Supplementary methods

For all analyses in this study `BlobTools v1.0` (Laetsch et al., 2017) was used.

Reference genomes were retrieved from ENSEMBL. Read datasets were simulated as Illumina HiSeq2500 paired-end reads at coverages listed in Table 2 using `ART v2.5.8` (Huang et al., 2012) (`-l 150 -m 500 -s 10`). Reads were concatenated into the two libraries and shuffled using `BBmap shuffle v37.02`.

All assemblies were performed using `CLC assembler v5.0.0.142510-160525-215357` specifying read libraries as paired-end with an insert size ranging from 300 to 700b (`-p fb ss 300 700`). To assess the recovered number of conserved genes, assemblies were evaluated using BUSCO (Simão et al., 2015).

All mapping files were created by mapping read libraries assemblies using `BWA mem v0.7.15-r1140` (Li, 2013) and `samtools v1.5` (Li et al., 2009). BAM files were converted to COV format using BlobTools `map2cov`.

Similarity searches were performed against NCBI nt (retrieved 2017-06-13) using `BLASTn 2.6.0+` (Camacho et al., 2009) and against Uniprot Reference Proteomes (retrieved 2017-07-07) using `Diamond blastx v0.9.5` (Buchfink et al., 2015). Parameters for BLASTn were `-evalue 1e-25 -outfmt '6 qseqid staxids bitscore std'` for all searches. Diamond blastx was run with the parameters `--sensitive --evalue 1e-25 --outfmt 6` and results were annotated with NCBI TaxIDs using the BlobTools module `taxify` and a UniProt ID mapping file, filtered to only include NCBI TaxIDs.

For the evaluation of the impact of parameters of sequence similarity searches on BlobTools taxonomic assignment, additional parameters were used for BLASTn and Diamond blastx searches. The search parameter ID and parameters were as follows:

- `MTS1`: `-max-target-seqs 1` (BLASTn and Diamond blastx)
- `MTS10`: `-max-target-seqs 10` (BLASTn and Diamond blastx)
- `HSP1`: `-max_hsps 1` (BLASTn and Diamond blastx)
- `CUL10`: `-culling_limit 10` (BLASTn)

Taxonomy restriction for BLASTn searches was achieved by retrieving GI lists from the NCBI nucleotide portal for the TaxIDs 9604 ('Hominidae'), 561 ('Escherichia'), 6239 ('Caenorhabditis elegans'), 28384 ('other sequences'), and 286 ('Pseudomonas') and supplying them with the parameter `-negative_gilist`. Taxonomy restriction for Diamond blastx searches was carried out by retrieving subtree TaxIDs for the relevant groups from NCBI taxonomy portal and removing the associated sequences from Uniprot Reference Proteomes.

Evaluation of sequence similarity searches was carried out by comparing the "true" taxonomy of sequences in the combined assembly of both simulated read libraries (`blobtools.assembly.A_B.fasta`) to the taxonomy inferred by BlobTools based on sequence similarity searches. The "true" taxonomy of each sequence was inferred by mapping the simulated read libraries against the assembly. Unambiguous taxonomy (cases where only reads originating from one reference genome map to a sequence in the assembly) could be assigned to 98.07% of sequences (16,289 out of 16,610 sequences) in the assembly, totalling 99.89% (158,001,623 out of 158,178,224 b) of assembled span.

Similarity search results were supplied to BlobTools to construct BlobDBs. Tabular output of the BlobDBs was generated using BlobTools `view` (`--hits -r superkingdom -r phylum -r order`). Taxrule 'bestsumorder' (`-x bestsumorder`) was specified in cases where results from both BLASTn searches against NCBI nt and Diamond blastx searches against Uniprot Reference Proteomes were used as input (always in this order). For each BlobDB, accuracy was calculated at the taxonomic ranks of order (Rhabditida, Primates, Pseudomonadales, and Enterobacterales). Tables were evaluated using the script `analyse_blobtools_tables.py`.

For each taxon, counts of bases in the assembly were classified as true/false positives/negatives as follows:

- True positives (TP): sum of bases in sequences where BlobTools taxonomic annotation of taxon is equal to "true" taxonomy.
- False positives (FP): sum of bases in sequences where BlobTools taxonomic annotation indicated the taxon, but the "true" taxonomy differed.
- True negatives (TN): sum of bases in sequences where both BlobTools taxonomic annotation and "true" taxonomy was different to taxon.
- False negatives (FN): sum of bases in sequences where BlobTools taxonomic annotation of taxon failed to reflect "true" taxonomy.

Precision and recall was calculated using the formulae:

$$\text{Precision} = \frac{TP}{(TP + FP)}$$

$$\text{Recall} = \frac{TP}{(TP + FN)}$$

F-score was calculated as the harmonic mean of precision and recall, based on the formula:

$$\text{F-score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

References

Dominik R Laetsch, Georgios Koutsovoulos, Tim Booth, Jason Stajich, and Sujai Kumar. Drl/blobtools: Blobtools v1.0. Jul 2017. doi: 10.5281/zenodo.833879.

Weichun Huang, Leping Li, Jason R Myers, and Gabor T Marth. Art: a next-generation sequencing read simulator. *Bioinformatics*, 28(4):593–594, feb 2012. doi: 10.1093/bioinformatics/btr708. URL `http://dx.doi.org/10.1093/bioinformatics/btr708`.

Felipe A Simão, Robert M Waterhouse, Panagiotis Ioannidis, Evgenia V Kriventseva, and Evgeny M Zdobnov. Busco: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, 31(19):3210–3212, oct 2015. doi: 10.1093/bioinformatics/btv351. URL `http://dx.doi.org/10.1093/bioinformatics/btv351`.

Heng Li. Aligning sequence reads, clone sequences and assembly contigs with bwa-mem. mar 2013. URL `https://arxiv.org/abs/1303.3997`.

Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, aug 2009. doi: 10.1093/bioinformatics/btp352. URL `http://dx.doi.org/10.1093/bioinformatics/btp352`.

Christiam Camacho, George Coulouris, Vahram Avagyan, Ning Ma, Jason Papadopoulos, Kevin Bealer, and Thomas L Madden. Blast+: architecture and applications. *BMC Bioinformatics*, 10:421, dec 2009. doi: 10.1186/1471-2105-10-421. URL `http://dx.doi.org/10.1186/1471-2105-10-421`.

Benjamin Buchfink, Chao Xie, and Daniel H Huson. Fast and sensitive protein alignment using diamond. *Nat Methods*, 12(1):59–60, jan 2015. doi: 10.1038/nmeth.3176. URL `http://dx.doi.org/10.1038/nmeth.3176`.