

# Computing Oak Divergence Using MUMMER Alignments

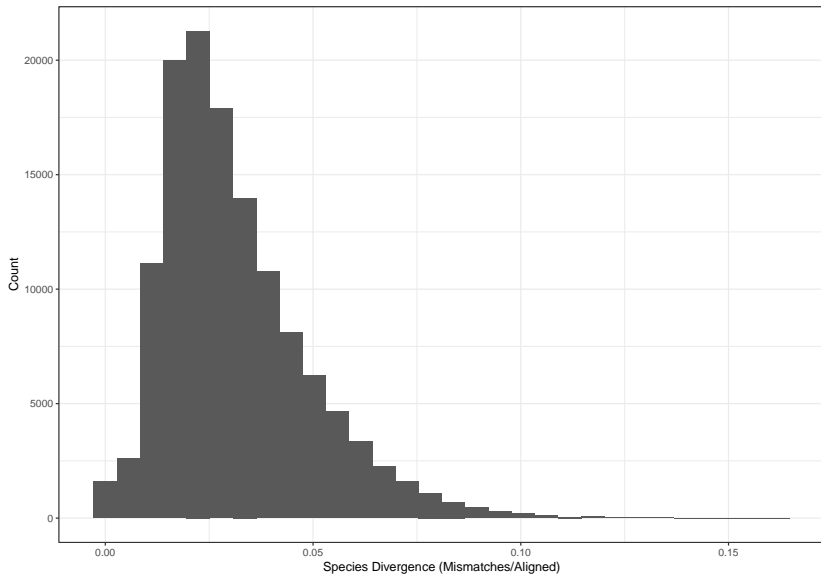
Jesse Garcia

10/18/2018

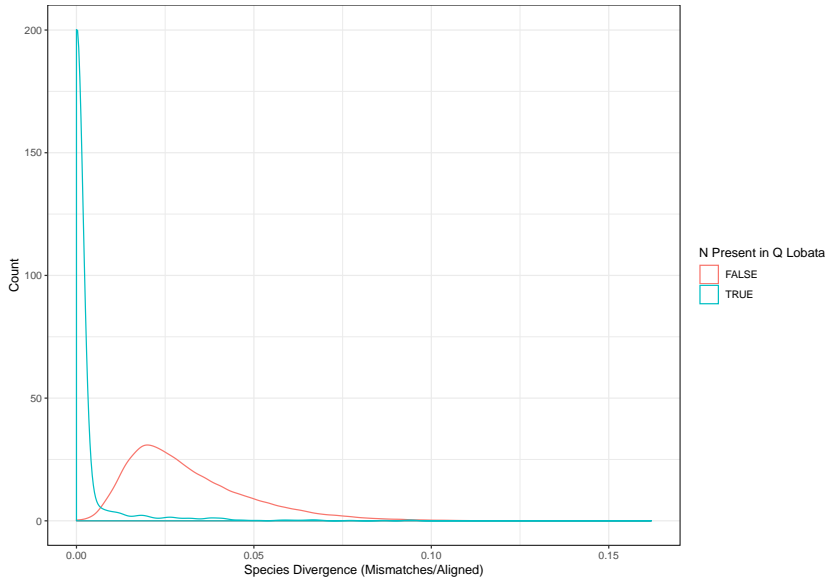
# Processing of Data

1. Used .coords file and Sorel's "chromosome name to scaffold and orientation table" to convert 2017 Q Lobata alignment coordinates to 2018 Q Lobata alignment coordinates
2. Removed all alignments that weren't on chromosomes
3. Used remaining alignment coordinates to extract nucleotide sequences from Q Robur and Q Lobata
4. Nucleotide sequences were differing lengths because of indels, so globally aligned sequence pairs using NW algorithm. Insertions were indicated with "\_". Now Sequences are the same length and have 1-1 alignment.
5. Currently 128,513 Alignment pairs out of 129,162 finished computing (99.5%). 30 Pairs were too large for NW align.

There's some variability of divergence across alignment coordinates



I think runs of N's contribute to this



Aligned N's Count as matches. Only last 6 bases are nucleotides.

[illegible][illegible]

```
## [1] 724
```

## NNNCAAATCT

## Divergence Estimates by N

Total Divergence for everything: 0.0275034

Table 1: Table continues below

NPresent	Divergence	Alignments	TotalLength	NumberOfMatches
FALSE	0.02758	126917	498540988	484790888
TRUE	0.002726	1596	1554477	1550239

---

NumberOfMismatches

---

13750100

4238

---

## Next Steps

1. Remove Repeats, SNPs near Runs of Ns, and CDS.
2. Sorel also has GATK called variants where Q Robur and Q Lobata were aligned to a reference genome. I think Computing divergence for this should be easier because scripts exist.