High-speed microbial community profiling

Daniel H Haft & Andrey Tovchigrechko

A precomputed database of lineage-restricted reference genes yields a fast and accurate tool that uses sequence similarity alone to compute clade abundances from shotgun metagenomic data sets.

As next-generation sequencing forces down the price of obtaining vast quantities of raw DNA sequence data, the companion costs of sample collection, nucleic acid extraction and data analysis grow larger in comparison. The changing relative costs may affect whether the best overall design for a new scientific study is to sequence just one marker, the 16S ribosomal RNA, or perform shotgun sequencing on the whole microbiome to prepare for a broad range of analyses. In this issue, Segata *et al.*[1] introduce MetaPhlAn, a tool that dramatically reduces the computational burden when using shotgun metagenomic sequence data, rather than 16S rRNA, to estimate clade abundances for clinical or environmental samples. This new tool may steer more scientists to choose whole-microbiome shotgun sequencing as their favored approach for metagenomic studies.

No gene is more widely studied than *rrs*, which encodes the ribosomal small-subunit structural RNA, called 16S rRNA in bacteria and many archaea. The number of complete and high-quality draft microbial reference genomes, now nearing 3,000, is dwarfed by the catalog of known *rrs* sequences. Many bacteria remain uncultured, unsequenced and uncharacterized, and are known only by their 16S rRNA. But studying diversity based only on 16S rRNA has serious drawbacks. First, sequencing requires PCR amplification of variable regions situated between 'universal primers'. Unfortunately, the necessary PCR step may introduce bias and distort the perceived structure of the microbial community. Second, 16S rRNA is not usually interesting in itself but rather as a proxy for the rest of its genome, a lineage-specific 'bag of genes' that consists of, for example, enzymes, transporters, regulators, adhesion proteins, mobile elements and virulence factors. But 16S rRNA is an imperfect proxy: strains that are indistinguishable from each other with respect to this marker often differ critically elsewhere in their genomes, seen famously for the many strains of *Escherichia coli*. If there is no corresponding reference genome, then the bag of genes that the 16S rRNA should represent is unknown and the corresponding set of protein functions cannot be inferred.

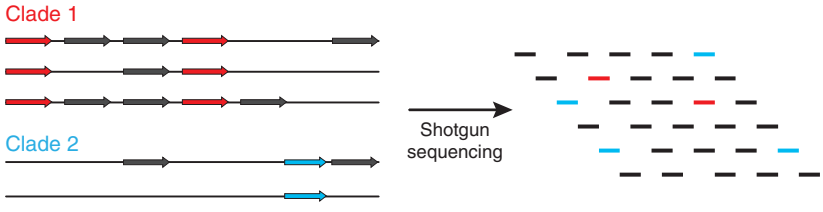

Figure 1 | MetaPhlAn simplifies assignment of taxonomies to whole-metagenome shotgun sequencing reads. Genes in red are restricted to, yet universal across, the reference genomes of Clade 1, whereas genes in blue are restricted to and universal across Clade 2. Shotgun metagenomic DNA sequences are colored if they match core genes from Clade 1 or Clade 2. Reads in black, which do not match a clade-specific gene, are ignored by MetaPhlAn.

Daniel H. Haft and Andrey Tovchigrechko are at the J. Craig Venter Institute, Rockville, Maryland, USA.
e-mail: dhaft@jcvi.org

16S rRNA protocols, commonly executed on Roche's 454 platform, may produce longer reads than, for example, whole-metagenome shotgun sequencing performed on an Illumina sequencer, but the latter technology produces much more total data. This larger metagenomic data set supports uses that 16S data cannot, which argues in favor of the shotgun sequencing approach. Metagenomic assembly often produces complete genes from still-uncultured species, or even whole operons, thereby putting those genes in context. Reads or assemblies may run through automated annotation pipelines that quantify protein function for a microbial community. Years from now, archives of short reads may be reexamined to quantify markers whose significance was not understood when the sequencing was performed. Each of these applications complements the fast community profiling that MetaPhlAn performs.

MetaPhlAn achieves its speed by following one of the most ancient principles in bioinformatics: know which data to ignore. It searches against a reduced reference database culled from the contents of 2,887 prokaryotic reference genomes. Any gene that shows high DNA sequence similarity across distantly related taxa is discarded because it would confound efforts to make unambiguous phylogenetic placements based on sequence similarity. Likewise, any gene found sporadically within a clade is discarded. What remain—about 5% of genes—are 'perfect' markers that are universal for their taxonomic clade yet absent outside it as scored by BLAST (**Fig. 1**). The MetaPhlAn database provides this precomputed set of markers, so sequence similarity detected by BLAST translates instantly to phylogenetic placement: there is no gene calling, no frameshift correction

and no multiple sequence alignment with phylogenetic trees to interpret.

A variety of tools already exist that assign taxonomies to individual reads. But for taxonomic profiling of, say, large clinical microbiome samples, as from the Human Microbiome Project, most of these may be too slow. Segata *et al.*[1] report benchmarking MetaPhlAn against six popular methods, all alignment based, nucleotide word frequency based (also called 'compositional') or hybrid (combining both approaches). All six methods attempt to assign every read. The next best to MetaPhlAn in speed, PhyloPythiaS[2], was 50-fold slower, whereas PhymmBL[3], which was comparable to MetaPhlAn in accuracy, was several times slower yet. A competing fast method, MetaPhyler[4], is based on BLAST searches against a manually curated list of lineage-specific versions of 31 widely distributed genes, and it similarly avoids classifying most reads. It was not explicitly compared to MetaPhlAn.

Two additional recent methods, both based on statistical mixture models, illustrate evolving approaches to abundance estimation. Taxy[5] examines oligonucleotide distributions for a sample as a whole and makes clade abundance estimates for a metagenome in just minutes while making no individual read assignments at all, but its use was described only for coarse-grained divisions, to the phylum or class level. GRAMMy[6], which can build its analysis on per-read classifications made by other pipelines, claims that its mixture model greatly reduces the amount of input data needed to estimate taxonomic abundances to a desired accuracy, an alternative approach to reducing computational costs.

Methods based on 16S rRNA gene sequencing still maintain a cost advantage, and they perform better at community profiling if a significant fraction of the analyzed community has no closely related reference genome. But continual advances in tool development, falling sequencing costs and new reference genomes from underrepresented lineages make metagenomic shotgun sequencing increasingly attractive.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

PUBLISHED ONLINE 10 JUNE 2012; HTTP://WWW.NATURE.COM/DOIFINDER/10.1038/NMETH.2080

1. Segata, N. *et al. Nat. Methods* 811–814 (2012).
2. McHardy, A.C., Martín, H.G., Tsirigos, A., Hugenholtz, P. & Rigoutsos, I. *Nat. Methods* **4**, 63–72 (2007).
3. Brady, A. & Salzberg, S. *Nat. Methods* **6**, 673–676 (2009).
4. Liu, B., Gibbons, T., Ghodsi, M., Treangen, T. & Pop, M. *BMC Genomics* **12** (suppl. 2), S4 (2011).
5. Meinicke, P., Aßhauer, K.P. & Lingner, T. *Bioinformatics* **27**, 1618–1624 (2011).
6. Xia, L.C., Cram, J.A., Chen, T., Fuhrman, J.A. & Sun, F. *PLoS ONE* **6**, e27992 (2011).

Connecting ecology and conservation through experiment

Nick M Haddad

An experimental infrastructure consisting of environmentally controlled and spatially linked habitat patches permits studies on terrestrial animal dispersal at an unprecedented scale for an experiment with such strict control.

To understand the central trade-off in experiments in spatial ecology, consider how to create an experiment, a true experiment with full replication and control, for populations of bears or birds. It is much easier to imagine a controlled experiment for populations of protozoans in bottles or for plants in small grassland plots. Yet these experimental settings lack the realism and complexity of landscapes traversed by larger organisms. There is a strong trade-off in ecology between the spatial extent of studies and the ability to exert experimental control[1] (**Fig. 1**). In this issue of *Nature Methods*, Legrand *et al.*[2] report a new experimental infrastructure, called the Metatron, that is remarkable for the control it achieves over a relatively large area.

The Metatron consists of 48 enclosed habitat patches, each 100 square meters in area

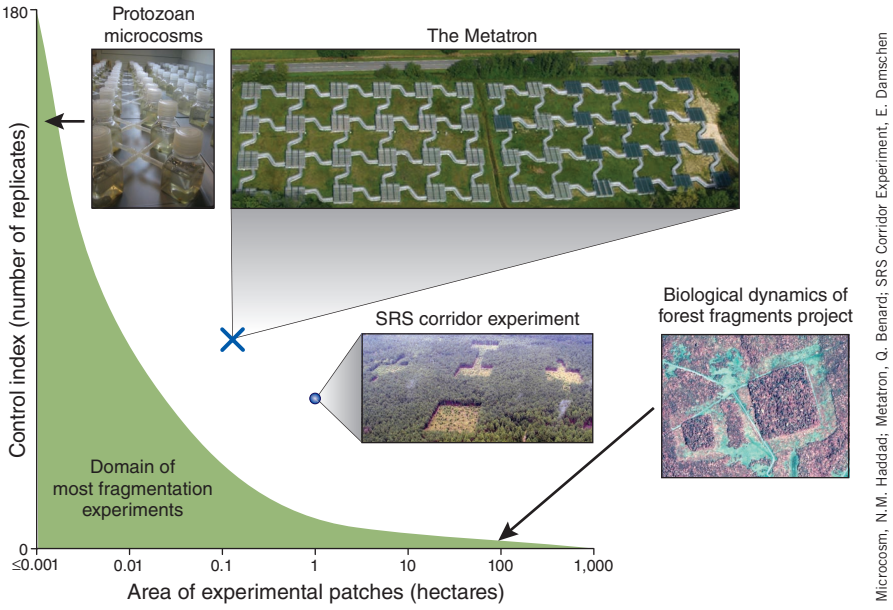

Figure 1 | The relationship between the size of study areas and the degree of experimental control in spatial ecology studies. Figure modified from ref. 1. Nearly all experiments from spatial ecology fall within the shaded area. The Metatron is remarkable for its combination of large patches and strict control. SRS, Savannah River Site.

Nick M. Haddad is in the Department of Biology, North Carolina State University, Raleigh, North Carolina, USA.
e-mail: nick_haddad@ncsu.edu