AEM
Journals.ASM.org

Comparative Analysis of *Lactobacillus plantarum* WCFS1 Transcriptomes by Using DNA Microarray and Next-Generation Sequencing Technologies

Milkha M. Leimena,[a,b] Michiel Wels,[a,e] Roger S. Bongers,[e] Eddy J. Smid,[a,c] Erwin G. Zoetendal,[a,b] and Michiel Kleerebezem[a,b,d,e]

TI Food and Nutrition (TIFN), Wageningen, the Netherlands[a]; Laboratory of Microbiology,[b] Laboratory of Food Microbiology,[c] and Host-Microbe Interactomics Group,[d] Wageningen University, Wageningen, the Netherlands; and NIZO food research B.V., Ede, the Netherlands[e]

RNA sequencing is starting to compete with the use of DNA microarrays for transcription analysis in eukaryotes as well as in prokaryotes. The application of RNA sequencing in prokaryotes requires additional steps in the RNA preparation procedure to increase the relative abundance of mRNA and cannot employ the poly(T)-primed approach in cDNA synthesis. In this study, we aimed to validate the use of RNA sequencing (direct cDNA sequencing and 3′-untranslated region [UTR] sequencing) using *Lactobacillus plantarum* WCFS1 as a model organism, employing its established microarray platform as a reference. A limited effect of mRNA enrichment on genome-wide transcript quantification was observed, and comparative transcriptome analyses were performed for *L. plantarum* WCFS1 grown in two different laboratory media. Microarray analyses and both RNA sequencing methods resulted in similar depths of analysis and generated similar fold-change ratios of differentially expressed genes. The highest overall correlation was found between microarray and direct cDNA sequencing-derived transcriptomes, while the 3′-UTR sequencing-derived transcriptome appeared to deviate the most. Overall, a high similarity between patterns of transcript abundance and fold-change levels of differentially expressed genes was detected by all three methods, indicating that the biological conclusions drawn from the transcriptome data were consistent among the three technologies.

Understanding the influence of environmental conditions on genome-wide gene expression levels requires the accurate quantification of all expressed (m)RNAs. Microarrays provide an effective method for the analysis of thousands of transcripts in a parallel manner, and they allow the measurement of the genome-wide transcriptome of an organism in a single experiment (12, 51). It is especially suited for the transcriptome comparison of two biological conditions (34). However, background and saturation problems (42) and the low reproducibility of results between laboratories (16) during microarray analyses could limit the use of microarrays for transcriptome interpretation.

The rapid development of next-generation sequencing (NGS) technology for transcriptome analysis, which is known as RNA sequencing, is promoting the use of this method as a replacement for DNA microarrays. RNA sequencing using NGS technology has the advantage of low per-base costs through massive parallel *de novo* sequencing. This is starting to make RNA sequencing a cost-effective alternative for transcriptome analysis, and it is especially suited for samples from biological material with unknown genetic content. RNA sequencing enables the direct determination of the identity and abundance of a transcript, which facilitates the identification of novel transcripts (4, 24) and allows the detection of rare transcripts at considerable sequencing depth (43). The RNA sequencing approach was initially described for eukaryotic cells, such as yeast (26), mouse embryonic stem cells and embryoid bodies (6), human cell lines (36), and plants (11, 42). The main principle of RNA sequencing in eukaryote cells includes the selective conversion of mRNA into double-stranded cDNA fragments by poly(T) (or random)-primed reverse transcription and strand duplication, followed by the direct sequencing of the double-stranded cDNA and quantitative mapping of the identified reads to the genome to estimate the level of gene expression (21, 46). To assess the robustness of the RNA sequencing methodology compared to that of microarrays, several studies were conducted using RNA of eukaryote cells, such as human liver and kidney (22) and mouse hippocampi (39). These comparative studies revealed a good correlation between the levels of transcripts measured by microarrays and RNA sequencing. Moreover, these studies favored RNA sequencing in terms of its higher reproducibility and higher accuracy of detection of the fold change in expression level (22, 39). However, these conclusions were contradicted by a well-defined study that used synthetic RNA samples and demonstrated that microarray quantification correlated better with actual transcript levels and was more sensitive than RNA sequencing, while both methods performed equally well with respect to reproducibility and relative transcript ratio determination (47). In addition to the expressed sequence tag (EST) sequencing, an alternative sequencing-based transcriptome approach was described by Eveland et al. (11), in which the 3′-untranslated region (3′-UTR) of mRNAs in *Zea mays* was sequenced. This 3′-UTR sequencing method offers the possibility to determine differential expression between closely related genes. To date, no studies have been reported that assess the robustness of 3′-UTR sequencing for transcriptome analysis or its comparison to alternative transcriptome analysis methods.

Although RNA sequencing technologies have been implemented and validated in eukaryotes, it is still quite challenging to

Received 15 February 2012 Accepted 29 March 2012
Published ahead of print 6 April 2012
Address correspondence to Michiel Kleerebezem, michiel.kleerebezem@nizo.nl.
Supplemental material for this article may be found at http://aem.asm.org/.

doi:10.1128/AEM.00470-12

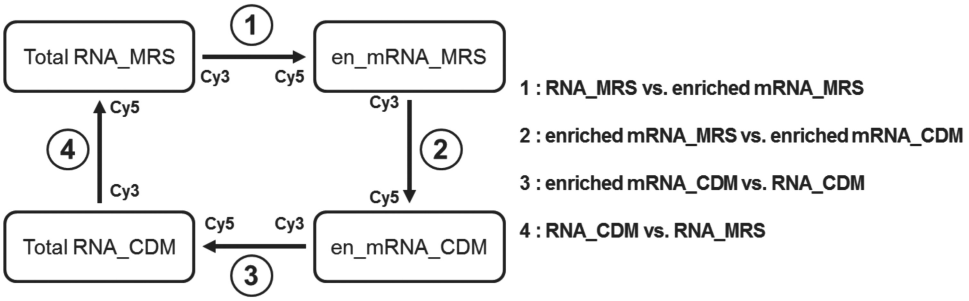
Total RNA_MRS
en_mRNA_MRS
Total RNA_CDM
en_mRNA_CDM
1
2
3
4
Cy3
Cy5
1 : RNA_MRS vs. enriched mRNA_MRS
2 : enriched mRNA_MRS vs. enriched mRNA_CDM
3 : enriched mRNA_CDM vs. RNA_CDM
4 : RNA_CDM vs. RNA_MRS

FIG 1 Hybridization scheme of total RNA and enriched mRNA of *L. plantarum* WCFS1 grown in CDM and MRS. Each arrow represents a single hybridization. Samples at the base of the arrow were labeled with Cy3 label and samples at the arrowhead with Cy5.

employ these methods for prokaryote transcription analysis. This is not surprising, since the prokaryote RNA pool contains a large amount of rRNA and tRNA, which may constitute more than 95% of the total RNA (31), while the selective reverse transcription of mRNA by poly(T) priming is not possible (9, 42). Moreover, prokaryote transcriptional profiles are considered to be much more dynamically regulated and less stable than those of eukaryotes. To increase the relative abundance of mRNA in total prokaryote RNA material, several methods have been developed, including rRNA capture, the selective degradation of processed RNA, the selective polyadenylation of mRNA, and the antibody capture of subsets of mRNAs that interact with Hfq proteins (32). Due to the lack of a poly(A) tail in prokaryote mRNA, alternative priming approaches during reverse transcription (RT) are commonly based on random oligonucleotide priming (hexamers or longer) and sometimes employ multiplexed gene-specific oligonucleotides (28, 50) or based on a combination of gene-specific priming and 5′-end RNA sequencing (48). Alternatively, oligo(dT) priming can be employed following the artificial polyadenylation of mRNAs (13). The development of mRNA enrichment and priming methods allowed the successful use of RNA sequencing approaches for the investigation of transcriptome changes under different growth conditions of *Burkholderia cenocepacia* (50) and *Bacillus anthracis* (28).

In this study, we aimed to validate the use of different RNA sequencing techniques using a model prokaryote organism while employing an established microarray platform as a reference. The transcriptomes of *Lactobacillus plantarum* WCFS1 (grown in two different laboratory media) were compared using custom-made oligonucleotide microarrays and RNA sequencing approaches. The microarray was also employed to evaluate the impact on the transcriptome of mRNA enrichment by RNA capture methods. This study includes the comparison of two RNA sequencing approaches, direct cDNA and 3′-UTR cDNA sequencing, to evaluate their applicability in prokaryote transcriptome analyses. Our analyses show that the depth of analysis for both RNA sequencing methodologies was similar to that observed for the microarray, leading to a coverage of >95% of all genes encoded in the *L. plantarum* WCFS1 genome. The best transcriptome correlation was found between microarray and direct cDNA sequencing analyses, while the 3′-UTR sequencing method appeared to deviate the most. Overall, patterns of transcript abundance and fold-change levels of differentially expressed genes were similar for all three methods.

MATERIALS AND METHODS

Bacterial strain and growth conditions. *L. plantarum* WCFS1 (19) was grown in chemically defined medium (CDM) (38) and de Man Rogosa Sharpe (MRS) medium (8) at 37°C without agitation. Cells were harvested by centrifugation for 10 min at 4,570 × *g* and 4°C using a Heraeus Multifuge 3 S-R centrifuge (DJB Labcare Ltd., England) at an optical density at 600 nm (OD_{600}) of approximately 1.0, which corresponds to the mid-logarithmic phase of growth for both media.

Total RNA isolation and mRNA enrichment. Total RNA was extracted from the cell pellets according to the Macaloid-based RNA isolation protocol (52). Extraction was followed by RNA purification using the RNAeasy minikit (Qiagen), including an on-column DNase I (Roche, Germany) treatment as described previously (52). The enrichment of mRNA was performed by the selective removal of 16S and 23S rRNA using oligonucleotide probes attached to magnetic beads according to the manufacturer's protocol (Microbexpress; Ambion, Applied Biosystems, Niewerkerk a/d Ijssel, the Netherlands) (44). Total RNA and enriched mRNA yields were quantified spectrophotometrically (NanoDrop 1000; Nanodrop Technologies, Wilmington, DE), and total RNA quality was assessed by a microfluidics-based electrophoresis system (Experion RNA StdSens; Bio-Rad Laboratories Inc.).

DNA microarray-based transcriptome analysis. The microarray used was a custom-designed *L. plantarum* WCFS1, 8×15K Agilent oligonucleotide microarray (GPL13984) containing (maximally) three different probes per annotated gene that were spotted in duplicate (30). Both total RNA and enriched mRNA were subjected to cDNA synthesis using a random nonamer primed approach as has been described before (33). Cy3- and Cy5-labeled cDNAs were prepared using a Cyscribe postlabeling kit (Amersham Biosciences, United Kingdom) according to the manufacturer's protocol. Cy5/Cy3 dye swaps were performed for the cDNA samples according to the scheme shown in Fig. 1. Labeled cDNA mixtures were subsequently concentrated in a Hetovac VR-1 (Heto Lab Equipment A/S, Birkerod, Denmark) to a final volume of 25 μl (if needed), incubated at 98°C for 3 min, and cooled at room temperature for 5 min. After the addition of 25 μl 2× GEX HI-RPM hybridization buffer (Agilent Technologies, Palo Alto, CA), 40 μl of each mixture was applied to an Agilent 8×15K array (Agilent Technologies, Palo Alto, CA). The hybridization and scanning of the microarray slides were performed as described previously (23). Slides were scanned with a ScanArray Express 4000 scanner (Perkin Elmer, Wellesley, MA), and the image was analyzed and processed using ImaGene version 7.5 software (BioDiscovery Inc., Marina Del Rey, CA). Both total RNA and mRNA-enriched data sets were normalized and corrected by the local fitting of an M-A plot applying the Lowess algorithm (49) and interslide scaling, which are available in MicroPrep (41), and different transcriptomes were compared using CyberT (3), taking into account the dye swaps of each of the conditions as described previously (23). The microarray data have been deposited in NCBI's Gene.

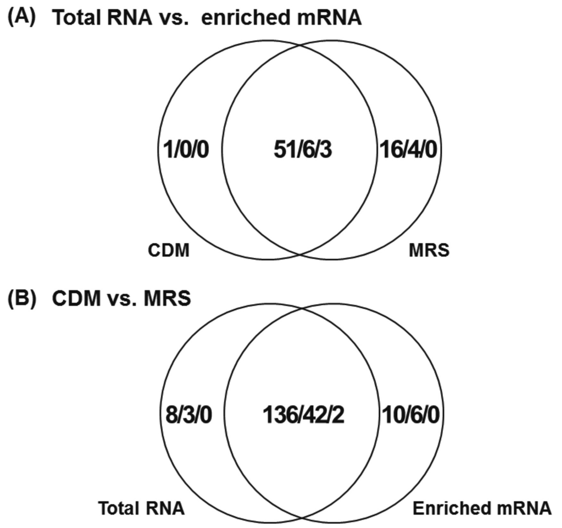

FIG 2 Venn diagram showing the number of upregulated/downregulated/oppositely regulated genes in the enriched mRNA sample obtained from bacterial cells grown in CDM and MRS (A) or in the RNA obtained from cells growing in CDM, either total RNA or after mRNA enrichment (B).

RNA sequencing-based transcriptome analysis. Double-stranded cDNA was synthesized using enriched mRNA of CDM- or MRS-grown *L. plantarum* WCFS1 using the SuperScript double-stranded cDNA synthesis kit (11917-010; Invitrogen) with the addition of SuperScript III reverse transcriptase (18080-044; Invitrogen) and random primers (48190-011; Invitrogen) as described previously (50). This was followed by RNase A (Roche, Germany) treatment, phenol-chloroform extraction, and ethanol precipitation. Double-stranded cDNA was quantified using the NanoDrop 1000 spectrophotometer (Nanodrop Technologies, Wilmington, DE), and the quantity and purity were verified by GATC Biotech using an Agilent 2100 Bioanalyzer (Agilent Technologies Inc., Waldbronn, Germany). Sequencing libraries were constructed from double-stranded cDNA samples according to the Illumina Genome Analyzer II protocol (46), followed by direct cDNA sequencing (GATC Biotech, Konstanz, Germany). In addition, enriched mRNA samples of *L. plantarum* WCFS1 from both CDM and MRS were used for 3′-UTR library preparation. To this end, the enriched mRNA samples were poly(A) tailed using poly(A) polymerase, treated with tobacco acid pyrophosphatase (TAP), and ligated at the 5′ end to an RNA adaptor (GATC Biotech, Konstanz, Germany). First-strand cDNA synthesis was performed using an oligo(dT) adapter primer and Moloney murine leukemia virus (M-MLV) H-reverse transcriptase. The resulting cDNAs were PCR amplified to 20 to 30 ng/μl in 18 cycles using a high-fidelity DNA polymerase. PCR products were purified using the NucleoSpin Extract II kit (Macherey-Nagel GmbH & Co. KG., Germany) and were examined using a Shimadzu MultiNA microchip electrophoresis system (Shimadzu Corporation, Japan). Both direct cDNA and 3′-UTR sequencing were performed simultaneously using a single flow cell of the Illumina Genome Analyzer II (GATC Biotech, Konstanz, Germany) at 8 pM. Sequence data were cleaned for the poly(A) (for 3′-UTR sequencing only) and low-complexity regions using seqclean (http://compbio.dfci.harvard.edu/tgi/software/), with a length threshold of 20. The mapping and quantification of the cleaned sequences to an *in silico* transcriptome reference was performed by GATC Biotech. The cDNA reference was created using the annotation of the *L. plantarum* WCFS1 genome obtained from UCSC Genome Bioinformatics (http://genome.ucsc.edu/). To map the 3′-UTR-derived sequences to the appropriate gene-specific transcripts, additional mappings of 100-, 200-, and 300-bp 3′-extended transcriptional unit predictions were employed (45). These multiple mappings were performed to increase the frequency of read assignments to genes, because the position of a 3′-UTR at the end of a transcript in *L. plantarum* WCFS1 is unknown. The visualization of the mapped transcript was performed using the UCSC Genome Browser (http://microbes.ucsc.edu/).

Comparative data analyses of direct cDNA sequencing versus 3′-UTR sequencing and microarray versus both RNA sequencings. The signal intensity data obtained by microarrays and the number of read counts of direct cDNA and 3′-UTR sequencings were quantile normalized (5) using the CLC-Bio Genomic Workbench software to adjust the data range. Normalized read counts of direct cDNA sequencing were plotted against those of the 3′-UTR sequencing using a scatter plot to investigate the read distribution between the two data sets. Rank-based analysis using the Spearman correlation coefficient was applied to investigate the correlation between two sequencing techniques for CDM and MRS culture-derived RNA samples. For the comparison using the microarray, which was utilized as the benchmark technology, normalized microarray signal intensities were used for the comparison of the normalized read counts from both RNA sequencing techniques using Spearman correlation analysis. The analysis of the differentially expressed genes (DEG) based on a $\log_2$ fold-change ratio of CDM/MRS between microarray and both RNA sequencing techniques was performed using a parametric method, Pearson correlation, assuming that the relative expression of the DEG should be conserved within all techniques irrespective of the difference in absolute gene expression values or the various dynamic ranges of the different techniques. Only those genes that showed a >2-fold absolute fold-change ratio for all techniques and displayed significant (FDR-adjusted *P* values of <0.05) differential expression according to the microarray analysis were used. Spearman and Pearson correlation analyses were performed using the PASW Statistic 17.0 software suite (SPSS Inc., Chicago, IL).

Microarray data accession number. The microarray data have been deposited in NCBI's Gene Expression Omnibus (10) and are accessible through GEO series accession number GSE35754.

RESULTS AND DISCUSSION

Microarray transcript profiles for total RNA and mRNA samples. In this study, *L. plantarum* WCFS1 was grown in two laboratory media (CDM and MRS) to represent different environmental conditions. Microarray analysis was performed using both total RNA and enriched mRNA samples. The effect of mRNA enrichment on the transcriptome profile was evaluated by comparing normalized signal intensities per gene in the total RNA to those of mRNA-enriched transcriptome data sets by Spearman correlation analysis. A highly similar ranking of gene expression values in total RNA versus mRNA-enriched samples was detected, as illustrated by the high Spearman correlation coefficients of 0.957 ($P < 0.01$) and 0.953 ($P < 0.01$) for the RNA samples derived from CDM- and MRS-grown cultures, respectively. Only 81 genes were differentially quantified with FDR-adjusted *P* values of <0.05 for total RNA versus enriched mRNA samples for both growth conditions (Fig. 2), indicating that mRNA enrichment has only a limited effect on overall transcript quantification. Notably, of these 81 genes, 60 were differentially quantified in the RNA samples from both growth conditions and were consistently observed at a higher level in the mRNA-enriched sample, suggesting that the enrichment procedure selectively and consistently enriches a small but specific RNA subset. Their fold-change ratio generally varied from 2- to 10-fold, and in the few cases where the fold change exceeded a factor of 10, the genes were among the least expressed within the data set. The majority of these differentially quantified genes were related to hypothetical protein functions (Fig. 3). In addition, the limitation of the mRNA enrichment method used (Microbexpress; Ambion), which does not target tRNA removal, resulted in the differential quantification of some tRNAs in the mRNA-enriched fraction (17).

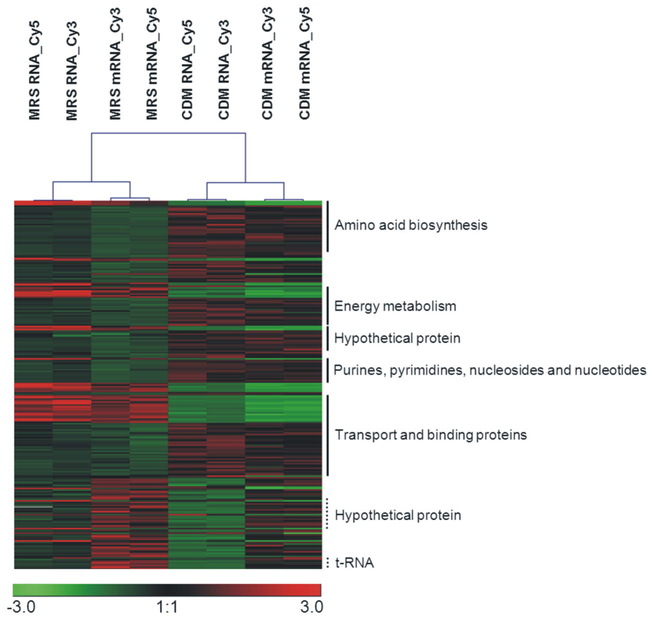

FIG 3 Cluster analysis of 240 genes, 60 of total RNA versus mRNA enrichment and 180 of CDM versus MRS (with a >2-fold change and FDR-adjusted *P* values of <0.05) of *L. plantarum* WCFS1. Functional categories enriched with the gene data sets in different growth conditions are indicated with continuous lines, while dotted lines indicate clusters of genes that displayed differential quantification due to the mRNA enrichment procedure. Very similar clustering results were obtained when the complete transcriptome data sets were used (data not shown).

Variation in gene expression levels caused by the different growth conditions (CDM versus MRS) was observed for a total of 207 genes (FDR-adjusted *P* values of <0.05) from both total RNA and mRNA-enriched analyses. Of these 207 genes, 180 genes were shared between the differential genes identified in the total RNA and mRNA-enriched samples, of which 178 showed conserved up- or downregulation as a consequence of the difference in growth medium (Fig. 2). Average linkage hierarchical clustering with Pearson correlation distance (35) confirmed a more pronounced separation of CDM and MRS profiles relative to the separation seen for total RNA versus mRNA enrichment profiles (Fig. 3).

This finding shows that the transcriptome variation caused by different growth conditions exceeds the variation caused by the enrichment procedure, indicating that mRNA enrichment will have only a limited effect on the biological interpretation of transcriptome data, which are validated for a well-defined culture under well-defined conditions, with the anticipation of similar performance in complex ecosystems. The genes displaying significant differential expression between cultures grown on CDM and MRS predominantly belonged to specific functional categories that appear to reflect the different medium compositions, such as transport and binding proteins (in particular for amino acids, peptides, and amines), amino acid biosynthesis (in particular for histidine and aspartate), energy metabolism, and the synthesis of purines, pyrimidines, nucleosides, and nucleotides (Fig. 3; also see Fig. S1 in the supplemental material). The limited amount of nucleotides (18) and specific amino acids available in CDM relative to MRS apparently requires an alternative pallet of transport functions to import those components, which could be concluded consistently from the arrays irrespective of the RNA source (enriched mRNA versus total RNA) used.

RNA sequencing-based transcriptome analysis: direct cDNA sequencing versus 3′-UTR sequencing. Direct cDNA sequencing and 3′-UTR sequencing were performed using mRNA-enriched samples of *L. plantarum* WCFS1 grown in CDM or MRS. The number of sequence reads recovered varied between 17.5 and 28.1 million per sample (see Table S1 in the supplemental material), with an average trimmed length of 36 bp. Of all sequence reads obtained, 93 to 98% could be assigned to the *L. plantarum* WCFS1 genome. Sequence reads that could be mapped to the genome were subsequently aligned to the coding sequences (CDS) based on the current annotation of protein-encoding genes of *L. plantarum* WCFS1 (19). The majority of the direct cDNA sequencing reads that mapped to the genome could be aligned to the CDS (14.6 to 18.5 million). In contrast, the sequences obtained by 3′-UTR sequencing mapped with much lower frequency to the CDS (<20%) (see Table S1 in the supplemental material). A possible explanation for the strongly reduced CDS mapping of the short reads (~36 bp) obtained by 3′-UTR sequencing is the preferential sequencing of the genetic regions downstream of the protein-coding region that is intrinsic to this method. Unfortunately, there is no accurate prediction of the 3′ end of the transcript sequences for the *L. plantarum* genome. To overcome the low CDS mapping, an *in silico* approach was chosen that included a stepwise 3′ extension of the CDS with 100, 200, and 300 bp. *In silico* predictions indicated that approximately 75% of the predicted terminator sequences in the *L. plantarum* WCFS1 genome were encompassed within the 100-bp extension (7), while an additional 12 and 6% of the predicted terminators were encountered within the 200- and 300-bp extended 3′-UTRs, respectively (see Fig. S2 in the supplemental material). Analogously, the *in silico* 3′ extension of the CDS of the *L. plantarum* WCFS1 genome by 100 bp enabled an 80 and 130% increase in the gene-specific mapping of the CDM and MRS 3′-UTR transcript sequence data sets, respectively. Notably, larger 3′ extensions of the gene sequences with 200 and 300 bp led to significantly smaller increases of CDS-specific transcript mapping (~25 and ~30%, respectively), supporting the prediction that 75% of the terminators are within the first 100 bases downstream of the CDS (see Fig. S2A). Moreover, 200- and 300-bp 3′ extensions of gene sequences included a significantly higher fraction of the transcript sequences that were erroneously mapped to downstream genes, which is a consequence of the overlap of these extensions with downstream genes (see Fig. S2B). Based on these analyses, 100-bp 3′ extensions were incorporated in the gene-specific mapping of 3′-UTR transcript sequence mapping to the *L. plantarum* WCFS1 genome, which improved the number of reads mapped to the CDS from below 20% to approximately 35%.

As anticipated, the distribution of the mapped sequences to the protein-encoding CDS was markedly different between direct cDNA sequencing and 3′-UTR sequencing. While the reads obtained from 3′-UTR sequencing predominantly mapped at the 3′ end of the (extended) genes (see Table S2 in the supplemental material), the reads obtained from direct cDNA sequencing appeared to distribute relatively equally over the entire CDS. Many prokaryotic genes are transcribed in operons that generate polycistronic transcripts that cover several genes, which are commonly functionally related (20, 45). Analogously, most of the 3′-UTR sequence data sets (~70%) consistently mapped to the last gene of such polycistronic transcripts (Fig. 4). This indicates

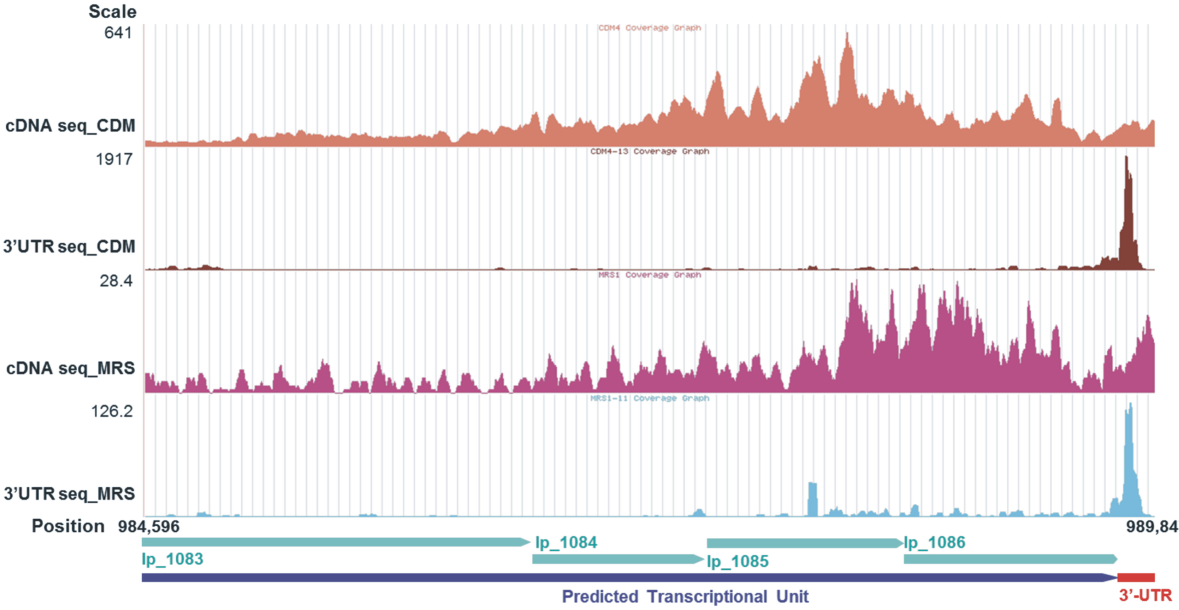

FIG 4 Mapping of *L. plantarum* WCFS1 transcripts from direct cDNA sequencing and 3′-UTR sequencing of MRS- and CDM-grown cultures based on a predicted transcriptional unit (38). Scaling differences of the *y* axis range are indicative for the upregulated transcription level observed in cells obtained from CDM-grown cultures.

that the accurate functional interpretation of 3′-UTR sequence data sets requires the correct prediction of transcriptional units (including operons) to precisely encompass all functions expressed.

Both sequencing techniques showed comparable transcript coverage, where >95% of all annotated genes (3,135 genes) of the *L. plantarum* WCFS1 genome were at least covered by a single sequence read. A similar read distribution was observed for direct cDNA sequencing and 3′-UTR sequencing (i.e., similar proportion between the area above and below the continuous line) (Fig. 5), which indicates similar gene expression patterns. Notably, several genes were apparently overestimated by 3′-UTR sequencing (Fig. 5, upper left), which may be due to either a technical artifact from the application of the poly(A) tail, an artifact in the data interpretation from the 100-bp extension of the mapping, or the existence of some internal promoters (7).

RNA sequencing validation by comparative analysis with the microarray-derived transcriptomes. Since microarrays can be considered an established transcriptome methodology, the data obtained from the microarrays were employed as a reference to evaluate the overall validity of direct cDNA and 3′-UTR sequencing. Only the genes that gave a value for all methods (2,962 genes) were used for the comparison of the transcriptomes obtained by microarray and direct cDNA or 3′-UTR sequencing. Normalized signal intensity values per gene obtained by microarray analysis were plotted against normalized CDS read assignment frequencies derived from both RNA sequencing methods.

Both microarray and RNA sequencing transcriptome data sets were normalized using quantile normalization as a quick and simple method to create an even distribution of microarray probe intensities and RNA sequencing read counts (5). Additional normalization approaches, such as RPKM (reads per kilobase of exon

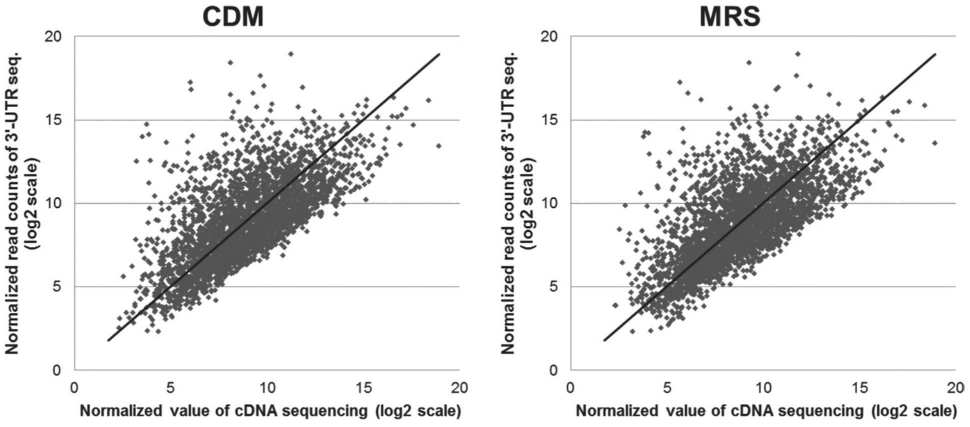

FIG 5 Comparison of normalized signal intensity between direct cDNA sequencing and 3′-UTR sequencing for bacteria grown in CDM (Spearman coefficient, 0.686; $P < 0.01$) and MRS (Spearman coefficient, 0.678; $P < 0.01$).

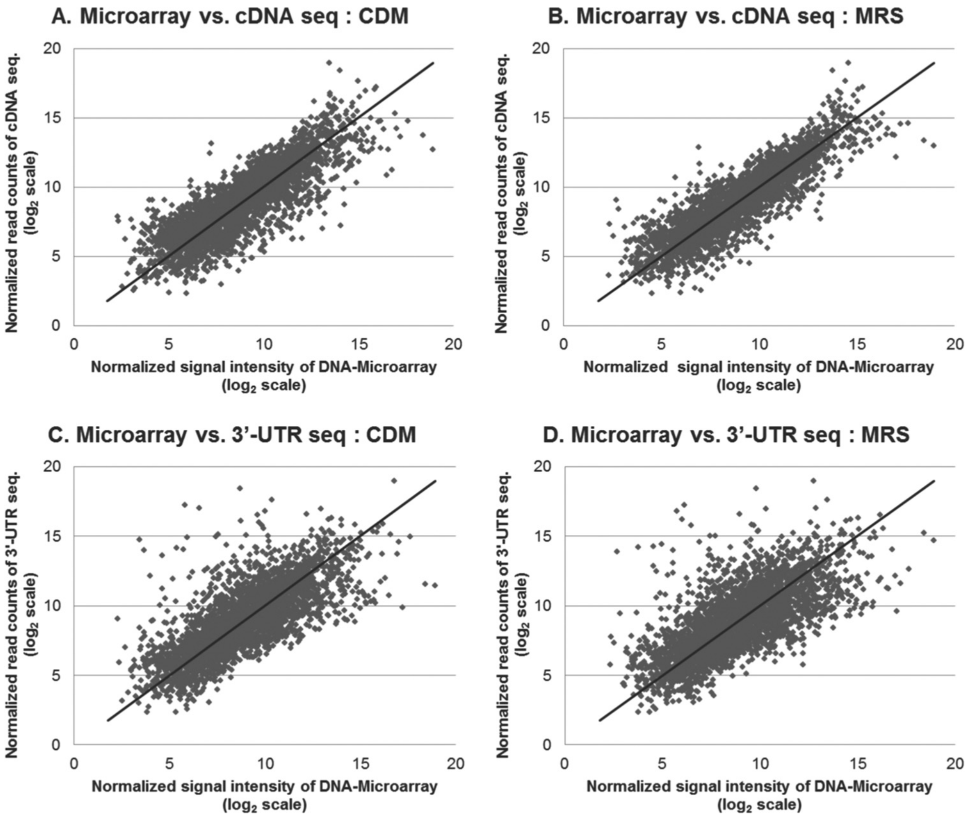

FIG 6 Comparison between normalized signal intensity level of microarray and normalized read counts of direct cDNA sequencing (A and B) and 3′-UTR sequencing (C and D) in transcriptome data sets from bacteria grown in CDM and MRS.

model per million mapped reads) (25) or FPKM (fragments per kilobase of transcript per million fragments mapped) (40), which take into consideration the influence of transcript length toward the gene expression quantification of RNA sequencing reads, could give more accurate gene expression quantification, especially of direct cDNA sequencing. Although overall transcriptome comparisons were done without considering the transcript length, high comparability was shown between microarray and direct cDNA sequencing as well as between microarray and 3′-UTR sequencing (Fig. 6).

Direct cDNA sequencing displayed a higher correlation to the microarray than the 3′-UTR (Fig. 6). This was especially apparent in transcripts with relatively high 3′-UTR sequencing assignments compared to the array signal intensity (Fig. 6C and D, upper left). These results indicate that direct cDNA sequencing generates transcriptome results that resemble those obtained by microarray transcriptome profiling, and that the 3′-UTR estimated expression levels appear to be higher for subsets of genes than for the other two methods. These conclusions were also supported by rank-based Spearman correlation analysis, showing higher correlation values between microarray and direct cDNA sequencing data sets (CDM, 0.835 [$P < 0.01$] and 0.762 [$P < 0.01$]; MRS, 0.881 [$P < 0.01$] and 0.707 [$P < 0.01$]; for direct cDNA sequencing and 3′-UTR sequencing, respectively).

The application of 3′-UTR sequencing as a method for prokaryote transcriptome analysis has not yet been well established and may require additional normalization or processing steps to obtain an appropriate quantitative representation of the transcript levels that can be compared to microarray-derived transcriptome data sets. To evaluate whether the lower correlation between 3′-UTR sequencing and array-based transcript data sets was caused by a biased positioning of the sequence reads within an operon, the expression values of the last genes in operons was also assigned to each upstream gene within the same predicted operon. However, this data transformation step to accommodate polycistronic operon transcripts in the 3′-UTR data did not improve the correlation with the array-derived data sets (data not shown). This suggests that the lower correlation of these data sets arises from a bias in the 3′-UTR extension or sequencing technology employed.

The most relevant comparative analysis of the three methods employed here undoubtedly relates to the comparisons of the biological conclusions they may generate. To this end, the ability of the three technologies to consistently identify the same genes (1, 22) that are differentially expressed (DEG) after growth on CDM and MRS. The sequence-based transcriptome quantification was determined by the ratio of sequence reads assigned to a gene in data sets obtained from CDM and MRS samples, while the differential expression per gene in the microarray data set was calculated using CyberT (3). In total, 538 DEG with an expression fold change of >2 were detected within the DNA microarray, while 442 and 466 DEG with an expression fold change of >2 were detected by direct cDNA sequencing and 3′-UTR sequencing, respectively. Among the latter groups of genes, 233 and 204 DEG were shared between the microarray-based analysis and direct cDNA and 3′-UTR sequencing, respectively. Moreover, 172 genes were identified to have an absolute fold change of >2 for all tech-

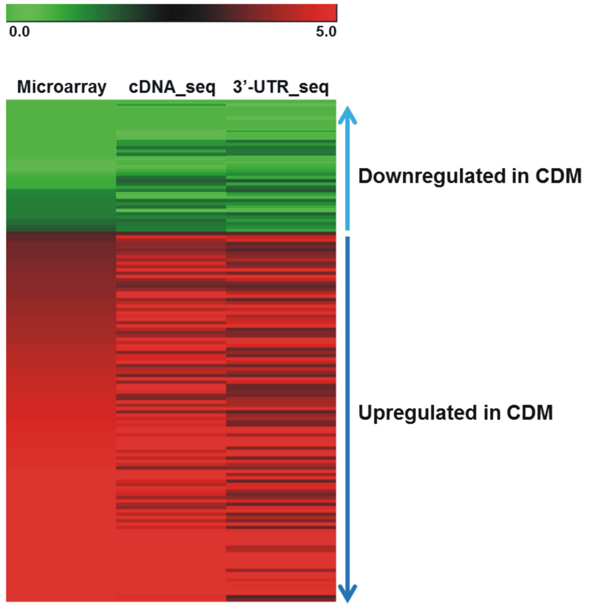

FIG 7 Comparison of 152 transcript levels (40 downregulated in CDM and 112 upregulated in CDM) that were consistently classified among the DEG gene sets as determined by microarray transcriptomes or direct cDNA and 3′-UTR transcriptome sequencing. Data are sorted according to their fold change within the reference data sets (DNA microarray technology).

niques with the same up- or downregulation pattern, among which 152 genes were considered to be significantly differentially expressed according to the DNA microarray technology (FDR, <0.05) that was used as the reference technology. Therefore, this comparative analysis of differentially expressed genes establishes a good consistency of the biological outcomes generated by the three transcriptome technologies, which is characterized by similar fold changes of expression for most shared DEG. The heat map analysis of the differential expression data confirmed that 3′-UTR sequencing deviates slightly more from the microarray than direct cDNA sequencing (Fig. 7). This is also reflected by the somewhat lower Pearson correlation for the comparison of the microarray to 3′-UTR sequencing (0.852; $P < 0.01$) relative to the comparison of the microarray to direct cDNA sequencing (0.897; $P < 0.01$). Notably, the highest Pearson correlation was obtained for the two sequencing-based technologies (0.951; $P < 0.01$), which might be due to the saturated hybridization signals in the array data sets (see Fig. S3 in the supplemental material) (14).

Since this analysis of DEG was performed with microarray data as a reference, DEG that display differential expression only according to the transcriptome sequencing analyses may have been missed. The DEG analysis of the direct cDNA and 3′-UTR sequencing data sets revealed 50 additional genes with a differential expression value of >2 in both sequencing-based data sets. Of these genes, 40 appeared not to reach significance of regulation in the array data set (FDR, >0.05) but displayed a conserved direction of differential expression according to the array analyses, albeit it with a <2 absolute fold-change ratio. Moreover, many of the probes associated with 36 of these 40 genes revealed saturated hybridization signals in the array data sets (see Fig. S3 in the supplemental material), suggesting that they were inaccurately measured by the array due to falling outside the dynamic range of the array technology (14). This observation implies that RNA sequencing exceeds the depth of analysis of the traditional array technologies, especially for genes that are expressed at a high level.

Unlike microarray data, RNA sequencing count data generally are not well represented as a continuous distribution (27). Therefore, normalization procedures that are successfully applied for microarray data might not be optimal for RNA sequencing data sets. Data normalization based on parametric approaches was implemented in several analyses platforms, such as edgeR (29), baySeq (15), and DESeq (2), which allow the lowering of both biological and technical variability for replicated count data. Moreover, nonparametric approaches, like the noise modeling employed in NOISeq, allow the evaluation of low expression counts without any need for replicates (37). Overall, it is very encouraging that the data presented establish that the three transcriptome methods generate a very similar biological view of the transcriptional behavior of a well-defined culture under well-defined conditions.

Concluding remarks and outlook toward undefined ecosystem metatranscriptome sequencing. The present study provides a validation of RNA sequencing techniques in prokaryotes, using a well-studied bacterium under well-defined conditions and employing DNA microarray technology as the reference transcriptome methodology. Such a validation of sequence-based transcriptomics methodology is required to confidently apply sequence-based transcriptome methods to samples derived from complex microbial communities with unknown composition and that live in poorly defined growth conditions. Such ecosystem metatranscriptomic analyses cannot be performed using DNA microarrays due to sequence variations among the coding capacities among (close) relatives of similar phylogenetic origins, which makes the quantification of transcripts on the basis of hybridization signals highly unreliable. This study also demonstrates that 3′-UTR sequencing is complicated by the processing of the sequence data that do not map to coding regions of genes and therefore can be anticipated to present considerable uncertainties during the biological (i.e., genes and functions) interpretation of 3′-UTR RNA sequencing data sets obtained from complex microbial communities with unknown genetic content. Taken together, the results presented in this study indicate that direct cDNA sequencing technology is a promising approach for the generation of metatranscriptome data sets of an unknown microbial community, and it offers good possibilities for biological interpretation with a set of representative microbial genomes as a mapping platform.

ACKNOWLEDGMENTS

We thank Yadhu Kumar and Nastasja Trunk of GATC-Biotech (Konstanz, Germany) for providing bioinformatics support and additional information regarding the RNA sequencing procedures.

This project was supported by the Netherlands Bioinformatics Centre (NBIC).

REFERENCES

1. **Agarwal A, et al.** 2010. Comparison and calibration of transcriptome data from RNA-Seq and tiling arrays. BMC Genomics **11**:383–399.
2. **Anders S, Huber W.** 2010. Differential expression analysis for sequence count data. Genome Biol. **11**:R106.
3. **Baldi P, Long AD.** 2001. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes. Bioinformatics **17**:509–519.

4. **Bloom JS, Khan Z, Kruglyak L, Singh M, Caudy AA.** 2009. Measuring differential gene expression by short read sequencing: quantitative comparison to 2-channel gene expression microarrays. BMC Genomics **10**: 221–231.
5. **Bolstad BM, Irizarry RA, Astrand M, Speed TP.** 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics **19**:185–193.
6. **Cloonan N, et al.** 2008. Stem cell transcriptome profiling via massive-scale mRNA sequencing. Nat. Methods **5**:613–619.
7. **de Hoon MJL, Makita Y, Nakai K, Miyano S.** 2005. Prediction of transcriptional terminators in *Bacillus subtilis* and related species. PLoS Comput. Biol. **1**:e25.
8. **De Man JC, Rogosa M, Sharpe ME.** 1960. A medium for the cultivation of lactobacilli. J. Appl. Microbiol. **23**:130–135.
9. **Deutscher MP.** 2006. Degradation of RNA in bacteria: comparison of mRNA and stable RNA. Nucleic Acids Res. **34**:659–666.
10. **Edgar R, Domrachev M, Lash AE.** 2002. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. **30**:207–210.
11. **Eveland AL, McCarty DR, Koch KE.** 2008. Transcript profiling by 3′-untranslated region sequencing resolves expression of gene families. Plant Physiol. **146**:32–44.
12. **Everett KR, Rees-George J, Pushparajah IPS, Janssen BJ, Luo Z.** 2010. Advantages and disadvantages of microarrays to study microbial population dynamics-a minireview. New Zealand Plant Protect. **63**:1–6.
13. **Frias-Lopez J, et al.** 2008. Microbial community gene expression in ocean surface waters. Proc. Natl. Acad. Sci. U. S. A. **105**:3805–3810.
14. **Garcia de la Nava J, van Hijum S, Trelles O.** 2004. Saturation and quantization reduction in microarray experiments using two scans at different sensitivities. Stat. Appl. Genet. Mol. Biol. **3**:1–16.
15. **Hardcastle TJ, Kelly KA.** 2010. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics **11**:422.
16. **Irizarry RA, et al.** 2005. Multiple-laboratory comparison of microarray platforms. Nat. Methods **2**:345–350.
17. **Kang Y, et al.** 2011. Transcript amplification from single bacterium for transcriptome analysis. Genome Res. **21**:925–935.
18. **Kilstrup M, Hammer K, Ruhdal Jensen P, Martinussen J.** 2005. Nucleotide metabolism and its control in lactic acid bacteria. FEMS Microbiol. Rev. **29**:555–590.
19. **Kleerebezem M, et al.** 2003. Complete genome sequence of *Lactobacillus plantarum* WCFS1. Proc. Natl. Acad. Sci. U. S. A. **100**:1990–1995.
20. **Kozak M.** 1983. Comparison of initiation of protein synthesis in procaryotes, eucaryotes, and organelles. Microbiol. Rev. **47**:1–45.
21. **Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN.** 2010. RNA-seq gene expression estimation with read mapping uncertainty. Bioinformatics **26**:493–500.
22. **Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y.** 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Res. **18**:1509–1517.
23. **Meijerink M, et al.** 2010. Identification of genetic loci in *Lactobacillus plantarum* that modulate the immune response of dendritic cells using comparative genome hybridization. PLoS One **5**:e10632.
24. **Morozova O, Hirst M, Marra MA.** 2009. Applications of new sequencing technologies for transcriptome analysis. Annu. Rev. Genomics Hum. Genet. **10**:135–151.
25. **Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B.** 2008. Mapping and quantifying mammalian transcriptomes by RNA-seq. Nat. Methods **5**:621–628.
26. **Nagalakshmi U, et al.** 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science **320**:1344–1349.
27. **Oshlack A, Robinson MD, Young MD.** 2010. From RNA-seq reads to differential expression results. Genome Biol. **11**:220.
28. **Passalacqua KD, et al.** 2009. Structure and complexity of a bacterial transcriptome. J. Bacteriol. **191**:3203–3211.
29. **Robinson MD, McCarthy DJ, Smyth GK.** 2010. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics **26**:139–140.
30. **Serrano LM, et al.** 2007. Thioredoxin reductase is a key factor in the oxidative stress response of *Lactobacillus plantarum* WCFS1. Microb. Cell Fact. **6**:29.
31. **Siezen RJ, Wilson G, Todt T.** 2010. Prokaryotic whole-transcriptome analysis: deep sequencing and tiling arrays. Microb. Biotechnol. **3**:125–130.
32. **Sorek R, Cossart P.** 2010. Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. Nat. Rev. Genet. **11**:9–16.
33. **Stevens MJA, Molenaar D, de Jong A, De Vos WM, Kleerebezem M.** 2010. Sigma54-mediated control of the mannose phosphotransferase system in *Lactobacillus plantarum* impacts on carbohydrate metabolism. Microbiology **156**:695–707.
34. **Stoughton RB.** 2005. Applications of DNA microarrays in biology. Annu. Rev. Biochem. **74**:53–82.
35. **Sturn A, Quackenbush J, Trajanoski Z.** 2002. Genesis: cluster analysis of microarray data. Bioinformatics **18**:207–208.
36. **Sultan M, et al.** 2008. A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science **321**: 956–960.
37. **Tarazona S, Garcia-Alcalde F, Dopazo J, Ferrer A, Conesa A.** 2011. Differential expression in RNA-seq: a matter of depth. Genome Res. **21**: 2213–2223.
38. **Teusink B, et al.** 2005. In silico reconstruction of the metabolic pathways of *Lactobacillus plantarum*: comparing predictions of nutrient requirements with those from growth experiments. Appl. Environ. Microbiol. **71**:7253–7262.
39. **'t Hoen PA, et al.** 2008. Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. Nucleic Acids Res. **36**:e141.
40. **Trapnell C, et al.** 2012. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. **7**:562–578.
41. **van Hijum SAFT, Garcia de la Nava J, Trelles O, Kok J, Kuipers OP.** 2003. MicroPreP: a cDNA microarray data pre-processing framework. Appl. Bioinformatics **2**:241–244.
42. **van Vliet AH.** 2010. Next generation sequencing of microbial transcriptomes: challenges and opportunities. FEMS Microbiol. Lett. **302**:1–7.
43. **Wang Z, Gerstein M, Snyder M.** 2009. RNA-seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. **10**:57–63.
44. **Warnecke F, Hess M.** 2009. A perspective: metatranscriptomics as a tool for the discovery of novel biocatalysts. J. Biotechnol. **142**:91–95.
45. **Wels MWW.** 2008. Unraveling the regulatory network of *Lactobacillus plantarum* WCFS1. Wageningen University, Wageningen, the Netherlands.
46. **Wilhelm BT, Landry JR.** 2009. RNA-seq-quantitative measurement of expression through massively parallel RNA-sequencing. Methods **48**: 249–257.
47. **Willenbrock H, et al.** 2009. Quantitative miRNA expression analysis: comparing microarrays with next-generation sequencing. RNA **15**:2028–2034.
48. **Wurtzel O, et al.** 2010. A single-base resolution map of an archaeal transcriptome. Genome Res. **20**:133–141.
49. **Yang YH, et al.** 2002. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res. **30**:e15.
50. **Yoder-Himes DR, et al.** 2009. Mapping the *Burkholderia cenocepacia* niche response via high-throughput sequencing. Proc. Natl. Acad. Sci. U. S. A. **106**:3976–3981.
51. **Zhou J, Thompson DK.** 2002. Challenges in applying microarrays to environmental studies. Curr. Opin. Biotechnol. **13**:204–207.
52. **Zoetendal EG, et al.** 2006. Isolation of RNA from bacterial samples of the human gastrointestinal tract. Nat. Protoc. **1**:954–959.