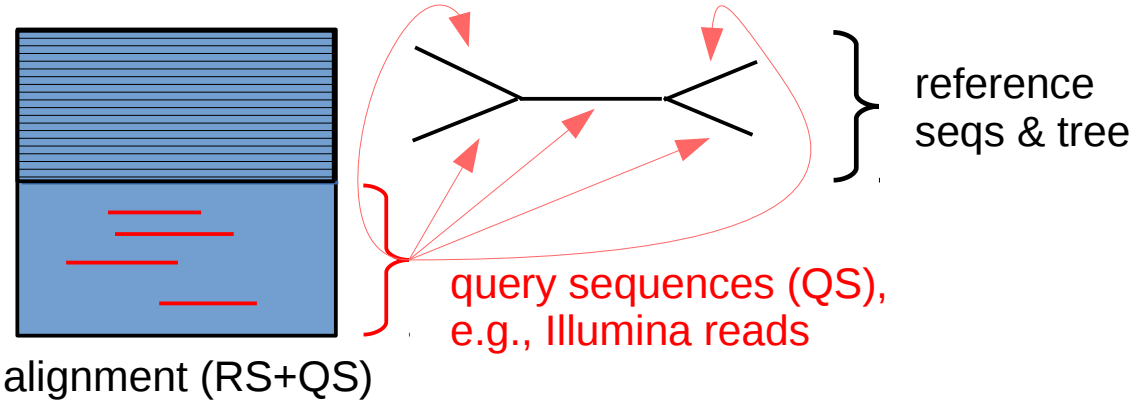
RAxML: phylogenetic inference

- Tools for large-scale ML and Bayesian phylogenetic inference
  - RAxML (Stamatakis 2006,2014) → "classic"
  - ExaML (Stamatakis & Aberer 2013, Kozlov et al 2015) → "phylogenomics"
  - RAxML-NG (Kozlov et al, manuscript in prep.) → "all-in-one" & 2x-3x faster
  - ExaBayes (Aberer et al 2014) → Bayesian inference
- Focus on high performance
  - Low-level optimization of likelihood kernels, efficient parallelization & load balancing, checkpointing ...
- Scalability
  - Good on phylogenomic datasets: 50 taxa x 300M sites → ~2 hours @ 4096 cores
  - Challenging on single-gene datasets: ~500K *SSU rRNA* → ~2 weeks @ 32 cores
  - We have ideas how to improve the latter → planned for 2018/19

EPA: Evolutionary placement

- Idea

  Place anonymous query sequences (QS) onto an annotated phylogenetic tree → reference tree



- Implementations
  - RAxML-EPA → part of standard RAxML (Berger et al 2011)
  - EPA-NG → a new, dedicated, more efficient version (Barbera et al, in prep.)
- Scalability
  - Millions of queries →feasible even with old RAxML-EPA (Mahe et al 2017)
  - Large reference tree (100K+ seqs) → challenging, but should improve with EPA-NG (evaluation in progress)

SATIVA: Mislabel identification

- Idea
  - Semi-automatic detection of putatively mislabeled sequences (Kozlov et al 2016)
  - Use RAxML and EPA to find sequences whose taxonomic annotations are in conflict with their placement in the ML phylogenetic tree
- Evaluation
  - Simulated and empirical *SSU rRNA* sequences
  - Works pretty well for individual mislabels
  - Inherent taxonomy ↔ phylogeny conflicts are still problematic
- Scalability
  - Basically limited by RAxML and EPA → will improve with the new implementations
  - Largest dataset analyzed: ~500K sequences → several weeks wall-time on a cluster with up to 4160 cores

Other tools & projects

- PUmPER
  - Extend an existing phylogenetic tree by incrementally adding new GenBank sequences (Izquierdo-Carrasco 2014)
  - Stephen will probably know better
- mPTP
  - Species delimitation with (multi-rate) Poisson Tree Processes (Kapli et al 2016, Zhang et al 2013)
  - Very fast :)
- UniEuk project
  - Similar(?) goals, but focus on protists
  - http://unieuk.org/

Software availability

- ExaML: https://github.com/stamatak/ExaML
- RAxML-NG: https://github.com/amkozlov/raxml-ng
- ExaBayes: https://github.com/aberer/exabayes
- EPA-NG: https://github.com/Pbdas/epa-ng
- SATIVA: https://github.com/amkozlov/sativa
- MPTP: https://github.com/Pas-Kapli/mptp