PROGRESS

APPLICATIONS OF NEXT-GENERATION SEQUENCING

Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity

Rotem Sorek and Pascale Cossart

Abstract | Transcriptome-wide studies in eukaryotes have been instrumental in the characterization of fundamental regulatory mechanisms for more than a decade. By contrast, in prokaryotes (bacteria and archaea) whole-transcriptome studies have not been performed until recently owing to the general view that microbial gene structures are simple, as well as technical difficulties in enriching for mRNAs that lack poly(A) tails. Deep RNA sequencing and tiling array studies are now revolutionizing our understanding of the complexity, plasticity and regulation of microbial transcriptomes.

Transcriptomics is a powerful tool for understanding gene structures and RNA-based regulation in any organism. In contrast to the classical 'single gene' reductionist approach, in which biological phenomena are studied using a small set of model genes, transcriptome research allows a bird's-eye view of selected phenomena in all genes simultaneously. In eukaryotes, such as humans, mice, flies and yeast, transcriptomes were initially studied by sequencing millions of expressed sequence tags[1] and, more recently, by cDNA sequencing using ultra high-throughput technologies[2]. Analysis of sequenced cDNAs has altered traditional views on many RNA-based regulatory mechanisms: for example, alternative splicing is now known to affect most human genes[3], and *cis*-antisense transcripts have been shown to overlap more than 10% of all protein-coding genes in metazoan genomes[4–7].

Although whole-transcriptome studies have been highly productive in eukaryotes for more than a decade, the transcriptomes of bacteria and archaea have been largely overlooked until recently. One reason is that prokaryotic transcriptomes were regarded as simple compared with eukaryotic transcriptomes; prokaryotic transcripts, except for rare cases, lack introns and are not alternatively spliced or edited. Another major reason is that mRNA enrichment is more challenging in prokaryotes, as prokaryotic mRNAs lack the 3′-end poly(A) tail that marks mature mRNAs in eukaryotes. As >95% of cellular RNA is composed of ribosomal RNA and tRNA[8], transcriptome sequencing of non-enriched total RNA would yield mostly non-mRNA sequences.

With the enormous increase in sequencing capacity through new sequencing technologies, in combination with specialized mRNA enrichment and tiling array techniques, it has recently become feasible to interrogate whole prokaryotic transcriptomes. The first studies to explore such transcriptomes have revealed many surprising findings, including a plethora of non-coding RNAs (ncRNAs), novel untranslated regulatory elements and alternative operon structures. This Progress article discusses recent advances in prokaryotic transcriptomics, including methodological and conceptual developments. For the sake of clarity, we do not specifically discuss differential expression, although many previous differential gene expression studies in prokaryotes have been successfully performed using non-tiling, low-resolution microarrays[9]. We also do not discuss general considerations and challenges of RNA–seq that are not specific to prokaryotic transcriptome applications, as these have been considered elsewhere[2]. We focus on how transcriptome studies — using deep sequencing and high-resolution tiling arrays — can advance our understanding of the structures and regulatory roles of functional elements in prokaryotic genomes.

Transcriptomic approaches

Genome-wide transcriptome analyses of prokaryotes have so far been conducted by two main techniques: RNA–seq[10–15] and genomic tiling arrays[14–20]. The following section provides a brief description of these approaches and the considerations taken when they are applied to prokaryotes.

RNA–seq. New sequencing technologies, such as the Roche 454, Illumina Genome Analyzer and Applied Biosystems SOLiD platforms, allow the cost-effective direct sequencing of whole transcriptomes to a great depth[2]. Total RNA is first extracted from the organism and converted into cDNA by reverse transcription (RT). Because prokaryotic mRNAs lack the poly(A) tail that is typically used for RT priming in eukaryotic RNA–seq applications, alternative priming approaches are used. These include random hexamer priming[10,11,13], oligo(dT) priming from artificially polyadenylated mRNAs[21] and priming from a specific RNA probe ligated to mRNAs[12,15].

An optional, but important, step before RT is the enrichment for mRNAs. Successful reduction of the combined rRNA and tRNA fraction from 95% to 50% of the sample can result in tenfold enrichment of the mRNA output (from 5% to 50%). This is therefore important for sufficient transcript coverage. The various methods that have been developed for mRNA enrichment in prokaryotic samples are detailed in BOX 1. Another consideration is the selection of an appropriate cDNA library construction protocol: some protocols generate strand-specific libraries, which provide valuable information about the orientation of the transcripts, but these protocols are labour intensive and usually require an inefficient RNA–RNA ligation step[2].

Box 1 | **RNA enrichment**

mRNA can constitute as little as 1–5% of total RNA in the prokaryotic cell, so mRNA enrichment is recommended before transcriptome sequencing. Although mRNA enrichment is not an absolute necessity, it can substantially increase transcript coverage and therefore increase the resolution of the resulting transcriptome maps. The downside of enrichment might be unanticipated biases in the sequenced transcriptome, but data is currently lacking with regard to such possible biases. The lower coverage in non-enriched samples could be easily compensated for by sequencing more cDNA, although this might increase the cost of the experiment. Several methods, which take advantage of the unique features of prokaryotic RNA, have been applied for mRNA enrichment.

Ribosomal RNA capture

Probes that correspond to conserved positions along the 16S and 23S bacterial rRNAs are attached to magnetic beads (see the figure, part **a**). rRNA in the RNA sample hybridizes to the probes and is magnetically pulled down and removed. This approach, using the MICROBExpress kit (Ambion, Austin, Texas), was applied in several recent studies[10,11,13,51,56]. In some cases it removed most of the rRNA molecules[13,51]. However, as specific probes are used to capture the rRNA, the depletion efficiency varies between organisms. For example, archaeal rRNA cannot be captured using the MICROBExpress approach.

Degradation of processed RNA

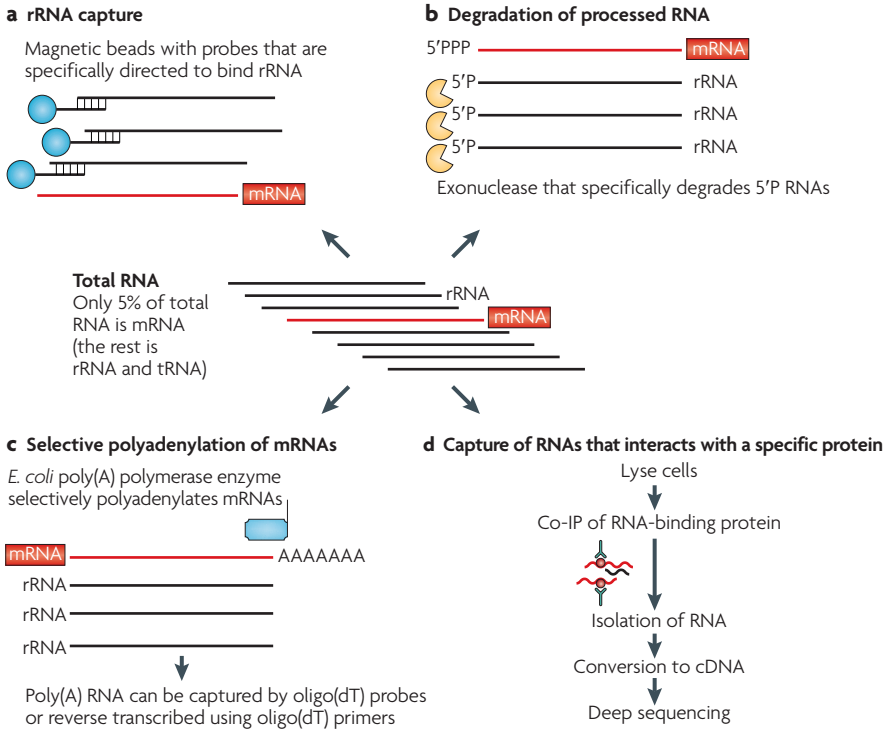
Most bacterial and archaeal mRNAs carry a 5′ triphosphate (5′PPP), which is analogous to the cap structure in eukaryotic RNA (part **b**). Processed RNA molecules, such as rRNAs and tRNAs, carry a 5′ monophosphate (5′P). The mRNA-only prokaryotic mRNA isolation kit (Epicentre, Madison, Wisconsin) uses a 5′-to-3′ exonuclease that degrades 5′P RNA molecules, leaving the mRNAs intact. Preliminary analyses indicate that this approach can remove 10–20% of rRNA for Gram-positive and Gram-negative bacteria (R.S. and P. Hugenholtz, unpublished observations).

Selective polyadenylation of mRNAs

This method uses the *Escherichia coli* poly(A) polymerase based on the observation that this enzyme preferentially polyadenylates mRNAs but not rRNAs[57,58] (part **c**). Following polyadenylation, mRNAs can be captured using oligo(dT) probes or reverse transcribed using oligo(dT) primers. This approach, a modification of which is incorporated into the MessageAmp II-Bacteria kit (Ambion, Austin, Texas), enabled the reduction of rRNA reads to 51% of the sample in a recent metatranscriptomic study[21].

Antibody capture of RNAs that interact with a specific protein

This approach targets a specific subset of RNAs (part **d**). Co-immunoprecipitation (Co-IP) was used to isolate RNAs that are associated with Hfq, a protein that mediates between small RNA and their mRNA targets in bacteria[12,36]. Sittka *et al.* reported that Hfq targeting reduced the fraction of rRNAs and tRNAs to about half of all sequenced cDNA reads[12].



The output of RNA–seq is usually composed of millions of short (25–200 bp) sequence reads that represent fragments of RNAs. To generate a transcriptome map, these reads are computationally mapped to the reference genome, and expressed regions are determined based on their continuous coverage by RNA–seq reads.

Tiling arrays. Genomic tiling arrays, which usually represent both strands of the genome at high densities, have been used to study the transcriptomes of *Listeria monocytogenes*[18], *Bacillus subtilis*[20], *Halobacterium salinarum*[19] and *Mycoplasma pneumoniae*[14], as well as specific genomic features in *Escherichia coli*[16] and *Caulobacter crescentus*[17]. Following cDNA synthesis, the library is hybridized to the array and expression is inferred using signal intensities. In contrast to RNA–seq, mRNA enrichment is not obligatory, and the experimental procedures are well established. However, the array-based approach necessitates hundreds of thousands of probes and is limited by background noise and cross hybridization, and therefore it requires extensive normalization. Once normalized, the data can be used to infer contiguous transcription, as can be done using RNA–seq. The ideal tiling array would contain probes that start at every single base in the genome; however, most tiling arrays have a lower density, mainly owing to the cost associated with the large number of probes needed. Therefore, the transcriptome maps that result from tiling arrays are usually of a lower resolution than the maps produced by RNA–seq, which have single-base-pair resolution. Nevertheless, the relatively small size of prokaryotic genomes makes the tiling array technique appealing for future transcriptome studies in other prokaryotes.

As both RNA–seq and tiling array techniques require a reference genome, they are usually confined to species that have already been sequenced. However, progress was recently made in studying the transcriptomes of species for which no reference genome is available (see 'Metatranscriptomics' below).

Unexpected transcriptome complexity

Several recent key studies have used various combinations of the methods described above to study the transcriptomes of *Burkholderia cenocepacia*[13], *L. monocytogenes*[18], *Bacillus anthracis*[10], *B. subtilis*[20], *H. salinarum*[19], *M. pneumoniae*[14], *Sulpholobus solfataricus*[15] and *Salmonella* spp.[11,12] grown in pathogenic and

non-pathogenic conditions. The results of these studies are beginning to re-shape our understanding of the complexity of the bacterial transcriptome. The following section describes various ways in which whole-transcriptome studies are providing insights into functional genomic elements and their regulatory roles in bacteria.

Gene structures. Most sequenced prokaryotic genomes are annotated nearly exclusively by gene-prediction software that can identify protein-coding genes and a small set of ncRNAs and small RNAs (sRNAs)[22,23]. Such annotations are error prone, fail to detect small genes, and leave most ncRNAs and UTRs unannotated[23]. Accordingly, transcript coverage analysis enabled the identification of conserved, small peptides that were shorter than the threshold size used to computationally define ORFs in the *Salmonella* spp., *L. monocytogenes* and *S. solfataricus* genome annotations[12,15,18]. Also, in *B. anthracis*[10], transcript coverage analysis validated the expression of dozens of genes for which the the annotation was previously deemed unreliable (FIG. 1a). In addition, transcriptome sequencing can refine the predicted structures of authentic genes: Wurtzel *et al.* detected 162 genes in *S. solfataricus* for which the actual transcription start site (TSS) occurred downstream of the predicted one, which reflects the tendency of automated gene prediction to select the largest possible ORF[15] (FIG. 1b). Therefore, transcriptome analysis is beneficial for improving the annotation of any sequenced prokaryotic genome.

Untranslated regulatory regions. The untranslated portions of prokaryotic mRNAs are known to contain important regulatory elements, including riboswitches[24] and binding sites for small regulatory RNAs[25]. Although riboswitches are thought to regulate up to 2% of bacterial genes[26], these RNA regulatory elements are difficult to detect computationally without knowledge of similar sequences and structures that are already known to act as riboswitches[27]. A transcriptome analysis can globally map UTRs across the entire genome, as contiguous expression extending into the flanking intergenic region of a protein-coding gene is indicative of a 5′ or 3′ UTR (FIG. 1c). When such contiguous expression is interrupted in one growth condition and not in another one, this is a strong indication of the presence of a riboswitch. Using this technique, 40 out of 42 predicted riboswitches and 13 candidate

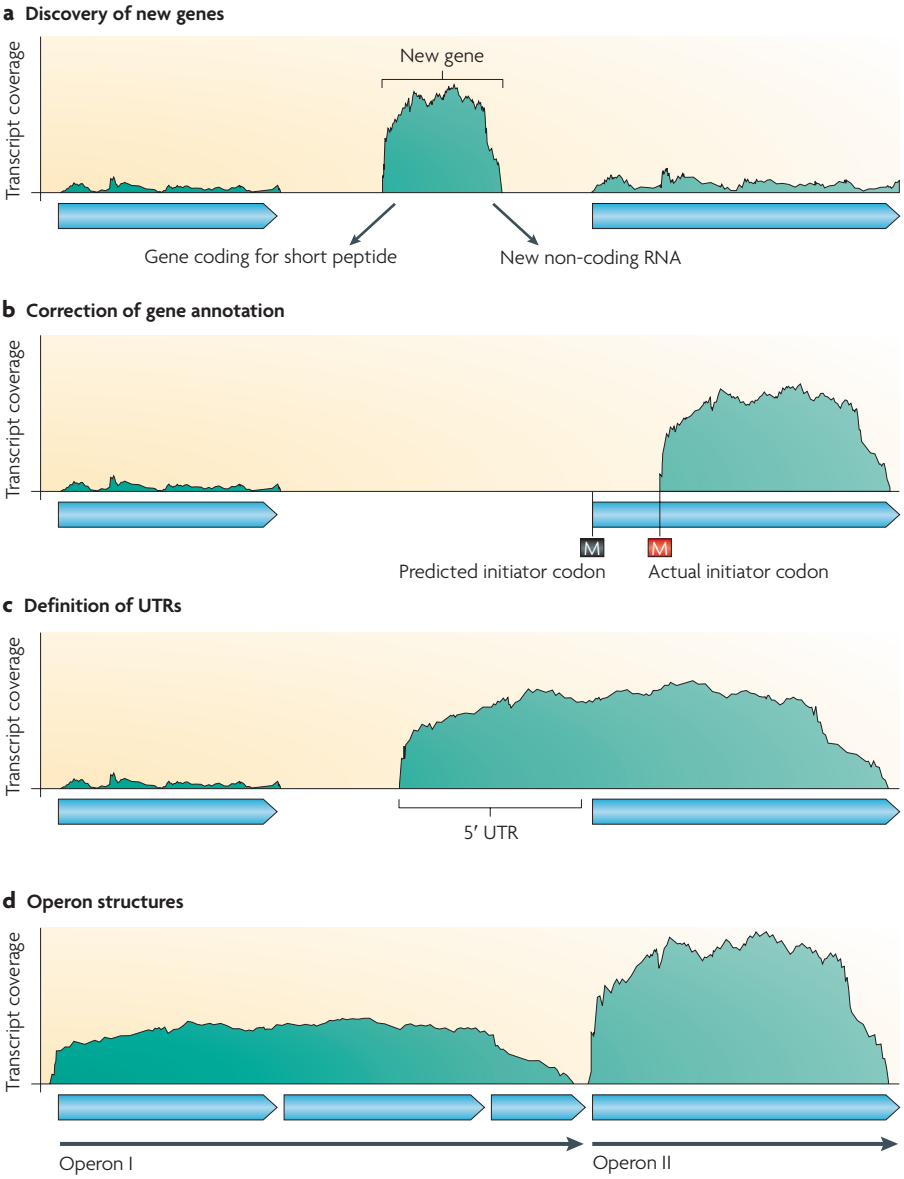


Figure 1 | **Contribution of transcriptomics to annotation of functional elements.** In all panels the *x* axis represents a schematic genomic region and the *y* axis represents transcript coverage as derived from RNA–seq or tiling array experiments. The light blue arrows depict annotated genes. Transcriptomic information can be used to improve genome annotation by enabling: the discovery of new genes (**a**); the correction of gene structure annotations (**b**; the black M depicts the predicted first methionine and the red M is the first methionine in the corrected annotation); the detection of UTRs and transcription start sites (**c**); and the determination of operon relationships (**d**).

novel riboswitches were detected in the transcriptome of *L. monocytogenes*[18]. Usually, 5′ UTRs in bacteria are shorter than 30 bp; therefore, the presence of 37 5′ UTRs longer than 100 bp in *B. anthracis* might indicate that they have functional capacity[10]. Similarly, in *Salmonella* Typhi, 25 genes were found to have long 5′ UTRs, two of which reside in a pathogenicity island (a genomic region in which virulence genes are concentrated), which suggests a role for these UTRs in virulence regulation[11].

Detection of the TSSs of genes also enables the characterization of promoter sequences[16]. For example, McGrath *et al.* used a high-density array that was specifically designed to detect the TSS positions in *C. crescentus*. They detected 769 TSSs and subsequently identified 27 regulatory promoter motifs[17]. Interestingly, 7% of the genes in that organism were found to be transcribed from multiple start sites, and some of them showed cell cycle-dependent promoter switching[17]. Although the

Glossary

Expressed sequence tag
A fragment of cDNA that is generated using random shotgun sequencing of the transcriptome.

Genomic tiling array
A DNA microarray that uses a set of overlapping oligonucleotide probes that cover the whole genome or a proportion of the genome at high resolution.

Polycistronic mRNA
An mRNA (also known as a polycistron) that encodes several polypeptides. Polycistronic transcripts are common in bacteria.

Quorum sensing
A mechanism used by many bacteria to detect a critical bacterial cell density. Some genes are only expressed at high cell density. Cell densities are proportional to the concentration of small molecules or peptides (autoinducers) that are secreted by the bacteria in the medium. These molecules coordinate the expression of specific genes — for example, virulence genes in pathogenic bacteria.

Riboswitch
An RNA element that is located at the 5′ end of an mRNA and that can adopt alternative structures. When a riboswitch binds a metabolite, metal or even a tRNA, the transcription of the downstream gene or, in some cases, the translation of the gene is inhibited.

RNA–seq
An approach for whole-transcriptome profiling in which a population of RNA is converted to cDNA and subjected to high-throughput sequencing. Sequences are mapped to the genome to generate a high-resolution transcriptome map.

occurrence of multiple promoters in prokaryotes had been documented previously, this study was the first attempt to quantify this phenomenon in a genome-wide manner.

Although attention has so far mainly focused on determining the TSSs and 5′ UTRs of genes in prokaryotes, it seems likely that future transcriptome studies will also reveal a regulatory role for the 3′ UTRs in specific genes or organisms. Long 3′ UTRs might affect the expression of genes that are located on the opposite strand[18]. Interestingly, 3′ UTRs of substantial sizes have been found in archaeal transcripts[15,19,28] and have recently been reported to have roles in translation regulation[29].

Operon structures. Genes in prokaryotic genomes are often arranged in operons, with one polycistronic mRNA encompassing several co-transcribed genes. Predicting the structure of operons from the genome sequence is not trivial, and is generally based on the occurrence of a short distance between consecutive genes, conservation of gene order in related organisms, and on identification of functional association between adjacent genes. Various operon-prediction tools are available online, but these have limited accuracy[30]. Based on a contiguous expression signal obtained from RNA–seq or dense tiling arrays and on results obtained from several growth conditions, the first experimentally determined operon maps are now available for bacteria[14,18] and archaea[15,19] (FIG. 1d). These maps show that although most (60–70%) of the genes in bacteria are transcribed as part of a polycistron, fewer genes in archaea (30–40%) show such polycistronic associations.

One striking result in this field stems from the recent transcriptome analysis of *M. pneumoniae* grown in 173 different conditions[14]. This analysis revealed extensive context-dependent modulation of operon structures in response to different conditions — that is, a gene encoded within a polycistron in one condition can be transcribed as a monocistronic transcript in another condition. Such versatile operon behaviour was observed for more than 40% of transcripts in *M. pneumoniae*, and a similar rate was documented in an archaeal transcriptome[19]. These reports reinforce the view of operons as alternating (rather than static) structures that increase the regulatory capacity of prokaryotic transcriptomes in a manner that is functionally analogous to alternative promoters or alternative splicing in eukaryotic transcriptomes.

The availability of transcriptome-based experimental tools to determine operon structures is expected to push forward our understanding of the versatile regulation and evolution of polycistronic transcripts in prokaryotes. For example, future studies might use whole-transcriptome analyses of closely related species to study the conservation of alternative operon structures, how operon structures evolve, and what genomic changes are needed to disrupt, or create, a co-transcriptional relationship between genes.

Non-coding RNAs. In recent years it has become clear that sRNAs have key roles in prokaryotic physiology. These sRNAs, which are usually between 50 and 500 bp long, have been shown to regulate important biological processes, such as virulence, stress response and quorum sensing[31–33]. Most characterized sRNAs regulate the translation or the stability of their mRNA targets through base-pairing with the target 5′ UTR, and are therefore functionally analogous to eukaryotic miRNAs[25]. Base-pairing of sRNA and mRNA usually results in mRNA degradation or in inhibition of translation, and this effect is frequently mediated by the RNA chaperone Hfq[34]. In *E. coli*, in which sRNAs have been extensively studied, more than 80 functional sRNAs have been experimentally validated[35], but owing to the general ineffectiveness of gene-prediction software in detecting ncRNAs, little is known about sRNA abundance in other bacteria and archaea.

Whole-transcriptome analysis now allows the global interrogation of sRNA abundance in any species primarily by detecting expression from non-protein-coding regions (FIG. 1a). In this way, 13 sRNAs that were induced during niche switching in the opportunistic pathogen *B. cenocepacia* were recently discovered[13]. Similarly, Perkins *et al.* detected 55 intergenic regions that are likely to encode new sRNAs in *Salmonella* Typhi Ty2 (REF. 11), and the number of known sRNAs in *L. monocytogenes* has been more than doubled to 50 sRNAs by a tiling array-based study[18]. Two of the *L. monocytogenes* sRNAs were shown to be involved in virulence, as their deletion mutants showed altered pathogenic capabilities. Interestingly, some of the sRNAs in *L. monocytogenes* show base-pairing potential with other sRNAs, which suggests that there are multilevel regulatory networks that involve interactions among several sRNAs[18].

sRNAs can also be selectively enriched by immunoprecipitation of specific RNA-binding proteins, such as Hfq[36]. By selectively sequencing RNAs bound by Hfq, Vogel and colleagues have characterized sRNAs with functions that are mediated by this RNA-binding protein in *Salmonella* Typhimurium[12]. This one study has increased the number of experimentally validated sRNAs in *S.* Typhimurium by more than twofold to a total of 64 sRNAs. Together, these studies highlight whole-transcriptome analysis as a key technique for sRNA discovery.

Antisense transcription. A *cis*-antisense locus is a genomic locus at which two partially overlapping genes are transcribed from opposite strands of the DNA. RNA transcribed from the sense gene might interact with the antisense RNA, thereby regulating its transcription, translation or degradation[37]. High-throughput sequencing of cDNA in flies, mice and humans has shown that antisense transcripts overlap 5–25% of all protein-coding genes[4–7]. Until recently, *cis*-antisense was thought to be extremely rare in prokaryotes, as only 10 cases of chromosomally encoded *cis*-antisense transcripts had been reported in the entire prokaryotic domain, with ~20 additional cases found in

plasmids, phages and transposons[38]. These few prokaryotic *cis*-antisense cases were shown to regulate important processes, such as replication, stress response and iron transport[38].

Over the past year, whole-transcriptome analyses have revolutionized our appreciation of the abundance of *cis*-antisense transcripts in both bacteria and archaea. Hundreds of antisense transcripts were detected in multiple genomes, and between 3% and 13% of all protein-coding genes in *Synechocystis* spp.[39], *Vibrio cholerae*[40], *S.* Typhimurium[12], *B. subtilis*[20], *M. pneumoniae*[14], and *S. solfataricus*[15] were shown to be overlapped by *cis*-antisense transcripts. In a few cases, for example in *L. monocytogenes* and *Synechocystis* spp., further experimental analysis of specific transcripts suggested their involvement in the downregulation of their sense counterparts[18,39] (FIG. 2). Some antisense transcripts are long, spanning more than one ORF and apparently functioning as ncRNAs, and in some cases (as highlighted in *L. monocytogenes*), the overlapping portion of the transcript is the 5′ UTR or the 3′ UTR of a flanking protein-coding gene[18] (FIG. 2; discussed below). Therefore, in contrast to what was generally thought just two years ago, chromosomally encoded *cis*-antisense transcripts might be a common form of regulation in bacterial and archaeal genomes. Future experimental analyses of such transcripts in various genomes are now needed to establish their functional significance and mechanisms of action.

Functional plasticity of RNA elements. One of the most exciting outcomes of recent prokaryotic transcriptome studies is the increasing understanding that functional RNA elements can adopt different roles in different contexts (defined here as 'plasticity'). One example is the lysine riboswitch in *L. monocytogenes*, which lies in the 5′ UTR of the lysine transporter gene (FIG. 2a)[18]. In the presence of lysine, this element forms a transcriptional terminator, so that the downstream transporter is not expressed. In the absence of lysine, riboswitch refolding allows transcription of the lysine transporter gene. Interestingly, the same element acts as a terminator at the 3′ end of the upstream gene. Additional riboswitch elements, including the cobalamin, tRNA-binding box (T-box), metal-sensing box (M-box) and *S*-adenosylmethionine binding (SAM) riboswitches, were also shown to operate as both 5′ and 3′ UTR elements in *L. monocytogenes*[18], and similar functional plasticity was suggested to be associated with a specific T-box element in *B. anthracis*[10]. Surprisingly, some riboswitches were recently shown to also act as small regulatory RNAs — that is, in conditions in which the downstream genes were not expressed, the small transcript could diffuse, pair with the 5′ UTR of another mRNA and regulate its translation[41].

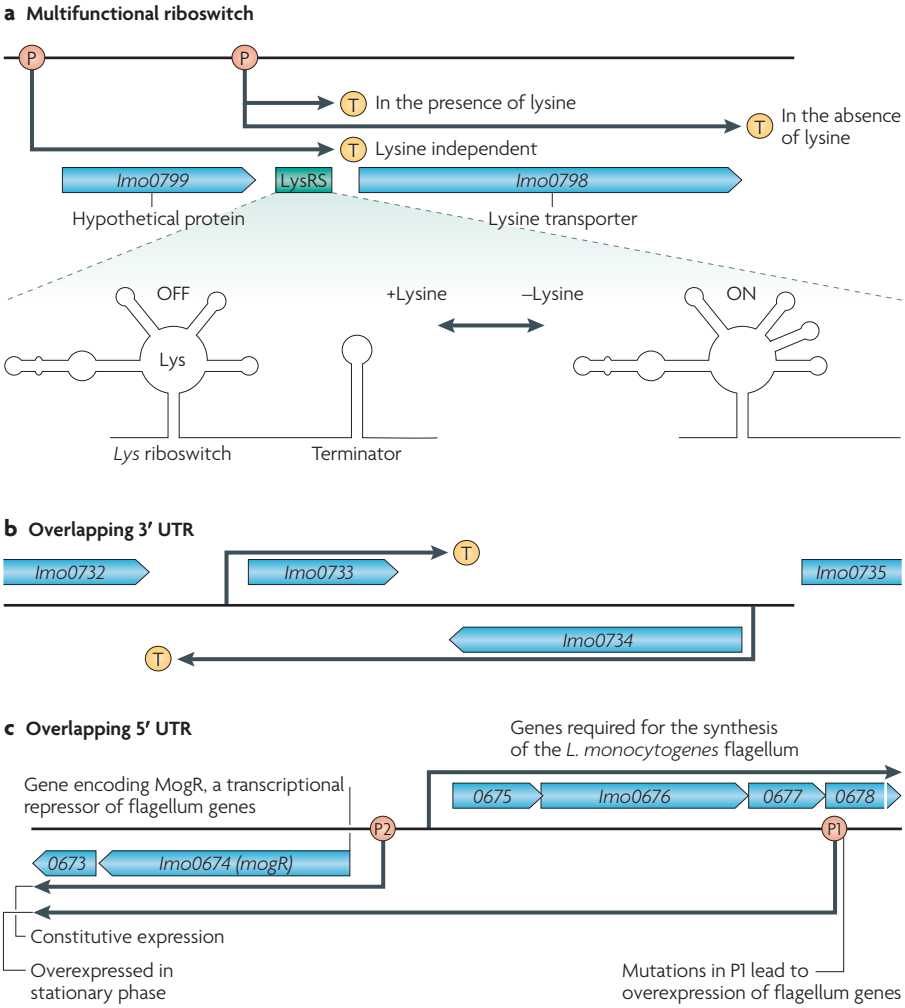


Figure 2 | **Multifunctional RNA elements in *Listeria monocytogenes*. a** | The lysine riboswitch LysRS lies in the 5′ UTR of a lysine transporter gene (*lmo0798*). In the presence of lysine the LysRS tertiary structure is stabilized, and a Rho-independent terminator within this structure prevents the transcription of the downstream gene. Without lysine, transcription of the lysine transporter mRNA proceeds. LysRS also functions as the 3′ end terminator of the upstream gene (*lmo0799*). **b** | An overlapping 3′ UTR. The mRNA of gene *lmo0734* (protein of unknown function) also serves as an antisense transcript for *lmo0733* (protein of unknown function). The functional consequence of this interaction is unknown. **c** | An overlapping 5′ UTR. The *lmo0674* mRNA (a transcriptional repressor of flagellum genes; also known as *mogR*), can be transcribed from two promoters. Transcription from the upstream promoter (P1) generates a long 5′ UTR that overlaps three genes that are required for the synthesis of the flagellum in *Listeria monocytogenes*. The short transcript (transcribed from P2) is constitutively expressed, but the long transcript is induced in stationary phase. The long RNA, when transcribed, hybridizes with the transcript coding for the three flagellum genes, which leads to its degradation. Because the long transcript also produces the MogR protein that represses transcription of the flagellum genes, the same transcript is involved in transcriptional and post-transcriptional regulation of flagellum biogenesis. The flagellum biogenesis system in *L. monocytogenes* is also regulated by multiple additional factors that are coded in other genomic loci. This figure is based on data from REF. 18. P, promoter; T, terminator.

Another type of transcript functional plasticity has been detected in *L. monocytogenes* from the identification of overlapping 5′ and 3′ UTRs. Some genes were recently shown to encode long 3′ UTRs[18], the transcription of which ends inside or after the gene located on the opposite strand. Therefore, the genes on the opposite

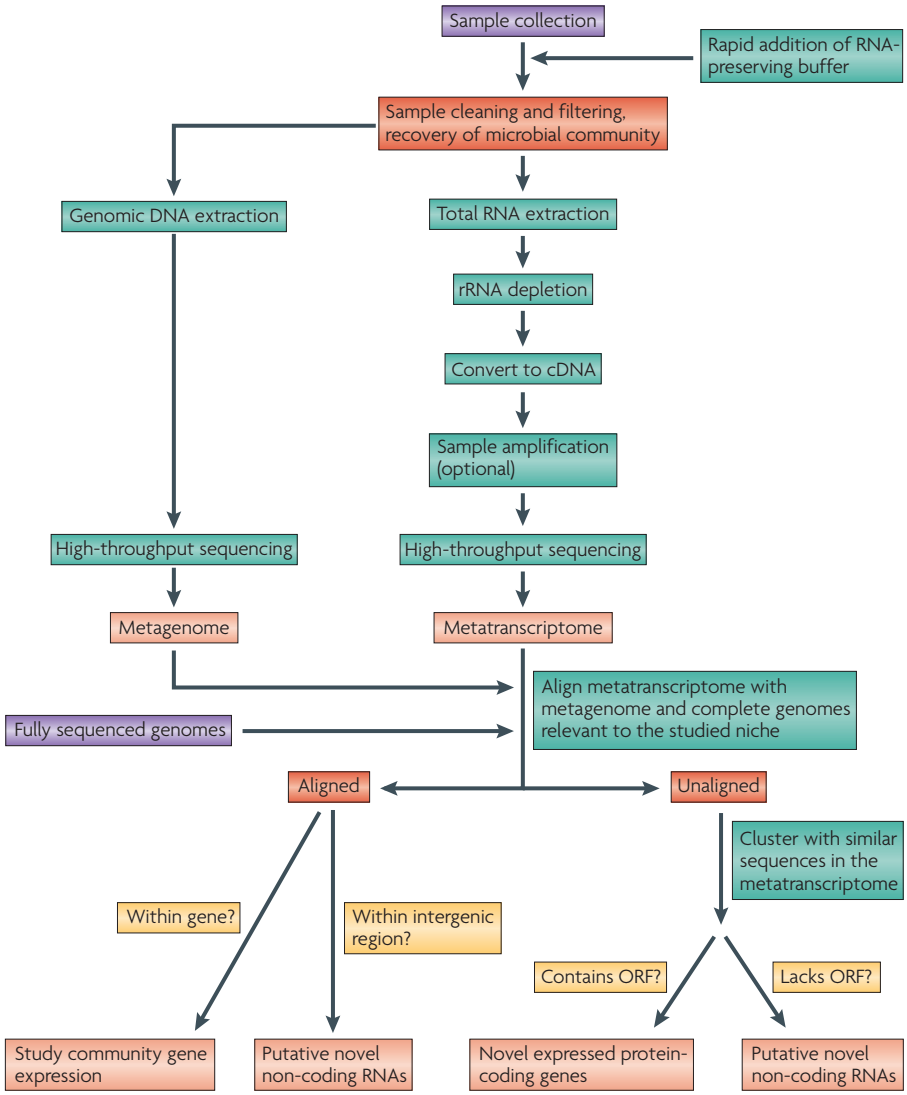


Figure 3 | **Metatranscriptomics: a flow diagram of the steps involved in metatranscriptome sequencing and analysis.** Inputs are in purple; methodological steps are in green; outputs are in orange and red; and analysis questions are in yellow.

strand might serve as RNA regulators of the convergent gene (FIG. 2b) in addition to encoding a protein. Furthermore, a set of genes with long overlapping 5′ UTRs has been identified for which transcription starts from a promoter located inside the ORF of the opposite strand (FIG. 2c). Although the examples presented here derive mostly from analysis of the transcriptomes of *L. monocytogenes*, we believe that future detailed transcriptomic studies might reveal similar transcript complexities in other prokaryotes. If such plasticity turns out to be the rule rather than the exception for prokaryotic transcripts, the general perception of the relative simplicity of bacterial and archaeal transcripts will have to be amended.

Metatranscriptomics

Traditional techniques for studying a prokaryotic species rely on the ability to grow it in a pure culture. However, it is estimated that 99% of all prokaryotic species cannot be readily grown in laboratory conditions[42]. A widely accepted approach for exploring the uncharted domain of uncultivated bacteria is metagenomics, in which the genomic DNA of a microbial community is recovered from the environment and sequenced[43]. Metagenomics has been used successfully to assess species diversity in the soil[44], ocean[45–47] and other niches[48].

Although metagenomics provides a 'snapshot' of the gene content of microbial communities in a given environment, it cannot distinguish between expressed and non-expressed genes. In addition, owing to the fragmented nature of recovered DNA sequences, the discovery of novel ncRNAs and riboswitches in metagenomic data is even harder than in single genomes. The emerging field of metatranscriptomics is now beginning to tackle these challenges.

In metatranscriptomics, total RNA is extracted from a microbial community, converted into cDNA and sequenced[47] (FIG. 3). Early metatranscriptomics studies involved low-throughput sequencing of cDNA clones derived from marine bacterioplankton communities to characterize abundantly expressed transcripts[49]. More recent studies have used high-throughput sequencing with the Roche 454 platform to characterize community gene expression in the soil or ocean water[21,50–52]. To allow the normalization of transcript abundance to the corresponding gene copy number in the DNA pool of the community, both genomic DNA and cDNA from the same sample are sequenced[21].

The observation that a large fraction of sequenced marine community cDNA shares no homology with protein-coding genes has prompted DeLong and colleagues to look for ncRNAs within these cDNAs[53]. By comparing the cDNAs with sequenced metagenomes of marine bacteria, they identified ~40,000 sequences that mapped to intergenic regions. Some of these sequences corresponded to previously characterized ncRNAs and various riboswitches; however, most reads mapped to uncharacterized intergenic regions and were hence suspected to be novel ncRNAs. Clustering of these cDNAs revealed >50 putative novel ncRNAs, each represented by at least 100 cDNA reads. Many of these ncRNAs had conserved RNA secondary structures. Potential base pairing with protein-coding genes flanking the expressed intergenic regions suggested that some of the novel ncRNAs are regulators of carbon metabolism and energy production[53]. Therefore, metatranscriptomics provides a revolutionary means of discovering functional ncRNAs in as-yet-uncharacterized organisms. Considering the tremendous diversity of uncultivated prokaryotes, future metatranscriptomics studies have the potential to identify large numbers of new, niche-specific ncRNA families among localized microbial communities.

Conclusions and outlook

Although the field of microbial transcriptomics is still in its infancy, it is already clear that it provides invaluable means

for understanding RNA-based regulatory mechanisms in a genome-wide manner. With whole-transcriptome analysis it is now realistic to study the involvement of elements, such as ncRNAs, riboswitches and *cis*-antisense regulators, in the physiology and pathogenicity of any prokaryote. In addition, owing to the effectiveness of transcriptome analyses in refining the annotation of well-studied genomes, we predict that this tool will become a standard component of genome-annotation procedures.

The picture emerging from the recent transcriptome studies in prokaryotes is of a regulatory complexity and redundancy that exceeds by far what was originally anticipated from genetic and transcriptional studies of these unicellular organisms. Future transcriptome analyses of additional prokaryotes might show that networks of regulatory sRNAs, context-dependent functional switching in *cis*-acting RNA elements, and long antisense transcripts are the rule rather than the exception for bacteria and archaea. As with other phenomena characterized with high-throughput methods, detailed experimental interrogation of these phenomena will be required to establish their functional relevance for bacterial regulation, physiology and pathogenicity.

One of the current limitations of transcriptomic methods is that they require millions of cells as a starting material. This hinders the interpretation of some of the regulatory phenomena observed. For example, in the case of *cis*-antisense transcription, it is not clear whether one bacterium expresses the sense and antisense transcripts at the same time or whether sense and antisense transcription is mutually exclusive, with some bacteria in the population expressing the sense and some the antisense. Therefore, a major future challenge will be the determination of transcriptome maps for single cells, which will open an avenue for understanding regulatory heterogeneity in microbial cultures. Single-cell transcriptomics will also allow studies of the transcriptome of individual cells from unculturable species. Furthermore, developments in single-cell transcriptomics will allow us to take a close look at stochasticity in gene expression, which is gaining increasing attention as a potential contributor to microbial adaptability to different environments[54]. Tang *et al.* recently demonstrated such a single-cell RNA–seq approach on a mouse blastomer[55], but considerable adaptations are needed before this method can be applied to prokaryotes.

After billions of years of evolution, prokaryotes have developed a huge diversity of regulatory mechanisms, many of which are probably uncharacterized. Now that the powerful tool of whole-transcriptome analysis can be used to study the RNA of bacteria and archaea, a new set of unexpected RNA-based regulatory strategies might be revealed.

Rotem Sorek is at the Department of Molecular Genetics, Weizmann Institute of Science, Rehovot 76100, Israel.

Pascale Cossart is at the Institut Pasteur, Unité des Interactions Bactéries-Cellules, Paris F-75015, France; the Institut National de la Santé et de la Recherche Médicale U604, Paris F-75015, France; and the Institut National de la Recherche Agronomique USC2020, Paris F-75015, France.

Correspondence to R.S.
e-mail: rotem.sorek@weizmann.ac.il

doi:10.1038/nrg2695
Published online 24 November 2009

1. Strausberg, R. L. & Riggins, G. J. Navigating the human transcriptome. *Proc. Natl Acad. Sci. USA* **98**, 11837–11838 (2001).
2. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Rev. Genet.* **10**, 57–63 (2009).
3. Xing, Y. & Lee, C. Alternative splicing and RNA selection pressure — evolutionary consequences for eukaryotic genomes. *Nature Rev. Genet.* **7**, 499–509 (2006).
4. Misra, S. *et al.* Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review. *Genome Biol.* **3**, research0083 (2002).
5. Kiyosawa, H., Yamanaka, I., Osato, N., Kondo, S. & Hayashizaki, Y. Antisense transcripts with FANTOM2 clone set and their implications for gene regulation. *Genome Res.* **13**, 1324–1334 (2003).
6. Yelin, R. *et al.* Widespread occurrence of antisense transcription in the human genome. *Nature Biotech.* **21**, 379–386 (2003).
7. He, Y., Vogelstein, B., Velculescu, V. E., Papadopoulos, N. & Kinzler, K. W. The antisense transcriptomes of human cells. *Science* **322**, 1855–1857 (2008).
8. Neidhardt, F. C. *et al. Escherichia coli and Salmonella: Cellular and Molecular Biology* (ed. Neidhardt, F. C.) 13–16 (ASM Press, Washington, DC, 1996).
9. Bryant, P. A., Venter, D., Robins-Browne, R. & Curtis, N. Chips with everything: DNA microarrays in infectious diseases. *Lancet Infect. Dis.* **4**, 100–111 (2004).
10. Passalacqua, K. D. *et al.* The structure and complexity of a bacterial transcriptome. *J. Bacteriol.* **191**, 3203–3211 (2009).
11. Perkins, T. T. *et al.* A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus *Salmonella* Typhi. *PLoS Genet.* **5**, e1000569 (2009).
12. Sittka, A. *et al.* Deep sequencing analysis of small noncoding RNA and mRNA targets of the global post-transcriptional regulator, Hfq. *PLoS Genet.* **4**, e1000163 (2008).
13. Yoder-Himes, D. R. *et al.* Mapping the *Burkholderia cenocepacia* niche response via high-throughput sequencing. *Proc. Natl Acad. Sci. USA* **106**, 3976–3981 (2009).
14. Guell, M. *et al.* Transcriptome complexity in a genome-reduced bacteria. *Science* (in the press).
15. Wurtzel, O. *et al.* A single-base resolution map of an archaeal transcriptome *Genome Res.* 2 Nov 2009 (doi:10.1101/gr.100396.109).
16. Selinger, D. W. *et al.* RNA expression analysis using a 30 base pair resolution *Escherichia coli* genome array. *Nature Biotech.* **18**, 1262–1268 (2000).
17. McGrath, P. T. *et al.* High-throughput identification of transcription start sites, conserved promoter motifs and predicted regulons. *Nature Biotech.* **25**, 584–592 (2007).
18. Toledo-Arana, A. *et al.* The *Listeria* transcriptional landscape from saprophytism to virulence. *Nature* **459**, 950–956 (2009).
19. Koide, T. *et al.* Prevalence of transcription promoters within archaeal operons and coding sequences. *Mol. Syst. Biol.* **5**, 285 (2009).
20. Rasmussen, S., Nielsen, H. B. & Jarmer, H. The transcriptionally active regions in the genome of *Bacillus subtilis*. *Mol. Microbiol.* **73**, 1043–1057 (2009).
21. Frias-Lopez, J. *et al.* Microbial community gene expression in ocean surface waters. *Proc. Natl Acad. Sci. USA* **105**, 3805–3810 (2008).
22. McHardy, A. C., Goesmann, A., Puhler, A. & Meyer, F. Development of joint application strategies for two microbial gene finders. *Bioinformatics* **20**, 1622–1631 (2004).
23. Overbeek, R., Bartels, D., Vonstein, V. & Meyer, F. Annotation of bacterial and archaeal genomes: improving accuracy and consistency. *Chem. Rev.* **107**, 3431–3447 (2007).
24. Coppins, R. L., Hall, K. B. & Groisman, E. A. The intricate world of riboswitches. *Curr. Opin. Microbiol.* **10**, 176–181 (2007).
25. Waters, L. S. & Storz, G. Regulatory RNAs in bacteria. *Cell* **136**, 615–628 (2009).
26. Mandal, M., Boese, B., Barrick, J. E., Winkler, W. C. & Breaker, R. R. Riboswitches control fundamental biochemical pathways in *Bacillus subtilis* and other bacteria. *Cell* **113**, 577–586 (2003).
27. Hammann, C. & Westhof, E. Searching genomes for ribozymes and riboswitches. *Genome Biol.* **8**, 210 (2007).
28. Brenneis, M., Hering, O., Lange, C. & Soppa, J. Experimental characterization of *cis*-acting elements important for translation and transcription in halophilic archaea. *PLoS Genet.* **3**, e229 (2007).
29. Brenneis, M. & Soppa, J. Regulation of translation in haloarchaea: 5′- and 3′-UTRs are essential and have to functionally interact *in vivo*. *PLoS ONE* **4**, e4484 (2009).
30. Mao, F., Dam, P., Chou, J., Olman, V. & Xu, Y. DOOR: a database for prokaryotic operons. *Nucleic Acids Res.* **37**, D459–D463 (2009).
31. Bejerano-Sagie, M. & Xavier, K. B. The role of small RNAs in quorum sensing. *Curr. Opin. Microbiol.* **10**, 189–198 (2007).
32. Masse, E., Salvail, H., Desnoyers, G. & Arguin, M. Small RNAs controlling iron metabolism. *Curr. Opin. Microbiol.* **10**, 140–145 (2007).
33. Toledo-Arana, A., Repoila, F. & Cossart, P. Small noncoding RNAs controlling pathogenesis. *Curr. Opin. Microbiol.* **10**, 182–188 (2007).
34. Aiba, H. Mechanism of RNA silencing by Hfq-binding small RNAs. *Curr. Opin. Microbiol.* **10**, 134–139 (2007).
35. Livny, J. & Waldor, M. K. Identification of small RNAs in diverse bacterial species. *Curr. Opin. Microbiol.* **10**, 96–101 (2007).
36. Zhang, A. *et al.* Global analysis of small RNA and mRNA targets of Hfq. *Mol. Microbiol.* **50**, 1111–1124 (2003).
37. Lavorgna, G. *et al.* In search of antisense. *Trends Biochem. Sci.* **29**, 88–94 (2004).
38. Brantl, S. Regulatory mechanisms employed by *cis*-encoded antisense RNAs. *Curr. Opin. Microbiol.* **10**, 102–109 (2007).
39. Georg, J. *et al.* Evidence for a major role of antisense RNAs in cyanobacterial gene regulation. *Mol. Syst. Biol.* **5**, 305 (2009).
40. Liu, J. M. *et al.* Experimental discovery of sRNAs in *Vibrio cholerae* by direct cloning, 5S/tRNA depletion and parallel sequencing. *Nucleic Acids Res.* **37**, e46 (2009).
41. Loh, E. *et al.* A *trans*-acting riboswitch controls expression of the virulence regulator PrfA in *Listeria monocytogenes*. *Cell* **139**, 770–779 (2009).
42. Pace, N. R. A molecular view of microbial diversity and the biosphere. *Science* **276**, 734–740 (1997).
43. Tringe, S. G. & Rubin, E. M. Metagenomics: DNA sequencing of environmental samples. *Nature Rev. Genet.* **6**, 805–814 (2005).
44. Tringe, S. G. *et al.* Comparative metagenomics of microbial communities. *Science* **308**, 554–557 (2005).

45. DeLong, E. F. *et al.* Community genomics among stratified microbial assemblages in the ocean's interior. *Science* **311**, 496–503 (2006).
46. Rusch, D. B. *et al.* The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol.* **5**, e77 (2007).
47. DeLong, E. F. The microbial ocean from genomes to biomes. *Nature* **459**, 200–206 (2009).
48. Warnecke, F. *et al.* Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature* **450**, 560–565 (2007).
49. Poretsky, R. S. *et al.* Analysis of microbial gene transcripts in environmental samples. *Appl. Environ. Microbiol.* **71**, 4121–4126 (2005).
50. Leininger, S. *et al.* Archaea predominate among ammonia-oxidizing prokaryotes in soils. *Nature* **442**, 806–809 (2006).
51. Gilbert, J. A. *et al.* Detection of large numbers of novel sequences in the metatranscriptomes of complex marine microbial communities. *PLoS ONE* **3**, e3042 (2008).
52. Urich, T. *et al.* Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. *PLoS ONE* **3**, e2527 (2008).
53. Shi, Y., Tyson, G. W. & DeLong, E. F. Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. *Nature* **459**, 266–269 (2009).
54. Kaern, M., Elston, T. C., Blake, W. J. & Collins, J. J. Stochasticity in gene expression: from theories to phenotypes. *Nature Rev. Genet.* **6**, 451–464 (2005).
55. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods* **6**, 377–382 (2009).
56. Mao, C., Evans, C., Jensen, R. V. & Sobral, B. W. Identification of new genes in *Sinorhizobium meliloti* using the Genome Sequencer FLX system. *BMC Microbiol.* **8**, 72 (2008).
57. Amara, R. R. & Vijaya, S. Specific polyadenylation and purification of total messenger RNA from *Escherichia coli*. *Nucleic Acids Res.* **25**, 3465–3470 (1997).
58. Wendisch, V. F. *et al.* Isolation of *Escherichia coli* mRNA and comparison of expression using mRNA and total RNA on DNA microarrays. *Anal. Biochem.* **290**, 205–213 (2001).

Acknowledgements
We thank O. Wurtzel and A. Levy for stimulating comments. R.S. is supported, in part, by the Israel Science Foundation Focal Initiatives in Research in Science and Technology (FIRST) program (grant 1615/09), the Crown Human Genome Center, the Y. Leon Benoziyo Institute for Molecular Medicine and the M.D. Moross Institute for Cancer Research at the Weizmann Institute of Science, as well as the Alon Fellowship. P.C. is a Howard Hughes International Scholar. She has received financial support for her work on RNA from the Agence Nationale de la Recherche, the European Union (BacRNA 2005-018618) and recently from the European Research Council.

Competing interests statement
The authors declare no competing financial interests.

DATABASES
Entrez Gene: http://www.ncbi.nlm.nih.gov/gene
lmo0674 | *lmo0798*
UniProtKB: http://www.uniprot.org
Hfq

FURTHER INFORMATION
Rotem Sorek's homepage:
http://www.weizmann.ac.il/molgen/Sorek
Pascale Cossart's homepage:
http://www.pasteur.fr/recherche/unites/uibc/welcome.html

ALL LINKS ARE ACTIVE IN THE ONLINE PDF