**Anchored Phylogenomics Methods**

**Paired-Read Merging** (Merge.java) **--** Typically, 20% of sequenced library fragments had an insert size between 125bp and 175bp. Since 100bp paired-end sequencing was performed, this means that the majority of the paired reads did not overlap. The overlapping reads were identified and merged following the methods of Rokyta et al. 2012. In short, for each degree of overlap for each read we computed the probability of obtaining the observed number of matches by chance, and selected degree of overlap that produced the lowest probability, with a p-value less than 10-10 required to merge reads. When reads are merged, mismatches are reconciled using base-specific quality scores, which were combine to form the new quality scores for the merged read (see Rokyta et al. 2012 for details). Reads failing to meet the probability criterion were kept separate but still used in the assembly. The merging process produces three files one containing merged reads and two containing the unmerged reads.

**Assembly** (Assembler.java) **--** The reads were assembled into contigs using an assembler that makes use of both a divergent reference assembly approach to map reads to the probe regions and a de-novo assembly approach to extend the assembly into the flanks. The reference assembler uses a library of spaced 20-mers derived from the conserved sites of the alignments used during probe design (20 mers were derived from Anolis sequences). A preliminary match was called if at least 17 of 20 matches exist between a spaced kmer and the corresponding positions in a read. Reads obtaining a preliminary match were then compared to an appropriate reference sequence used for probe design to determine the maximum number of matches out of 100 consecutive bases (all possible gap-free alignments between the read and the reference ware considered). The read was considered mapped to the given locus if at least 55 matches were found. Once a read is mapped, an approximate alignment position was estimated using the position of the spaced 20-mer, and all 60-mers existing in the read were stored in a hash table used by the de-novo assembler. The de-novo assembler identifies exact matches between a read and one of the 60-mers found in the hash table. Simultaneously using the two levels of assembly described above, the three read files were traversed repeatedly until an entire pass through the reads produced no additional mapped reads.

For each locus, mapped reads were then clustered into clusters using 60-mer pairs observed in the reads mapped to that locus. In short, a list of all 60mers found in the mapped reads was compiled, the 60-mers were clustered if found together in at least two reads. The 60-mer clusters were then used to separate the reads into clusters for contig estimation. Relative alignment positions of reads within each cluster were then refined in order to increase the agreement across the reads. Up to one gap was also inserted per read if needed to improve the alignment. Note that given sufficient coverage and an absence of contamination, each single-copy locus should produce a single assembly cluster. Low coverage (leading to a break in the assembly), contamination, and gene duplication, can all lead to an increased number of assembly clusters. A whole-genome duplication, for example would increase the number of clusters to two per locus.

Consensus bases were called from assembly clusters as follows. For each site an unambiguous base was called if the bases present were identical or if the polymorphism of that site could be explained as sequencing error, assuming a binomial probability model with the probability of error equal to 0.1 and alpha equal to 0.05. If the polymorphism could not be explained as sequencing error, the ambiguous base was called that corresponded to all of the observed bases at that site (e.g. 'R' was used if 'A' and 'G' were observed). Called bases were soft-masked (made lowercase) for sites with coverage lower than 5. A summary of the assembly results is presented in Supplemental Table X (Pyron-etal\_AssemblySummary.xlsx).

**Contamination Filtering** (IdentifyGoodSeqsViaReadsMapped.r, GatherALLConSeqsWithOKCoverage.java) --In order to filter out possible low-level contaminants, consensus sequences derived from very low coverage assembly clusters (<100 reads) were removed from further analysis. After filtering, consensus sequences were grouped by locus (across individuals) in order to produce sets of homologs.

**Orthology** (GetPairwiseDistanceMeasures.java, plotMDS5.r) -- Orthology was then determined for each locus as follows. First, a pairwise distance measure was computed for pairs of homologs. To compute the pairwise distance between two sequences, we computed the percent of 20-mers observed in the two sequences that were found in both sequences. Note that the list of 20-mers was constructed from consecutive 20-mers as well as spaced 20-mers (every third base), in order to allow increased levels of sequence divergence. Using the distance matrix, we clustered the sequences using a Neighbor-Joining algorithm, but allowing at most one sequence per species to be in a given cluster. Clusters containing fewer than 30 sequences were removed from downstream processing.

**Alignment** (MAFFT) -- Sequences in each orthologous set were aligned using MAFFT v7.023b (Katoh, 2013), with --genafpair and --maxiterate 1000 flags utilized.

**Alignment Trimming** (TrimAndMaskRawAlignments.java) -- The alignment for each locus was then trimmed/masked using the following procedure. First, each alignment site was identified as "good" if the most common character observed was present in 50% of the sequences. Second, 20 bp regions of each sequence that contained < 14 good sites were masked. Third, sites with more than 22 masked or missing bases were removed from the alignment.

**References:**

Lemmon, A. R., S. Emme and E. M. Lemmon. 2012. Anchored hybrid enrichment for massively high-throughput phylogenetics. **Systematic Biology**. 61: 721-744.

Prum, R. O, J. S. Berv, A. Dornburg, D. J. Field, J. P. Townsend, E. C. Lemmon, and A. R. Lemmon. 2015. A Fully Resolved, Comprehensive Phylogeny of Birds (Aves) using Targeted Next Generation DNA Sequencing. **Nature**. In Press.

Rokyta, D. R., A. R. Lemmon, M. J. Margres, and K. Arnow. 2012. The venom-gland transcriptome of the eastern diamondback rattlesnake (*Crotalus adamanteus*). **BMC Genomics**. 13:312.

Katoh, Standley 2013 ([*Molecular Biology and Evolution* **30**:772-780](http://mbe.oxfordjournals.org/content/30/4/772))

MAFFT multiple sequence alignment software version 7: improvements in performance and usability.

(outlines version 7)