Dual RNA-seq of pathogen and host

Alexander J. Westermann, Stanislaw A. Gorski and Jörg Vogel

Abstract | A comprehensive understanding of host–pathogen interactions requires a knowledge of the associated gene expression changes in both the pathogen and the host. Traditional, probe-dependent approaches using microarrays or reverse transcription PCR typically require the pathogen and host cells to be physically separated before gene expression analysis. However, the development of the probe-independent RNA sequencing (RNA-seq) approach has begun to revolutionize transcriptomics. Here, we assess the feasibility of taking transcriptomics one step further by performing 'dual RNA-seq', in which gene expression changes in both the pathogen and the host are analysed simultaneously.

Pathogen-associated molecular patterns (PAMPs).General small molecular motifs that are present on microorganisms and engage host innate immune receptors, in particular Toll-like receptors. Examples of PAMPs include lipopolysaccharide, peptidoglycan and flagellin.

Institute for Molecular Infection Biology, University of Würzburg, D-97080, Germany. Correspondence to J.V. e-mail: joerg.vogel@uni-wuerzburg.de

doi:10.1038/nrmicro2852

Eukaryotic host cells are subject to infection by agents of varying complexity, from viruses to bacteria to eukaryotic parasites such as fungi and protozoa. Infection initiates a dynamic cascade of events that culminates in altered gene expression patterns in both interacting organisms. These changes lead to the adaptation and persistence of the pathogen or to its clearance from the host by the immune response. An unbiased and global understanding of the transcriptomes of both host and pathogen can provide new insights into this process by identifying new virulence factors in the pathogen, or new pathways in the host cell that respond to the exposure to specific pathogens or pathogen-associated molecular patterns (PAMPs).

In many cases, the introduction of new technologies to a field can overcome previously existing limitations and obstacles and can lead to significant leaps in our understanding of a biological process. For example, the introduction of microarrays two decades ago enabled the study of changes in the expression levels of many genes simultaneously. In principle, this technology also allows one to comprehensively monitor gene expression in both the pathogen and the host during their interaction[1,2]. However, the technical difficulties associated with simultaneously determining two often very different transcriptomes, including issues such as probe selection, cross-hybridization and the required design and cost of custom chips, make microarray-based studies challenging and expensive when they are applied to determining both the host and pathogen transcriptomes; thus, the majority of these studies have focused on either the pathogen or the host at any one time.

The recently developed RNA sequencing (RNA-seq) technique provides a conceptually novel approach to the study of transcriptomes and would, in principle, allow the host and pathogen transcriptomes to be analysed in parallel. The major benefit of such an approach is the potential to monitor gene expression in two organisms to a high level of accuracy and depth. Given the sensitivity of this approach, it could potentially be used to sequence the transcriptomes of a small number of cells at the initial site of infection, a feat which is yet to be achieved in infection biology. Most importantly, a dual approach would allow the monitoring of genes from both host and pathogen at different time points throughout the infection process — that is, from initial contact through to invasion and, finally, the manipulation of the host. It thus enables the temporal determination of responses and changes in the cellular networks in both organisms. A dual approach could be particularly important in infection biology, as it can also be easily applied to different pathogens that use distinct infection methods and have different life cycles. Compared to microarray-based approaches[3], RNA-seq has several benefits, ranging from the economical (it does not require the design of new chips for experiments analysing different pathogens or hosts and is therefore a species-independent platform) to the technical (such as the significantly increased sensitivity, dynamic range and discriminatory power). However, such an approach, which we refer to as 'dual RNA-seq', has yet to be achieved for mixed bacterium–eukaryote model systems. Following a brief overview of gene expression studies in infection biology, we focus here on the potential for and feasibility of using dual RNA-seq to reveal the complex interplay between a bacterial pathogen and its mammalian host during infection.

Host or pathogen transcriptomics

The first transcriptomic studies became possible with the development of cDNA microarrays[4,5]. This entailed the use of immobilized gene-specific DNA probes, which hybridize to their corresponding labelled cDNA.

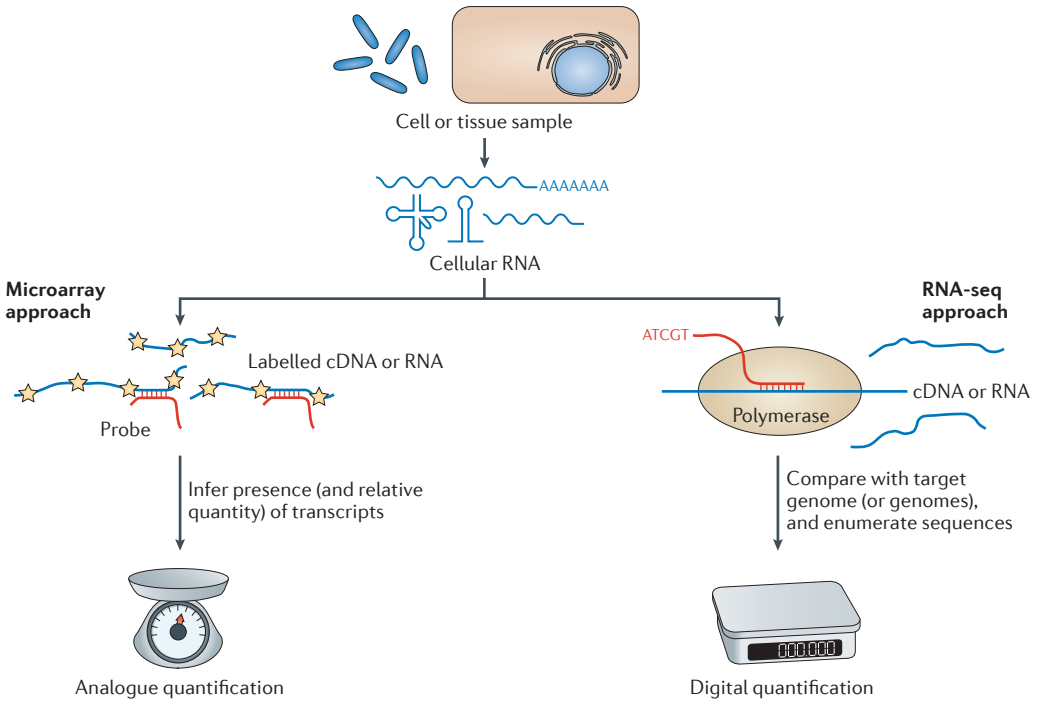


Figure 1 | **Fundamental differences between probe-dependent and probe-independent approaches to gene expression analysis.** In a probe-dependent method such as microarray analysis, the relative abundance of a given labelled transcript is inferred from the fluorescent signal that is retained following hybridization to immobilized probes. Signal intensity has both a lower (sensitivity) and upper (saturation) threshold. By contrast, direct counting of sequencing reads for a given transcript by the probe-independent RNA sequencing (RNA-seq) method in theory has an infinite dynamic range.

The resulting change in fluorescence provides a readout of the relative abundance of the transcript (FIG. 1). The probes for early microarrays were designed based on cDNA libraries of open reading frames, which restricted the analysis to known or predicted mRNAs. However, this limitation was overcome by the advent of high-resolution tiling arrays, which contain probes that, in principle, could represent an entire genome with single-base resolution and thus extend the repertoire of detectable transcripts to include, for example, antisense and other non-coding RNA species[6–8].

Microarray-based studies provided the first global analyses of gene expression changes in pathogens such as *Vibrio cholerae*[9], *Borrelia burgdorferi*[10], *Chlamydia trachomatis*[11], *Chlamydia pneumoniae* (also known as *Chlamydophila pneumoniae*)[12] and *Salmonella enterica*[13,14], revealing the strategies that are used by these microorganisms to adapt to the host. Tiling arrays uncovered the gene expression changes in *Listeria monocytogenes* grown under various *in vitro* and *in vivo* conditions[15], as well as identifying the small non-coding RNA transcriptomes and new virulence genes in streptococci[16–18].

Mammalian host cell transcriptomes are considerably larger and more complex than those of their pathogens, but microarrays were again seminal in providing the first information on global gene expression changes within host cells during viral infection[19] and following interferon stimulation[20]. Furthermore, microarrays yielded the first global insights into the host innate immune response to PAMPs[21–23] as well as the effects of bacterial infection on the expression of various host factors (reviewed in REF. 3).

The power of array-based analyses notwithstanding, there are some major caveats to this approach. One is that, owing to probe cross-reactivity between host and pathogen cDNAs, either cross-hybridizing clones must be eliminated (as reported, for example, in REF. 24) or the RNA of the pathogen and the host must be analysed separately. However, to obtain RNA from one of the two interacting organisms, that of the other is usually sacrificed; for example, the eukaryotic RNA is lost during lysis of the host cell in the course of isolating RNA from intracellular bacteria. Another caveat is the immense cost of tiling arrays. Until recently, our understanding of the transcriptome was limited to mRNAs, tRNA and rRNAs being the relevant classes of transcripts for which expression was worth monitoring, and these RNAs could easily be profiled with arrays that cost in the range of a few hundred Euros. However, we now know that these RNA classes constitute only part of the functional transcriptome (see below). To analyse the full transcriptome at high resolution, including the many transcripts from non-coding regions, hybridization-based methods require arrays with hundreds of millions of probes, the costs of which may exceed those of an RNA-seq experiment. In addition, these high-density arrays further exacerbate the problem of cross-hybridization,

Tiling arrays
DNA microarray chips on which probe sequences are tiled (overlapping) and comprise a subset of, or the whole, genome at high resolution.

Small non-coding RNA
A short transcript (~50–500 nucleotides) that regulates gene expression in bacteria, often by base-pairing with mRNAs.

Box 1 | **Currently available next-generation sequencing platforms**

The deep-sequencing platforms that are currently available can be divided into two groups: most commonly, cDNA samples need to be amplified by PCR before sequencing, but some platforms are sufficiently sensitive to omit the amplification step and sequence cDNA directly. Below, we briefly describe the technical principles of the major representatives from each group and discuss their individual strengths and limitations. For more detailed information, readers are referred to two excellent recent in-depth reviews[93,94].

The Illumina (Solexa), Life Technologies (SOLiD), and Roche (454) sequencing platforms have been widely used for genome-wide transcriptomic studies (see Supplementary information S1 (table)) and are all amplification-based methods. SOLiD and 454 sequencing rely on emulsion-based PCR, whereas the Illumina platform using Solexa chemistry makes use of solid-phase amplification to generate the cDNA library. However both of these amplification strategies are prone to biases (such that certain sequences are preferred over others) and are also susceptible to the introduction of mutations.

In addition to template preparation, the three major next-generation sequencing platforms vary in several other respects. As of August 2012, the respective company webpages state the following performance parameters: the HiSeq 2000 from Illumina generates 600 Gb per single run, the maximum read length is 2 × 100 bases and the run time is ~11 days; the SOLiD 4 machine generates 55–70 Gb per run of mappable data for paired-end runs, with a read length of up to 50 bases and a run time of 11–13 days; and the GS FLX Titanium XL+ from Roche typically has a throughput of 700 Mb per run, generates reads of up to 1,000 bases in length and has a run time of only 23 hours.

Life Technologies was the first to develop a post-light sequencing system: the Ion Torrent platform. Rather than measuring a fluorescent signal (as used by previous approaches), the Ion Torrent platform takes advantage of a semiconductor that senses the pH changes resulting from nucleotide incorporation into the nascent DNA strand[95]. Ion Torrent systems, just like the SOLiD and 454 systems, require emulsion-based PCR amplification of the starting material.

So-called 'third-generation' technologies are capable of single-molecule sequencing (SMS). Helicos BioSciences and Pacific Biosciences have developed deep-sequencing platforms that require less starting material than amplification-based approaches and so obviate the need for an amplification step. Rather, a single cDNA molecule is directly sequenced. Although SMS omits PCR-introduced biases, the method is not free from errors. For instance, it was found that SMS is prone to producing reads that include randomly introduced gaps evoked by so-called dark bases, the incorporation of which does not lead to a quantifiable fluorescence signal[96]. Recently, a hybrid RNA-sequencing approach has been introduced that combines amplification-based and amplification-free sequencing techniques to correct for errors in SMS reads[97].

An alternative SMS method will be provided by Oxford Nanopore Technologies. The working principle is based on a bacterial transmembrane protein that forms a hydrophilic channel a few nanometres in diameter within a membrane (that is, the nanopore). A given template strand can be ratcheted through the pore in a stepwise manner. At the end of each cycle, overlapping nucleotide triplets are read at the pore's centre, and the nucleotide sequence of the template is successively deduced. Oxford Nanopore Technologies was set up to commercialize this technology in 2008 (REF. 98). It is noteworthy that nanopore sequencing is claimed to be compatible with direct sequencing of RNA.

whereas the ever-growing capacity of deep sequencing directly translates into higher data output without such problems.

Some probe-independent, tag-based methods such as SAGE[25] or CAGE[26] (serial or cap analysis of gene expression, respectively) have partly overcome these problems. The sequencing of small cDNA fragments ('tags') of 13–15 bp in length from the 5′ end (CAGE) or 3′ end (SAGE) of mRNAs provides more accurate quantification, as it enumerates individual transcripts in a digital manner, resulting in an almost unlimited dynamic range. Initial limitations of these tag-based approaches, stemming from the difficulty in unequivocally mapping such short sequences onto the genome, are now being overcome by the longer tags that are generated by SuperSAGE, which generates ~26 bp tags instead of the 13–15 bp in the conventional approach[27]. Tag-based techniques have been applied to the field of infection biology, particularly with eukaryotic pathogens (reviewed in REF. 28).

Nonetheless, none of the aforementioned probe- and tag-based methods can routinely discriminate between different mRNA isoforms, uncover unannotated non-coding RNA species or define transcript borders and splice junctions with high resolution. Background signals attributable to cross-hybridization severely limit the dynamic range of microarrays to ~3 logs[29]. Most importantly, however, sophisticated RNA sample preparation is required before a cDNA library can be generated. RNA-seq, which overcomes many of the above problems and is rapidly becoming the method of choice for studying transcriptomes, promises to facilitate a new type of dual gene expression analysis in pathogen and host.

What is RNA-seq?

RNA-seq is essentially massively parallel sequencing of RNA (or, in fact, the corresponding cDNA) and has heralded the second technical revolution in transcriptomics (reviewed in REF. 30). It is based on next-generation sequencing (NGS) platforms that were initially developed for high-throughput sequencing of genomic DNA (BOX 1). Typically, all the RNA molecules in a sample are reverse transcribed into cDNA, and depending on the platform to be used, the cDNA molecules may (amplification-based sequencing) or may not (single-molecule sequencing (SMS)) be amplified before deep sequencing. After the sequencing reaction has taken place, the obtained sequence stretches (reads) are mapped onto a reference genome to deduce the structure and/or expression state of any given transcript in the sample (FIG. 1).

In 2008, the first genome-wide RNA-seq experiments were carried out in mice and humans and yielded typically around 5–15 million reads per lane[31,32] (BOX 2). For mammalian gene expression profiling, reads from several lanes were often pooled, generating data sets of ~30 million–100 million reads[32–34]. Read lengths were typically short (25–32 bp). Subsequently, RNA-seq was applied to numerous bacterial species[35–38]. These studies were based on ~5 million reads per sample and increased read lengths of ~36–40 bp. Today, the latest machines can generate more than 1 billion reads of >150 bp in a single run; in-depth gene expression profiling in humans is sometimes based on >500 million reads[39]. However, the upper limits of sequencing resolution have not yet been reached. In particular, as we enter the age of 'third-generation sequencing' (for example, single-molecule sequencing and/or direct sequencing of RNA; BOX 1), both read number and read length are expected to increase even further.

In addition to potentially providing full-genome coverage, RNA-seq provides several advantages over microarray- and tag-based approaches. As is true for all sequence-based approaches, RNA-seq is a digital quantification method and thus has a high (and theoretically

Box 2 | **The current state of transcriptomics: how deep is 'deep'?**

It is of key importance to estimate the sequencing depth required to effectively sample the transcriptome of interest using RNA-seq. However, this depth varies greatly depending on the purpose of the proposed study (for example, gene expression versus gene discovery studies) and is not easily predicted because the level of transcriptional activity varies from genome to genome. In addition, sequencing-depth requirements are expected to change in the future, taking into account recent and ongoing advancements in deep-sequencing technologies (BOX 1). Some guide numbers that have been put forward over the years are given in Supplementary information S2 (table). The level of sequencing depth that is attainable for gene expression studies has already increased by two orders of magnitude since these studies were first reported in 2008.

Toung and colleagues[39] found that, for human B cells, a sequencing depth of ~25 million reads (of which ~80% could be mapped) resulted in detection of >80% of all expected transcripts. With 100 million reads, 90% of all expected transcripts were identified, but further increasing the read number had only a modest effect on the number of new transcripts detected (~1% additional transcripts per 100 million additional reads). However, 100 million reads was still too few to be able to accurately quantify expression. Rather, these and other authors recommend that 500 million–700 million reads are obtained for accurate gene expression quantification in mammals[99]. However, there is evidence that some commonly used algorithms are incompatible with high read numbers, thereby rendering an analysis of differential expression even more difficult when the sequencing depth exceeds a certain threshold[88]. In practice, the respective expression levels determine the required sequencing depth (in a simplistic example for mammalian systems, ~200 million sequences per sample are needed for very poorly expressed genes, and a minimum of only 8 million–10 million sequences per sample are required for abundant genes[100]).

infinite) dynamic range. Initial experiments suggested that the linear dynamic range for RNA-seq was at least four orders of magnitude[40], and it is now approaching six orders of magnitude[41], which is comparable to the upper limit of changes in gene expression in eukaryotic cells[42]. In addition, RNA-seq is extremely sensitive and can identify novel transcripts. For example, a single RNA-seq study of mouse myoblasts identified almost 4,000 previously unknown transcripts[43]. Furthermore, the single-nucleotide resolution provided by RNA-seq allows gene structure to be refined through accurate determination of transcript borders, alternative splicing and processing events.

An important step was the development of strand-specific RNA-seq (see below for details), which preserves information about the directionality of a transcript. This is especially important given the prevalence of non-coding and antisense transcripts throughout both the pathogen and host genomes, and for the characterization of operons in bacteria. These benefits suggest that RNA-seq has the potential to revolutionize the study of changes in gene expression during host–pathogen interactions, and that it is likely to provide the basis for new molecular insights into the mechanisms of pathogenesis and the corresponding immune response.

RNA-seq of pathogens. A major benefit of RNA-seq is that it provides an unbiased approach and can be used not only to detect which genes are expressed but also to provide high-resolution data on potentially transcribed sequences upstream and downstream of the annotated coding region. The pioneering RNA-seq studies generally described the extent and nature of the transcriptome of important pathogens in the absence of a host (reviewed in REFS 44,45), improving the annotation of the pathogen genomes, providing extensive information on transcription start sites (TSSs) and the location of the 5′ and 3′ UTRs of known genes, and reporting new ORFs as well as many hitherto unknown small non-coding RNAs. Furthermore, as RNA-seq provides transcript information without prior knowledge of mRNA sequences, it has also proved to be important for identifying co-regulated genes, therefore enabling the organization of pathogen genomes into operons. Likewise, genome annotation has been carried out for several eukaryotic pathogens, such as *Candida albicans*[46], *Trypanosoma brucei*[47,48] and *Plasmodium falciparum*[49]. More recently, comparative RNA-seq studies have identified differences in gene expression between closely related species (for example, pathogenic and non-pathogenic *Listeria* spp.[50]).

The regulation of gene expression involves multiple steps during which cellular transcripts can be modified or processed. This is exemplified by a study of *Helicobacter pylori*; this particular study introduced a novel differential RNA-seq approach to characterize the transcriptome of this pathogen, involving genome-wide discrimination of newly generated primary transcripts (most mRNAs and small non-coding RNAs, representing the TSSs in the sample) and processed RNA species (rRNAs and tRNAs)[38]. Differential RNA-seq involves selective pretreatment of isolated total RNA with an exonuclease that degrades processed RNAs (containing a 5′ monophosphate), but leaves primary mRNAs (with a 5′ triphosphate) undigested. Analyses of RNA isolated during exponential phase and during acid stress, which the bacterium normally experiences in the stomach of infected individuals, identified 1,900 TSSs and enabled the grouping of the ~1,700 protein-coding genes into 337 operons and a further 126 suboperons. *H. pylori* was also found to express a plethora of non-coding RNAs, including antisense RNAs to 46% of all genes.

The differential RNA-seq approach has also been applied to different *Chlamydia* spp.[51,52] to study the two important differentiated states (elementary bodies and reticulate bodies) that occur during infection by these organisms. As methods for genetic and molecular manipulation of these obligate intracellular pathogens were lacking until recently, the extensive annotation of the *C. trachomatis* and *C. pneumoniae* transcriptomes was an important step towards identifying stage-specific candidate genes that may be involved in the invasion and infection processes.

Clearly, the environment in an animal host will differ substantially from that in *in vitro* models, as will gene expression. Therefore, to understand how a pathogen copes with the complex within-host environment, *V. cholerae* bacteria were isolated from the caecum of infected rabbits or the intestine of infected mice and subjected to RNA-seq[53]. This identified a core set of 39 transcripts (out of 478) that showed altered expression in both animal models compared with in *in vitro* culture. Although the use of different NGS platforms for the two samples limits data interpretation, we note that

the core set contains well-characterized virulence factors such as cholera toxin and the type IV pilus TCP (toxin-co-regulated pilus). Furthermore, coupling of the RNA-seq data with metabolic data also identified several transcripts that may be affected by the conditions in the bacterium's specific niche. Altogether, RNA-seq is clearly living up to expectations as the method that will gradually replace microarrays for the transcriptomic analysis of pathogens.

RNA-seq of host transcriptomes. The first studies to apply RNA-seq to genome-wide trancriptomics started to uncover the sheer level of complexity of eukaryotic gene expression. Analyses of the transcriptomes from different mouse tissues (including liver, skeletal muscle and total brain)[33], different cell states of embryonic stem cells versus embryoid bodies[54], and various human cell types[31,39] detected novel transcripts; these included non-coding RNAs, many of which are normally expressed at low levels (and hence had not been detected by previous approaches). These analyses also led to increased accuracy in annotation of the 5′ and 3′ gene boundaries, and provided hints about the extent of alternative splicing and the potential number of different isoforms in various cell types.

The subsequent use of RNA-seq focused on assessing the degree of RNA processing and the types of RNA modification, both of which may also play an important part in the infection process. The specific analysis of RNA corresponding to alternative splicing sites — which can be directly identified by the presence of exon–exon boundaries that are not highly represented in many microarray studies — in a variety of human tissues demonstrated that ~95% of multi-exon genes undergo alternative splicing, resulting in ~100,000 intermediate to highly abundant splicing events[55,56]. Similarly, the global analysis of RNA-editing sites has revealed extensive modification of transcripts in humans. A comparison of human B cell RNA with the genome sequence identified an unexpectedly high number of editing sites (>22,600), including those in microRNAs (miRNAs) and other non-coding RNAs; some mRNAs contained more than ten edited sites[57]. The majority of these sites generate A-to-G substitutions and occur within introns, although a considerable number were also identified in the 3′ UTRs of genes; further analysis of the 3′ UTR sites suggested that 20% of them alter putative target sites for miRNAs. This suggests that RNA editing has a more prominent role in regulating human gene expression than was previously appreciated and that RNA-seq could be used to elucidate its role in infection.

In relation to infection biology, there have been few published RNA-seq studies concerning the response of the mammalian host cell to infection. Exceptions include a study into the eukaryotic miRNA response to bacteria (for example, on infection of macrophages or HeLa cells by *Salmonella enterica*) upon the enrichment of the miRNA fraction[58]. However, the routine application of RNA-seq and especially dual RNA-seq to the response of the mammalian host promises to provide exciting new insights into the infection process.

microRNA
A short (~22 nucleotide) processed RNA that guides post-transcriptional repression of mRNAs in animals and plants.

Towards dual RNA-seq

To date, transcriptomic experiments have predominantly focused on either the host or the pathogen. A deeper understanding of the infection process will require the simultaneous analysis of both interaction partners; although this is yet to be achieved genome wide with RNA-seq, several transcriptomic studies have used either tag- or hybridization-based methods to simultaneously detect the response of both the host and the pathogen. For example, separate host and pathogen microarrays have been used to simultaneously study mRNA changes in both *Aspergillus fumigatus* and human airway epithelial cells[59]. Specialized microarrays containing both pathogen and host gene probes (as described in REF. 3) identified the changes in gene expression that are associated with virulent *Escherichia coli* CP9 infection in a mouse model of local infection[24], and also characterized the response of different mouse tissues to infection by *Plasmodium berghei*[60].

The application of probe-independent methods such as SAGE to *Leishmania major*-infected human macrophages enabled the simultaneous characterization of the parasite and host cell transcriptomes at distinct developmental stages of the parasite[61]. Approximately half of the uniquely mapped tags accounted for the host, and the other half, for the parasite (3,814 tags versus 3,666 tags). On the host side, multiple key immune response genes (such as cytokines) showed a differential expression pattern, whereas many parasite genes that were preferentially expressed during the intracellular-development stage appeared to be among the most highly regulated. Of note, initial work in plants used SuperSAGE, resulting in a greater ability to differentiate transcripts from the two genomes. Taking this approach with rice leaves (*Oryza sativa*) infected with the fungal pathogen *Magnaporthe grisea*, ~0.6% of a total ~12,000 analysed tags mapped exclusively to the pathogen's genome; curiously, half of these tags corresponded to the hydrophobin gene, which is required for appressorium formation before penetration of the host[62]. The same approach was able to detect changes in gene expression during the hypersensitive response caused by the *Phytophthora infestans* elicitin (INF1) in the non-model organism *Nicotiana benthamiana*[62]. Although these studies provide some information regarding the dynamic changes in both transcriptomes, they are unlikely to provide the full picture. For example, the SuperSAGE method identified only a small fraction (~0.6%) of tags from the pathogen[62]. In addition, note that CAGE, SAGE and SuperSAGE base their profiling on the partial sequencing of transcripts, from the 3′ end[25,62] or 5′ end[26], and are therefore intrinsically restricted in their information output.

Given the increased sensitivity and depth of sequencing that is now available, RNA-seq appears to be evermore promising for the study of infected mammalian cells (FIG. 2). Indeed, a recent study used RNA-seq for dual transcriptomics of the fungus *C. albicans* and its host mouse dendritic cells[63]. Although the sequencing depth was low (~120 million reads of 36 bp in total for five time points from rRNA-depleted libraries), a

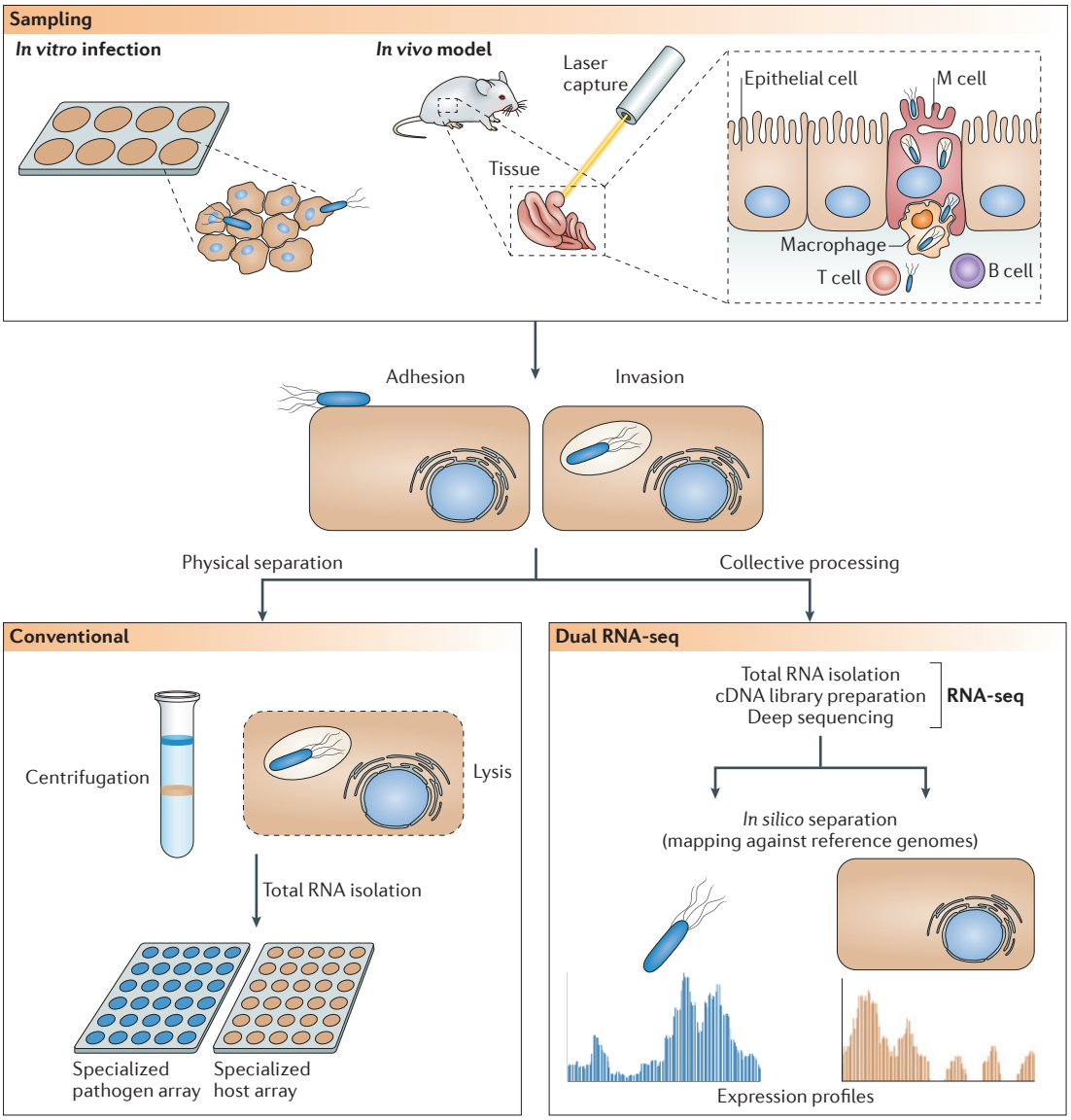


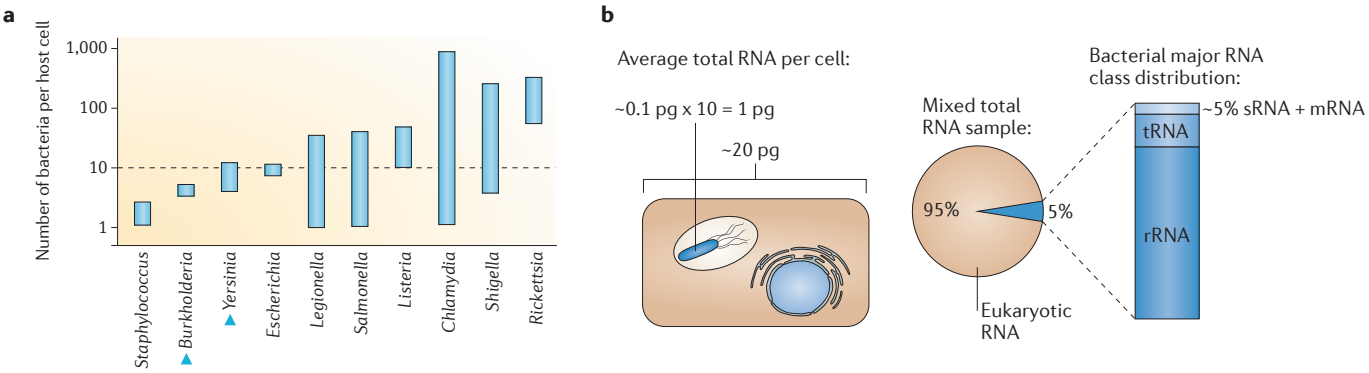
Figure 2 | **A paradigm shift in parallel host–pathogen transcriptomics.** The substantial host background levels and the potential for cross-hybridization when using probe-based methods (such as microarrays) typically require that the host and pathogen be physically separated.The launch of a species-independent platform such as RNA sequencing (RNA-seq) is therefore highly promising, as it could enable the different organisms to be analysed collectively. In this approach, the discrimination between host and pathogen would take place only at the bioinformatics stage.

handful of candidate genes from an interspecies regulatory network could be inferred through complementation with previous knowledge from the literature. However, the scope of this study was to generate an interspecies computational model of molecular host–pathogen interactions, rather than an in-depth characterization of the global response to infection, so the authors focused on previously annotated virulence genes in the pathogen and known immunity-associated genes in the host when selecting candidates. Thus, although this study represents the first successful application of dual RNA-seq to a eukaryotic interaction model, in order to obtain a more complete and unbiased picture of infection — especially for models that are based on bacterial pathogens — it is vital to sequence more deeply. However, as discussed below, depth is only one aspect to be taken into account when planning a host–pathogen dual RNA-seq experiment.

Dual RNA-seq: a gedankenexperiment

To establish the feasibility of dual RNA-seq and to estimate the sequencing depth that would be required for the accurate representation of both the bacterial pathogen and the mammalian host, one needs to consider potential limiting factors.

Different RNA contents. The human genome is 3,000 Mb in size, whereas the genomes of typical pathogens such as *E. coli* or *S. enterica* are ~5 Mb. This size difference translates into different amounts of cellular RNA; eukaryotic



c

Separate RNA-seq

Organism	Total RNA	Enriched RNA
Pathogen	4–10 million reads	2–5 million reads
Host	>2,000 million reads	≥100 million reads

Dual RNA-seq

Organism	Total RNA	Enriched RNA
Pathogen	≥ 400 million reads	≥200 million reads
Host	>2,000 million reads	≥100 million reads

Figure 3 | **Estimation of the minimal sequencing depth required for dual host–pathogen RNA-seq.** **a** | Overview of the reported copy numbers of selected adhesive or invasive model pathogens; an approximation is ten bacteria per host cell. Blue triangles indicate adherent pathogens, and other pathogens are invasive. **b** | Whereas a single mammalian cell typically contains around 20 pg of total RNA, the bacterial cellular RNA content does not exceed a few hundred femtograms. Given an average copy number of ten bacterial cells per host cell, the ratio of bacterial to eukaryotic RNA in samples derived from infected cells was calculated as ~1:20. **c** | The required sequencing depth for dual RNA-seq. The calculation is based on a total of 1 million non-rRNA reads being derived from the pathogen[35–37] and a minimum of 100 million poly(A)+ reads being derived from a human host[39]. Poly(A)+ or rRNA-depleted libraries are referred to as enriched, whereas libraries for which neither poly(A)-based enrichment nor rRNA depletion are carried out are termed total. Note that for dual RNA-seq more reads are required for the pathogen than for the host on enrichment of RNA owing to the lower reported efficiency of rRNA depletion in bacteria. For details, see main text. sRNA, small non-coding RNA.

cells contain in the range of 10–20 pg of total RNA, which is ~100–200 times the ~0.1 pg present in a bacterial cell. In practice, this excess is reduced, as in most cases a single infected host cell will contain multiple bacteria. The intracellular copy number for a pathogen varies from species to species (FIG. 3; see Supplementary information S3 (table)), but assuming that on average each infected mammalian cell is associated with ten bacteria, the relative difference in total RNA content will be decreased to 10–20-fold in most infection models.

Extraction of total RNA. The first step in an RNA-seq experiment is to isolate the total RNA, which should be done as rapidly as possible. However, this may not be practical under certain conditions, such as when infected samples need to undergo time-consuming cell sorting. Thus, cells must be fixed to maintain transcriptome integrity during these steps, but fixation may cause partial fragmentation of the RNA[64]. This can be tolerated, especially when the isolates will be further fragmented during cDNA library generation (see below). However, whether differential fragmentation of bacterial and eukaryotic RNA introduces a bias in cDNA synthesis or sequencing is currently unknown.

It is also important to remove genomic DNA in order to reduce sources of background noise. Nowadays, however, many cDNA library preparation protocols ligate sequencing-specific linkers directly to the RNA molecule on a routine basis, thereby depleting genomic DNA indirectly and obviating the need for a rigorous DNase treatment. Therefore, in a scenario in which only minute amounts of RNA can be isolated, DNase treatment may be omitted.

Enrichment of specific RNA species. One tremendous challenge for dual RNA-seq is the heterogeneity of the RNA that is present in the eukaryotic and bacterial transcriptomes. Although the core transcript classes (rRNA, tRNA and mRNA) are present in both domains, their specific properties may differ (TABLE 1). There are also several RNA species that are specific to eukaryotes, such as miRNAs[65], long non-coding RNAs (lncRNAs)[66], small nuclear RNAs (snRNAs)[67] and small nucleolar RNAs (snoRNAs)[68]; reciprocally, bacteria contain many small non-coding RNAs[69].

The ubiquitous rRNA, which represents the most abundant class of RNA in both eukaryotic and bacterial cells, provides little additional information. Depletion of rRNA is one way to increase information content in the sample and is discussed below. As another method, early studies directly enriched for certain RNAs of interest, such as polyadenylated, non-polyadenylated or small non-coding RNAs. However, it is important to bear in mind that enrichment for specific transcripts will, by

Long non-coding RNAs
Heterogeneous non-coding RNAs (> 200 nucleotides) that lack protein-coding capability and are found in eukaryotes.

Small nuclear RNAs
Short RNAs that are involved in precursor mRNA processing.

Small nucleolar RNAs
RNAs that typically guide ribose methylation and pseudouridylation in other RNA molecules.

Table 1 | **Major RNA classes in bacteria and eukaryotes**

Transcript class	Proportion	Subcellular localization	Transcript features
Bacterial cell			
rRNA	~80%	Not applicable	• Cleaved from precursor transcripts • 120 nt (5S), 1,500 nt (16S), 2,900 nt (23S)
tRNA	14–15%	Not applicable	• Cleaved from precursor transcripts • 75–95 nt
mRNA	4–5%	Not applicable	• Polycistronic • Uncapped • Variable polyadenylation state (<50 As)
tmRNA	<1%	Not applicable	• Processing similar to that of tRNA • 230–400 nt
sRNA	Varies	Not applicable	• Processed or unprocessed • ~50–300 nt
Eukaryotic cell			
rRNA	~80%	Cytoplasm	• Processed or unprocessed • 120 nt (5S), 160 nt (5.8S), 1,874 nt (18S), 4,718 nt (28S)
tRNA	~15%	Cytoplasm	• Cleaved from precursor transcripts • 75–95 nt
snRNA	~5%	Nucleus	• U-rich • ~150 nt
snoRNA		Nucleolus	• U-rich • ~60–300 nt • Two major classes (H/ACA and C/D box)
scaRNA		Cajal bodies	• Similar to snRNAs
mRNA		Cytoplasm, nucleus	• Processed from pre-mRNA via splicing • 5′ capped • 3′ polyadenylated (~250 As)
miRNA		Cytoplasm, nucleus	• Processed • 21–23 nt
siRNA		Cytoplasm, nucleus	• Processed • 20–25 nt
piRNA		Cytoplasm, nucleus	• Processed • 24–31 nt
lncRNA, lincRNA		Nucleus, cytoplasm, subcompartments	• Heterogenic

lincRNA, long intergenic non-coding RNA; lncRNA, long non-coding RNA; miRNA, microRNA; nt, nucleotides; piRNA, PIWI-interacting RNA; sRNA, small non-coding RNA; scaRNA, small Cajal body-specific RNA; siRNA, small interfering RNA; snRNA, small nuclear RNA; snoRNA, small nucleolar RNA; tmRNA, transfer–messenger RNA.

definition, result in a loss of information from the global transcriptome, and this may be particularly relevant in dual RNA-seq, as many RNA species differ substantially in structure and modifications between eukaryotes and bacteria. For example, eukaryotic mRNAs are transcribed from monocistronic genes, and they also acquire a 5′ methylguanine cap and a poly(A) tail, both of which are important for translation and increase mRNA stability. Eukaryotic mRNAs have half-lives in the range of many hours[70,71]. By contrast, bacterial mRNAs are often transcribed from polycistronic genes; they do not acquire a 5′ cap structure, but contain a 5′ triphosphate modification; and they rarely contain a poly(A) tail, but when they do, it serves as a tag for degradation. Moreover, bacterial mRNAs have a much shorter half-life than their eukaryotic counterparts, in the range of a few minutes[6]. Consequently, selection of the polyadenylated fraction would isolate transcripts with very different fates in different organisms: stable transcripts in the eukaryotic host versus transcripts undergoing degradation in the bacterial pathogen. Therefore, enrichment for specific RNAs is not recommended when carrying out dual RNA-seq. Overcoming the transcript heterogeneity within and between organisms, and the resulting library complexity, to provide a complete and reliable view of the transcriptional landscape during infection will be a major challenge for dual RNA-seq.

Depletion of rRNA. Both the bacterial and eukaryotic cellular RNA pools consist predominantly (>80%) of rRNA, whereas mRNA constitutes only a minor fraction (<5%). Many transcriptomic studies have tried to increase the information content by depleting rRNA. This may have its limitations, however, so it is important to decide whether this step is necessary before generating the cDNA library. Numerous protocols and commercial kits for rRNA depletion exist. Some kits actively deplete either the bacterial or eukaryotic

rRNAs with the help of sequence-specific oligonucleotides bound to magnetic beads, whereas rRNA can also be removed indirectly by reverse transcribing the sample using a pool of 'not-so-random' primers that is devoid of annealing sites in rRNA[72]. However, commercial rRNA removal kits often give away little information of the exact procedure and components. Furthermore, these kits frequently have different efficiencies[73–75] and, as a result, may add biases. With respect to dual RNA-seq, one has to further consider that eukaryotic and bacterial rRNAs need to be treated in succession and with different kits. However, each individual depletion step not only removes rRNA but also decreases the final yield of non-rRNA transcripts.

Instead, increasing the read numbers for non-rRNA transcripts can be achieved simply by increasing the overall sequencing depth. Importantly, both the pioneering differential RNA-seq study of *H. pylori*[38] and the follow-up studies in many other organisms abstained from rRNA depletion but still achieved sufficient sequencing coverage to determine gene expression changes. For dual RNA-seq experiments using samples that have not been depleted for rRNA, the sequence coverage required is currently at the upper limit of sequencing capacities (FIG. 3); nonetheless, ideally rRNA should not be depleted.

Post-transcriptional RNA modifications. An RNA transcript can deviate from its genomic DNA template as a result of post-transcriptional processing events, splicing, polyadenylation or RNA editing[76]. Moreover, >100 naturally occurring RNA modifications have been described in all three major RNA species (rRNA, tRNA and mRNA) and several minor species[77]. Owing to differences in the chemical structure of modified ribonucleosides compared to the four standard ribonucleosides[78], the former can have a negative impact on cDNA synthesis[79]. Chemical RNA modifications can not only cause reverse transcriptase stalling, but also leave signatures in the form of mismatches and conspicuous patterns in the resulting deep-sequencing data[80], or generally hamper the mapping process. With respect to dual RNA-seq, mismatches resulting from RNA modifications could, in rare circumstances, be erroneously mapped onto the non-parental genome. It is worth mentioning that, when regarded carefully, such mismatches can in fact be useful sources of information that can help towards the identification of sites with possible RNA base modifications[81,82].

cDNA library construction. The method required for construction of the cDNA library is dependent on the sequencing platform to be used. However, there are three recommendations that hold true for any of the currently popular platforms. The first relates to sample fragmentation. As the generation of more uniform fragments is desired when mixed samples are analysed, the RNA should be fragmented. Fragmentation can be achieved mechanically, chemically or enzymatically, but it should be done at the RNA level, because this typically results in more even coverage, rather than at the cDNA level, which has a strong bias towards the 3′ end[30]. The second recommendation is that the cDNA library protocol preserves strand information, to allow the identification of antisense transcription, for example. Several options exist to ensure strand specificity of RNA-seq (reviewed in REF. 83). One is the use of a deoxy-UTP (dUTP) second-strand-marking protocol[84,85]: actinomycin D is added to the reverse transcription reaction to specifically inhibit DNA-dependent DNA synthesis (and thus reduce the effect of genomic contamination); dUTP is then incorporated into the second strand of cDNA; and this strand undergoes subsequent selective destruction. In bacteria, strand specificity has also been achieved by omitting second-strand synthesis[86] or by ligating 5′ adaptors before cDNA synthesis[38]. In principle, any strategy is compatible with both eukaryotic and bacterial RNA and should therefore work for dual RNA-seq. The third recommendation concerns the addition of bar-code sequences for multiplexing. These sequences are typically integrated either at the level of RNA by direct ligation or at the level of cDNA as part of the PCR primer sequence. However, there is evidence that the enzymes usually used for adaptor ligation have a strong bias for specific sequences[87]. As the sequence biases for eukaryotic RNA–bacterial RNA mixtures are expected to exceed those for homogeneous samples, we strongly recommend introducing bar-codes during PCR amplification and not during adaptor ligation.

Reproducibility and internal references. Another issue that is frequently ignored for RNA-seq experiments is that of replicates. The technical reproducibility for RNA-seq has been claimed to be high[32,33] but should be checked for each data set[88], especially when the coverage is low[89]. To provide a reference that allows for inter-experimental comparisons, biological samples should always be supplemented with artificially synthesized or *in vitro*-transcribed RNAs (so-called spike-in RNAs)[90]. The sequence of any spike-in RNA must be confirmed bioinformatically to be absent from all the genomes under investigation. Spike-in RNA features such as length, GC content or concentration have to be empirically determined for any assay system. Just as for traditional methods of assessing gene expression, the large number of data points generated by RNA-seq means that a high degree of accuracy is needed to reduce the false-discovery rate. Thus, biological replicates are absolutely essential for robust results.

Estimating sequencing depth for dual RNA-seq. In both bacteria and eukaryotes, mRNA and regulatory non-coding RNAs constitute around 5% of the total RNA. As outlined above, assuming that there are ten intracellular bacteria per host cell, the ratio of bacterial to eukaryotic RNA in the infected cell is ~1:20, meaning that bacterial mRNAs and small non-coding RNAs correspond to ~0.25% of the total RNA sample (FIG. 3). Previous bacterial RNA-seq studies have suggested that a minimum of 2–5 million reads from an rRNA-depleted library are required for accurate coverage[35–37]. In all of these studies,

Bar-code sequences
Short unique sequence tags (~4–6 nucleotides) that are incorporated into cDNA fragments and used to tag a specific sequence as belonging to a particular cDNA library.

rRNA was removed using the MICROBExpress kit (Life Technologies). However, even with the enrichment of non-rRNA species provided by this protocol, more than half of the bacterial reads correspond to ribosomal transcripts (for example, see REFS 36,75). This suggests that, using the assumption that the minimal requirement is 2 million reads, ≥1 million bacterial non-rRNA reads are required to detect pathogen gene expression in the host background. As the bacterial small non-coding RNA and mRNA make up 0.25% of the total RNA, this means that 400 million total reads would be required to attain the 1 million small non-coding RNA and mRNA reads from the pathogen. However, for reasons that are not quite understood, rRNAs and tRNAs have lower conversion rates into cDNA than do mRNAs and other RNA species[38]; although this again may depend on the studied pathogen and the cDNA protocol, it would be in favour of covering the more informative RNA species when sequencing samples without rRNA depletion.

In the case of the host, human gene detection studies require ~100 million rRNA-depleted reads, and for accurate gene expression quantification this increases to ~500 million reads derived from a poly(A)+ library from human B cells[39]. The standard protocol for cDNA library construction in mammals[33] enriches for mRNAs by using oligo(dT)-conjugated magnetic beads (Life Technologies), following which the enriched RNA fraction is essentially free of rRNA. Likewise, eukaryotic rRNA removal kits are highly efficient[73,74]. When neither poly(A) enrichment nor rRNA depletion are carried out before sequencing, using the assumption that cellular housekeeping non-coding RNA classes (rRNA and tRNA) together constitute ~95% of the total RNA (so mRNAs and regulatory non-coding RNAs, at 5% of the total, are diluted by 20-fold in comparison to when rRNA and tRNAs are depleted), the number of reads required increases by a factor of 20. Given this increase in the number of reads required, achieving the same sequencing depth as that in the above study of human B cells[39] will require ≥2,000 million reads for gene detection (BOX 2; FIG. 3).

In conclusion, an estimated minimum of about 2,000 million reads from total RNA and 200 million reads from rRNA-depleted samples seems to be required to simultaneously monitor gene expression in both host and pathogen. In the described example, depending on whether or not rRNA is depleted, the different rRNA removal efficiencies for pathogen and host sequences mean that obtaining sufficient sequence coverage of either transcriptome is a limiting parameter. Omitting the rRNA depletion step would require two dedicated flow cells on the currently popular HiSeq 2000 machine (Illumina), with the sequencing capacity available to date. By contrast, when rRNA is depleted, a minimum of two lanes seems to be required, albeit with the above-mentioned caveats. Therefore, although dual RNA-seq on total RNA would be costly, from a technical perspective it appears to already be feasible. Of note, the above estimates are conservative, as they are based on read lengths of 36 bp for bacteria[35] and 50 bp for humans[39] and do not take into account the fact that longer read lengths are already routine.

Although these calculations describe the best-case scenario and rely on reported sequencing-depth numbers, which vary widely depending on the model system or cDNA library generation protocol used, they provide rough estimates for the feasibility of the dual RNA-seq approach. For instance, the selection of bacterial strain and host cell type defines the magnitude of the pathogen–host transcriptome, the number of individual pathogens per host cell and, thus, the fraction of bacterial RNA within the mixed sample. Furthermore, the actual demands and aims of an experiment determine the required depth (BOX 2): in studies designed to refine transcript boundaries and splice junctions or to accurately quantify gene expression on either side, considerably more reads will be required than in experiments that merely aim to detect known transcripts.

Dual RNA-seq of pathogen and host appears to be within practical reach. Mixed RNA samples can be processed collectively and eventually converted into uniform-length cDNA fragments. The discrimination between host and pathogen expression profiles occurs only *in silico* by simultaneous read mapping onto the respective reference genomes. Ideally, sequencing power alone would determine the coverage of the genomes and thus define the amount of information that can be extrapolated from the data set. Today, when the goal is not simply detecting host or pathogen transcripts but rather carrying out quantitative expression profiling, mixed cDNA libraries would have to be sequenced repeatedly (that is, over several lanes), and the resulting data subsets pooled. However, as deep-sequencing technologies continue to improve, dual RNA-seq is likely to become the gold standard to investigate host–pathogen interactions.

Conclusions and perspectives

The application of RNA-seq has provided an opportunity for unparalleled access to the transcriptomes of hosts and pathogens, increasing the annotation of genomes and the detection of new genes. It has further enabled a switch from hypothesis-based methods for investigating the mechanisms involved in infection to an unbiased, discovery-based approach that has a higher likelihood of identifying novel principles. For example, analysis of the transcriptomes of pathogens under different environmental conditions or of cells with different genetic backgrounds has begun to identify new virulence genes and the factors involved in their regulation. However, the future development of dual RNA-seq to simultaneously determine the transcriptomes of host and pathogen has the potential to provide further insights into the host–pathogen interaction that currently cannot be obtained by sequencing of the individual players. It is conceivable that the close temporal congruence in determining the host and pathogen transcriptomes will reveal not only new molecular strategies that lead to infection or the clearance of the pathogen, but also the response of each cell type to the deleterious presence of the other cell. Dual RNA-seq also promises to enable future research to move away from revealing the 'average' cellular response to infections, as it is likely that different cell types have

Box 3 | **Dual RNA-seq of single cells**

Neither cell cultures nor tissues are homogenous, and the infection process is likewise highly diverse. Therefore, infection studies at the population level simply reflect the average behaviour of all population constituents. To resolve the specific and frequently greatly differing responses of individual cells, gene expression would ideally be carried out at the single-cell level.

Single-cell RNA-seq has been achieved for eukaryotic cells alone[101–107] and seems to be within reach for bacterial cells[108]. For eukaryotic single-cell RNA-seq, ~5–10 pg of total RNA was obtained from a single mammalian cell[103,106]; reverse transcription and PCR-based amplification to 20–200 ng of cDNA was sufficient for RNA-seq. By contrast, a single *Burkholderia* sp. cell yielded <2 pg of total RNA[108]. To overcome this limitation, the researchers used a method termed total transcript amplification, in which circularized single-stranded DNA molecules were used to produce vast amounts (~10 μg) of double-stranded DNA.

The read depth requirements for single-cell analyses seem to differ from those for population RNA-seq experiments (BOX 2). One study estimates that ≤40 million sequencing reads per cell covers an entire mammalian transcriptome, including the detection of new genes and the refinement of known genes, whereas 100 million reads would be ideal for studies investigating alternative splicing[102]. Another study aiming to generate a high-resolution map of embryonic stem cells sampled ~250,000 reads per cell, adding up to 10 million reads in total. The authors conclude that sequencing of more individual cells at moderate depth increases the accuracy of the aggregate data[106].

Nanodevices for manipulating single cells are already available. Single cells have been collected under a dissection microscope with mouth pipettes[101,102,104,105], with the help of a semi-automated cell picker and commercially available cell capture plates[106], by laser capture[108] or manually[107]. New microfluidics devices for single-cell analyses (reviewed in REF. 109) — including proteinous picolitre-scale microcavities referred to as lobster traps[110] — as well as laser capture microdissection (reviewed in REF. 111) promise to lift single-cell transcriptomics to an as-yet-unreached level.

Thus, the present limitations for single-cell dual RNA-seq are primarily associated with the processing of mixed RNA samples and the generation of high-quality cDNA libraries. For instance, single-cell RNA-seq in eukaryotic cells[101,102,104–106] used oligo(dT) primers for reverse transcription, to indirectly deplete rRNA. With bacterial RNA, however, this would enrich RNA decay intermediates (see main text). The minute amount of bacterial RNA material also precludes the depletion of rRNA. A reasonable solution to this may be 'not-so-random' primers (see main text), although this remains to be tested. To circumvent length biases that occur during PCR amplification, it would be ideal to omit the amplification step and to directly combine single-cell analyses with single-molecule sequencing of cDNA or even with direct sequencing of RNA, which would additionally prevent the influence of conversion efficiency (that is, the efficiency of converting RNA into sequence reads). In contrast to RNA-seq from bulk samples, and owing to the fact that most genes are expressed in the range of about ten copies per cell[112], single-cell RNA-seq can be greatly affected by the conversion efficiency, as many transcripts might easily be lost completely. At present, the RNA amounts that are required for single-molecule sequencing of cDNA are still on a microgram scale[113], exceeding the RNA content of single cells by several orders of magnitude. Likewise, direct sequencing of RNA has not been achieved for single cells. However, assuming that progress will be made at a pace comparable to that of the past few years, single-molecule sequencing of cDNA or even RNA in single cells may be feasible in the near future.

different susceptibilities to, and modes of dealing with, pathogens and infection; likewise, there may be heterogeneity within the bacterial population with respect to pathogenicity and virulence. Each pathogen is likely to interact with a variety of cell types, or with cells that are of the same type but in different cellular states owing to prior cellular interactions, making dual RNA-seq at the single-cell level (BOX 3) an important ultimate goal. The synchronized determination of transcriptomes will enable us to assess the importance of stochasticity and cell type-specific interactions, a feat that would not be possible by determining the transcriptome of either host or pathogen alone.

Dual RNA-seq will require high sequencing depth in order to provide accurate representations of the host and pathogen genomes; this is highly likely to be attainable in the future given the potential for near-infinite sequencing power. As outlined above, the technical requirements to achieve this are about to be fulfilled, and currently dual RNA-seq on the population level appears to be costly but feasible. The latest sequencing platforms can generate an output of up to several hundred gigabases per experimental run, suggesting that the ~200–2,000 million reads required for dual RNA-seq can be achieved. As sequencing depth increases, the most important technical issue may be not the acquisition of the data, but perhaps its processing and storage (reviewed in REF. 91). The files generated by RNA-seq are several orders of magnitude larger than those from arrays[92]. This means that computational processes become time consuming simply owing to the shear mass of data.

Ongoing progress and improvement in third-generation sequencing technologies will further facilitate dual transcriptomics. In particular, nanopore sequencing is highly promising, as it is potentially compatible with direct RNA sequencing without cDNA intermediates. These devices can apparently generate long and accurate reads and, according to the manufacturer, will be comparable to current platforms from a financial perspective but at the same time will be considerably faster (~20–400 bases per second). As a consequence of this ever-growing sensitivity, we will be able to resolve gene expression events during infections in much finer detail by scaling down the number of cells required for assessment. Ultimately, single-cell RNA-seq of pathogen-containing eukaryotic cells, extracted from infected tissue, should become technically feasible.

1. Jenner, R. G. & Young, R. A. Insights into host responses against pathogens from transcriptional profiling. *Nature Rev. Microbiol.* **3**, 281–294 (2005).
2. Hossain, H., Tchatalbachev, S. & Chakraborty, T. Host gene expression profiling in pathogen-host interactions. *Curr. Opin. Immunol.* **18**, 422–429 (2006).
3. Rappuoli, R. Pushing the limits of cellular microbiology: microarrays to study bacteria–host cell intimate contacts. *Proc. Natl Acad. Sci. USA* **97**, 13467–13469 (2000).
4. Fodor, S. P. *et al.* Multiplexed biochemical assays with biological chips. *Nature* **364**, 555–556 (1993).
5. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470 (1995).
6. Selinger, D. W., Saxena, R. M., Cheung, K. J., Church, G. M. & Rosenow, C. Global RNA half-life analysis in *Escherichia coli* reveals positional patterns of transcript degradation. *Genome Res.* **13**, 216–223 (2003).
7. Yamada, K. *et al.* Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science* **302**, 842–846 (2003).
8. Bertone, P. *et al.* Global identification of human transcribed sequences with genome tiling arrays. *Science* **306**, 2242–2246 (2004).
9. Merrell, D. S. *et al.* Host-induced epidemic spread of the cholera bacterium. *Nature* **417**, 642–645 (2002).
10. Revel, A. T., Talaat, A. M. & Norgard, M. V. DNA microarray analysis of differential gene expression in *Borrelia burgdorferi*, the Lyme disease spirochete. *Proc. Natl Acad. Sci. USA* **99**, 1562–1567 (2002).
11. Belland, R. J. *et al.* Genomic transcriptional profiling of the developmental cycle of *Chlamydia trachomatis*. *Proc. Natl Acad. Sci. USA* **100**, 8478–8483 (2003).
12. Maurer, A. P., Mehlitz, A., Mollenkopf, H. J. & Meyer, T. F. Gene expression profiles of *Chlamydophila pneumoniae* during the developmental cycle and iron depletion-mediated persistence. *PLoS Pathog.* **3**, e83 (2007).
13. Eriksson, S., Lucchini, S., Thompson, A., Rhen, M. & Hinton, J. C. Unravelling the biology of macrophage infection by gene expression profiling of intracellular *Salmonella enterica*. *Mol. Microbiol.* **47**, 103–118 (2003).
14. Hautefort, I. *et al.* During infection of epithelial cells *Salmonella enterica* serovar Typhimurium undergoes a time-dependent transcriptional adaptation that results in simultaneous expression of three type 3 secretion systems. *Cell. Microbiol.* **10**, 958–984 (2008).
15. Toledo-Arana, A. *et al.* The *Listeria* transcriptional landscape from saprophytism to virulence. *Nature* **459**, 950–956 (2009).
16. Perez, N. *et al.* A genome-wide analysis of small regulatory RNAs in the human pathogen group A *Streptococcus*. *PLoS ONE* **4**, e7668 (2009).
17. Kumar, R. *et al.* Identification of novel non-coding small RNAs from *Streptococcus pneumoniae* TIGR4 using high-resolution genome tiling arrays. *BMC Genomics* **11**, 350 (2010).
18. Zheng, X. *et al.* Identification of genes and genomic islands correlated with high pathogenicity in *Streptococcus suis* using whole genome tiling microarrays. *PLoS ONE* **6**, e17987 (2011).
19. Der, S. D., Zhou, A., Williams, B. R. & Silverman, R. H. Identification of genes differentially regulated by interferon α, β, or γ using oligonucleotide arrays. *Proc. Natl Acad. Sci. USA* **95**, 15623–15628 (1998).
20. Zhu, H., Cong, J. P., Mamtora, G., Gingeras, T. & Shenk, T. Cellular gene expression altered by human cytomegalovirus: global monitoring with oligonucleotide arrays. *Proc. Natl Acad. Sci. USA* **95**, 14470–14475 (1998).
21. Hedges, J. F., Lubick, K. J. & Jutila, M. A. γδ T cells respond directly to pathogen-associated molecular patterns. *J. Immunol.* **174**, 6045–6053 (2005).
22. Kerns, H. M., Jutila, M. A. & Hedges, J. F. The distinct response of γδ T cells to the Nod2 agonist muramyl dipeptide. *Cell. Immunol.* **257**, 38–43 (2009).
23. Tross, D., Petrenko, L., Klaschik, S., Zhu, Q. & Klinman, D. M. Global changes in gene expression and synergistic interactions induced by TLR9 and TLR3. *Mol. Immunol.* **46**, 2557–2564 (2009).
24. Motley, S. T. *et al.* Simultaneous analysis of host and pathogen interactions during an *in vivo* infection reveals local induction of host acute phase response proteins, a novel bacterial stress response, and evidence of a host-imposed metal ion limited environment. *Cell. Microbiol.* **6**, 849–865 (2004).
25. Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. Serial analysis of gene expression. *Science* **270**, 484–487 (1995).
26. Shiraki, T. *et al.* Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl Acad. Sci. USA* **100**, 15776–15781 (2003).
27. Matsumura, H., Kruger, D. H., Kahl, G. & Terauchi, R. SuperSAGE: a modern platform for genome-wide quantitative transcript profiling. *Curr. Pharm. Biotechnol.* **9**, 368–374 (2008).
28. Kronstad, J. W. Serial analysis of gene expression in eukaryotic pathogens. *Infect. Disord. Drug Targets* **6**, 281–297 (2006).
29. Canales, R. D. *et al.* Evaluation of DNA microarray results with quantitative gene expression platforms. *Nature Biotech.* **24**, 1115–1122 (2006).
30. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Rev. Genet.* **10**, 57–63 (2009).
31. Sultan, M. *et al.* A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science* **321**, 956–960 (2008).
32. Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M. & Gilad, Y. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* **18**, 1509–1517 (2008).
33. Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L. & Wold, B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5**, 621–628 (2008).
34. Morin, R. *et al.* Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. *Biotechniques* **45**, 81–94 (2008).
35. 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).
36. 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).
37. Oliver, H. F. *et al.* Deep RNA sequencing of *L. monocytogenes* reveals overlapping and extensive stationary phase and sigma B-dependent transcriptomes, including multiple highly transcribed noncoding RNAs. *BMC Genomics* **10**, 641 (2009).
38. Sharma, C. M. *et al.* The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature* **464**, 250–255 (2010).
A good example of how to use strand-specific RNA-seq of total RNA to analyse gene expression and annotate a bacterial transcriptome with respect to both coding and non-coding information and operon structure.
39. Toung, J. M., Morley, M., Li, M. & Cheung, V. G. RNA-sequence analysis of human B-cells. *Genome Res.* **21**, 991–998 (2011).
An in-depth transcriptomic study on human B cells that defines the coverage requirements for different biological applications such as gene discovery and expression quantification.
40. Wilhelm, B. T. & Landry, J. R. RNA-Seq—quantitative measurement of expression through massively parallel RNA-sequencing. *Methods* **48**, 249–257 (2009).
41. Xiong, J. *et al.* Transcriptome analysis of the model protozoan, *Tetrahymena thermophila*, using deep RNA sequencing. *PLoS ONE* **7**, e30630 (2012).
42. Holland, M. J. Transcript abundance in yeast varies over six orders of magnitude. *J. Biol. Chem.* **277**, 14363–14366 (2002).
43. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotech.* **28**, 511–515 (2010).
44. Croucher, N. J. & Thomson, N. R. Studying bacterial transcriptomes using RNA-seq. *Curr. Opin. Microbiol.* **13**, 619–624 (2010).
45. Filiatrault, M. J. Progress in prokaryotic transcriptomics. *Curr. Opin. Microbiol.* **14**, 579–586 (2011).
46. Bruno, V. M. *et al.* Comprehensive annotation of the transcriptome of the human fungal pathogen *Candida albicans* using RNA-seq. *Genome Res.* **20**, 1451–1458 (2010).
47. Siegel, T. N., Hekstra, D. R., Wang, X., Dewell, S. & Cross, G. A. Genome-wide analysis of mRNA abundance in two life-cycle stages of *Trypanosoma brucei* and identification of splicing and polyadenylation sites. *Nucleic Acids Res.* **38**, 4946–4957 (2010).
48. Kolev, N. G. *et al.* The transcriptome of the human pathogen *Trypanosoma brucei* at single-nucleotide resolution. *PLoS Pathog.* **6**, e1001090 (2010).
49. Sorber, K., Dimon, M. T. & DeRisi, J. L. RNA-Seq analysis of splicing in *Plasmodium falciparum* uncovers new splice junctions, alternative splicing and splicing of antisense transcripts. *Nucleic Acids Res.* **39**, 3820–3835 (2011).
50. Wurtzel, O. *et al.* Comparative transcriptomics of pathogenic and non-pathogenic *Listeria* species. *Mol. Syst. Biol.* **8**, 583 (2012).
A pioneering study that uses RNA-seq to explore differences and similarities between related bacterial transcriptomes.
51. Albrecht, M., Sharma, C. M., Reinhardt, R., Vogel, J. & Rudel, T. Deep sequencing-based discovery of the *Chlamydia trachomatis* transcriptome. *Nucleic Acids Res.* **38**, 868–877 (2010).
52. Albrecht, M. *et al.* The transcriptional landscape of *Chlamydia pneumoniae*. *Genome Biol.* **12**, R98 (2011).
53. Mandlik, A. *et al.* RNA-seq-based monitoring of infection-linked changes in *Vibrio cholerae* gene expression. *Cell Host Microbe* **10**, 165–174 (2011).
A description of the pathogen transcriptomes from bacteria isolated from either rabbit or mouse hosts.
54. Cloonan, N. *et al.* Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nature Methods* **5**, 613–619 (2008).
55. Wang, E. T. *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**, 470–476 (2008).
56. Pan, Q., Shai, O., Lee, L. J., Frey, B. J. & Blencowe, B. J. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genet.* **40**, 1413–1415 (2008).
57. Peng, Z. *et al.* Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. *Nature Biotech.* **30**, 253–260 (2012).
58. Schulte, L. N., Eulalio, A., Mollenkopf, H. J., Reinhardt, R. & Vogel, J. Analysis of the host microRNA response to *Salmonella* uncovers the control of major cytokines by the *let-7* family. *EMBO J.* **30**, 1977–1989 (2011).
59. Oosthuizen, J. L. *et al.* Dual organism transcriptomics of airway epithelial cells interacting with conidia of *Aspergillus fumigatus*. *PLoS ONE* **6**, e20527 (2011).
60. Lovegrove, F. E. *et al.* Simultaneous host and parasite expression profiling identifies tissue-specific transcriptional programs associated with susceptibility or resistance to experimental cerebral malaria. *BMC Genomics* **7**, 295 (2006).
61. Guerfali, F. Z. *et al.* Simultaneous gene expression profiling in human macrophages infected with *Leishmania major* parasites using SAGE. *BMC Genomics* **9**, 238 (2008).
62. Matsumura, H. *et al.* Gene expression analysis of plant host–pathogen interactions by SuperSAGE. *Proc. Natl Acad. Sci. USA* **100**, 15718–15723 (2003).
63. Tierney, L. *et al.* An interspecies regulatory network inferred from simultaneous RNA-seq of *Candida albicans* invading innate immune cells. *Front. Microbiol.* **3**, 85 (2012).
The first study to describe the parallel analysis of a eukaryotic host and a eukaryotic pathogen via RNA-seq.
64. Cox, M. L. *et al.* Investigating fixative-induced changes in RNA quality and utility by microarray analysis. *Exp. Mol. Pathol.* **84**, 156–172 (2008).
65. Eulalio, A., Schulte, L. & Vogel, J. The mammalian microRNA response to bacterial infections. *RNA Biol* **9**, 742–750 (2012).
66. Wang, K. C. & Chang, H. Y. Molecular mechanisms of long noncoding RNAs. *Mol. Cell* **43**, 904–914 (2011).
67. Chen, J. & Wagner, E. J. snRNA 3′ end formation: the dawn of the Integrator complex. *Biochem. Soc. Trans.* **38**, 1082–1087 (2010).
68. Bratkovic, T. & Rogelj, B. Biology and applications of small nucleolar RNAs. *Cell. Mol. Life Sci.* **68**, 3843–3851 (2011).
69. Papenfort, K. & Vogel, J. Regulatory RNA in bacterial pathogens. *Cell Host Microbe* **8**, 116–127 (2010).
70. Clark, M. B. *et al.* Genome-wide analysis of long noncoding RNA stability. *Genome Res.* **22**, 885–898 (2012).
71. Yang, E. *et al.* Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes. *Genome Res.* **13**, 1863–1872 (2003).

72. Armour, C. D. *et al.* Digital transcriptome profiling using selective hexamer priming for cDNA synthesis. *Nature Methods* **6**, 647–649 (2009).
73. Chen, Z. & Duan, X. Ribosomal RNA depletion for massively parallel bacterial RNA-sequencing applications. *Methods Mol. Biol.* **733**, 93–103 (2011).
74. Huang, R. *et al.* An RNA-Seq strategy to detect the complete coding and non-coding transcriptome including full-length imprinted macro ncRNAs. *PLoS ONE* **6**, e27288 (2011).
75. Giannoukos, G. *et al.* Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. *Genome Biol.* **13**, R23 (2012).
76. Knoop, V. When you can't trust the DNA: RNA editing changes transcript sequences. *Cell. Mol. Life Sci.* **68**, 567–586 (2011).
77. Cantara, W. A. *et al.* The RNA Modification Database, RNAMDB: 2011 update. *Nucleic Acids Res.* **39**, D195–D201 (2011).
78. Ishitani, R., Yokoyama, S. & Nureki, O. Structure, dynamics, and function of RNA modification enzymes. *Curr. Opin. Struct. Biol.* **18**, 330–339 (2008).
79. Motorin, Y., Muller, S., Behm-Ansmant, I. & Branlant, C. Identification of modified residues in RNAs by reverse transcription-based methods. *Methods Enzymol.* **425**, 21–53 (2007).
80. Findeiss, S., Langenberger, D., Stadler, P. F. & Hoffmann, S. Traces of post-transcriptional RNA modifications in deep sequencing data. *Biol. Chem.* **392**, 305–313 (2011).
81. Iida, K., Jin, H. & Zhu, J. K. Bioinformatics analysis suggests base modifications of tRNAs and miRNAs in *Arabidopsis thaliana*. *BMC Genomics* **10**, 155 (2009).
82. Ebhardt, H. A. *et al.* Meta-analysis of small RNA-sequencing errors reveals ubiquitous post-transcriptional RNA modifications. *Nucleic Acids Res.* **37**, 2461–2470 (2009).
83. Levin, J. Z. *et al.* Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nature Methods* **7**, 709–715 (2010).
A detailed overview of the current approaches used to ensure the maintenance of strand-specific information in RNA-seq experiments.
84. Parkhomchuk, D. *et al.* Transcriptome analysis by strand-specific sequencing of complementary DNA. *Nucleic Acids Res.* **37**, e123 (2009).
85. Borodina, T., Adjaye, J. & Sultan, M. A strand-specific library preparation protocol for RNA sequencing. *Methods Enzymol.* **500**, 79–98 (2011).
86. Croucher, N. J. *et al.* A simple method for directional transcriptome sequencing using Illumina technology. *Nucleic Acids Res.* **37**, e148 (2009).
87. Munafo, D. B. & Robb, G. B. Optimization of enzymatic reaction conditions for generating representative pools of cDNA from small RNA. *RNA* **16**, 2537–2552 (2010).
88. Tarazona, S., Garcia-Alcalde, F., Dopazo, J., Ferrer, A. & Conesa, A. Differential expression in RNA-seq: a matter of depth. *Genome Res.* **21**, 2213–2223 (2011).
An exploration of the sequencing depths that are required for accurate gene expression profiling in mammals.
89. McIntyre, L. M. *et al.* RNA-seq: technical variability and sampling. *BMC Genomics* **12**, 293 (2011).
90. Jiang, L. *et al.* Synthetic spike-in standards for RNA-seq experiments. *Genome Res.* **21**, 1543–1551 (2011).
91. Garber, M., Grabherr, M. G., Guttman, M. & Trapnell, C. Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature Methods* **8**, 469–477 (2011).
92. Malone, J. H. & Oliver, B. Microarrays, deep sequencing and the true measure of the transcriptome. *BMC Biol.* **9**, 34 (2011).
93. Metzker, M. L. Sequencing technologies — the next generation. *Nature Rev. Genet.* **11**, 31–46 (2010).
A review of the working principles, performances and costs of popular NGS platforms, with an outlook on third-generation sequencing techniques.
94. Zhang, J., Chiodini, R., Badr, A. & Zhang, G. The impact of next-generation sequencing on genomics. *J. Genet. Genomics* **38**, 95–109 (2011).
95. Rothberg, J. M. *et al.* An integrated semiconductor device enabling non-optical genome sequencing. *Nature* **475**, 348–352 (2011).
96. Sam, L. T. *et al.* A comparison of single molecule and amplification based sequencing of cancer transcriptomes. *PLoS ONE* **6**, e17305 (2011).
97. Koren, S. *et al.* Hybrid error correction and *de novo* assembly of single-molecule sequencing reads. *Nature Biotech.* **30**, 693–700 (2012).
98. Clarke, J. *et al.* Continuous base identification for single-molecule nanopore DNA sequencing. *Nature Nanotechnol.* **4**, 265–270 (2009).
99. Blencowe, B. J., Ahmad, S. & Lee, L. J. Current-generation high-throughput sequencing: deepening insights into mammalian transcriptomes. *Genes Dev.* **23**, 1379–1386 (2009).
100. Antoniou, E. & Taft, R. Gene expression in mouse oocytes by RNA-Seq. *Methods Mol. Biol.* **825**, 237–251 (2012).
101. Lao, K. Q. *et al.* mRNA-sequencing whole transcriptome analysis of a single cell on the SOLiD system. *J. Biomol. Tech.* **20**, 266–271 (2009).
The first study to successfully combine single-cell analysis with RNA-seq.
102. Tang, F. *et al.* RNA-Seq analysis to capture the transcriptome landscape of a single cell. *Nature Protoc.* **5**, 516–535 (2010).
103. Tang, F., Lao, K. & Surani, M. A. Development and applications of single-cell transcriptome analysis. *Nature Methods* **8**, S6–S11 (2011).
An overview of the requirements for single-cell RNA-seq.
104. Tang, F. *et al.* Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis. *Cell Stem Cell* **6**, 468–478 (2010).
105. Tang, F. *et al.* Deterministic and stochastic allele specific gene expression in single mouse blastomeres. *PLoS ONE* **6**, e21208 (2011).
106. Islam, S. *et al.* Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. *Genome Res.* **21**, 1160–1167 (2011).
107. Ramsköld, D. *et al.* Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. *Nature Biotech.* 22 Jul 2012 (doi:10.1038/nbt.2282).
108. Kang, Y. *et al.* Transcript amplification from single bacterium for transcriptome analysis. *Genome Res.* **21**, 925–935 (2011).
The first study to monitor differential gene expression in a single bacterial cell by Sanger sequencing.
109. Brouzes, E. Droplet microfluidics for single-cell analysis. *Methods Mol. Biol.* **853**, 105–139 (2012).
110. Connell, J. L. *et al.* Probing prokaryotic social behaviors with bacterial "lobster traps". *mBio* **1**, e00202-10 (2010).
111. Decarlo, K., Emley, A., Dadzie, O. E. & Mahalingam, M. Laser capture microdissection: methods and applications. *Methods Mol. Biol.* **755**, 1–15 (2011).
112. Ozsolak, F. *et al.* Amplification-free digital gene expression profiling from minute cell quantities. *Nature Methods* **7**, 619–621 (2010).
113. Linnarsson, S. Recent advances in DNA sequencing methods - general principles of sample preparation. *Exp. Cell Res.* **316**, 1339–1343 (2010).

Acknowledgements
The authors acknowledge support from the German Research Foundation (DFG) Priority Program SPP1258 (DFG grant Vo875/4-2) and from the German Ministry of Education and Research (BMBF) (grant 01GS0806). A.J.W. is the recipient of an Elite Advancement Ph.D. stipend from the Universität Bayern e.V., Germany.

Competing interests statement
The authors declare no competing financial interests.

FURTHER INFORMATION
Jörg Vogel's homepage:
http://www.imib-wuerzburg.de/research/vogel/research

SUPPLEMENTARY INFORMATION
See online article: S1 (table) | S2 (table) | S3 (table)

ALL LINKS ARE ACTIVE IN THE ONLINE PDF