GENOME WATCH

Finding a needle in a haystack

Lia Chappell

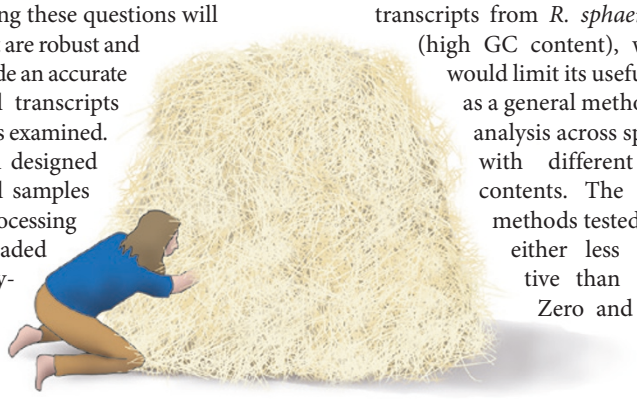
This month's Genome Watch highlights some of the technical challenges that need to be overcome to gain further insight into microbial metatranscriptomes.

Host-associated microbial communities have a number of important roles in human health, and the interrogation of the genomic content of these complex communities has been the subject of many recent studies[1,2]. However, a static snapshot of the genomes of these microorganisms is insufficient to provide a full understanding of the systems and processes that influence human health. Expression studies have the power to fill in the missing gaps, helping us to understand how these organisms interact with each other and their hosts. Transcriptomics is usually the most tractable of the possible approaches for expression studies because, unlike proteins or metabolites, nucleic acids can be amplified.

High-throughput sequencing of transcriptomes (RNA-seq) has facilitated the identification of the complete set of RNA transcripts in individual microorganisms, as well as in several more complex communities; such a set of community sequences is often referred to as a metatranscriptome. Some of the most exciting questions in this area are yet to be addressed, such as those that would require the application of this technique to clinical samples, and will push the limits of RNA-seq technology. Answering these questions will require protocols that are robust and reproducible to provide an accurate representation of all transcripts present in the samples examined. Crucially, a protocol designed to cope with clinical samples must be capable of processing fragmented or degraded RNA, rather than relying on the isolation of full-length transcripts. Furthermore, an effective ribosomal RNA depletion step is essential for the enrichment of mRNA transcripts in the sequence data. This problematic issue is analogous to finding a needle in a haystack, as rRNA represents >98% of the sequence data from total RNA.

A team of researchers at the Broad Institute (Cambridge, Massachusetts, USA) recently published a study[3] aimed at developing an RNA-seq approach to tackle these issues. They first tested a range of commercially available rRNA depletion methods on cultured bacteria, examining three species with differing GC contents: *Prochlorococcus marinus* (30% GC), *Escherichia coli* (51% GC) and *Rhodobacter sphaeroides* (69% GC). A strand-specific approach was used to enable the identification of antisense transcripts.

They found that the performance of the various rRNA depletion approaches varied widely. Ribo-Zero (from Epicentre) performed best, enriching non-rRNA transcripts by up to 40-fold and increasing the proportion of reads that mapped to coding DNA to 97–98%. Crucially, the relative abundance of these reads also corresponded well to that detected in total RNA from the same sample. The duplex-specific nuclease (DSN) method (which preferentially degrades the most abundant transcripts) came in second best owing to its inefficiency in the enrichment of mRNA transcripts from *R. sphaeroides* (high GC content), which would limit its usefulness as a general method for analysis across species with different GC contents. The other methods tested were either less effective than Ribo-Zero and DSN at removing rRNA, or resulted in transcriptome profiles that varied from the relative abundances of transcripts in the undepleted samples.



Next, the authors applied the Ribo-Zero depletion method to the analysis of clinical samples and examined the DNA and RNA content of two human stool samples, one of which contained partially degraded RNA. To determine the species content of the samples, they mapped the DNA reads to 649 bacterial reference genomes. Of these, 19 genomes were abundant and were used as templates to map the RNA-seq reads. The most abundant species was *Prevotella copri*, and this species constituted almost half of the DNA reads. Consistent with this, up to 87% of the RNA reads in the samples originated from *P. copri*. However, the authors note that the performance of the Ribo-Zero method falls off when the input amount of total RNA is below 5 μg, so further improvements may be necessary for the analysis of samples containing low levels of RNA.

This detailed methodological analysis should lay the foundation for many exciting studies from more challenging sample types. The continued development of metatranscriptomics will give us more power to investigate the previously unseen world of host-associated microbial communities.

Lia Chappell is at the Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK.
e-mail: microbes@sanger.ac.uk

doi:10.1038/nrmicro2821

1. Nelson, K. E. *et al*. A catalog of reference genomes from the human microbiome. *Science* **328**, 994–999 (2010).
2. Qin, J. *et al*. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65 (2010).
3. Giannoukos, G. *et al*. Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. *Genome Biol.* **13**, r23 (2012).

Competing interests statement
The author declares no competing financial interests.