PHYLOGENETIC APPROACH FOR DDRAD-SEQ READS

Jeronymo Dalapicolla

For the phylogenetic approach, we performed the *ipyrad* pipeline (http://ipyrad.readthedocs.io/) in seven steps. In the first step the raw data was demultiplexed in individuals according to the barcode list I provided, without mismatches in barcodes (*max_barcode_mismatch*). In the second step I edited the reads, eliminating barcodes, adapters and the reads with more than 5 pb with low quality Q>20 (*max_low_qual_bases*) and with low quality scores for Illumina, Phed below 33 (*phred_Qscore_offset*). We allowed reads of variable sizes with 110 pb of minimum size (*filter_min_trim_len*). We grouped the reads as homologous in the third step if they presented a similarity ≥ 90% (*clust_threshold*), and the clusters were *de novo* aligned with Muscle (Edgar 2004). In the step 4, the heterozygosity indices and the sequencing error rates were calculated and used for the fifth step, the consensus step. A consensus sequence was created for each allele from the aligned reads, considering the values of the parameters calculated in the fourth step. In the fifth step the data was also filtered, allowing up to 5 Ns (uncalled bases) per allele consensus sequences (*max_Ns_consens*), and the number of alleles per locus was calculated. In step 6 the consensus sequences are aligned again with Muscle (Edgar 2004) using the parameters of step 3, and in step 7 the outputs were created for the subsequent analyzes with some filters: maximum of 6 indels per locus (*max_Indels_locus*); maximum of 50% of heterozygous sites per locus (*max_shared_Hs_locus*), and up to 2 unique alleles were allowed in an individual (*max_alleles_consens*); the we chose an amount of missing data values per loci (*min_samples_locus*), see Material and Methods to details about this number.

REFERENCES

Edgar, R.C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research* 32, 1792–7.