phasing

March 14, 2018

1 Phasing allele sequences

After you mapped your reads against the reference library during the reference-based assembly step, you are ready to phase your reads into the two different allele sequences (in case of diploid organisms). This step is simple to execute, since the function only requires the path to the reference-based assembly output and the user-set minimal read depth for generating the consensus sequence:

```
In [1]: %%bash
        source activate secapr_env
        secapr phase_alleles -h
```

```
usage: secapr phase_alleles [-h] --input INPUT --output OUTPUT
                            [--min_coverage MIN_COVERAGE]

Phase remapped reads form reference-based assembly into two separate alleles.
Then produce consensus sequence for each allele.

optional arguments:
  -h, --help            show this help message and exit
  --input INPUT         Call the folder that contains the results of the
                        reference based assembly (output of reference_assembly
                        function, containing the bam-files).
  --output OUTPUT       The output directory where results will be safed.
  --min_coverage MIN_COVERAGE
                        Set the minimum read coverage. Only positions that are
                        covered by this number of reads will be called in the
                        consensus sequence, otherwise the program will add an
                        ambiguity at this position.
```

We can run the command simply like this:

```
secapr phase_alleles --input ../../data/processed/remapped_reads/ --output ../.
```

We can also choose to phase only the selected loci that were produced with the `secapr locus_selection` function:

```
secapr phase_alleles --input ../../data/processed/selected_loci --output ../../data
```

1.1 Producing allele alignments

Now all we need to do is to run the `secapr align_sequences` function in order to align the extracted allele sequences of all samples for each locus. We can run the command like this:

```
secapr align_sequences --sequences ../../data/processed/allele_sequences/joined_all
```

Or like this if we want to instead build allele alignments from only the selected loci:

```
secapr align_sequences --sequences ../../data/processed/allele_sequences_selected_l
```

1.1.1 Adding missing sequences

Before using these alignments for phylogenetic analyses it usually is a good idea to make sure that all taxa contain the same number of sequences. As of right now, some alignments may be missing one of the two allele sequences for some samples, because not enough reads were present that were supporting both haplotypes (controlled by the `--min_coverage` flag int he `phase_alleles` command. In order to add missing sequences as dummy sequences containing n's we can use the `secapr add_missing_sequences` function:

```
secapr add_missing_sequences --input ../../data/processed/alignments/selected_loci_
```

Previous page