Appl. Environ. Microbiol. doi:10.1128/AEM.06470-11

COMPARATIVE ANALYSIS OF *LACTOBACILLUS PLANTARUM* WCFS1 TRANSCRIPTOMES USING DNA MICROARRAY AND NEXT GENERATION SEQUENCING TECHNOLOGIES

Running Title: Comparison of transcriptomic technologies for bacteria

Milkha M. Leimena[a,b], Michiel W.W. 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), P.O. Box 557, 6700 AN Wageningen, the Netherlands[a]; Laboratory of Microbiology[b], Laboratory of Food Microbiology[c], and Host-Microbe Interactomics Group[d], Wageningen University, Dreijenplein 10, 6703 HB, Wageningen, the Netherlands; NIZO Food Research B.V., P.O. Box 20, 6710 BA, Ede, the Netherlands[e]

[#] Correspondent footnote. Mailing address: NIZO Food Research B.V., P.O. Box 20, 6710 BA, Ede, the Netherlands. Phone: +31318659 629. Fax: +31318650400. E-mail: michiel.kleerebezem@nizo.nl

Abstract

RNA sequencing is starting to compete with the use of DNA microarrays for transcription analysis in eukaryotes as well as in prokaryotes. 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'-UTR sequencing) using *Lactobacillus plantarum* WCFS1 as a model organism, employing its established microarray platform as a reference. Limited impact 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 depth of analysis and generated similar fold-change ratio 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 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 between the three technologies.

Introduction

Understanding the influence of environmental conditions on genome-wide gene expression levels requires accurate quantification of all expressed (m)RNAs. Microarrays provide an effective method for analysis of thousands of transcripts in a parallel manner, and allow 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 low reproducibility of results between laboratories (16) during microarray analyses could limit the usage of microarray 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, which is especially suited for samples from biological material with an 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 (ds) cDNA fragments by poly-T (or random) primed reverse transcription and strand duplication, followed by direct sequencing of the ds 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 microarrays, several studies were conducted using RNA of eukaryote cells, such as human liver and kidney (22), and mouse hippocampi (37). These comparative studies revealed a good correlation between the level 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, 37). 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 as compared to 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 employ these methods for prokaryote transcription analysis. This is not surprising since the prokaryote RNA pool contains a high amount of rRNA and tRNA, which may constitute more than 95% of the total RNA (31), while 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 as compared to those of eukaryotes. To increase the

84 relative abundance of mRNA in total prokaryote RNA material, several methods have
85 been developed, including rRNA capture, selective degradation of processed RNA,
86 selective poly-adenylation of mRNA, and antibody capture of subsets of mRNAs that
87 interact with Hfq proteins (32). Due to the lack of a poly-A tail in prokaryote mRNA,
88 alternative priming approaches during reverse transcription (RT) are commonly based on
89 random oligonucleotide priming (hexamers or longer), and sometimes employ
90 multiplexed gene specific oligonucleotides (28, 50), or a combination with gene specific
91 priming of the 5'-end RNA-seq (48). Alternatively, oligo-dT priming can be employed
92 following artificial poly-adenylation of mRNAs (13). The development of mRNA
93 enrichment and priming methods allowed the successful use of RNA sequencing
94 approaches for the investigation of transcriptome changes under different growth
95 conditions of *Burkholderia cenocepacia* (50) and *Bacillus anthracis* (28).
96 In this study, we aimed to validate the use of different RNA sequencing techniques using
97 a model prokaryote organism, while employing an established microarray platform as a
98 reference. The transcriptomes of *Lactobacillus plantarum* WCFS1 (grown in two
99 different laboratory media) were compared using custom-made oligonucleotide
100 microarrays and RNA sequencing approaches. The microarray was also employed to
101 evaluate the impact on the transcriptome of mRNA enrichment by RNA capture methods.
102 This study includes the comparison of two RNA sequencing approaches, direct cDNA-
103 and 3'-UTR cDNA- sequencing to evaluate their applicability in prokaryote
104 transcriptome analyses. Our analyses show that the depth of analysis for both RNA
105 sequencing methodologies was similar to that observed for the microarray, leading to a
106 coverage of >95% of all genes encoded in the *L. plantarum* WCFS1 genome. The best
107 transcriptome correlation was found between microarray and direct cDNA sequencing

analyses, while the 3'-UTR sequencing method appeared to deviate 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) (39) and de Man Rogosa Sharpe (MRS) medium (8) at 37°C without agitation. Cells were harvested by centrifugation for 10 minutes at 4570×g and 4°C using a Heraeus Multifuge 3 S-R Centrifuge (DJB Labcare Ltd., England), at an optical density (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 mini kit (Qiagen, USA), including an on-column DNAseI (Roche, Germany) treatment as described previously (52). 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 (MICROBExpressTM, Ambion, Applied Biosystem, Niewerkerk a/d Ijssel, The Netherlands) (44). Total RNA and enriched mRNA yields were quantified spectrophotometrically (NanoDrop 1000; Nanodrop Technologies, Wilmington, USA) and total RNA quality was assessed by microfluidics-based electrophoresis system (Experion RNA Stdsens, Biorad Laboratories Inc., USA).

DNA Microarray based transcriptome analysis

The microarray used was a custom designed *L. plantarum* WCFS1, 8×15-K 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 in Figure 1. Labeled cDNA mixtures were subsequently concentrated in a Hetovac VR-1 (Heto Lab Equipment A/S, Birkerod, Denmark) to a final volume of 25µl (if needed), incubated at 98°C for 3 minutes, and cooled at room temperature for 5 minutes. After addition of 25µl 2×GEX HI-RPM hybridization buffer (Agilent technologies, Palo Alto, Ca, USA), 40µl of each mixture was applied on an Agilent 8×15K array (Agilent technologies, Palo Alto, Ca, USA). 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 the ImaGene Version 7.5 software (BioDiscovery Inc., Marina Del Rey, CA, USA). Both total RNA and mRNA enriched datasets were normalized and corrected by local fitting of an M-A plot applying the Lowess algorithm (49) and interslide scaling, as 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 Expression Omnibus (10) and are accessible through GEO series accession number GSE35754.

RNA sequencing based transcriptome analysis

Double stranded cDNA was synthesized using enriched mRNA of CDM or MRS grown *L. plantarum* WCFS1 using the SuperScriptTM Double Stranded cDNA Synthesis kit

(Invitrogen, 11917-010), with the addition of SuperScriptTM III Reverse Transcriptase (Invitrogen, 18080-044) and random primers (Invitrogen, 48190-011) 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, USA), 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 of both CDM and in 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 M-MLV H-reverse transcriptase. The resulting cDNAs were PCR-amplified to 20-30 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 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 concentration. 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. Mapping and quantification of the cleaned sequences to an *in*

183 *silico* transcriptome reference was performed by GATC Biotech. The cDNA reference
184 was created using the annotation of *L. plantarum* WCFS1 genome obtained from UCSC
185 Genome Bioinformatics (http://genome.ucsc.edu/). To map the 3'-UTR derived
186 sequences to the appropriate gene specific transcripts, additional mappings of 100 bps,
187 200 bps and 300 bps 3'-extended transcriptional unit predictions were employed (45).
188 These multiple mappings were performed to increase the frequency of read assignments
189 to genes, because the position of a 3'-UTR at the end of a transcript in *L. plantarum*
190 WCFS1 is unknown. Visualization of the mapped transcript was performed using the
191 UCSC Genome Browser (http://microbes.ucsc.edu/).
192
193 ***Comparative data analyses of direct cDNA sequencing versus 3'-UTR sequencing and***
194 ***microarray versus both RNA sequencings***
195 The signal intensity data obtained by microarrays and the number of read counts of direct
196 cDNA and 3'-UTR sequencings were quantile normalized (5) using the CLC-Bio
197 Genomic Workbench software (http://www.clcbio.com/) to adjust the data range.
198 Normalized read counts of direct cDNA sequencing were plotted against those of the 3'-
199 UTR sequencing using a scatter plot to investigate the reads distribution between the two
200 datasets. Rank based analysis using Spearman correlation coefficient was applied to
201 investigate the correlation between two sequencing techniques for CDM and MRS culture
202 derived RNA samples. For the comparison with the microarray, which was utilized as the
203 benchmark technology, normalized microarray signal intensities were used for the
204 comparison of the normalized read counts from both RNA sequencing techniques using
205 Spearman correlation analysis. Analysis of the differentially expressed genes (DEG)
206 based on log2 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 varying dynamic range of the different techniques. Only those genes that showed >2-fold absolute fold change ratio for all techniques and displayed significant (FDR-adjusted *P*-values < 0.05) differential expression according to the microarray analysis were used. Spearman and Pearson correlation analyses were performed using the PASW Satistic 17.0 software suite (SPSS Inc., Chicago, USA).

Results and Discussion

Microarray transcript profiles for total RNA and mRNA samples

In this study, *L. plantarum* WCFS1 were grown in two laboratory media (CDM or 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 versus mRNA enriched transcriptome datasets 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 <0.05 when comparing total RNA versus enriched mRNA samples for both growth conditions (Figure 2), indicating that mRNA enrichment has only a limited impact 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 very lowly expressed within the dataset. The majority of these differentially quantified genes were related to hypothetical protein functions (Figure 3). In addition, the limitation of the mRNA enrichment method used (MICROBExpressTM, Ambion), which does not target tRNA removal, resulted in differential quantification of some tRNAs in the mRNA enriched fraction (17).

239 Variation in gene expression level caused by the different growth conditions (CDM
240 versus MRS) was observed for a total of 207 genes (FDR-adjusted p-values <0.05) from
241 both total RNA and mRNA enriched analyses. Of these 207 genes, 180 genes were shared
242 between the differential genes identified in the total RNA and mRNA enriched samples,
243 of which 178 showed conserved up- or down-regulation as a consequence of the
244 difference in growth medium (Figure 2). Average linkage hierarchical clustering with
245 Pearson correlation distance (35) confirmed a more pronounced separation of CDM and
246 MRS profiles relative to the separation seen for total RNA versus mRNA enrichment
247 profiles (Figure 3).
248 This finding shows that the transcriptome variation caused by different growth conditions
249 exceeds the variation caused by enrichment procedure, indicating that mRNA enrichment
250 will only have a limited impact on the biological interpretation of transcriptome data,
251 which is validated for a well-defined culture under well-defined conditions, with the
252 anticipation towards similar performance in the complex ecosystems. The genes
253 displaying significant differential expression in cultures grown on CDM compared to
254 MRS predominantly belonged to specific functional categories that appear to reflect the
255 different medium composition, such as transport and binding proteins (in particular for
256 amino acid, peptides and amines), amino acid biosynthesis (in particular for histidine and
257 aspartate), energy metabolism, and synthesis of purines, pyrimidines, nucleosides and
258 nucleotides (Figure 3; Figure S1). The limited amount of nucleotides (18) and specific
259 amino acids available in CDM relative to MRS apparently requires an alternative pallet of
260 transport functions to import those components, which could consistently be concluded
261 from the arrays irrespective of the RNA source (enriched mRNA versus total RNA) used.
262

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 to 28.1 million per sample (Table S1) with an average trimmed length of 36bp. 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%) (Table S1). A possible explanation for the strongly reduced CDS-mapping of the short reads (~36bp) obtained by 3'-UTR sequencing is most probably due to 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 step-wise 3'-extension of the CDS with 100bp, 200bp and 300bp. *In silico* predictions indicated that approximately 75% of the predicted terminator sequences in the *L. plantarum* WCFS1 genome were encompassed within the 100bp extension (7), while an additional 12% and 6% of the predicted terminators were encountered within the 200bp and 300bp extended 3'-UTRs, respectively (Figure S2). Analogously, *in silico* 3'-extension of the CDS of the *L. plantarum* WCFS1 genome by 100bp enabled an 80% and 130% increase in the gene-specific mapping of the CDM and MRS 3'-UTR transcript

sequence datasets, respectively. Notably, larger 3'-extension of the gene sequences with 200bp and 300bp led to significantly smaller increases of CDS-specific transcript mapping (~25% and ~30%, respectively), supporting the prediction that 75% of the terminators within the first 100 bases downstream of the CDS (Figure S2A). Moreover, 200bp and 300bp 3'-extension 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 (Figure S2B). Based on these analyses, 100bp 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 (Table S2), 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 datasets (~70%) consistently mapped to the last gene of such polycistronic transcripts (Figure 4). This indicates that accurate functional interpretation of 3'-UTR sequence datasets 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 (3135 genes) of the *L. plantarum* WCFS1 genome were at least covered

311 by a single sequence read. Similar read distribution was observed for direct cDNA
312 sequencing and 3'-UTR sequencing (similar proportion between the area above and
313 below the continuous line; figure 5), which indicates similar gene expression patterns.
314 Notably, several genes were apparently overestimated by 3'-UTR sequencing, (upper left
315 area of figure 5), which may be due to either a technical artefact by the application of
316 poly-A tail, an artefact in the data interpretation by 100bp extension of the mapping, or
317 the existence of some internal promoters (7).
318
319 ***RNA sequencing validation by comparative analysis with the Microarray-derived***
320 ***transcriptomes***
321 Since microarrays can be considered as an 'established' transcriptome methodology, the
322 data obtained from the microarrays were employed as a reference to evaluate the overall
323 validity of direct cDNA and 3'-UTR sequencing. Only the genes that gave a value for all
324 methods (2962 genes) were used for comparison of the transcriptomes obtained by
325 microarray and direct cDNA or 3'-UTR sequencing. Normalized signal intensity values
326 per gene obtained by microarray analysis were plotted against normalized CDS-read
327 assignment frequencies derived from both RNA sequencing methods.
328 Both microarray and RNA sequencing transcriptome datasets were normalized using
329 quantile normalization, as a quick and simple method to create and even distribution of
330 microarray probe intensities and RNA sequencing read counts (5). Additional
331 normalization approaches, such as RPKM (reads per kilobase of exon model per million
332 mapped reads) (25) or FPKM (fragments per kilobase of transcript per million fragments
333 mapped) (40) approach, which take into consideration the influence of transcript length
334 towards the gene expression quantification of RNA sequencing reads, could give more

335 accurate gene expression quantification, especially of direct cDNA sequencing. Although
336 overall transcriptome comparisons was done without considering the transcript length,
337 high comparability was shown between microarray and direct cDNA sequencing as well
338 as between microarray and 3'-UTR sequencing (Figure 6).
339 Direct cDNA sequencing displayed a higher correlation to the microarray as compared to
340 the 3'-UTR (Figure 6). This was especially apparent in transcripts with relatively high 3'-
341 UTR sequencing assignments compared to the array signal intensity (upper left area of
342 figures 6C and 6D). These results indicate that direct cDNA sequencing generates
343 transcriptome results that resemble those obtained by microarray transcriptome profiling,
344 and that 3'-UTR appears to estimated expression levels are higher for subsets of genes as
345 compared to the other two methods. These conclusions were also supported by rank based
346 Spearman correlation analysis, showing higher correlation values between microarray and
347 direct cDNA sequencing datasets (CDM: 0.835 [$p<0.01$] and 0.762 [$p<0.01$]; MRS: 0.881
348 [$p<0.01$] and 0.707 [$p<0.01$], for direct cDNA sequencing and 3'-UTR sequencing,
349 respectively).
350 The application of 3'-UTR sequencing as a method for prokaryote transcriptome analysis
351 has not yet been well established and may require additional normalization or processing
352 steps to obtain an appropriate quantitative representation of the transcript levels that can
353 be compared with microarray derived transcriptome datasets. To evaluate whether the
354 lower correlation between 3'-UTR sequencing and array-based transcript datasets was
355 caused by a biased positioning of the sequence reads within an operon, the expression
356 values of the last genes in operons was also assigned to each upstream encoded gene
357 within the same predicted operon. However, this data transformation step to
358 accommodate polycistronic operon transcript in the 3'-UTR data did not improve the

359 correlation with the array derived datasets (data not shown). This may suggest that the
360 lower correlation of these datasets may arise from a bias in the 3'-UTR extension or
361 sequencing technology employed.
362 The most relevant comparative analysis of the three methods employed here undoubtedly
363 relates to the comparisons of the biological conclusions they may generate. To this end,
364 the ability of the three technologies to consistently identify the same genes (1, 22) that are
365 differentially expressed (DEG) when comparing growth on CDM and MRS. The
366 sequence-based transcriptome quantification was determined by the ratio of sequence
367 reads assigned to a gene in CDM and MRS obtained datasets, while the differential
368 expression per gene in the microarray dataset was calculated using CyberT (3). In total,
369 538 DEG with an expression fold-change >2 were detected within the DNA microarray;
370 while 442 and 466 DEG with an expression fold-change >2 were detected by direct
371 cDNA sequencing and 3'-UTR sequencing, respectively. Among the latter groups of
372 genes, 233 and 204 DEG were shared between the microarray-based analysis and direct
373 cDNA sequencing and 3'-UTR sequencing, respectively. Moreover, 172 genes were
374 identified to have absolute fold change >2 for all techniques with the same up- or down-
375 regulation pattern; among which 152 genes were considered to be significantly
376 differentially expressed according to the DNA microarray technology (FDR < 0.05) that
377 was used as the reference technology. Thereby, this comparative analysis of differentially
378 expressed genes establishes a good consistency of the biological outcomes generated by
379 the three transcriptome technologies, characterized by similar fold-change of expression
380 for most shared DEG. Heat map analysis of the differential expression data confirmed
381 that 3'-UTR sequencing deviates slightly more from the microarray than direct cDNA
382 sequencing (Figure 7). This is also reflected by the somewhat lower Pearson correlation

383 when comparing the microarray with 3'-UTR-sequencing (0.852; $p<0.01$) relative to the
384 comparison of the microarray with the direct cDNA sequencing (0.897; $p<0.01$). Notably,
385 the highest Pearson correlation was obtained of the two sequencing based technologies
386 (0.951; $p<0.01$), which might be due to the saturated hybridization-signals in the array
387 datasets (Figure S3) (14).
388 Since this analysis of DEG was performed taking microarray data as a reference, DEG
389 that display differential expression only according to the transcriptome sequencing
390 analyses may have been missed. DEG analysis of the direct cDNA and 3'-UTR
391 sequencing datasets revealed 50 additional genes with a differential expression value > 2
392 in both sequencing based datasets. Of these genes, 40 appeared not to reach significance
393 of regulation in the array dataset (FDR >0.05), but displayed conserved direction of
394 differential expression according the array analyses, albeit it with < 2 absolute fold
395 change ratio. Moreover, many of the probes associated with 36 of these 40 genes revealed
396 saturated hybridization-signals in the array datasets (Figure S3), suggesting that they were
397 inaccurately measured by the array due to falling outside the dynamic range of the array
398 technology (14). This observation implies that RNA sequencing may exceed the depth of
399 analysis in comparison to the 'traditional' array technologies, especially for genes that are
400 expressed at a high level.
401 Unlike microarray data, RNA sequencing count data is generally not well represented as
402 continues distribution (27). Therefore, normalization procedure which are successfully
403 applied for microarray data, might not be optimal for RNA sequencing dataset. Data
404 normalization based on parametric approaches was implemented in several analyses
405 platforms, such as edgeR (29), baySeq (15), and DESeq (2), which allow lowering of
406 both biological and technical variability of replicated count data. Moreover, non-

parametric approaches, like the noise modeling employed in NOISeq, allow the evaluation of low expression counts without any needs of replicates (38). 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 towards undefined ecosystem metatrancriptome 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 validation of sequence based transcriptomics methodology is required to confidently apply sequence based transcriptome methods in samples derived from complex microbial communities with unknown composition and that live in poorly defined growth conditions. Such ecosystem meta-transcriptomic analyses cannot be performed using DNA microarrays due to sequence variations among the coding capacities among (close) relatives of similar phylogenetic origin, which makes quantification of transcripts on 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 biological (genes and functions) interpretation of 3'-UTR RNA sequencing datasets obtained from complex microbial communities with an 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 meta-transcriptome datasets of

an unknown microbial community that offers good possibilities for biological interpretation taking a set of representative microbial genomes as a mapping platform.

Acknowledgement

The authors would like to thank Yadhu Kumar and Nastasja Trunk of GATC-Biotech (Konstanz, Germany) for assistance in providing the bioinformatics support and additional information regarding the RNA sequencing procedures. This project was supported by the Netherlands Bioinformatics Centre (NBIC).

438 **Reference**

439 1. **Agarwal, A., D. Koppstein, J. Rozowsky, A. Sboner, L. Habegger, L. W.**
440 **Hillier, R. Sasidharan, V. Reinke, R. H. Waterston, and M. Gerstein.** 2010.
441 Comparison and calibration of transcriptome data from RNA-Seq and tiling
442 arrays. BMC Genomics **11:**383-399.

443 2. **Anders, S., and W. Huber.** 2010. Differential expression analysis for sequence
444 count data. Genome biology **11:**R106.

445 3. **Baldi, P., and A. D. Long.** 2001. A Bayesian framework for the analysis of
446 microarray expression data: regularized t -test and statistical inferences of gene
447 changes. Bioinformatics (Oxford, England) **17:**509-519.

448 4. **Bloom, J. S., Z. Khan, L. Kruglyak, M. Singh, and A. A. Caudy.** 2009.
449 Measuring differential gene expression by short read sequencing: quantitative
450 comparison to 2-channel gene expression microarrays. BMC Genomics **10:**221-
451 231.

452 5. **Bolstad, B. M., R. A. Irizarry, M. Astrand, and T. P. Speed.** 2003. A
453 comparison of normalization methods for high density oligonucleotide array data
454 based on variance and bias. Bioinformatics (Oxford, England) **19:**185-193.

455 6. **Cloonan, N., A. R. R. Forrest, G. Kolle, B. B. A. Gardiner, G. J. Faulkner, M.**
456 **K. Brown, D. F. Taylor, A. L. Steptoe, S. Wani, G. Bethel, A. J. Robertson, A.**
457 **C. Perkins, S. J. Bruce, C. C. Lee, S. S. Ranade, H. E. Peckham, J. M.**
458 **Manning, K. J. McKernan, and S. M. Grimmond.** 2008. Stem cell
459 transcriptome profiling via massive-scale mRNA sequencing. Nature Methods
460 **5:**613-619.

461 7. **de Hoon, M. J. L., Y. Makita, K. Nakai, and S. Miyano.** 2005. Prediction of
462 transcriptional terminators in *Bacillus subtilis* and related species. PLoS
463 computational biology **1:**e25.

464 8. **De Man, J. C., M. Rogosa, and M. E. Sharpe.** 1960. A Medium for the
465 cultivation of Lactobacilli. Journal of Applied Microbiology **23:**130-135.

466 9. **Deutscher, M. P.** 2006. Degradation of RNA in bacteria: comparison of mRNA
467 and stable RNA. Nucleic Acids Res **34:**659-666.

468 10. **Edgar, R., M. Domrachev, and A. E. Lash.** 2002. Gene Expression Omnibus:
469 NCBI gene expression and hybridization array data repository. Nucleic acids
470 research **30:**207-210.

471 11. **Eveland, A. L., D. R. McCarty, and K. E. Koch.** 2008. Transcript profiling by
472 3'-untranslated region sequencing resolves expression of gene families. Plant
473 Physiol **146:**32-44.

474 12. **Everett, K. R., J. Rees-George, I. P. S. Pushparajah, B. J. Janssen, and Z.**
475 **Luo.** 2010. Advantages and disadvantages of microarrays to study microbial
476 population dynamics - a minireview. New Zealand Plant Protection **63:**1-6.

477 13. **Frias-Lopez, J., Y. Shi, G. W. Tyson, M. L. Coleman, S. C. Schuster, S. W.**
478 **Chisholm, and E. F. Delong.** 2008. Microbial community gene expression in
479 ocean surface waters. Proceedings of the National Academy of Sciences of the
480 United States of America **105:**3805-3810.

481 14. **Garcia de la Nava, J., S. van Hijum, and O. Trelles.** 2004. Saturation and
482 quantization reduction in microarray experiments using two scans at different
483 sensitivities. Stat Appl Genet Mol Biol **3:**1-16.

15. **Hardcastle, T. J., and K. A. Kelly.** 2010. baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. BMC Bioinformatics **11:**422.

16. **Irizarry, R. A., D. Warren, F. Spencer, I. F. Kim, S. Biswal, B. C. Frank, E. Gabrielson, J. G. N. Garcia, J. Geoghegan, G. Germino, C. Griffin, S. C. Hilmer, E. Hoffman, A. E. Jedlicka, E. Kawasaki, F. Martinez-Murillo, L. Morsberger, H. Lee, D. Petersen, J. Quackenbush, A. Scott, M. Wilson, Y. Yang, S. Q. Ye, and W. Yu.** 2005. Multiple-laboratory comparison of microarray platforms. Nature Methods **2:**345-350.

17. **Kang, Y., M. H. Norris, J. Zarzycki-Siek, W. C. Nierman, S. P. Donachie, and T. T. Hoang.** 2011. Transcript amplification from single bacterium for transcriptome analysis. Genome Research **21:**925-935.

18. **Kilstrup, M., K. Hammer, P. Ruhdal Jensen, and J. Martinussen.** 2005. Nucleotide metabolism and its control in lactic acid bacteria. FEMS Microbiol Rev **29:**555-590.

19. **Kleerebezem, M., J. Boekhorst, R. van Kranenburg, D. Molenaar, O. P. Kuipers, R. Leer, R. Tarchini, S. A. Peters, H. M. Sandbrink, M. W. Fiers, W. Stiekema, R. M. Lankhorst, P. A. Bron, S. M. Hoffer, M. N. Groot, R. Kerkhoven, M. de Vries, B. Ursing, W. M. de Vos, and R. J. Siezen.** 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., V. Ruotti, R. M. Stewart, J. A. Thomson, and C. N. Dewey.** 2010. RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics (Oxford, England) **26:**493-500.

22. **Marioni, J. C., C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad.** 2008. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. Genome Research **18:**1509-1517.

23. **Meijerink, M., S. van Hemert, N. Taverne, M. Wels, P. de Vos, P. A. Bron, H. F. Savelkoul, J. van Bilsen, M. Kleerebezem, and J. M. Wells.** 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., M. Hirst, and M. A. Marra.** 2009. Applications of new sequencing technologies for transcriptome analysis. Annu Rev Genomics Hum Genet **10:**135-151.

25. **Mortazavi, A., B. A. Williams, K. McCue, L. Schaeffer, and B. Wold.** 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods **5:**621-628.

26. **Nagalakshmi, U., Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder.** 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science (New York, N Y) **320:**1344-1349.

27. **Oshlack, A., M. D. Robinson, and M. D. Young.** 2010. From RNA-seq reads to differential expression results. Genome biology **11:**220.

28. **Passalacqua, K. D., A. Varadarajan, B. D. Ondov, D. T. Okou, M. E. Zwick, and N. H. Bergman.** 2009. Structure and complexity of a bacterial transcriptome. J Bacteriol **191:**3203-3211.

29. **Robinson, M. D., D. J. McCarthy, and G. K. Smyth.** 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics (Oxford, England) **26:**139-140.

30. **Serrano, L. M., D. Molenaar, M. Wels, B. Teusink, P. A. Bron, W. M. de Vos, and E. J. Smid.** 2007. Thioredoxin reductase is a key factor in the oxidative stress response of *Lactobacillus plantarum* WCFS1. Microbial Cell Factories **6:**29.

31. **Siezen, R. J., G. Wilson, and T. Todt.** 2010. Prokaryotic whole-transcriptome analysis: deep sequencing and tiling arrays. Microbial Biotechnology **3:**125-130.

32. **Sorek, R., and P. Cossart.** 2010. Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. Nature reviews Genetics **11:**9-16.

33. **Stevens, M. J. A., D. Molenaar, A. de Jong, W. M. De Vos, and M. Kleerebezem.** 2010. sigma54-Mediated control of the mannose phosphotransferase sytem in *Lactobacillus plantarum* impacts on carbohydrate metabolism. Microbiology (Reading, England) **156:**695-707.

34. **Stoughton, R. B.** 2005. Applications of DNA microarrays in biology. Annu Rev Biochem **74:**53-82.

35. **Sturn, A., J. Quackenbush, and Z. Trajanoski.** 2002. Genesis: cluster analysis of microarray data. Bioinformatics (Oxford, England) **18:**207-208.

36. **Sultan, M., M. H. Schulz, H. Richard, A. Magen, A. Klingenhoff, M. Scherf, M. Seifert, T. Borodina, A. Soldatov, D. Parkhomchuk, D. Schmidt, S. O'Keeffe, S. Haas, M. Vingron, H. Lehrach, and M.-L. Yaspo.** 2008. A global

553 view of gene activity and alternative splicing by deep sequencing of the human
554 transcriptome. Science (New York, N Y) **321:**956-960.
555 37. **t Hoen, P. A., Y. Ariyurek, H. H. Thygesen, E. Vreugdenhil, R. H. Vossen, R.**
556 **X. de Menezes, J. M. Boer, G. J. van Ommen, and J. T. den Dunnen.** 2008.
557 Deep sequencing-based expression analysis shows major advances in robustness,
558 resolution and inter-lab portability over five microarray platforms. Nucleic Acids
559 Res **36:**e141.
560 38. **Tarazona, S., F. Garcia-Alcalde, J. Dopazo, A. Ferrer, and A. Conesa.** 2011.
561 Differential expression in RNA-seq: a matter of depth. Genome Research
562 **21:**2213-2223.
563 39. **Teusink, B., F. H. van Enckevort, C. Francke, A. Wiersma, A. Wegkamp, E.**
564 **J. Smid, and R. J. Siezen.** 2005. In silico reconstruction of the metabolic
565 pathways of *Lactobacillus plantarum*: comparing predictions of nutrient
566 requirements with those from growth experiments. Appl Environ Microbiol
567 **71:**7253-7262.
568 40. **Trapnell, C., A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, H.**
569 **Pimentel, S. L. Salzberg, J. L. Rinn, and L. Pachter.** 2012. Differential gene
570 and transcript expression analysis of RNA-seq experiments with TopHat and
571 Cufflinks. Nature Protocols **7:**562-578.
572 41. **van Hijum, S. A. F. T., J. Garcia de la Nava, O. Trelles, J. Kok, and O. P.**
573 **Kuipers.** 2003. MicroPreP: a cDNA microarray data pre-processing framework.
574 Appl Bioinformatics **2:**241-244.
575 42. **van Vliet, A. H.** 2009. Next generation sequencing of microbial transcriptomes:
576 challenges and opportunities. FEMS Microbiol Lett **302:**1-7.

43. **Wang, Z., M. Gerstein, and M. Snyder.** 2009. RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews Genetics **10:**57-63.

44. **Warnecke, F., and M. Hess.** 2009. A perspective: metatranscriptomics as a tool for the discovery of novel biocatalysts. J Biotechnol **142:**91-95.

45. **Wels, M. W. W.** 2008. Unraveling the regulatory network of *Lactobacillus plantarum* WCFS1. Wageningen University, Wageningen.

46. **Wilhelm, B. T., and J. R. Landry.** 2009. RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. Methods **48:**249-257.

47. **Willenbrock, H., J. Salomon, R. Sokilde, K. B. Barken, T. N. Hansen, F. C. Nielsen, S. Moller, and T. Litman.** 2009. Quantitative miRNA expression analysis: comparing microarrays with next-generation sequencing. Rna **15:**2028-2034.

48. **Wurtzel, O., R. Sapra, F. Chen, Y. Zhu, B. A. Simmons, and R. Sorek.** 2010. A single-base resolution map of an archaeal transcriptome. Genome Research **20:**133-141.

49. **Yang, Y. H., S. Dudoit, P. Luu, D. M. Lin, V. Peng, J. Ngai, and T. P. Speed.** 2002. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic acids research **30:**e15.

50. **Yoder-Himes, D. R., P. S. Chain, Y. Zhu, O. Wurtzel, E. M. Rubin, J. M. Tiedje, and R. Sorek.** 2009. Mapping the *Burkholderia cenocepacia* niche response via high-throughput sequencing. Proc Natl Acad Sci U S A **106:**3976-3981.

51. **Zhou, J., and D. K. Thompson.** 2002. Challenges in applying microarrays to environmental studies. Current Opinion in Biotechnology **13:**204-207.

52. **Zoetendal, E. G., C. C. G. M. Booijink, E. S. Klaassens, H. G. H. J. Heilig, M. Kleerebezem, H. Smidt, and W. M. de Vos.** 2006. Isolation of RNA from bacterial samples of the human gastrointestinal tract. Nature Protocols **1:**954-959.

605 **List of Figures**

606

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

610

611 **FIG 2** Venn diagram showing the number of up-regulated/down-regulated/oppositely
612 regulated genes in the enriched mRNA sample obtained from CDM and MRS grown
613 bacterial cells (panel A) or in the RNA obtained from CDM growing cells, either total
614 RNA or after mRNA enrichment (panel B).

615

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

623

624 **FIG 4** Mapping of *L. plantarum* WCFS1 transcripts from direct cDNA sequencing and
625 3'-UTR sequencing of MRS and CDM grown cultures based on a predicted
626 transcriptional unit (38). Scaling differences of the Y-axis range are indicative for the up-
627 regulated transcription level observed in cells obtained from CDM grown cultures.

628

629 **FIG 5** Comparison of normalized signal intensity between direct cDNA sequencing and
630 3'-UTR sequencing, for bacteria grown in CDM (Spearman: 0.686; $p<0.01$) and MRS
631 (Spearman: 0.678; $p<0.01$) growth.
632
633 **FIG 6** Comparison between normalized signal intensity level of microarray and
634 normalized read counts of direct cDNA sequencing (A and B) and 3'-UTR sequencing (C
635 and D) in transcriptome datasets from bacteria grown in CDM and MRS.
636
637 **FIG 7** Comparison of 152 transcript levels (40 down-regulated in CDM and 112 up-
638 regulated in CDM) that consistently were classified among the DEG gene sets,
639 determined by microarray transcriptomes, or direct cDNA and 3'-UTR transcriptome
640 sequencing. Data are sorted according to their fold-change within the reference datasets
641 (DNA microarray technology).