGAACAGATCGGTAAGATCGGTACAG
ATATCCAAGATAACAAATGTTCTTGGGTAATCAACAAGGCATTGGAACTTG
CTTCCGCAGAACAAAGAAAGACTTTAGACGAAAATTACGGTAAGAAGGACT
CAGTCGCAGAAGCCAAATGCAAAAAGATTTTCAATGACTTGAAAATTGAAC
AGCTATACCACGAATATGAAGAGTCTATTGCCAAGGATTTGAAGGCCAAAA
TTTCTCAGGTCGATGAGTCTCGTGGCTTCAAAGCTGATGTCTTAACTGCGTT
CTTGAACAAAGTTTACAAGAGAAGCAAATAG

Gene sequence of *GGPPS* of *Taxus canadensis:*

ATGTTCGATTTCAACGAATACATGAAGAGTAAAGCCGTAGCCGTAGACGCA
GCATTAGACAAAGCAATACCATTAGAATACCCAGAAAAGATCCATGAATCC
ATGAGATATAGTTTGTTGGCTGGTGGTAAAAGAGTCAGACCAGCCTTGTGT
ATTGCTGCATGCGAATTAGTAGGTGGTTCTCAAGATTTGGCTATGCCTACAG
CATGTGCCATGGAAATGATACATACCATGTCATTGATCCACGATGACTTACC

ATGCATGGATAATGATGACTTTAGAAGAGGTAAACCTACTAACCATAAGGT
ATTCGGTGAAGATACAGCCGTTTTAGCTGGTGACGCATTGTTATCCTTTGCA
TTCGAACACATTGCTGTTGCAACTTCTAAGACAGTCCCATCAGATAGAACTT
TGAGAGTTATTTCCGAATTAGGTAAAACAATCGGTAGTCAAGGTTTGGTTGG
TGGTCAAGTTGTCGATATTACCTCTGAAGGTGACGCCAATGTTGACTTGAAA
ACCTTAGAATGGATTCATATACACAAGACTGCAGTTTTGTTAGAATGTTCTG
TAGTTTCAGGTGGTATATTAGGTGGTGCAACAGAAGATGAAATCGCCAGAA
TTAGAAGATATGCTAGATGCGTTGGTTTGTTGTTCCAAGTCGTAGATGACAT
ATTAGACGTCACAAAATCTTCAGAAGAATTGGGTAAAACCGCTGGTAAAGA
TTTGTTAACCGACAAAGCAACTTACCCTAAGTTGATGGGTTTAGAAAAAGCT
AAGGAATTTGCCGCTGAATTAGCCACTAGAGCTAAAGAAGAATTGTCCAGT
TTCGATCAAATCAAAGCAGCACCATTATTAGGTTTAGCAGACTACATCGCAT
TCAGACAAAATTAA

Gene sequence of *GGPPS* of *Sulfolobus acidocaldirus:*

GGATCCAAAACAATGAGTTATTTCGACAACTACTTCAACGAGATTGTTAACT
CCGTCAACGATATTATCAAGAATTACATAGCAGGGGACGTCCCGAAACTGT
ATGAGGCGAGTTATCATTTATTTACGTCAGGGGGTAAGCGTCTGAGGCCCCT
TATCTTAACGATCAGTAGTGACCTGTTTGGCGGGCAAAGAGAGAGGGCGTA
CTATGCTGGGGCTGCTATCGAGGTCTTACACACATTCACCCTAGTCCACGAC
GATATCATGGATCAAGATAATATAAGAAGGGGACTACCAACTGTCCACGTT
AAGTATGGGCTACCCCTAGCGATATTAGCCGGTGACCTATTGCATGCAAAG
GCGTTCCAGCTTTTGACTCAGGCCTTGAGAGGCCTGCCTAGCGAGACAATA
ATCAAAGCTTTCGATATATTTACCCGTTCCATCATAATAGTAAGCGAAGGAC
AAGCCGTAGATATGGAATTTGAGGACAGGATTGACATCAAGGAACAGGAGT
ATCTAGATATGATATCACGTAAAACTGCCGCCTTGTTCTCCGCCAGTAGCAG
TATTGGCGCTTTAATCGCGGGTGCCAACGACAACGACGTACGTTTGATGTCC
GACTTCGGCATGAATCTTGGTATAGCCTTTCAGATTGTGGACGACATTTTGG
GTCTGACCGCCGACGAAAAGGAGTTAGGGAAACCAGTTTTTAGCGACATTC
GTGAGGGTAAGAAGACGATTCTGATTATAAAAACATTGGAACTGTGCAAAG
AGGATGAGAAGAAGATAGTTCTGAGAGCGCTGGGTAACAAAAGCGCAAGC

AAGGAGGAGTTAATTAGCAGTGCGGATATCATAAAGAAATACTCATTAGAC
TATGCCTATAATCTGGCCGAAAAATACTATAAAAACGCCATTGATTCTTTAA
ATCAAGCTTCTAGTAAAACTGACATACCCGGCAAAGCCCTGAAGTACCTGG
CTGAATTTACAATCAGGAGGAGAAAATAAGCTAGC

Gene sequence of *GGPPS* of *Taxus Baccata:*

GGATCCAAAACAATGGCGTATACGGCCATGGCAGCTGGTACGCAGAGCCTA
CAATTAAGGACGGTCGCCAGCACCAAGAAGCAATTCAATGAGAAGTTGTTT
CAAGTTAACTCCCTTCAAGTCATTTCACGGAGTAAACTTTAATGTACCTTCT
CTGGGAGCAGCAAATTGTGAGATTATGGGTCACCTAAAGCTGGGCAGCTTA
CCTTACAAACAATGCAGTGTAAGCAGTAGGTCCACGAAAACCATGGCGCAG
CTAGTCGATCTAGCCGAGACCGAGAAAGCCGAGGGGAAAGATATTGAGTTC
GACTTCAATGAGTATATGAAGTCAAAAGCCGTTGCAGTCGACGCTGCCTTG
GACAAAGCGATCCCGCTTGAGTACCCCGAAAAGATCCACGAATCAATGAGA
TACTCCCTTCTGGCAGGGGGAAAAAGAGTTAGGCCTGCGTTGTGTATAGCG
GCGTGCGAGTTGGTTGGTGGAAGCCAGGATCTGGCAATGCCCACCGCTTGT
GCGATGGAGATGATTCATACCATGTCCCTAATTCATGACGACCTACCCTGCA
TGGACAATGATGATTTTAGAAGAGGGAAACCTACCAACCACAAGGTTTTCG
GAGAGGACACGGCCGTACTAGCAGGCGACGCACTACTTTCCTTTGCGTTTG
AACATATTGCGGTCGCAACCTCTAAAACGGTTCCGTCTGACAGAACACTGA
GGGTGATCAGCGAGTTGGGCAAGACGATAGGGTCCCAAGGTTTGGTGGGAG
GCCAGGTTGTGGACATTACGAGTGAAGGCGATGCCAACGTGGATTTGAAGA
CGCTAGAGTGGATACATATACATAAGACGGCCGTATTACTGGAGTGTAGTG
TCGTATCAGGGGGGATACTGGGTGGAGCAACAGAAGATGAGATCGCAAGA
ATTAGGCGTTATGCTAGGTGCGTAGGATTATTATTTCAAGTGGTCGATGACA
TACTAGACGTTACTAAGAGCTCTGAAGAGTTGGGTAAAACAGCGGGTAAAG
ACTTATTGACAGATAAAGCCACTTATCCAAAGTTAATGGGATTGGAGAAAG
CGAAAGAATTTGCTGCCGAATTGGCTACGAGGGCCAAGGAAGAACTGAGTT
CATTCGATCAAATAAAGGCAGCGCCCCTTTTAGGGCTTGCAGACTATATTGC
ATTCCGTCAAAATTAAGCTAGC

Gene sequence of *GGPPS* of *Taxus cuspidata:*

ATGGCGTATACGGCCATGGCAGCTGGTACGCAGAGCATGCAATTAAGGACG
GTCGCCAGCTACCAAGAATGCAATTCAATGAGAAGTTGTTTCAAGTTAACTC
CCTTCAAGTCATTTCACGGAGTAAACTTTAATTTTCCTTCTCTGGGAGCAGC
AAATTGTGAGATTATGGGTCACCTAAAGCTGGGCAGCTTACCTTACAAACA
ATGCAGTGTAAGCAGTAGGTCCACGAAAACCATGGCGCAGCTAGTCGATCT
AGCCGAGACCGAGAAAGCCAAAGGGAAAGATATTGAGTTCGACTTCAATG
AGTATATGAAGTCAAAAGCCGTTGCAGTCGACGCTGCCTTGGACAAAGCGA
TCCCGCTTGAGTACCCCGAAAAGATCCACGAATCAATGAGATACTCCCTTCT
GGCAGGGGGAAAAAGAGTTAGGCCTGCGTTGTGTATAGCGGCGTGCGAGTT
GGTTGGTGGAAGCCAGGATCTGGCAATGCCCACCGCTTGTGCGATGGAGAT
GATTCATACCATGTCCCTAATTCATGACGACCTACCCTGCATGGACAATGAT
GATTTTAGAAGAGGGAAACCTACCAACCACAAGGTTTTCGGAGAGGACACG
GCCGTACTAGCAGGCGACGCACTACTTTCCTTTGCGTTTGAACATATTGCGG
TCGCAACCTCTAAAACGGTTCCGTCTGACAGAACACTGAGGGTGATCAGCG
AGTTGGGCAAGACGATAGGGTCCCAAGGTTTGGTGGGAGGCCAGGTTGTGG
ACATTACGAGTGAAGGCGATGCCAACGTGGATTTGAAGACGCTAGAGTGGA
TACATATACATAAGACGGCCGTATTACTGGAGTGTAGTGTCGTATCAGGGG
GGATACTGGGTGGAGCAACAGAAGATGAGATCGCAAGAATTAGGCGTTATG
CTAGGTGCGTAGGATTATTATTTCAAGTGGTCGATGACATACTAGACGTTAC
TAAGAGCTCTGAAGAGTTGGGTAAAACAGCGGGTAAAGACTTATTGACAGA
TAAAGCCACTTATCCAAAGTTAATGGGATTGGAGAAAGCGAAAGAATTTGC
TGCCGAATTGGCTACGAGGGCCAAGGAAGAACTGAGTTCATTCGATCAAAT
AAAGGCAGCGCCCCTTTTAGGGCTTGCAGACTATATTGCATTCCGTC