README for Data Dryad repository files for:

Bornowski et al. (2020) Data from: Genome sequencing of four culinary herbs reveals terpenoid genes underlying chemodiversity in the Nepetoideae

Scaffold, gene, and transcript IDs were formatted as follows.

Scaffold names

* Scaffold names contain a numerical identifier, the starting and ending edge sequences, starting and ending vertices, Supernova output format version number, and output style
* Example:
  + >146 edges=2976109..2904591 left=5173 right=5182 ver=1.10 style=3

Gene and transcript names

* Gene names contain a numerical identifier, cluster identifier assigned by Trinity, and transcript isoform number concatenated by a ".t"
* Example:
  + >519.g1.t1
* Genes and transcripts joined by PASA2 have names concatenated by "\_"
* Example:
  + >456:1023.g46.t1\_1023.g47.t1

Files in this directory:

This directory contains files for each species' genome assembly and annotation files, where "\*\*\*" refers to a unique 3-letter abbreviation for each species, as follows:

* bas = basil (*Ocimum basilicum*)
* maj = marjoram (*Origanum majorana*)
* vul = oregano (*Origanum vulgare*)
* ros = rosemary (*Rosmarinus officinalis*)

1. Genome assemblies

Genomes were sequenced using 10x Genomics linked read technology and assembled with Supernova.

\*\*\*\_aa.min\_10k\_final.fa

* fasta genome assembly with minimum 10k scaffold size and no N scaffolds

2. All working gene models

Working gene models contain all loci and isoforms generated from the annotation pipeline and may include artifacts such as partial gene models.

\*\*\*\_aa.working\_models.cdna.fa

* cDNA sequences of all isoforms of the working gene set

\*\*\*\_aa.working\_models.cds.fa

* CDS of all isoforms of the working gene set

\*\*\*\_aa.working\_models.gff3

* GFF of all isoforms of the working gene set

\*\*\*\_aa.working\_models.pep.fa

* Peptide sequence of all isoforms of the working gene set

\*\*\*\_aa.working\_models.pep.func\_anno.txt

* Functional annotation of the working gene set

3. All high confidence gene models

High confidence gene models are a subset of the working gene models that were not partial models, did not contain an internal stop codon, were not transposable element-related, and had a PFAM domain match or a TPM > 0.

\*\*\*\_aa.gene\_models.hc.cdna.fa

* cDNA sequences of all high confidence gene models

\*\*\*\_aa.gene\_models.hc.cds.fa

* CDS of all high confidence gene models

\*\*\*\_aa.gene\_models.hc.gff3

* GFF of all high confidence gene models

\*\*\*\_aa.gene\_models.hc.pep.fa

* Peptide sequences of all high confidence gene models

\*\*\* \_aa.gene\_models.hc.pep.func\_anno.txt

* Functional annotation of the high confidence gene set

4. High confidence representative models

High confidence representative gene models are a subset of the high confidence gene model set. Each representative gene model is the isoform with the longest CDS at each locus.

\*\*\*\_aa.gene\_models.hc.repr.gff3

* GFF of high confidence representative gene models

\*\*\*\_aa.gene\_models.hc.repr.pep.fa

* Peptide sequences of high confidence representative gene models

5. FPKM matrices

Expression abundances of the high confidence and working gene sets were estimated using Cufflinks with RNA-Seq data generated from this study (NCBI SRA BioProject PRJNA592145).

\*\*\*\_Cufflinks\_FPKMs\_functions\_all\_hc\_genes.txt

* Expression abundances of the high confidence gene model set

\*\*\*\_Cufflinks\_FPKMs\_functions\_working\_genes.txt

* Expression abundances of the working gene model set