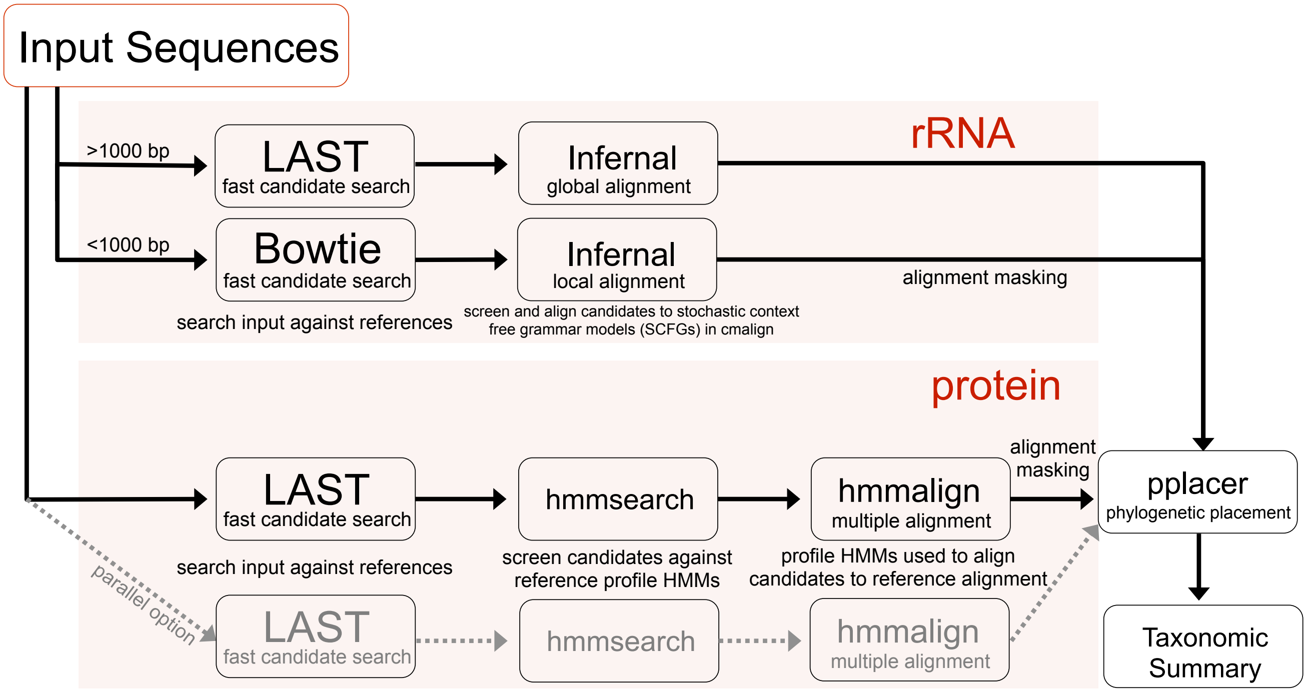
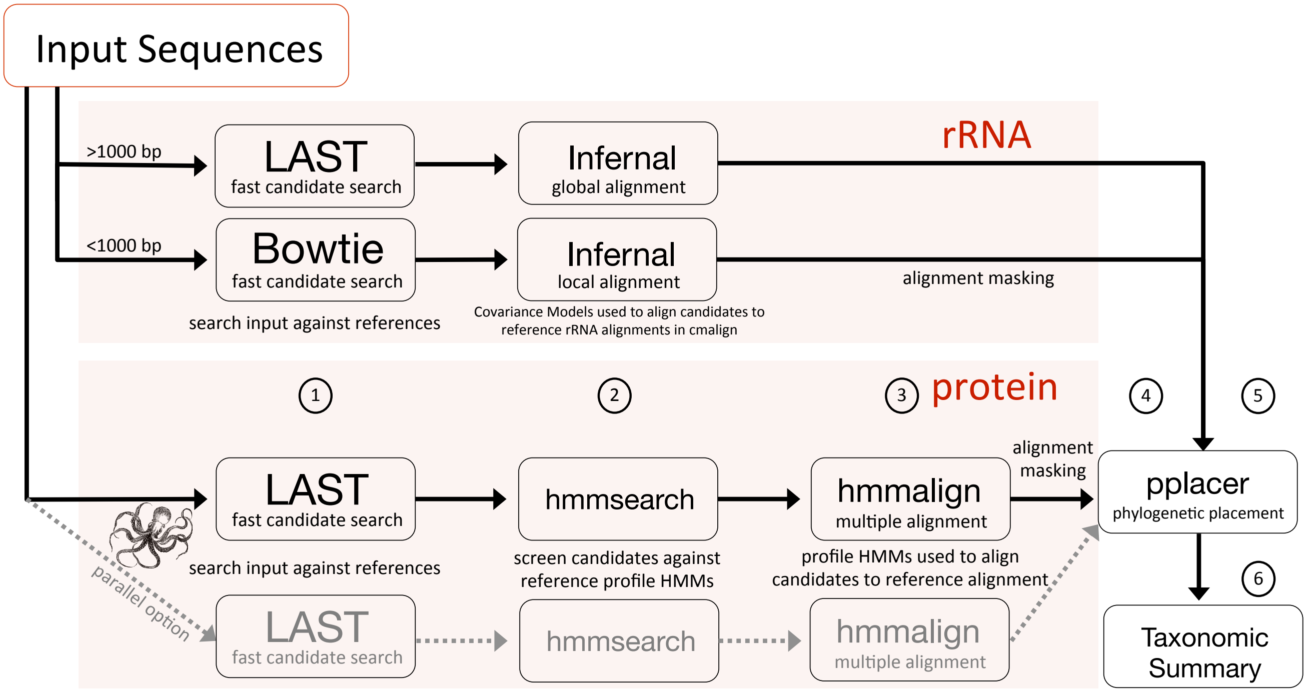

Input Sequences
rRNA
>1000 bp
LAST
fast candidate search
Infernal
global alignment
<1000 bp
Bowtie
fast candidate search
Infernal
local alignment
alignment masking
search input against references
screen and align candidates to stochastic context free grammar models (SCFGs) in cmalign
protein
alignment masking
LAST
fast candidate search
hmmsearch
hmmalign
multiple alignment
pplacer
phylogenetic placement
search input against references
screen candidates against reference profile HMMs
profile HMMs used to align candidates to reference alignment
parallel option
LAST
fast candidate search
hmmsearch
hmmalign
multiple alignment
Taxonomic Summary



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 pre-computed 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 relies on Covariance Model profiles (CMs; a class of Stochastic Context Free Grammar Models that utilize stem/loop information in rRNA secondary structure) and is carried out via the cmalign algorithm in the SSU-align software, using probability thresholds of $1x10^{-6}$ for sequences >1000 bp and $1x10^{-20}$ for sequences <1000 bp. Protein coding genes rely on profile Hidden Markov Models (computed via the HMMer software suite; Eddy 2010), with a threshold e-value set at 10. These profile alignments can be found in the /share/phylosift/markers/ directory as *.cm (rRNA) and *.hmm (protein coding genes) files.

3. Candidate protein-coding sequences that pass the hmmsearch filtering step must now undergo multiple alignment. For each marker gene, profile HMMs are used to align candidate input sequences to reference marker alignments. Alignments are subsequently trimmed by both the HMMer software package and PhyloSift scripts.

4. Alignments are then masked (PhyloSift discards lowercase and . characters following profile alignments of candidate sequences), and masked sequences which pass a minimum threshold (alignments containing >20 positions) are passed on 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)