---

2018.10.23

For the manuscript, scaffold, gene, and transcript IDs were formalized as follows.

## Sequence names:

1. All scaffolds were sorted based on their length, from the longest to the shortest

2. The 19 longest scaffolds (minimum length > 8 Mbp) were named following the format: e.g., Pseudomolecule\_01 to Pseudomolecule\_19

3. All the rest scaffolds were named with the format: e.g., Scaffold\_Un001 to Scaffold\_Un917, where "Un" means "not anchored"

## Gene names:

Gene names follow the format of rice genome annotation: "Tg01g00010", where

"Tg" stands for "tectona grandis,

"01" stands for the Scaffold\_01,

"g" stands for "gene",

"00010" stands for gene number with increments of 10 between adjacent two genes along a scaffold

Note: this means the closer their physical distance, the closer the ID number is.

## Transcript names:

Transcripts have the same format of new gene names, with each isoform followed by ".t1", ".t2"........

Files in this directory:

------

1. teak\_tectona\_grandis\_26Jun2018\_7GlFM\_fmt\_tp.fa

fasta sequences of the assembly

------

2. All working gene models

teak.working\_models\_HiC.cdna\_con\_sorted.fa

* cDNA sequences of all isoforms of the working gene set

teak.working\_models\_HiC.cds\_con\_sorted.fa

* CDS of all isoforms of the working gene set

teak.working\_models\_HiC.pep\_con\_sorted.fa

* Peptide sequence of all isoforms of the working gene set

teak.working\_models\_HiC\_fmtDes\_con\_sorted.gff

* GFF of all isoforms of the working gene set

------

3. All high-confidence gene models

teak\_hc\_models\_HiC\_con\_sorted\_modiGeneID.gff

* GFF of all high-confidence gene models

teak\_hc\_models\_HiC.cdna\_con\_sorted\_modiGeneID.fa

* cDNA sequences of all high-confidence gene models

teak\_hc\_models\_HiC.cds\_con\_sorted\_modiGeneID.fa

* CDS of all high-confidence gene models

teak\_hc\_models\_HiC.pep\_con\_sorted\_modiGeneID.fa

* Peptide sequences of all high-confidence gene models

------

4. Representative high-confidence gene models

teak\_repr\_hc\_models\_HiC\_con\_sorted\_modiGeneID.gff

* GFF of representative high-confidence gene models

teak\_repr\_hc\_models\_HiC.cdna\_con\_sorted\_modiGeneID.fa

* cDNA sequences of representative high-confidence gene models

teak\_repr\_hc\_models\_HiC.cds\_con\_sorted\_modiGeneID.fa

* CDS of representative high-confidence gene models

teak\_repr\_hc\_models\_HiC.pep\_con\_sorted\_modiGeneID.fa

* Peptide sequences of representative high-confidence gene models

------

5. Expression abundances of the working gene set were estimated using cufflinks

RNAseq experiment atlas from NCBI SRA BioProject PRJNA287604

File: teak\_working\_gene\_fpkm\_matrix\_con\_sorted.txt