National Mission on Himalayan Studies (NMHS)

PERFORMA FOR THE HALF YEARLY PROGRESS REPORT

(Reporting Period from April 2017 to September 2017)

1. Project Information

Project ID:	**NMHS/SG-2016/011**	**Sanction Date:**	**31st March,2016**
Project Title:	**Conservation strategies for *Taxus wallichiana* and *Ulmus wallichiana* by DNA markers and geospatial technologies.**		
BTG:			
PI and Affiliation (Institution):	**Dr. Pankaj Bhardwaj, Asst. Professor, Centre for Plant Sciences, Central University of Punjab, Bathinda.** Email: pankajihbt@gmail.com Phone No. 9501686709		
Name & Address of the Co-PI, if any:	**Dr. Puneeta Pandey, Asst.Professor, Centre for Environmental Sciences and Technology, Central University of Punjab, Bathinda** Email: puneetapandey@gmail.com		

Structured Abstract - detailing the current year progress [Word Limit 250 words]:	Samples for RNA isolation were collected from Saloni, Himachal Pradesh for *Taxus wallichiana* and from Rajori J&K for *Ulmus wallichiana* in liquid nitrogen. RNA was isolated using CTAB method with some modifications. Sequencing was performed on Illumina HiSeq 2000 platform. The raw reads were processed by using Trim Galore to remove adapter sequences and low quality bases. The cleaned reads were denovo assembled using Trinity followed by removal of sequence redundancy and generation of unigenes using CD-HIT at 95% sequence identity threshold. The completeness of the assembly was analysed using BUSCO version 2. Further, the raw reads were mapped back to the assembly using Bowtie2 for quality assessment. The non-redundant assembly was annotated by using the annotation pipeline Annocript. Further, the transcripts were assigned to gene families using the pipeline TRAPID. For the identification of transcription factors, PlantTFDB was used. SSRs are widely used markers for genetic studies because of the promising features like codominant nature, reproducibility, highly informative, cross-transferability between related species etc. Especially for wild tree species where SNP based approach is cumbersome due absence of reference genome and comparatively much input cost, these SSR markers are cost effective to decipher population and landscape genetic problems in such cases. *Taxus wallichiana* is an endangered species whose spread is constantly reducing due to human intervention and changing climatic scenario. Extensive population and landscape genetic information is required in order to plan conservative strategies for this species. For this reason we screened our transcripts for SSR identification and found that 6507 sequences contained 7041 SSRs out of which 534 were in compound form and 800 sequences contained more than 1 SSRs. For *Ulmus wallichiana*, we screened our transcripts for SSR identification and found that 14042 sequences contained 16570 SSRs out of which 1318 were in compound form and 2094 sequences contained more than 1 SSRs. We used BatchPrimer3 for designing primers for the identified SSRs. A total of 4958 and 9000 primer pairs were successfully designed for Taxus and Ulmus respectively from the SSR containing sequences. These primers will be

	employed in population and Landscape genetic analysis.

	Project Partner Name	Affiliations	Role & Responsibilities
1	**Dr. Pankaj Bhardwaj**	**Centre for Plant Sciences, School of Basic and Applied Sciences, Central University of Punjab, Bathinda**	Sample collection and nucleic acid isolation Transcriptome sequencing for *T. wallichiana* and *U. wallichiana* Assembly, annotation, Prediction of microsatellites, primer-designing and Characterization on the populations
2	**Dr. Puneeta Pandey**	**Centre for Environmental Science and Technology, Central University of Punjab, Bathinda**	Micro-level and macro-level spatial mapping of *T. wallichiana* and *U. wallichiana* Change detection studies to ascertain the changes in vegetation pattern and temperature variations in the last 40 years.

2. Project Site Details

Project site	Central University of Punjab, Bathinda
IHR states covered	
Lat./Long.	
Sitemaps	
Site photographs	

3. Project Activities Chart w.r.t. Timeframe [Gantt or PERT]

PROJECT ACTIVITIES	WORK UNDERTAKEN April 2017 to September 2018 Qtr 1	Qtr 2	Qtr 3	Qtr 4	OUTPUT
Sampling	Samples for isolation of RNA for *Ulmus wallichiana* was collected from Rajori J&K under senescent condition				RNA sample under senescent condition for Ulmus and under summer condition for Taxus.
RNA sequencing, Assembly and Functional Annotation	Sequencing was performed at Genotypic Technology Pvt. Ltd, Bangalore	Raw reads were processed and denovo assembled. Functional annotation was carried out using Annocript and Trapid.			7041 SSRs for Taxus and 16570 SSRs for Ulmus. 9000 designed primers for Ulmus and 4958 designed primer pairs for Taxus.

		Identification of SSRs containing sequences was done using MISA and Primer designing was carried using BatchPrimer3			

4. Project Beneficiary Groups

Beneficiary Groups [Capacity Building]	Target	Achieved
No. of Beneficiaries with income generation:		
No. of stakeholders trained, particularly women:		
No. of capacity building Workshops/ trainings:		
No. of Awareness & outreach programmes:		
No. of Research/ Manpower developed:		

5. Project Progress Summary (For the reporting period only)

Description	Total (Numeric)	Description
IHR States Covered	Himachal Pradesh, Uttrakhand & J&K (Partial)	
Project Site/ Field Stations Developed:	.... (attach photos) ... (attach maps)	
No. of Patents filed (Description):		
Article/ Review/ Research Paper/ Publication:		
New Methods/ Modellings Developed (description in 250 words):		
No. of Trainings (No. of Beneficiaries):		
Workshop:		
Demonstration Models (Site):	.... (attach maps about location & photos)	
Livelihood Options:		
Training Manuals:		
Processing Units:	.... (attach photos)	

Species Collection:		
Species identified:		
Database/ Images/ GIS Maps:		

6. Project Concluding Remark

Project Objectives	Project Output against each objective	Progress made against Monitoring Indicators (specified in Sanction Letter)	Remarks
Sample collection and nucleic acid isolation Transcriptome sequencing for *T. wallichiana* and *U. wallichiana*	RNA Sequencing done at Genotypic Technology pvt. Ltd. Bengaluru	Samples for isolation of RNA under senescence for Ulmus and Summer conditions for Taxus was carried out. Data processed, functional annotation done, SSRs identified and Primers designed for both Ulmus and Taxus.	

Methodology in brief	Samples for RNA isolation were collected from Saloni, Himachal Pradesh in liquid nitrogen. RNA was isolated using CTAB method with some modifications. Sequencing was performed on Illumina HiSeq 2000 platform. The raw reads were processed by using Trimmomatic to remove adapter sequences and low quality bases. The cleaned reads were denovo assembled using Trinity followed by removal of sequence redundancy and generation of unigenes using CD-HIT at 95% sequence identity threshold. The completeness of the assembly was analysed using BUSCO version 2. Further, the raw reads were mapped back to the assembly using Bowtie2 for quality assessment. The non-redundant assembly was annotated by using the annotation pipeline Annocript. Further, the transcripts were assigned to gene families using the pipeline TRAPID. For the identification of transcription factors, PlantTFDB was used. SSRs were identified using MISA and primers were designed using BatchPrimer3. Positional distribution of the SSRs in the transcripts was analysed by predicting the ORFs from the SSR containing sequences using orfPredictor (Min et al, 2005) followed by correlating the SSR start and end positions with the start and stop positions of the predicted ORFs.
Major Achievements	• SSRs identified • Primers designed
Brief conclusion, current year progress- during the reporting period (point wise)	We successfully sequenced transcriptomes of Taxus and Ulmus and screened them for SSRs containing sequences. We identified the SSR regions and designed primers for them. Further we functionally annotate the transcriptomes of both species.
Progress achieved (%)	
Remaining work to be done	• Marker characterization and their utilization in population and Landscape genetic analysis of *Taxus wallichiana* and *Ulmus wallichiana.* • Sampling from eastern Himalayan regions.

	• GIS mapping.

7. Next Reporting Plan and Projections (Month wise)

No	Month	
1	October	Designed markers to be synthesized
2	November	do
3	December	Characterization of the synthesized markers
4	January	Genotyping of the so far collected samples
5	February	do
6	March	do

8. Additional information, if any:

Submitted to:
Nodal Officer, NMHS-PMU
G.B.Pant National Institute of Himalayan Studies (NMHS)
Sustainable Development (GBPNIHESD), Kosi-Katarmal
Almora 263643, Uttrakhand
E-mail: nmhspmu2016@gmail.com

Submitted by:
Project PI (Signature)
Institutional (Seal):
Dated:

Please fill the NMHS Progress Report pro forma as applicable with respect to the reporting period and other requirements and submit via post/email. In case of any query, please contact at : nmhspmu2016@gmail.com

Annexure-I

Detailed Progress Report (min. 4-5 pages)
(For the reporting period only)

Each Object wise, please provide detailed description of the results and outcomes achieved with specific Measurable Deliverables during the reporting period (Also, supporting material/data/files/tables/figures depicting the results of the reporting material should be attached in quality and precision)

Details of work done

- **Sequencing, assembly and functional annotation**

Samples for RNA isolation were collected from Saloni, Himachal Pradesh for *Taxus wallichiana* and from Rajori J&K for *Ulmus wallichiana* in liquid nitrogen. RNA was isolated using CTAB method with some modifications. Sequencing was performed on Illumina HiSeq 2000 platform. The raw reads were processed by using Trimmomatic to remove adapter sequences and low quality bases. The cleaned reads were denovo assembled using Trinity followed by removal of sequence redundancy and generation of unigenes using CD-HIT at 95% sequence identity threshold. The completeness of the assembly was analyzed using BUSCO version 2. Further, the raw reads were mapped back to the assembly using Bowtie2 for quality assessment. The non-redundant assembly was annotated by using the annotation pipeline Annocript. Further, the transcripts were assigned to gene families using the pipeline TRAPID. For the identification of transcription factors, PlantTFDB was used.

Table 1: Read, contig and unigenes statistics of *Taxus wallichiana.*

Reads	
Total reads	48512357
Total bases (nt)	965875423584
Mean length (nt)	160
Assembled Contigs	
Number	158819
Median length (nt)	848
Average contig	1294.39
N50 statistics (nt)	1712
GC%	39.30
Unigenes	
Mean length (nt)	
Number	129869
Mean sequence length (nt)	1244
Median length (nt)	820
N50 statistics (nt)	1606
Longest/shortest contig	17362/500

Number of sequences > 1K (nt)	51352 (39.5%)
Number of sequences > 10K (nt	57 (0.0%)
Unigenes annotation	
Full length	11587 (8.9%)
Quasi full length	10660 (8.2%)
Partial	81224 (62%)
ORF with start codon	93033
ORF with stop codon	114956

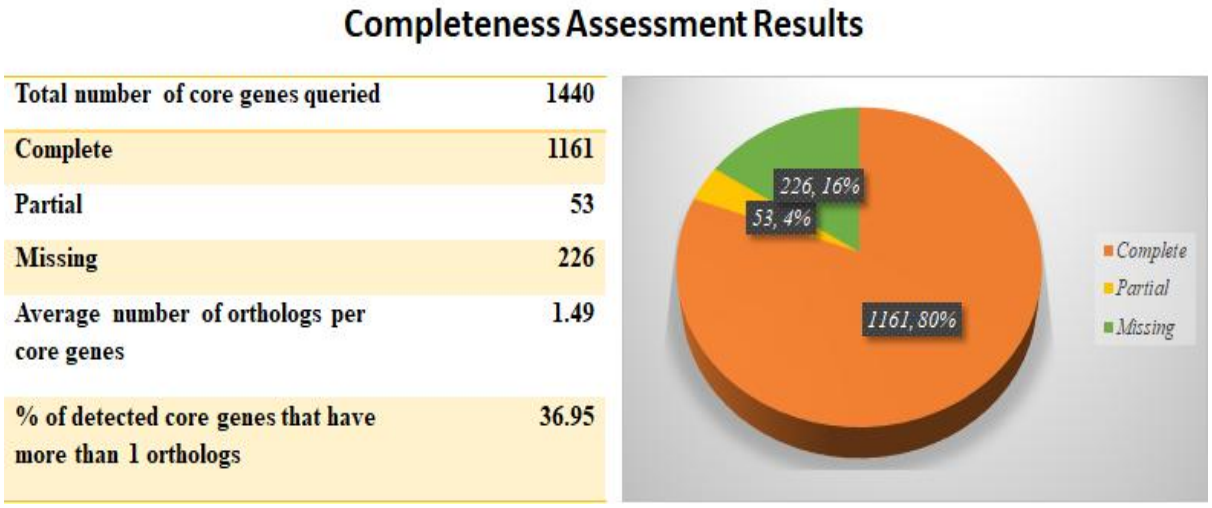


Figure 1: Busco analyzed completeness assessment results.

In case of *Taxus wallichiana* the Annocript resulted in a total of 35199 hits in Swiss-Prot corresponding to 342 organisms, 56946 hits in TrEMBL corresponding to 511 organisms, 44642 hits in Conserved Domain databases and 1056 hits in SILVA database. Functional annotation with GO database using Annocript yielded 2183 results in Biological Processes, 555 in Cellular components and 1723 in Molecular functions. Further, 58 results were obtained for pathway level 1, 207 in level 2 and 302 in level three pathway. 6219 sequences were observed to be long non-coding under 0.95 probability and 100 nucleotides of maximum length of ORF. A search in Pfam database using Annocript yielded 4134 hits among which 1424 transcripts (3.189%) were found to contain Protein kinase domain which constituted the largest found domain in our assembly. They are involved in transferring a gamma phosphate form nucleoside triphosphates to a serine/threonine or tyrosine residue of a protein substrate thereby causing a conformational change to affect the protein function. It belongs to Protein kinase-like (PK-like) superfamily. PPR repeat family is the second largest family in our assembly. This family is expanded in plant kingdom but with no known function. We found abundance of reverse transcriptases belonging to three families RVT_1/2/3. About 469 (1.05%) transcripts belong to Pentatricopeptide repeat (PPR) family. They are widespread in plants and their genes lack introns. Figure below represent the top hits for closer organism abundances in TrEMBL and SwissProt, GO annotations for biological processes, cellular components and molecular functions, pathway results and top conserved domain hits respectively.

Using TRAPID pipeline, out of the 129,869 transcripts in the assembly, 26398 (20.3%) were meta annotated as full length, 11587 (8.9%) as quasi full length, 10660 (8.2%) as partial and 81224 (62%) with no meta annotation information. Further, 93033 of our transcripts have the open reading frame with start codon while 114956 transcripts bear stop codon. Further, using Plaza Database, TRAPID assigned the

transcripts into gene families. TRAPID uses BLASTX and RapSearch as sequence similarity search tools. A total of 50628 (39%) transcripts were assigned to 6854 gene families of which 5983_HOM000056 constitutes the largest family accounting about 2315 transcripts. No GO or Protein domain annotation is reported from TRAPID for this assigned family. Thus it appears that these transcripts are unique to *Taxus wallichiana.* 1714 transcripts were found as single copy gene families.

In order to identify the transcription factors, we analyzed our transcripts in PlantTFDB (http://planttfdb.cbi.pku.edu.cn). The database hosts 26402 TFs inferred from **25** species with comprehensive gene and family level annotation. A total of 19 TF were predicted from our transcripts. These 19 transcription factors belong to seven TF families. GRAS constituted the largest family followed by ERF and Trihelix. GRAS TF occur throughout the plants and mainly act in developmental processes and signal transduction. ERF is involved in biological processes related to stress where it acts by responding to salt and water stress, ethylene, abscisic acid salicylic acid, and jasmonic acid and plays vital role in anatomical and developmental processes in diverse tissues like guard cells, seed, flower, etc. RAV is involved in ethylene activated signalling pathway and negatively regulates transcription, modulates drought and salt response. bHLH is involved in callose deposition in cell wall , pollen, anther wall and tapetum development. NF-YA is involved in various anatomical and developmental processes by acting as negative regulator of transcription and also by DNA binding action. C2H2 acts by binding to metal ions and DNA

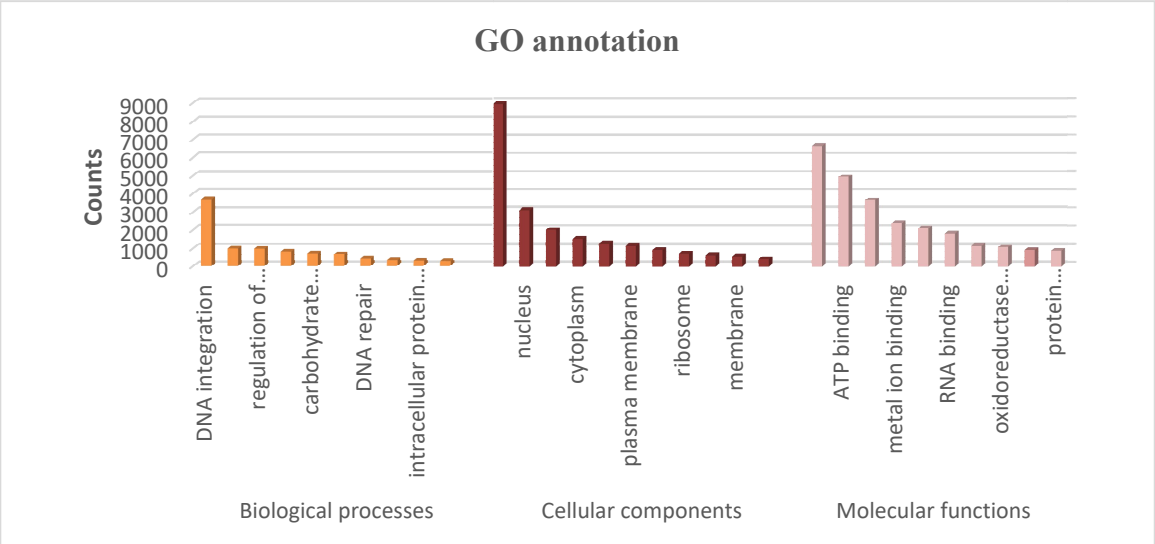


Figure: GO annotation of T. wallichiana

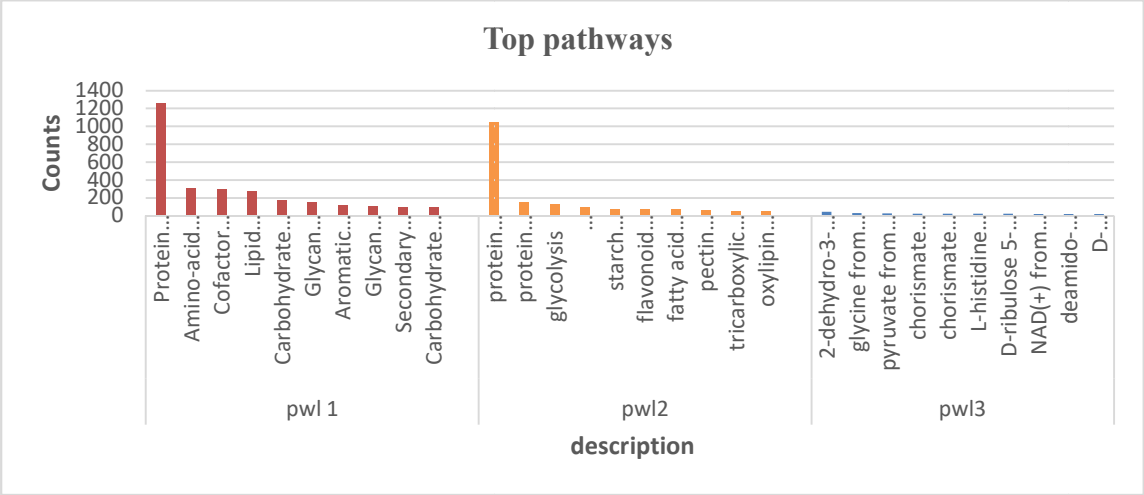


Figure: Top pathways obtained for T. wallichiana

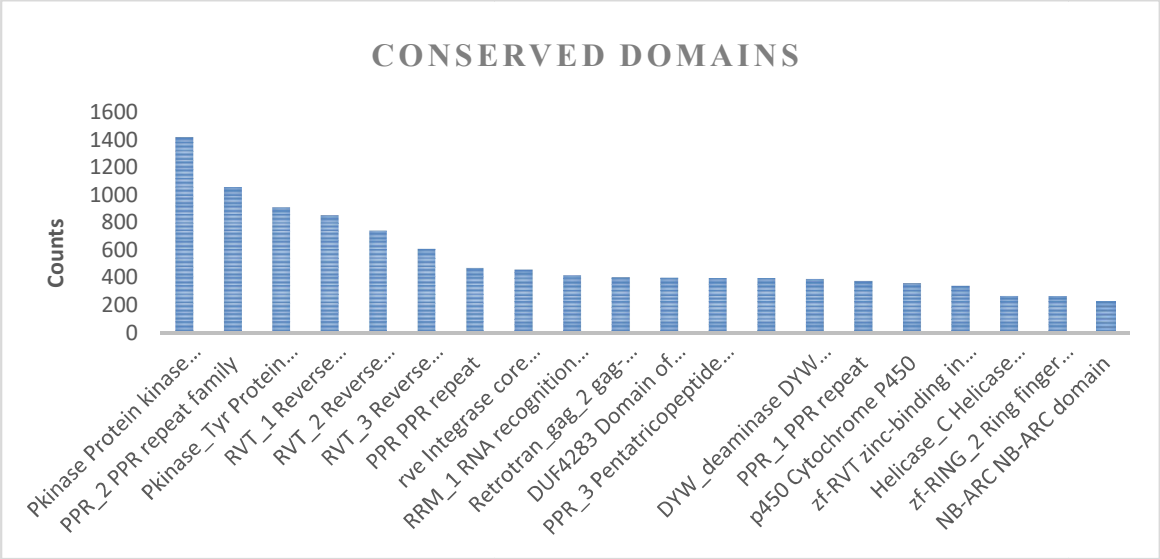
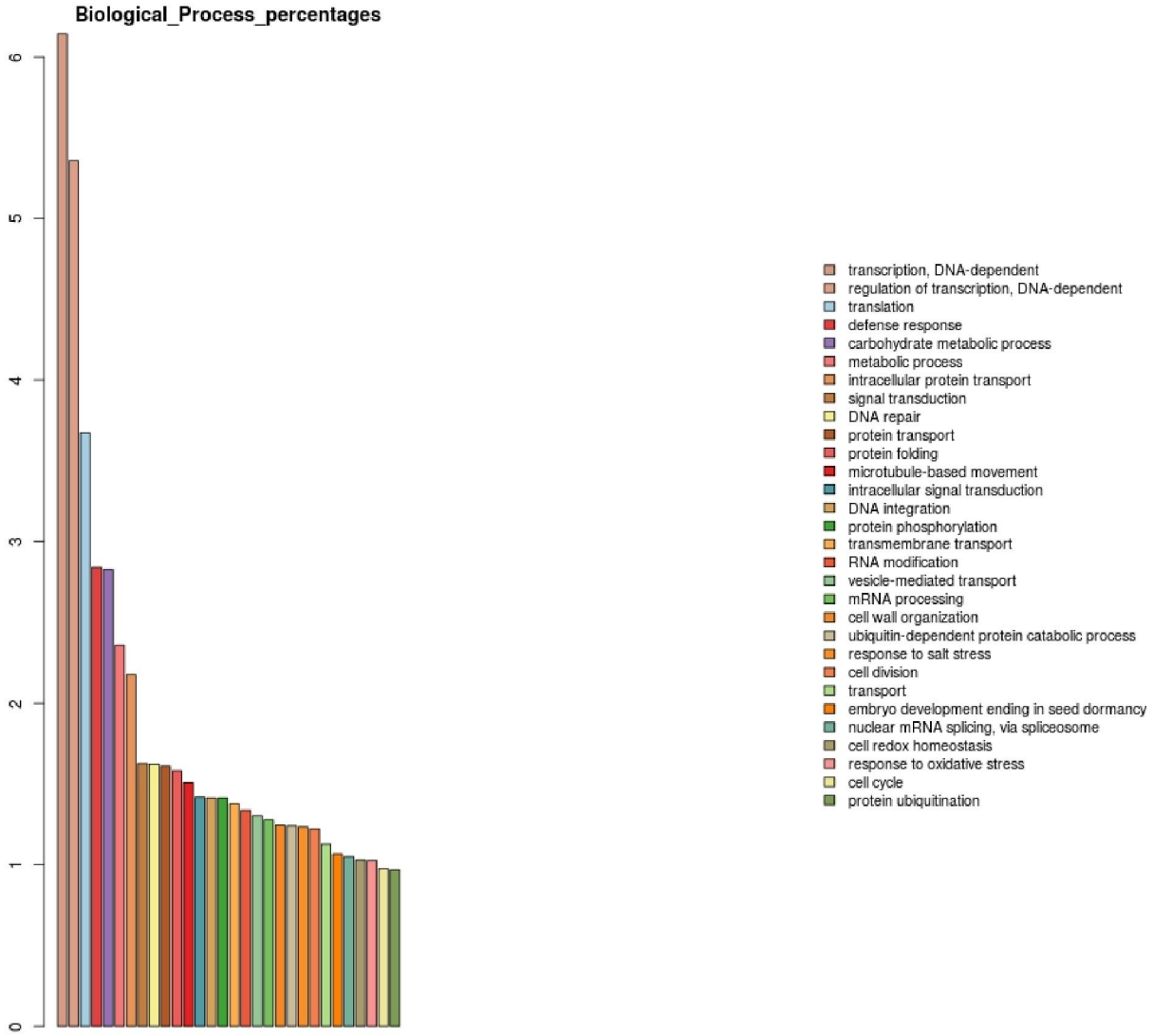
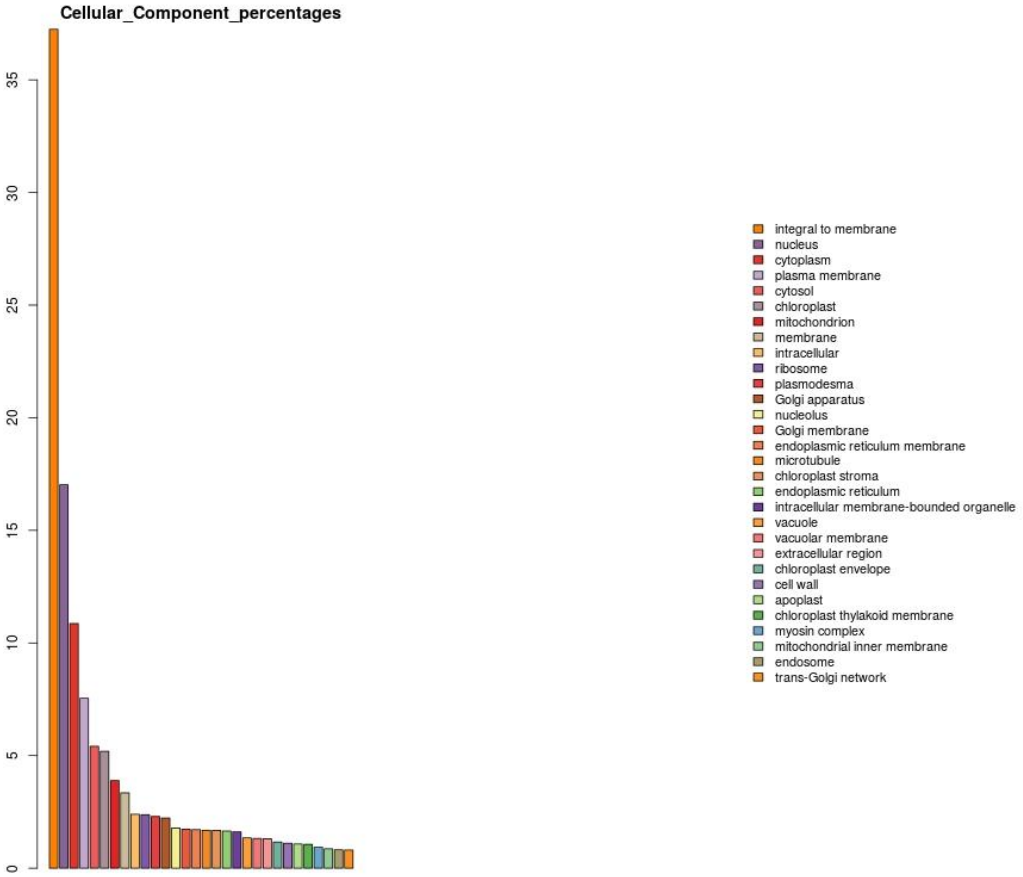
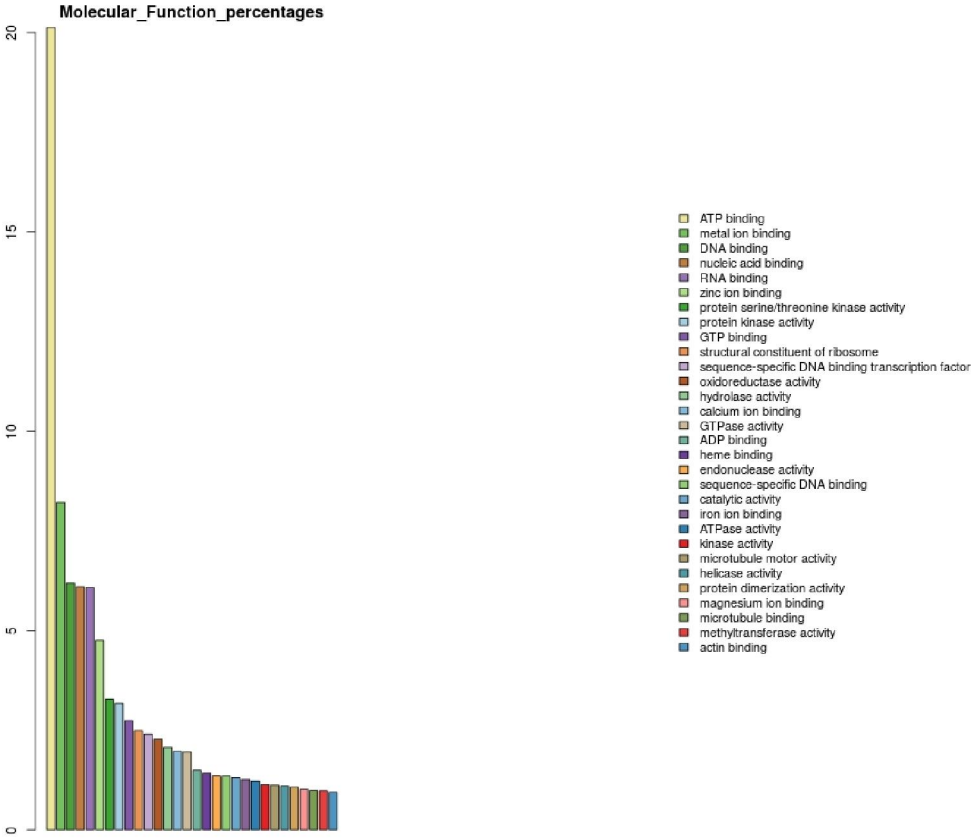


Figure: Top conserved domains for T. wallichiana

For *Ulmus wallichiana* the Annocript resulted in a total of 36476 hits in Swiss-Prot corresponding to 340 organisms, 51477 hits in TrEMBL corresponding to 619 organisms, 69989 hits in Conserved Domain databases and 579 hits in SILVA database. Functional annotation with GO database using Annocript yielded 2601 results in Biological Processes, 647 in Cellular components and 2041 in Molecular functions. Further, 60 results were obtained for pathway level 1, 214 in level 2 and 305 in level three pathway. 3273 sequences were observed to be long non-coding under 0.95 probability and 100 nucleotides of maximum length of ORF. A search in Pfam database using Annocript yielded 6621 hits. Below figures represent the annotation top hits for *Ulmus wallichiana* in various databases.


Biological_Process_percentages
0
1
2
3
4
5
6
transcription, DNA-dependent
regulation of transcription, DNA-dependent
translation
defense response
carbohydrate metabolic process
metabolic process
intracellular protein transport
signal transduction
DNA repair
protein transport
protein folding
microtubule-based movement
intracellular signal transduction
DNA integration
protein phosphorylation
transmembrane transport
RNA modification
vesicle-mediated transport
mRNA processing
cell wall organization
ubiquitin-dependent protein catabolic process
response to salt stress
cell division
transport
embryo development ending in seed dormancy
nuclear mRNA splicing, via spliceosome
cell redox homeostasis
response to oxidative stress
cell cycle
protein ubiquitination


Cellular_Component_percentages
0
5
10
15
20
25
30
35
integral to membrane
nucleus
cytoplasm
plasma membrane
cytosol
chloroplast
mitochondrion
membrane
intracellular
ribosome
plasmodesma
Golgi apparatus
nucleolus
Golgi membrane
endoplasmic reticulum membrane
microtubule
chloroplast stroma
endoplasmic reticulum
intracellular membrane-bounded organelle
vacuole
vacuolar membrane
extracellular region
chloroplast envelope
cell wall
apoplast
chloroplast thylakoid membrane
myosin complex
mitochondrial inner membrane
endosome
trans-Golgi network


Molecular_Function_percentages
0
5
10
15
20
ATP binding
metal ion binding
DNA binding
nucleic acid binding
RNA binding
zinc ion binding
protein serine/threonine kinase activity
protein kinase activity
GTP binding
structural constituent of ribosome
sequence-specific DNA binding transcription factor
oxidoreductase activity
hydrolase activity
calcium ion binding
GTPase activity
ADP binding
heme binding
endonuclease activity
sequence-specific DNA binding
catalytic activity
iron ion binding
ATPase activity
kinase activity
microtubule motor activity
helicase activity
protein dimerization activity
magnesium ion binding
microtubule binding
methyltransferase activity
actin binding

- **SSR marker development and primer designing**

SSRs are widely used markers for genetic studies because of the promising features like codominant nature, reproducibility, highly informative, cross-transferability between related species etc (Mason, 2015). Especially for wild tree species where SNP based approach is cumbersome due absence of reference genome for these non-model wild tree species and comparatively much input cost, these SSR markers are cost effective to decipher population and landscape genetic problems in such cases. *Taxus wallichiana* is an endangered species whose spread is constantly reducing due to human intervention and changing climatic scenario. Extensive population and landscape genetic information is required in order to plan conservative strategies for this species. For this reason we screened our transcripts for SSR identification and found that 6507 sequences contained 7041 SSRs out of which 534 were in compound form and 800 sequences contained more than 1 SSRs. Distribution of different repeat classes and abundance of identified motifs is shown in figure…These SSRs will prove useful for population and landscape genetic analysis of *Taxus wallichiana*. Among the dinucleotide repeats, TA/AT repeats were found in greater frequency (23.83%) followed by AG/GA (9.06%), CT/TC (8.3%), GT/TG (8.90%) and AC/CA (6.65%). GC/CG repeats were found to be in a very low frequency of 0.14%. We observed 60 different trinucleotide repeat motifs contained in 2618 SSRs among which TTC repeat comprise largest fraction of 2.98% followed by TCT with 2.20% and ATT with 2.07% frequency. A low percentage (5.4%) of tetra, penta and hexa nucleotide repeat motifs were observed in overall identified SSRs. CTGC, GGGAC and TTCCTC were among the dominant motifs in tetra, penta and hexa nucleotide repeat motifs. 7.5% of the SSRs were present in compound formation with a maximum of 100 bases between two SSRs. The largest SSR observed was in compound formation with AGG, TCC and CCT repeated 7 times with interrupting sequences between them. An overall density of I SSR/22.95 kb of the sequences was determined.

Positional distribution of the SSRs in the transcripts was analyzed by predicting the ORFs from the SSR containing sequences using orfPredictor followed by correlating the SSR start and end positions with the start and stop positions of the predicted ORFs. The sequences with SSRs in the coding sequence were functionally annotated. Differential polymorphism of SSRs present within 5'UTR, 3'UTR and CDS was analyzed using urea-PAGE on selected germplasms. 11 sequences were found without ORFs and among the remaining 6496 sequences with predicted ORFs, about 1781 (27.41%) sequences have SSR in their CDS region, 1711 (26.33%) sequences have SSR in 5`UTR while 2919 (44.93%) sequences have SSR in 3`UTR region and 85 (1.305) have SSR partially in CDS and partially in UTR regions. We used BatchPrimer3 v 2.0 () to design primer for the SSR markers. A total of 4958 primer pairs were successfully designed from the SSR containing sequences.

In case of *Ulmus wallichiana* we found that 14042 sequences contained 16570 SSRs out of which 1318 were in compound form and 2094 sequences contained more than 1 SSRs. Distribution of different repeat classes and abundance of identified motifs is shown in figure…These SSRs will prove useful for population and landscape genetic analysis of *Ulmus wallichiana*. Among the dinucleotide repeats, AG/CT

repeats were found in greater frequency (23.07%) followed by AC/GT (18.17%), AT/AT (13.20%), CG/CG (.46%). CG/CG repeats were found to be in a very low frequency. We observed 60 different trinucleotide repeat motifs contained in 6335 SSRs among which GAA repeat comprise largest fraction of 5.63% followed by GGC with 4.95% and TTC with 4.21% frequency. A low percentage (6.74%) of tetra, penta and hexa nucleotide repeat motifs were observed in overall identified SSRs. AAAC, AAAAT and AAAGAA were among the dominant motifs in tetra, penta and hexa nucleotide repeat motifs.