Available online www.jocpr.com

Journal of Chemical and Pharmaceutical Research, 2015, 7(3):800-805

Research Article

ISSN : 0975-7384
CODEN(USA) : JCPRC5

Comparative transcriptome analysis of non-model *Taxus* species for taxol biosynthesis

Prashant Saxena and Balasankar Karvadi*

Department of Bioinformatics, Sathyabama University, Chennai, Tamilnadu, India

---

ABSTRACT

Taxol is a known drug, approved by the food and drug administration in 1992 for the treatment of a wide range of cancers. Taxol is obtained from the taxus plant. Taxus plant has a very long life cycle and the production of taxol from the plant is a tedious process. In this study, we have tried to find new species of taxus which can be able to produce taxol and can overcome the burden of Taxus baccata plant. For this study we have used the next generation sequencing pipeline. In this study, we have collected three transcriptome data of taxus species namely Taxus X media, taxus cuspidata, and Taxus baccata. The transcriptome data contain lots of noise, to remove that noise quality check study were performed to generate high quality data. Once the high quality data were obtained trinity tool was used to generate the high quality reads. Blastx was used to find the homologous enzymes against the high quality reads. The mapping and annotation studies were performed to find out the function of the homologous enzymes. This comparative transcriptome analysis revealed that the enzymes present in the Taxus X media shows higher similarity with Taxus baccata and contains major enzymes which are involved in the taxol biosynthetic process. So the overall postulates of this study decipher that Taxus X media may be used as a new source for taxol production and it may overcome the burden of taxus plant.

Keywords: Taxol, Taxus species, NGS, Transcriptome.

INTRODUCTION

The human genome is composed of deoxyribonucleic acid (DNA), a molecule which is very long and occupies helical structure that contains the instructions needed to build and characterize cells. For these instructions to be accomplished, DNA must be transcribed into corresponding molecules of ribonucleic acid (RNA), referred to as transcripts. A transcriptome is a compilation of all the transcripts available in a given cell [1]. There are diverse kinds of RNA. The dominant type, known as messenger RNA (mRNA), plays a fundamental role in making proteins. In this process, mRNA transcribed from genes, which include the protein-coding region of the genome, is delivered to ribosomes, which are molecular machines present in the cell's cytoplasm. The ribosomes read, or "translate," the sequence of the chemical letters in mRNA to assemble building blocks called amino acids into proteins. The mRNA is transcribed from a gene and after that translated into a particular protein. Characterization of transcriptome on a global gene expression level is an optimal application of short-read sequencing. Commonly such analysis suggests complementary DNA (cDNA) library construction, Sanger sequencing of ESTs, and microarray analysis. A new approach knew as NGS (Next generation sequencing) has become a reliable method for increasing sequencing demand and coverage while reducing time and cost compared to the conventional Sanger's method [2].

Next-generation sequencing (NGS) technology has the broad range potential of mRNA sequencing to affirm the miraculous complexity of the transcriptomes. The transcriptome sequencing approach brings quick insight of a gene, help in identifying the gene of interest, analysis of gene expression and comparative genomic studies of an organism. Although these techniques are becoming cheaper, transcriptome sequencing is still remains an expensive

endeavour. The major challenges are the assembly of millions and billions of RNA-seq reads to construct the complete transcriptome poses with the wide range information [3].

Taxus is a genus of yews, small coniferous trees or shrubs in the yew from kingdom Plantae, division Pinophyta, class Pinopsida, order Pinales, family Taxaceae. They are relatively slow-growing and can be very long-lived, and reach heights of 1–40 m, with trunk diameters of up to 5 m [4]. They have reddish bark, lanceolate, flat, dark-green leaves 1–4cm long and 2–3mm broad, arranged spirally on the stem, but with the leaf bases twisted to align the leaves in two flat rows either side of the stem [5].

Taxus is a gymnosperm genus of yews, small coniferous trees or shrubs in the yew from kingdom Plantae, division Pinophyta, class Pinopsida, order Pinales, family Taxaceae. The life cycle of these plants is relatively very long, and they are very slow growing. Normal heights of these trees are 1 cm to 40cm and diameter is up to 5cm.

Taxus plants are very rich in economic value, have medical importance and produce a drug named Taxol [6] which is used for the treatment of a wide range of cancer including colon cancer and breast cancer. Some recent studies decipher that Taxol (Paclitaxel) [7] may also be useful in the treatment of alzheimer's disease [8].

Taxol is an anti-cancer chemotherapy drug. Taxol is classified as a "plant alkaloid," a "taxane" and an "antimicrotubule agent" [9]. Taxol (Paclitaxel) [7] is one of the natural diterpenoid from the bark of the yew (*Taxus brevifolia)* [10]. It has the capability to destroy the tumor cells by build-up the assembly of microtubules and constrain their depolymerisation. It is an approved drug by the food and drug administration (FDA) for the treatment of a wide range of cancers using chemotherapy.

For the commercial production of taxol very few species are being used till date, among them *Taxus baccata* is commonly used species, due to which the burden of taxol production from this plant increasing day by day and it is tough for the pharmaceutical companies to cope up with the market demand of taxol with very few known species. So to keep this challenge in our mind we have tried to find the other *Taxus* species which may also be able to produce taxol to decrease the burden of *Taxus baccata* plant and met with the demand of market.

EXPERIMENTAL SECTION

Data Collection

The transcriptome data of the three *taxus* species i.e. *Taxus baccata* (SRR065067), *Taxus cuspidata* (SRR032523)and *Taxus X media* (SRR534003)were collected from the DDBJ-DRA [11] (http://trace.ddbj.nig.ac.jp/DRASearch/) database. The transcriptome data contain83913444, 60173416 and 1225567350 bases respectively.

Pre-Processing of Reads

The transcriptome pre-processing analysis comprises of three stages namely- Primary analysis, Secondary analysis and Tertiary analysis. These steps together comprise the pre-processing of reads. QC is an important and effective measure for determining sample libraries' qualities.

Primary analysis of the transcriptome data includes scrutinizes the quality of the reads. Once we understand the quality of the reads we can make the decision over the data. If the quality of the obtained data is not satisfactory, various cleaning measures can be used to make the data useful. For this current study for the primary analysis of the data we have used two tools *i.e.* FASTQC and NGSQC Toolkit [12].

De-Novo Assembly of Reads

De-novo is referring to the start from beginning and de-novo assembly refers to the process of creating a transcriptome without the help of a reference genome. This approach is preferred over other methods to study the non-model organism. These assembled transcriptomes are able to decipher the peculiar protein which plays an important role in the biological process of the organism. In this study we have used DDBJ read annotation pipeline [13] [14] Trinity [15] tool for the assembly of the filtered reads.

Functional Annotations

The process of functional annotation refers to the annotation of the transcriptomes. Annotation includes methods to add biological information to the raw DNA sequence, identify the structural and functional elements, and integrate and display this information at a genomic level. Functional annotation of the transcriptomes was performed to identify the expressed genes present in the transcriptome of all three *taxus* species. The functional annotations of the

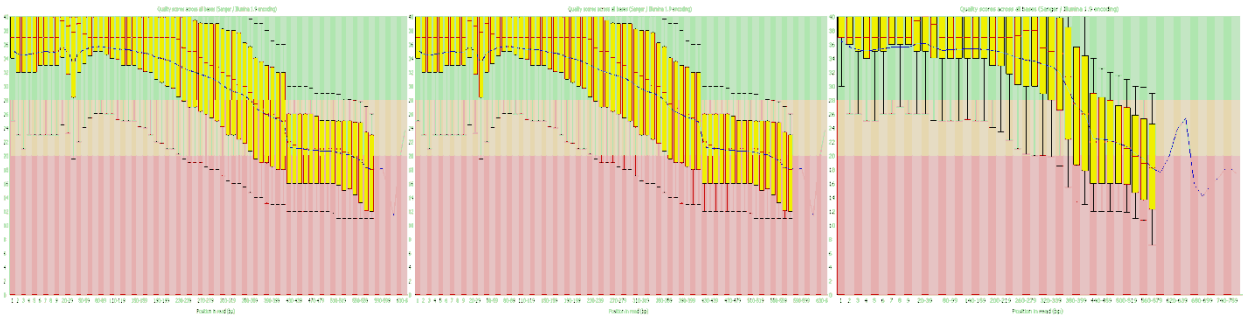
three transcriptomes were performed using Blast2Go PRO package. Blast2Go PRO [16] is a commercial package, developed by BioBam bioinformatics solutions.

RESULTS

The three transcriptome sequences of *Taxus* species were retrieved from the DDBJ-DRA database. Once the transcriptome data were downloaded the first step is the quality check of the data to verify the quality of the data.

Quality Check

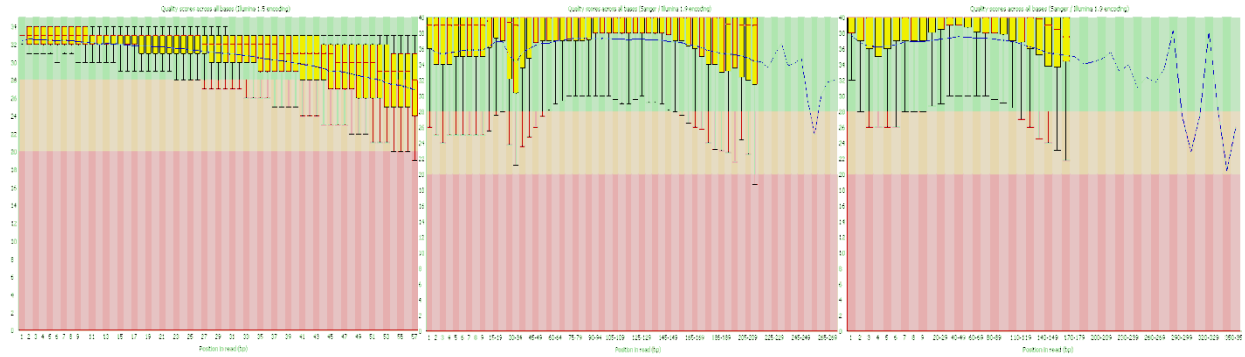
Transcriptome data of any organism contains lots of the noise in it. So the first step of the quality check is to identify the noise of the data and region in which noise resides.



(A) **(B)** **(C)**

Figure 1: Raw reads (A) *Taxus X media* (B) *Taxus cuspidata*(C) *Taxus baccata*

Once the noise of the data is been identified the next step is to clean the data and make this data useful. To make this data useful we have used NGSQC toolkit for the quality check. We used Illuqc and Trimming tool to increase the quality of the data and removal of unwanted data from the all three transcriptome sequence of *Taxus* species.



(A) **(B)** **(C)**

Figure 2: Filtered reads (A) *Taxus X media* (B) *Taxus cuspidata* (C) *Taxus baccata*

The graph shown in the figure is box-whisker graph, which is divided into three different colours. The green colour represents the very high quality reads, orange colour shows good quality reads while red colour shows the low quality reads. Figure-1 shows the low quality reads which contains a high amount of noise in the data while figure-2 contains the filtered reads which is noise free.

Table 1: Overall summary of quality check results

Species	Total Reads		Sequence Length		% GC Content	
	Before	After	Before	After	Before	After
Taxus X media	8170449	8132189	75	57	42	42
Taxus cuspidata	81146	77685	25-632	0-273	42	43
Taxus baccata	409750	397846	40-771	0-351	43	43

De-Novo Assembly

Transcriptome sequence of all three *Taxus* species was assembled using de-novo assembly approach, which can be used in absence of the reference sequence. We have used DDBJ read annotation pipeline Trinity tool to perform the de novo assembly of the transcriptomes. The k-mer size used for the de novo assembly was 25 and the minimum

contig size to be produced by trinity was set to be 200. The total contigs were 8353, 6901 and 20688 were obtained for *T X media, T cuspidata* and *T baccata* with N_{50} size 318, 868 and 1079 respectively.

Table 2: Contig assembly of *Taxus* transcriptome sequences

Measure	*T X media*	*T cuspidata*	*T baccata*
Total number of contigs	8353	6901	20688
Total contig size	2762900	5111480	17974328
Maximum contig size	2889	3586	7554
Minimum contig size	201	201	201
N_{50} contig size	318	868	1079

Functional Annotation

The annotation refers to the detailed classification of the particulardata, in this study we have annotated the biosynthetic pathway of the taxol. We have used Blast2Go package version 3.0. Functional annotation was performed in three steps *i.e.* BLAST to find homologus sequences, MAPPING to retrieve GO terms and ANNOTATION to select reliable functions.

BLAST (Basic local alignment search tool) is a database search algorithm, run a particular nucleotide/protein sequence against a database and returns the homologus sequence of it. In this study we have used Cloud Blast (A special feature of Blast2Go PRO package). Cloud blast is a very effective tool for large data set. We have used BLASTx tool against the nr database (Non redundant database), with the E-value 1e-5.

Mapping is the process to retrieve the GO terms. Gene ontology is a method by which we can give information about something we already know about. The Gene Ontology provides, ontologies for the defined terms specific for a gene product properties. The ontology gives the detail information divided into three domains i.e. cellular components, molecular function and biological processes.

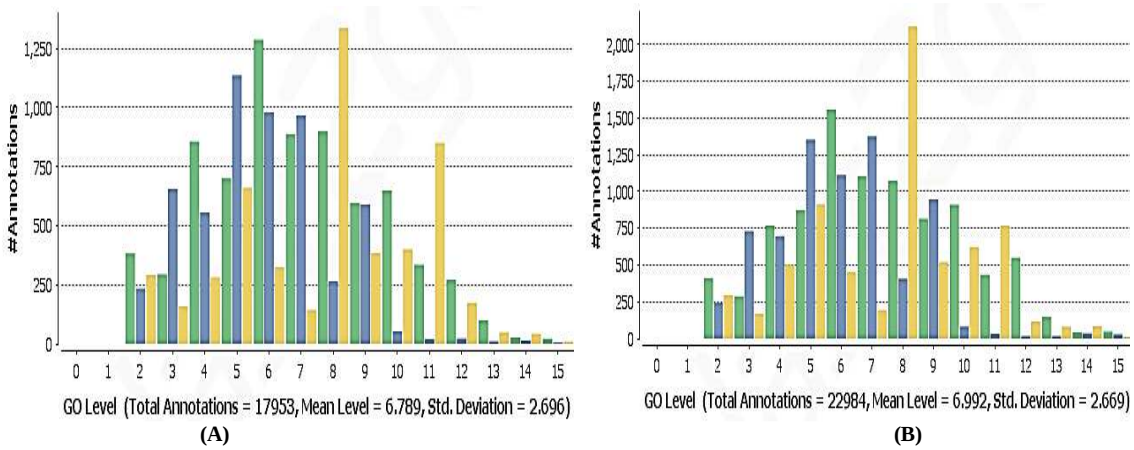
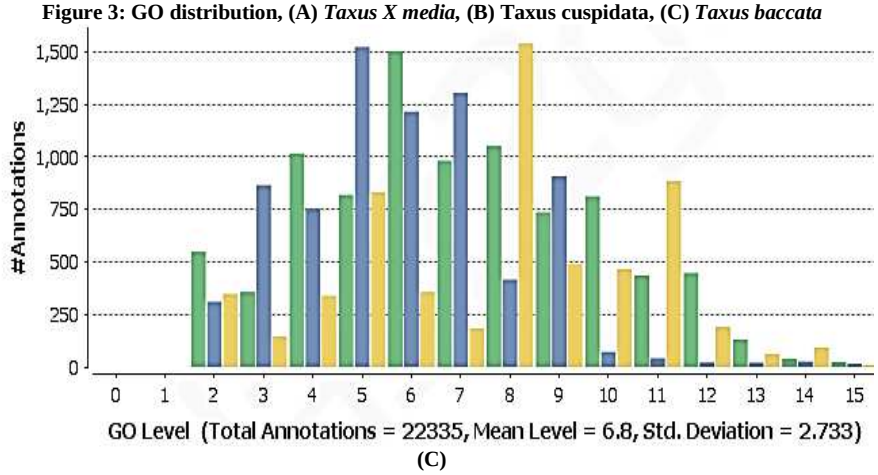


Figure 3: GO distribution, (A) *Taxus X media,* (B) Taxus cuspidata, (C) *Taxus baccata*



Three different in the graph explains the three different kind of data i.e. green tells about biological processes, blue tells about the molecular function and yellow tells about the cellular components. In this study we have considered

ontologies, only related to the biological processes mainly involved in the taxol biosynthesis to reach the aim of this study.

The biological processes involved various functions such as reproductive processes, response to a stimulus, cellular process, growth, immune system process and metabolic process.

Annotation is the process to gain the detail information about a particular gene or enzyme. In this study,we have considered the biological process to understand the metabolic process involved in the biosynthetic pathway of taxol.

Table 3: Comparative analysis of the enzymes among three *taxus* species

Enzymes	*Taxus X media*	*Taxus cuspidata*	*Taxus baccata*
Taxane 13 alpha hydroxylase	✓	✓	✓
Taxadiene 5n alpha partial	✓	✓	✓
Taxadienol acetyl partial	✓	*	✓
Taxadiene synthase	✓	*	✓
Taxadiene 5 alpha hydroxylase	✓	*	✓
Taxane 14 beta hydroxylase	✓	*	✓
10-deacetylbaccatin III 10 O acetyltransferase	✓	*	✓
Taxane 13 alpha hydroxylase	✓	*	✓
Geranylgeranyldiphosphate	✓	✓	✓

(- Not present)*

The table-3 explains about the presence and absence of the particular enzyme in the respective *Taxus* species. And from the table it is clear that the enzymes of the *Taxus baccata* and *Taxus X media* shows similarity while *Taxus cuspidata* do not show the similarity among them.

DISCUSSION

Functional annotation refers to the prediction of the function corresponds to the particular region of transcriptome to understand the overall function of the transcriptome. To achieve that we have classified the functional annotation part into three sub segment *i.e.* blast, mapping and annotation. The blast result shows the homologous sequence to the respective contigs, which will help in understanding the function of respective contigs. Among the 35942 contigs 25122 contigs were found homologous sequence. After the identification of homologous sequence, mapping was performed to identify the GO terms. 25122 contigs were analysed and they mapped against various databases *i.e.* Uniprotkb, TAIR, GR_PROTEIN, MGI, RSD, ASPGD, and SGN respectively. After mapping the contigs against respective databases, annotation studies were performed and their biological processes were studied. In this biological process various processes are involved such as single organism process, signalling, rhythmic process, response to stimulus, reproductive process, reproduction process, multicellular organismal process, multi organism process, metabolic process, biological reaction, cellular component process, cellular process, developmental process, growth, immune system process and localization process. Among these various processes, we considered metabolic process. These three species collaboratively have 8921 metabolic processes. Further analyses revealed that among these 8921 processes approximately 56 enzymes were found related to taxol biosynthetic pathway. Some conserved enzymes among these three species are Taxadiene partial, Taxadienol acetyl partial, Taxadiene 5n alpha partial, Taxadiene synthase, Taxane 14 beta hydroxylase, Taxadienol acetyl partial, Taxane 13 alpha hydroxylase.

CONCLUSION

This transcriptome analysis did not reveal all of the enzymes involved in taxol biosynthesis. The expression of genes related to taxol production is subjected to the environmental conditions. Some genes would be highly expressed when some elicitors are added or under certain stress conditions.This comparative transcriptome analysis of non-model plants revealed that the enzymes present in the *Taxus baccata* and *Taxus X media* shows higher similarity among them, while Taxus cuspidata also shares very few conserved enzymes. So the overall translations of this study decipher that *Taxus X* media may be able to produce the taxol which will help in the overcome the burden of taxus plant for the biosynthesis of taxol and may help in the finding of other source of taxol production.

REFERENCES

[1] Z Wang;M Gerstein;M Snyder.*Nat Rev Genet.*,**2009**, 10(1), 57-63.
[2] LJ Collins; PJ Biggs, *Genome Inform*,**2008**, 21, 3-14.
[3] DC Hao mail;G Guangbo;PXiao; YY Zhang;L Yang mail, *PLoS ONE*, **2011**, 6(6), 1-15.
[4] SR Strickler; ABombarely and LA Mueller, *Am. J. Bot*.**,2012**, 99(2), 257-266,
[5] Moir and Andy. *Quarterly Journal of Forestry*, **2013**, 2013, 187.

[6] ER Mustafa and NCoşkun, *ARKIVOC*,**2009**, (xii), 153-160.
[7] WPBaasandJF Ahmad, *Brain*, **2013**, 136(10), 2937–2951
[8] BHGuo; GY Kai; HB Jin and KX Tang, *African Journal of Biotechnology*,**2006**, 5 (1), 15-20.
[9] S Peltier;JM Oger;F Lagarce;W Couet;JP Benoît*Pharm Res*. **2006**, 23(6):1243-50.
[10] S Visalakchi and JMuthumary*Int. J.* Pharm. *Biol. Sci*, **2010**, 1(3), 1-9.
[11] E Kaminuma; J Mashima; Y Kodama et.al.,*Nucleic Acids Research*, **2010**, 38(Database issue), 33–38.
[12] MG Grabherr;BJ Haas; MYassour;JZ Levin; DA Thompson;I Amit et al, *Nat Biotechnol*, **2011**, 29(7):644-52.
[13] N Hiedeki et al., *DNA research*,**2013**, 20(4), 383-390.
[14] S Andrews (**2010** Data [Online]. Available online at:http://www.bioinformatics.babraham.ac.uk/projects/fastqc/
[15] A Conesa; S Götz; JM Garcia-Gomez; J Terol; M Talon and M Robles*Bioinformatics*,**2005**, 21, 3674-3676.
[16] A Conesa and SGötz*Int J Plant Genomics*,**2008**, 2008, 1-13.