Bioinformatics Testing for NuProbe Scientist I

04/05/2021

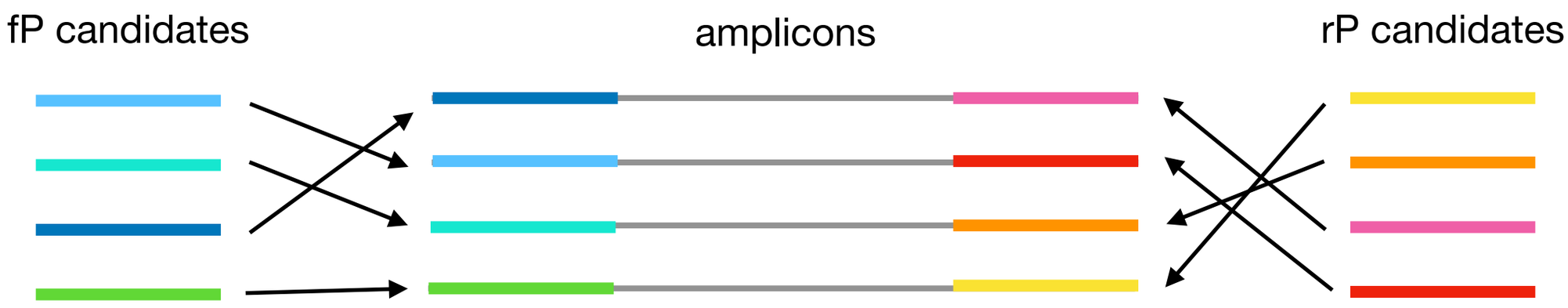
Task: find out primers causing non-specific amplification

Background:

When aligning PCR amplicons to reference sequences, the alignment software (e.g., Bowtie2) might fail to align some of the amplicons. Those amplicons are defined as “non-specific amplicons”, produced by non-specific amplification. In a multiplex PCR-based NGS panel, non-specific amplification would suppress on-target amplicons, resulting in lower uniformity and higher sequencing cost. Thus, it is important to find out the primers causing non-specific amplification.

Description:

In the attached bam file (input.bam), each sequence (or amplicon) is produced by a pair of primers: a forward primer (fP) and a reverse primer (rP). But we are only interested in the amplicons that are not aligned. ***For each unaligned amplicon, we need to find out which fP and rP produced it.*** A list of fP and rP candidates is provided (primers.xlsx). There is no adapter on either end of the amplicons, which means fP should start at the very beginning and rP should end at the very end of each amplicon.



Goals & requirements:

1. Group the unaligned amplicons by their sequences. For each unique sequence, find out the corresponding primer pair. You may refer to the primers by their index (shown in primers.xlsx). The example output file is output_NS_list.xlsx.
2. Calculate the number of unaligned amplicons under each primer pair. The example output file is output_NS_pairs.xlsx. You may also plot a heat map to show this visually (output_NS_pairs_log10.svg), but plotting is optional.
3. Python or Matlab is the preferred. If you would like to use coding languages other than those, please also include a sketch of your algorithm. Evaluation is based on the accuracy of the outputs, the efficiency of your algorithm and the elegance of your code. Please make your code user friendly and keep proper comments and documentations. Usage of Python packages (e.g., pysam, pandas) is not limited.

Hints:

1. For some unaligned amplicons, you might not be able to find the primer pair. In the example outputs, about 95% of the unaligned amplicons are assigned with a primer pair.
2. You don’t have to follow the formats of the output files.