PhymmBL expanded: confidence scores, custom databases, parallelization and more

To the Editor: PhymmBL[1] is a classification system designed for metagenomics experiments that assigns taxonomic labels to short DNA reads. Since the introduction of PhymmBL in 2009, we made extensive changes and added new features, of which we outline the most important ones here (**Supplementary Table 1**). We also describe results indicating that PhymmBL effectively classifies samples containing mixed eukaryotic and prokaryotic DNA.

PhymmBL combines two components: (i) composition-directed taxonomic predictions from Phymm and (ii) basic local alignment search tool (BLAST)-based homology results[2]. PhymmBL combines these to label each input sequence with its best guess as to the taxonomy of the source organism. Input sequences as short as 100 base pairs can be phylogenetically classified with PhymmBL more accurately than with any other existing method[1] including recently introduced methods (**Supplementary Note 1**).

PhymmBL predicts species, genus, family, order, class and phylum for each read, allowing users to arrange results according to levels of specificity relevant to their research goals. We describe how to configure and operate PhymmBL in a parallelized or grid environment in **Supplementary Note 2**. PhymmBL's open-source software runs on all UNIX-like systems, is written in Perl and C++ and can be downloaded free of charge (http://www.cbcb.umd.edu/software/phymmbl/; currently version 3.2).

To demonstrate PhymmBL's ability to classify eukaryotic DNA, we classified 2,278,901 short reads (average, 276 base pairs) from a permafrost-preserved woolly mammoth bone metagenome[3]. Before classification, we added a variety of genomes to PhymmBL's local database, representing plants, multicellular animals and protists (**Supplementary Table 2**). We also built models for the elephant (*Loxodonta africana*) genome (Elephant Genome Project), expecting that woolly mammoth reads would be labeled as elephant. Our goal was to examine whether PhymmBL could identify eukaryotic DNA as accurately as it had identified bacterial DNA.

The most abundant label (59.7%) was indeed elephant (**Fig. 1**). The next three most abundant predictions were *Flavobacterium johnsoniae* (2.4%), *Polaromonas naphthalenivorans* (1.6%) and *Polaromonas* sp. JS666 (1.0%), all three of which are known Arctic bacteria. These likely represent modern bacterial species present on the mammoth bone. PhymmBL therefore effectively separated eukaryotic from prokaryotic reads and accurately predicted the closest relative of the particular eukaryote that was sequenced, despite the presence of other potentially competing eukaryotic genomes in the local database. Details of the computational resources used by PhymmBL are available in **Supplementary Note 3**.

PhymmBL output now includes scores reflecting the software's confidence that its predictions are correct. A confidence score between 0 and 1 appears with each clade-level prediction for each input read. Polynomial functions estimate accuracy, mapping read length and raw score to these normalized confidence scores. For notes on use and algorithmic details of the confidence score computation, see **Supplementary Note 4**.

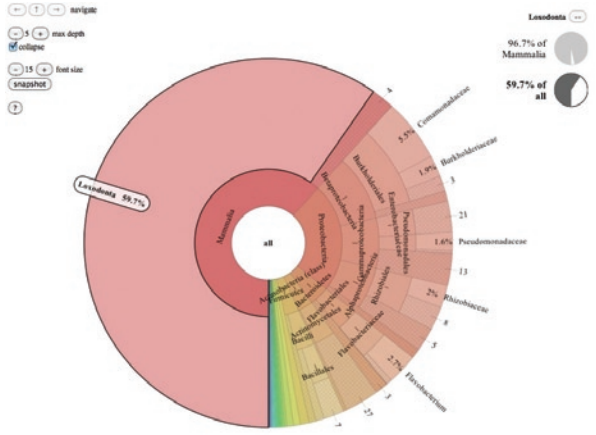

Figure 1 | Screenshot of predicted genus-level taxonomic distribution for the mammoth metagenome, focused on genus *Loxodonta*.

Users now can add an arbitrary amount of custom genomic data (as single genomes or in batches) to their PhymmBL database, eliminating exclusive reliance on RefSeq bacterial and archaeal genomes. The default classification database can be augmented with private (or even synthetic) genomes or with additional publicly available genomic data. Eukaryotic and viral sequences can also be added, expanding PhymmBL's classification mandate beyond prokaryotes.

For viewing PhymmBL output (**Fig. 1**), the freely available Krona package (http://krona.sourceforge.net/) provides an interactive viewer (B. Ondov, N. Bergman and A. Phillippy; personal communication). Krona offers an intuitive HTML5 interface that helps users explore predicted taxonomies for metagenomic read sets. Interacting with a dynamic radial visualization (**Fig. 1**) that changes to highlight designated areas of interest, users can view predictions for their sample population at different levels of phylogenetic granularity, with clade metadata and population statistics displayed alongside.

Note: Supplementary information is available on the Nature Methods website.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Arthur Brady & Steven Salzberg

Center for Bioinformatics and Computational Biology, University of Maryland, College Park, Maryland, USA.
e-mail: abrady@umiacs.umd.edu

1. Brady, A. & Salzberg, S.L. *Nat. Methods* **6**, 673–676 (2009).
2. Altschul, S.F. *et al. Nucleic Acids Res.* **25**, 3389–3402 (1997).
3. Poinar, H.N. *et al. Science* **311**, 392–394 (2006).