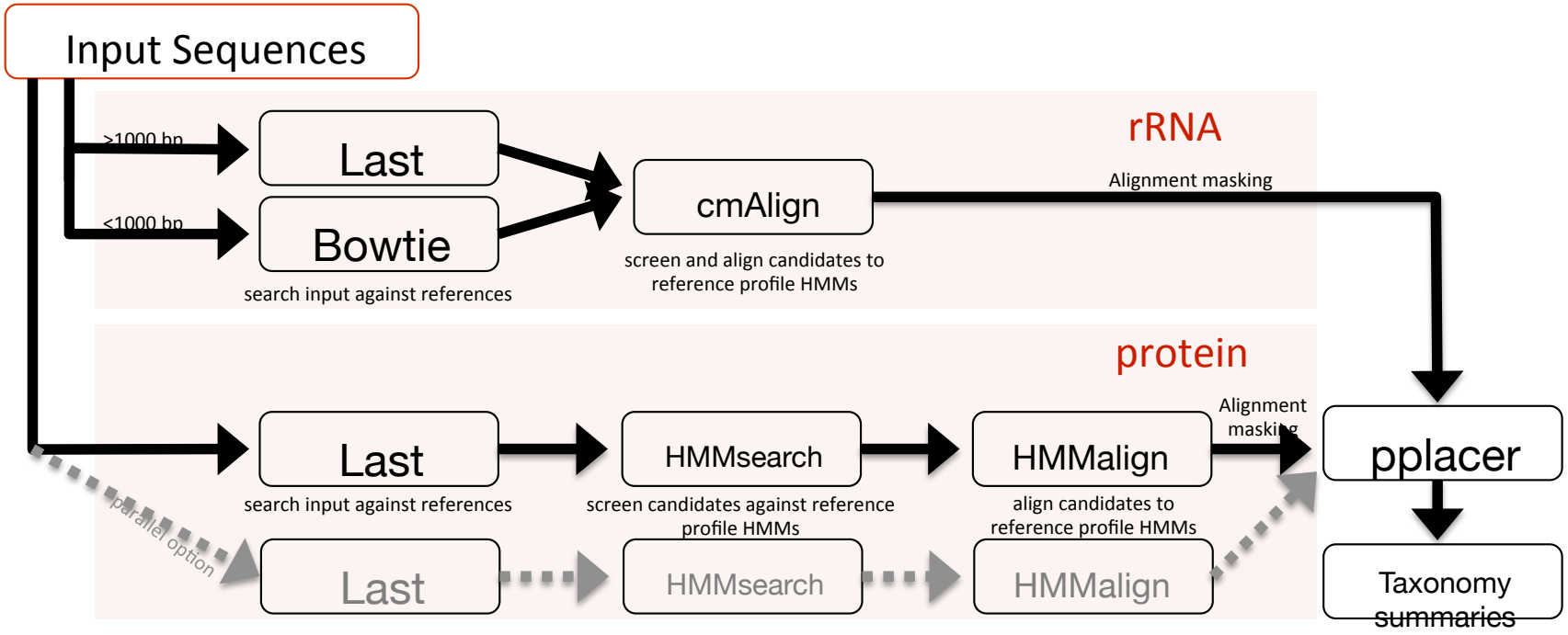
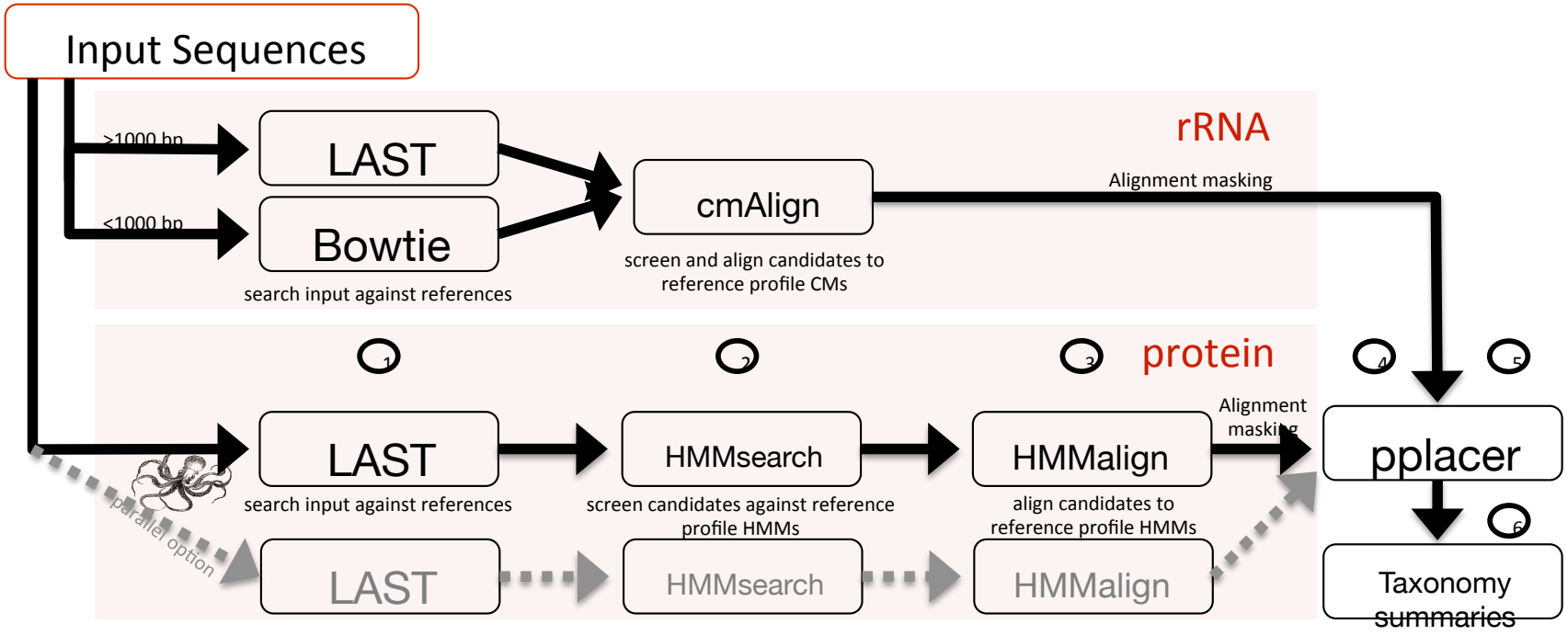

Input Sequences
>1000 bp
<1000 bp
Last
Bowtie
search input against references
cmAlign
screen and align candidates to reference profile HMMs
rRNA
Alignment masking
protein
Last
search input against references
HMMsearch
screen candidates against reference profile HMMs
HMMalign
align candidates to reference profile HMMs
Alignment masking
pplacer
parallel option
Last
HMMsearch
HMMalign
Taxonomy summaries



1. Input sequences are compared to unaligned reference marker sequences using adaptive seed searches implemented in the LAST algorithm (Kiełbasa et al. 2011). The LAST algorithm is ideal for mining rRNA and protein coding markers because it allows fast, sensitive searches of extremely large datasets (e.g. Illumina) and can additionally handle frameshift mutations and interpret quality information from Illumina FASTQ files.

2. Candidate marker sequences identified in LAST searches are next screened against profile alignments that have been precomputed for reference marker genes (housed in the local directory share/phylosift/markers). In order to take a stringent search approach towards short read data, PhyloSift relies on threshold e-values to accept or reject candidate sequences after initial LAST searches. For rRNA sequences, screening and alignment rely on Covariance Model profiles and are both carried out within one step (via the cmAlign algorithm in the SSU-align software, using an e-value of $1x10^{-6}$ for sequnces >1000 bp and $1x10^{-20}$ for sequences <1000 bp). Protein coding genes rely on profiles using Hidden Markov Models (via the HMMer software suite), with a threshold e-value set at 10.

3. Candidate protein-coding sequences that pass the HMMsearch filtering step are next aligned to reference profile HMMs for each marker gene.

4. Alignments are then masked (PhyloSift discards lowercase and . characters from alignments of candidate sequences), and masked sequences which pass a minimum threshold (alignments containing >20 positions) are passed on to pplacer for phylogenetic placement.

5. Marker gene alignments are concatenated and fed into pplacer, where candidate input sequences are placed onto a reference guide tree.

6. Tree placements are summarized in output files containing raw sequence placement information (sequence_taxa files) and taxon abundance information (taxa_summary files)