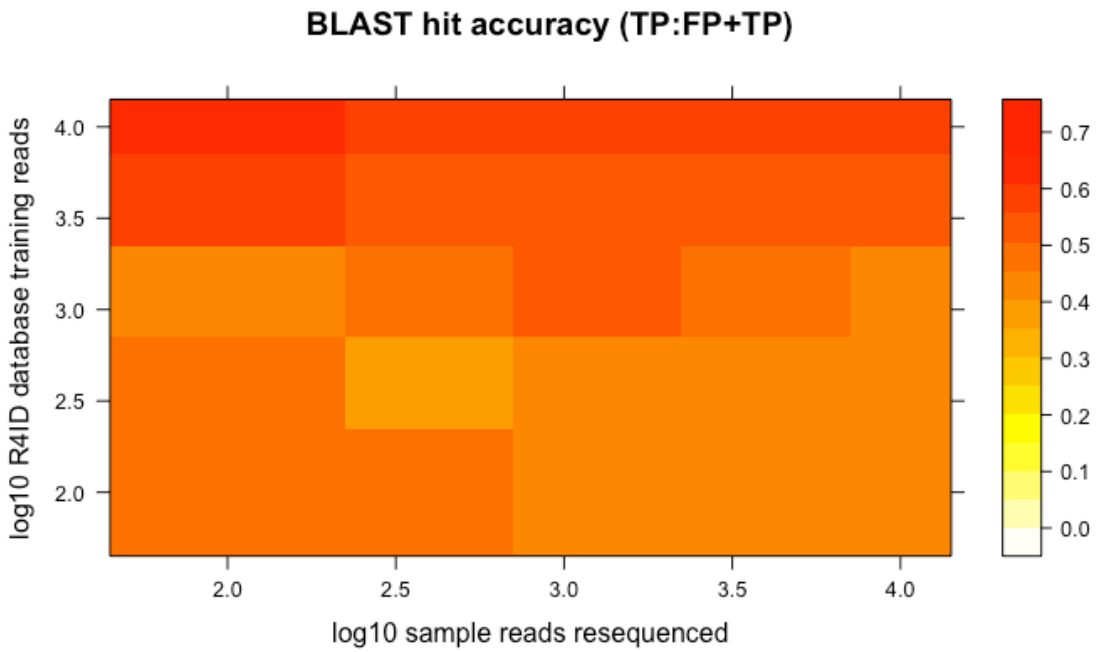
Rapid, raw-read reference and identification (R4IDs): A flexible platform for rapid generic species ID using long-read sequencing technology.

Joe Parker[1], Andrew Helmstetter[2], James Crowe[1], John Iacona[1], Dion Devey[1] & Alexander S. T. Papadopulos[3]**

[1]The Jodrell Laboratory, Royal Botanic Gardens, Kew, TW9 3AA UK
[2]CNRS Montpellier, France
[3]Molecular Ecology and Fisheries Genetics Laboratory, Environment Centre Wales, School of Biological Sciences, University of Bangor, Wales.
**Corresponding authors: joe.parker@kew.org and a.papadopulos@kew.org*

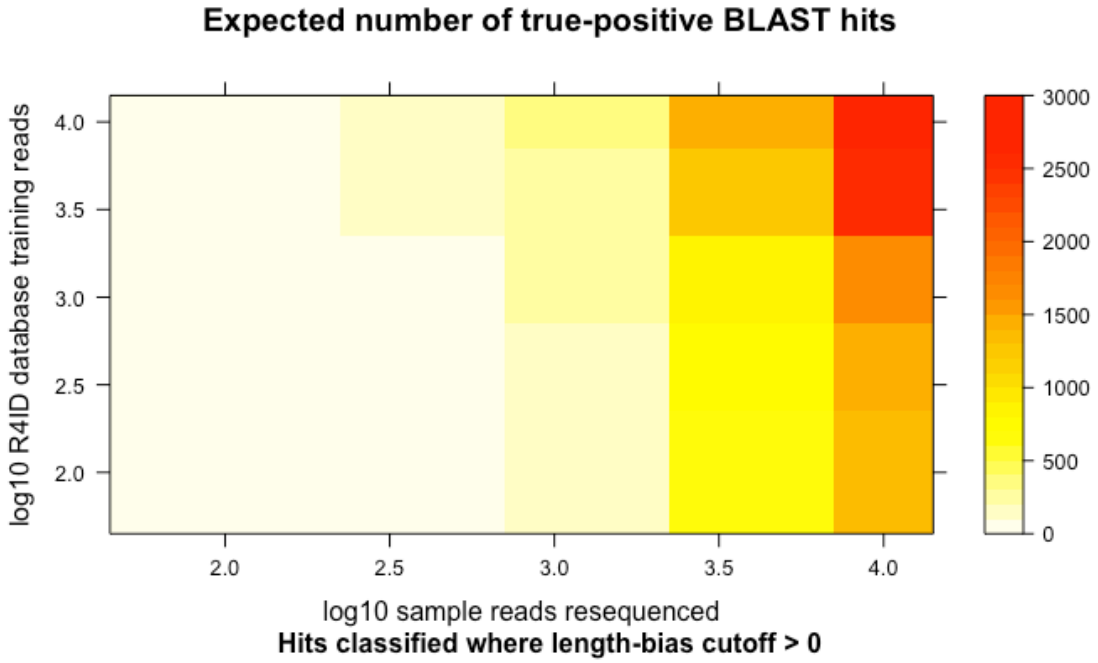
Supplementary Information

Supplementary Figures



Supplementary Figure 1: Expected true-positive : total-positive hit rates per-read; 20 replicates subsampled without replacement from R4ID training (number of reads; log10 y-axis) and unknown sample resequencing (number of reads; log10 x-axis) datasets.

a)



b)

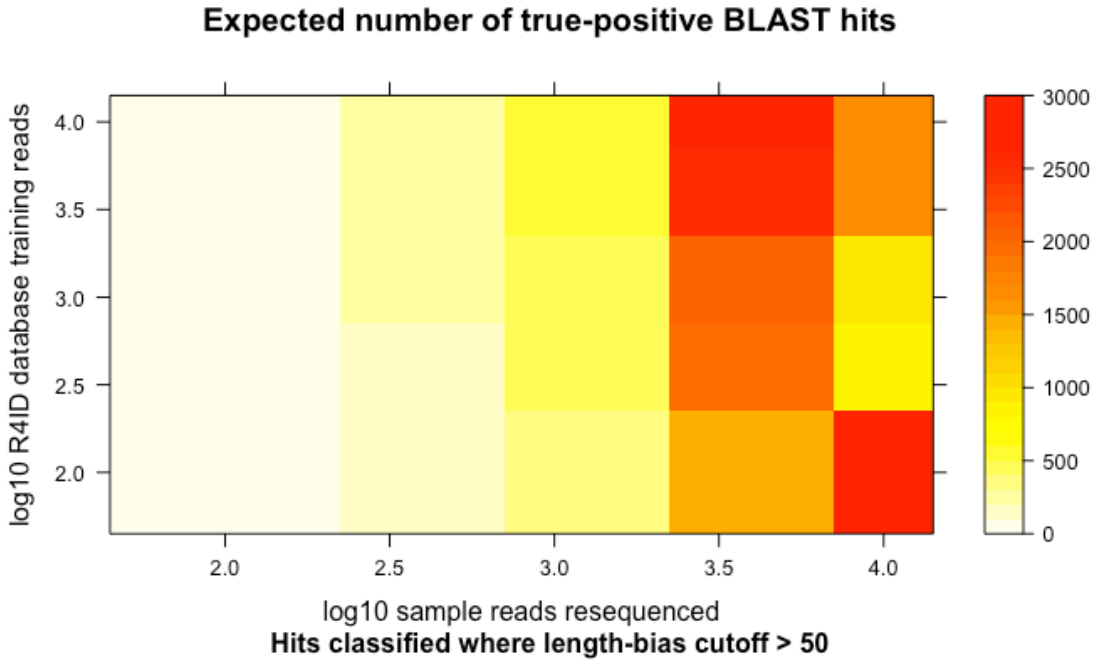
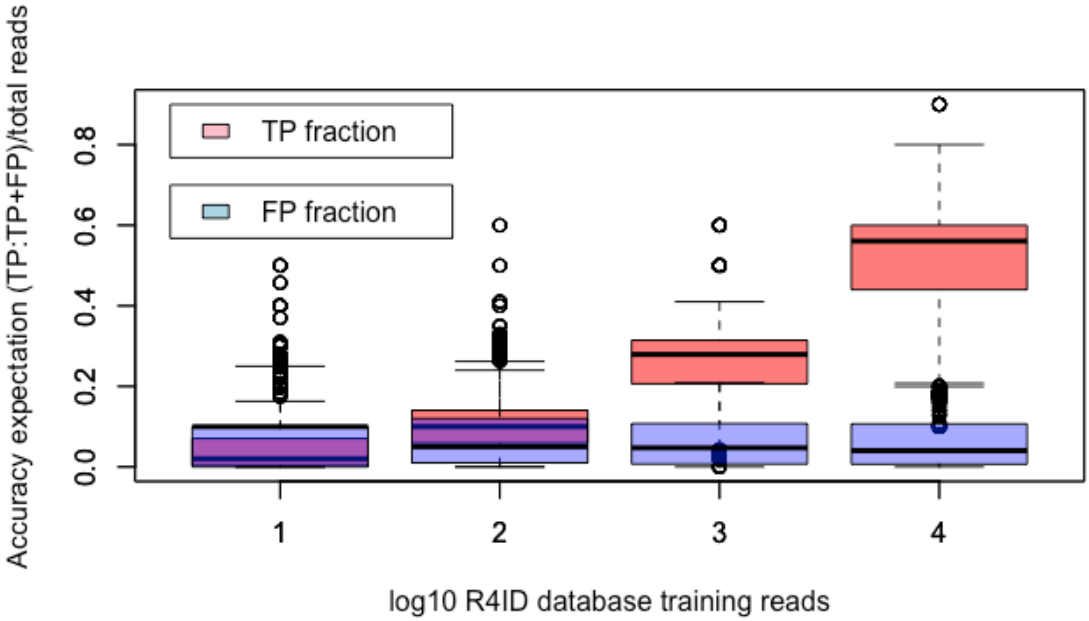


Table 1

Supplementary Figure 2: Expected true-positive hit counts (two-way positive hit rate * total BLAST hits) if BLAST alignment length biases of (a) greater than zero bp and (b) greater than 50bp are used for classification of individual unknown sample reads, based on 20 replicates resampling without replacement from R4ID training run (number of reads; y-axis, log10 scale) and resequencing ID runs (number of reads; x-axis, log10 scale) at 100, 500, 1000, 5000, and 10,000 reads.



Supplementary Figure 3: Expected true- and false-positive fractions among BLAST hits (pink and blue boxplots, respectively; y-axis) by R4IDs training dataset size (reads used to build BLAST database; x-axis, log10 scale) in 18 replicates simulated from published genome assemblies in the Camelinae.

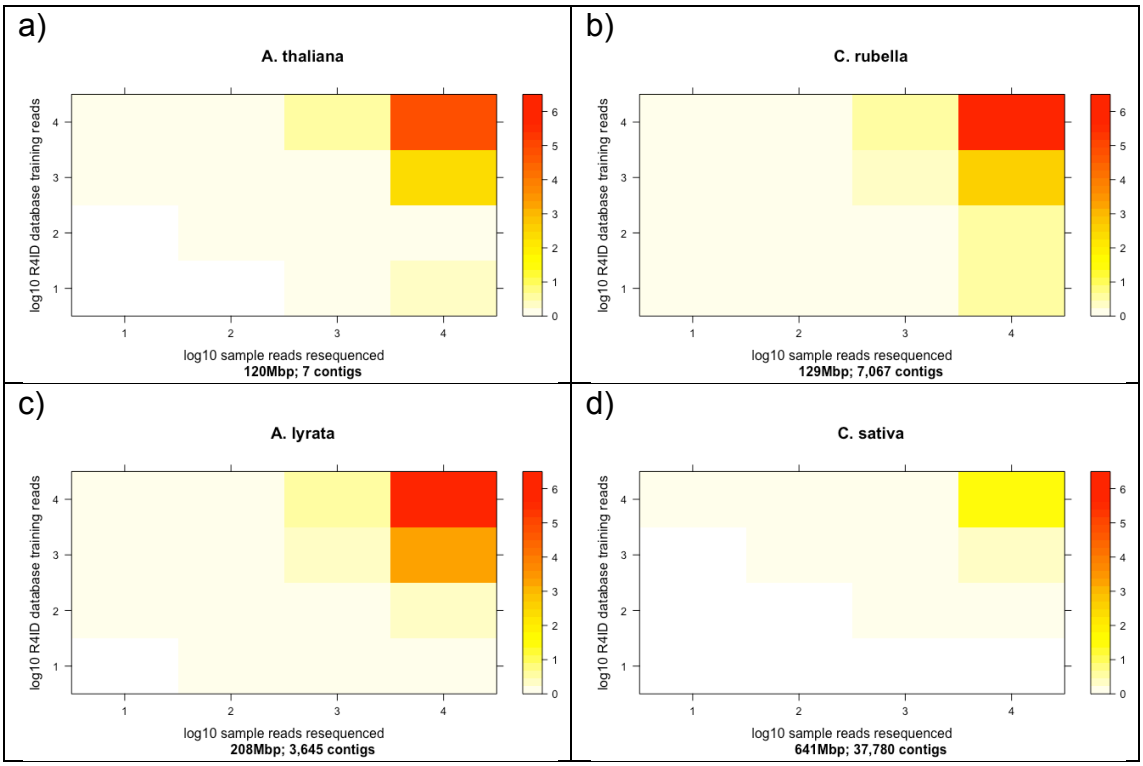


Table 2

Supplementary Figure 4: Relationship between species' genome size, sampling effort and hit rate. Species shown are (a) *A. thaliana* (assembly length: 120Mbp); (b) *C. rubella* (129Mbp); (c) *A. lyrata* (208Mbp); (d) *C. sativa* (641Mbp). For each, simulated 'correct' read expectations are shown (heatmap scale) as a function of sample resequencing read count (x-axis; log10) and R4ID training sequencing (y-axis; log10).