Analysis of the Mitragyna speciosa genome released by the Kratom Genome Project

Author: Bryan J. Jones Ph.D.[1]

FULL DISCLOSURE & TRANSPARENCY STATEMENT

The below analysis is my own, not that of the Kratom Genome Project. Funding was provided by the Kratom Genome Project for the analysis, not for any particular conclusions. An advanced copy of this report was provided to the Kratom Genome Project, and clarifying edits were made upon their comments. My original pre-edit version is also included in the same folder, as well as the info I was asked to clarify. All conclusions are my own independent findings based on the analysis performed.

SUMMARY

This independent verification of the genome released by the Kratom Genome Project (kratomdna.org) has verified that it is indeed *Mitragyna speciosa* (i.e. kratom). The raw data contains 21 million pairs of reads of high quality with a total of 6 billion bases read, ten times more data than previously available. The reads have a 37% GC content. The assembled partial genome is 301 million bases long, which is typical for a plant in this order. The genome has ~20X average coverage, which is decent but less than ideal, which means it will require more work to fill in the remaining gaps. The sequence of all annotated *M. speciosa* genes in NCBI databases, including ITS1 and ITS2, were identified in this genome (mostly 95%-100% identical) verifying the identity as *M. speciosa*.

RAW DATA (FASTQ FILES)

Raw sequencing data reported to be from *Mitragyna speciosa* (i.e. kratom) was retrieved from the Kratom Genome Project and independently analyzed for quality and overall characteristics. This data set contains a total of over 6 billion base reads. The fastq files show 21 million pairs of reads. Read lengths are 151 bp, with a few shorter reads. The Kratom Genome Project data set is approximately ten times larger than a previous sequencing data set (SRX2855178) available from NCBI (National Center for Biotechnology Information) which contains 659 million bases. The previous data set comes from The U.S. Food and Drug Administration's whole chloroplast genome sequencing effort, and was submitted in 2016.

FastQC analysis shows the reads from the Kratom Genome Project's data set are of high overall quality. The reads have an overall GC content of 37%, which is typical for plants in the Gentianales order. *Asclepias syriaca* (milkweed) has 36.9% GC, the medicinal plant *Rhazya stricta* has 32.9%, and *Catharanthus roseus* (madagascar periwinkle) has 33.6%. The analysis does show some per base sequence bias early in the reads due to overabundance of a few specific

[1] The Biotechnology Institute, University of Minnesota, 1479 Gortner Avenue, Saint Paul, Minnesota 55108, United States. Email: Bry@nJJon.es

kmers. This result is an expected bias due to the transposon technique used to create the sequencing library, and should not affect sequencing quality.

ASSEMBLY

The size of the assembled genome is 301,288,540 bases across 335,915 contigs. With the 6 billion bases read, this gives about 20X average coverage across genomic and nongenomic (chloroplast & mitochondria) DNA. 73.6% of the assembled genome has at least 5X coverage, while only 33.4% has 10X or greater coverage. This should be enough to further assemble into scaffolds, but is less than ideal. Typically, high quality completed genomes using this type of sequencing (MiSeq) will have at least 20-30 times coverage, more typically 50-100 times coverage (Ajay, 2011).

This is a typical genome size for plants in the Gentianales order. *A. syriaca* genome is 237 Mbp, *R. stricta* is 274 Mbp, and *C. roseus* is 523 Mbp. The 301 Mbp partial genome is the approximate size of the complete genome, the actual genome may be slightly bigger or smaller. Filling in gaps between contigs would increase this size, but some contigs or parts of contigs may be overlapping/redundant, which would reduce this number.

COMPARISON TO EXISTING SEQUENCES

ITS sequences from NCBI verify the identity of this genome as *M. speciosa*. A blastn search on a database created from this genome perfectly mapped the sequence containing ITS1, the 5.8S ribosomal RNA, and ITS2 (AB249645.1) to contig 56 at positions 8051-7444.

The largest annotated set of *M. speciosa* sequence data available from NCBI databases is the chloroplast genome. A blastn search found the sequence from the 155,600 bp plastid (ACCESSION:NC_034698.1) across 29 contigs from the new assembly: 125; 146; 275; 310; 368; 455; 661; 1,454; 1,549; 7,863; 25,398; 29,184; 71,593; 71,601; 74,814; 85,601;101,342; 108,704; 135,019; 150,624; 156,507; 158,438; 188,440; 191,639; 202,841; 206,287; 230,654; 266,741; and 306,777. Overall, the plastid sequence matched 99.2% with the reference sequence.

Other previously sequenced and annotated genes from *M. speciosa* are present in this genome, further verifying the identity as *M. speciosa* and verifying that it is close to complete. There are 38 genes previously released (in addition to the plastid), some from sequenced mRNA, others were sequenced from genomic DNA. Blastn searches found all of these genes present (at least in part) in this genome. Most are 95-100% identical to reference sequences, none were below 88%. As expected, this genomic DNA also shows the presence of introns that are not present in the previously sequenced genes from mRNA. These genes include enzymes like strictosidine synthase that are in the synthesis the pathway of active alkaloid compounds mitragynine (Charoonratana, 2013).

One would expect high similarity with existing sequences, but the genome from a different individual plant should have some differences (SNPs). This is what the data shows. The data presented here is indeed a match to the limited existing data available for *Mitragyna speciosa.*

FILES ANALYZED

Mitragyna_speciosa_R1.fastq.gz
SHA256sum: 4a9d0b4716134c0ea869ec6cbb0a6a653f9ed3ef0243b399f985e122ea3427ce
Mitragyna_speciosa_R2.fastq.gz
SHA256sum: a4384c44b5597090c53d20ae480159b6569fdfc15f3235eda1b439a45a899e86
Assembly: mitragyna_400bp_paired_assembly.fa
SHA265sum 8e160bfd4cd36dfe96f2c3ca551c10ab53af4b295d732ec6c3c1ad3e8e390d66

REFERENCES

Ajay SS, Parker SC, Abaan HO, Fajardo KV, Margulies EH. Accurate and comprehensive sequencing of personal genomes. Genome research. 2011 Sep 1;21(9):1498-505.

Charoonratana T, Wungsintaweekul J, Pathompak P, Georgiev MI, Choi YH, Verpoorte R. Limitation of mitragynine biosynthesis in *Mitragyna speciosa* (Roxb.) Korth. through tryptamine availability. Zeitschrift für Naturforschung C. 2013 Oct 1;68(9-10):394-405.