SESAME Supplementary Materials

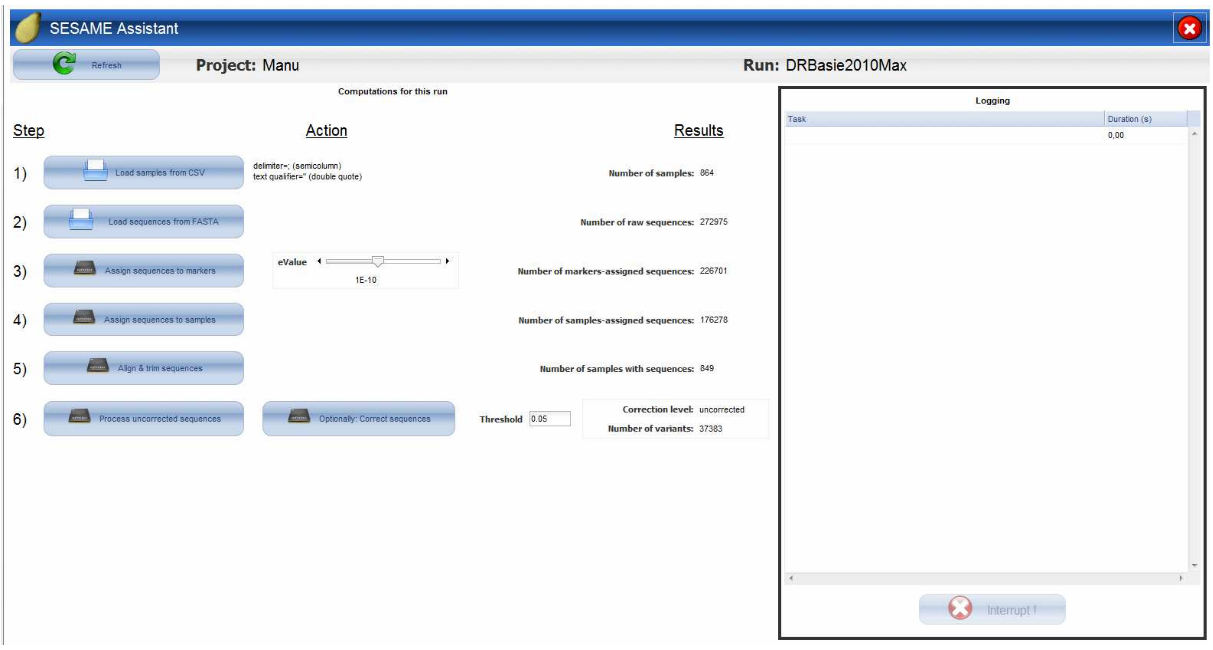
A ZIP file with a complete 454 dataset is found as a supplementary file

This test data set is composed of 864 samples from various rodent species. Each individual is identified uniquely by the combination of the reverse and forward tags as in Galan et al. 2010 (BMC Genomics) and the number of assigned reads per sample varied from 0 to 388. Only one marker was analysed for a 172 bp fragment of the exon 2 of the MHC class II gene DRB. A 454 run was performed on 2/8 of one GS FLX 454 sequencing plate.

Variant correction

Variant correction is based on the position-specific error rates that are calculated as 1 minus the proportion of the most frequent base in the standard sample on each alignment position. Then for real samples, the frequency of each base is calculated at each position. If the sum of proportions of minority bases (all but the most frequent) at a given position of an amplicon is less than the estimated error rate, and also less then 0.05 (default value) then minority bases are corrected to the most frequent base for this position. This procedure aims to reduce the sequencing noise, but avoids discarding weakly amplified alleles.

Assistant screenshot



Project: Manu **Run:** DRBasie2010Max

Computations for this run

Step	Action		Results
1)	Load samples from CSV	delimiter=; (semicolumn) text qualifier=" (double quote)	Number of samples: 864
2)	Load sequences from FASTA		Number of raw sequences: 272975
3)	Assign sequences to markers	eValue 1E-10	Number of markers-assigned sequences: 226701
4)	Assign sequences to samples		Number of samples-assigned sequences: 176278
5)	Align & trim sequences		Number of samples with sequences: 849
6)	Process uncorrected sequences	Optionally: Correct sequences Threshold 0.05	Correction level: uncorrected Number of variants: 37383

Task	Duration (s)
	0,00

Fig. 1. Screenshot of the Assistant.

Step 1: Amplicon information is read from a comma-separated values (CSV) file that contains sample name, marker name, “ploidy level” (in fact, the expected number of copies), primer and tag sequences, status (sample or standard), population name and species for each sample (individual - locus combination).

Step 2: DNA sequences are read from one single FASTA file.

Step 3: DNA sequences are BLASTed (with user adjustable e-value) against the reference sequences (previously loaded in the marker tab of the interface) to assign each read to a locus and determine read orientation.

Step 4: Reads are BLASTed against the list of tag sequences concatenated with primer sequences for sample assignation. Only reads with perfect match of both tags are assigned to sample, and thus all assigned reads cover the whole amplicon.
Step 5: For each amplicon, all reads and their reference sequence are aligned by MUSCLE or MAFFT. Based on this alignment, tags and primers are trimmed off.
Step 6: An optional position-specific noise correction of reads can be done if standard samples are provided by the user (typically a single cloned allele amplified and sequenced along with the samples).

When running, informations on the ongoing analyses are printed on the right side of the window.