Supplementary tables, figures and information for:

COLOMBOS v2.0: An ever expanding collection of bacterial expression compendia

Pieter Meysman[1,2], Paolo Sonego[3], Luca Bianco[3], Qiang Fu[4], Daniela Ledezma-Tejeida[5], Socorro Gama-Castro[5], Veerle Liebens[4], Jan Michiels[4], Kris Laukens[1,2], Kathleen Marchal[4,6,7], Julio Collado-Vides[5] and Kristof Engelen[3,4,*]

[1] Department of Mathematics and Computer Science, University of Antwerp, B-2020 Antwerp, Belgium
[2] Biomedical Informatics Research Center Antwerp (biomina), University of Antwerp / Antwerp University Hospital, B-2650 Edegem, Belgium
[3] Department of Computational Biology, Fondazione Edmund Mach, San Michele all'Adige, Trento (TN), 38010, Italy
[4] Department of Microbial and Molecular Sciences, KU Leuven, Leuven, B-3001, Belgium
[5] Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, 62210, Mexico
[6] Department of Plant Biotechnology and Bioinformatics, Ghent University, Gent, 9052, Belgium
[7] Department of Information Technology, IMinds, Ghent University, Gent, 9052, Belgium

* To whom correspondence should be addressed. Tel: +390461615646; Fax: +390461650218; Email: kristof.engelen@fmach.it

Comparison of RNA-seq and microarray data

We performed a dedicated experiment to evaluate the validity of combining data from both microarray and RNA-Seq methods in COLOMBOS, i.e. following the COLOMBOS homogenization strategy. In this experiment, we measured the exact same four RNA samples in parallel on Illumina MiSeq and on Affymetrix *E. coli* Genome 2.0 arrays (see Materials and Methods below for details). The four samples consisted of two biological replicates of two strains growing in exponential phase in minimal medium: the *Escherichia coli* K12 MG1655 *wild-type* (biological replicates designated WTB and WTC) and a Δ*ydcR (b1439)* mutant (biological replicates designated 23A and 23B). The microarray data were processed using the homogenization and normalization pipelines as described in the original COLOMBOS publication (1); the RNA-seq data were processed as described below in the Materials and Methods in this document. The three condition contrasts that end up in COLOMBOS for both the RNA-seq and the microarrays were defined as shown in Table ST1: the wild-type sample WTB was used as a common reference for the other three samples.

Figure S1, S2 and S3 show that for all the three comparisons, the bulk of the data lies on the red bisector, which represents 'perfect' correspondence between the two technologies. This is a visual indication that there is good correspondence between the log-ratios obtained from microarrays and those obtained from the RNA-seq data. More detailed analysis reveals that:

1. There is agreement between the two technologies with a Pearson correlation coefficient going from 0.39 (moderate positive relationship) to 0.70 (strong positive relationship).

2. The 'error noise' is in the same range for both platforms. One of the underlying assumptions of our normalization strategy is that the bulk of the genes do not change their expression between two conditions, i.e. that the log-ratios for a condition contrast should feature a prominent 'noise' distribution (which we assume Guassian) around 0, implying no change. (Note that during the normalization, we do not force the data to comply with this, instead we rely on this assumption to assess the appropriateness of the normalization *afterwards*.) We estimated this Guassian 'noise' distribution for all three contrasts in a robust way (i.e. taking outliers into account) using an iterative approach. The means were never significantly different from zero and the standard deviations for both the microarray and the RNA-seq data for all the contrasts are given in Table ST1. The strong similarity of these standard deviations between identical contrasts measured on the different platforms suggest that the measurement range of expression differences is comparable between the two technologies.

3. Taking into account point 1 and 2 above, a more informative way to assess the agreement between the two technologies than an overall correlation, is to see if genes showing obvious expression differences (i.e. away from the estimated noise distribution values) were consistently reported by both technologies. This is indeed the case in Figures S1, S2 and S3, where the genes with higher magnitudes of expression changes are generally situated in the bottom-left and top-right quadrants close to the bisector.

Our results are in line with several other studies (28-36) that reported substantial agreement between RNA-seq and microarray technologies, further giving validity to the idea that we can enrich the COLOMBOS compendia with data from both technologies.

Contrasts	Pearson correlation	σ microarray	σ RNA-seq
WTC vs WTB	0.70	0.2103	0.2221
23A vs WTB	0.59	0.1438	0.1644
23B vs WTB	0.39	0.1496	0.1575

Table ST1: Pearson correlation coefficient, and robust standard deviation estimates for both microarray and RNA-seq data for the three comparisons (*WTC vs WTB, 23A vs WTB, 23B2 vs WTB)*.

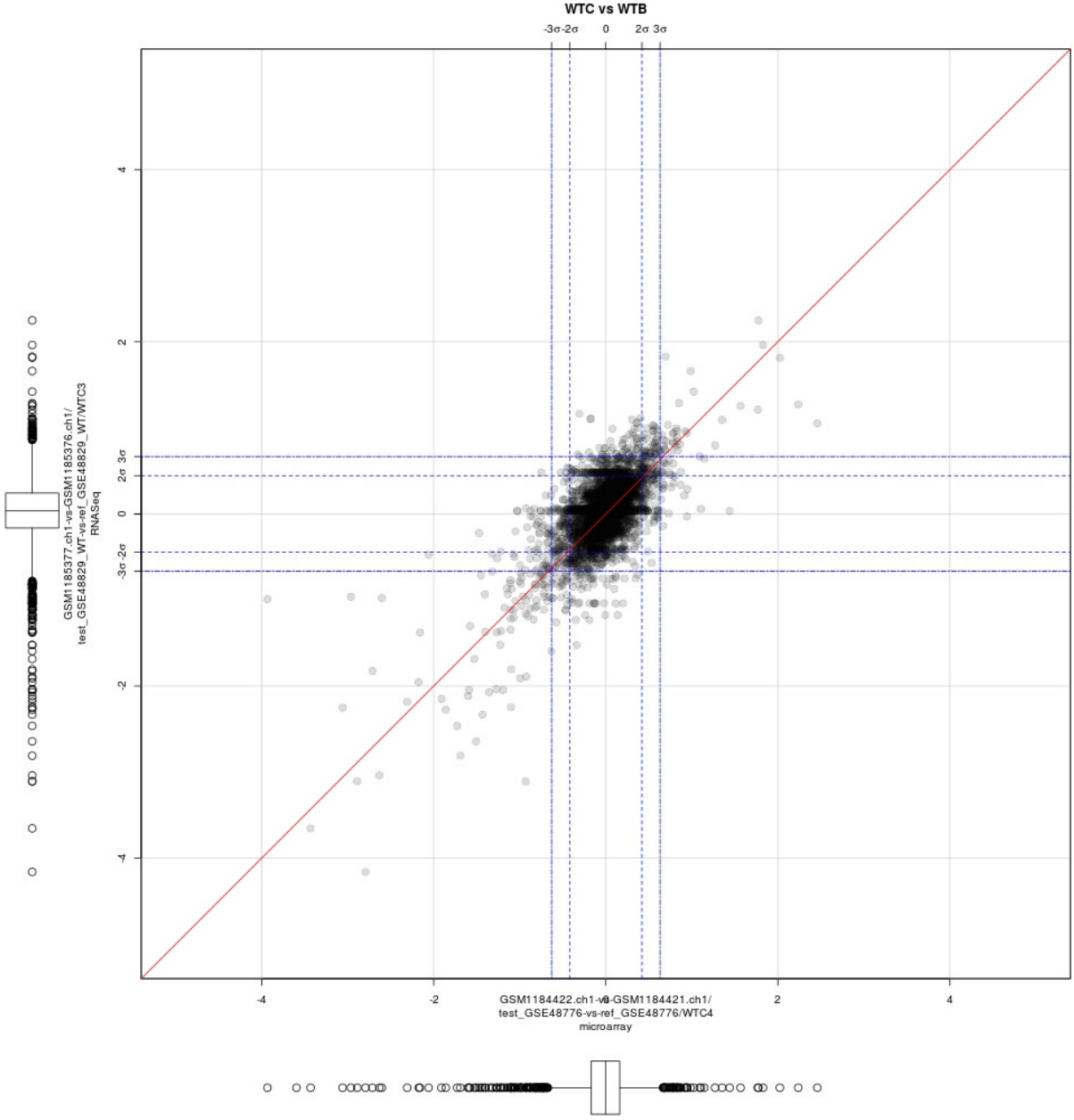

Figure S1: Comparison of the logratios between both technologies RNA-seq (Y-axis) and microarrays (X-axis) for the condition contrast comparing two biological replicates (*WTC vs WTB*). Red line: bisector; dashed blue line: 95% confidence interval of the estimated robust noise distribution; dashed-dotted blue line: 99% confidence interval of the estimated robust noise distribution; the box plots alongside the X and Y axis give further information on the shape of the logratio distributions for the microarrays and RNA-seq respectively.

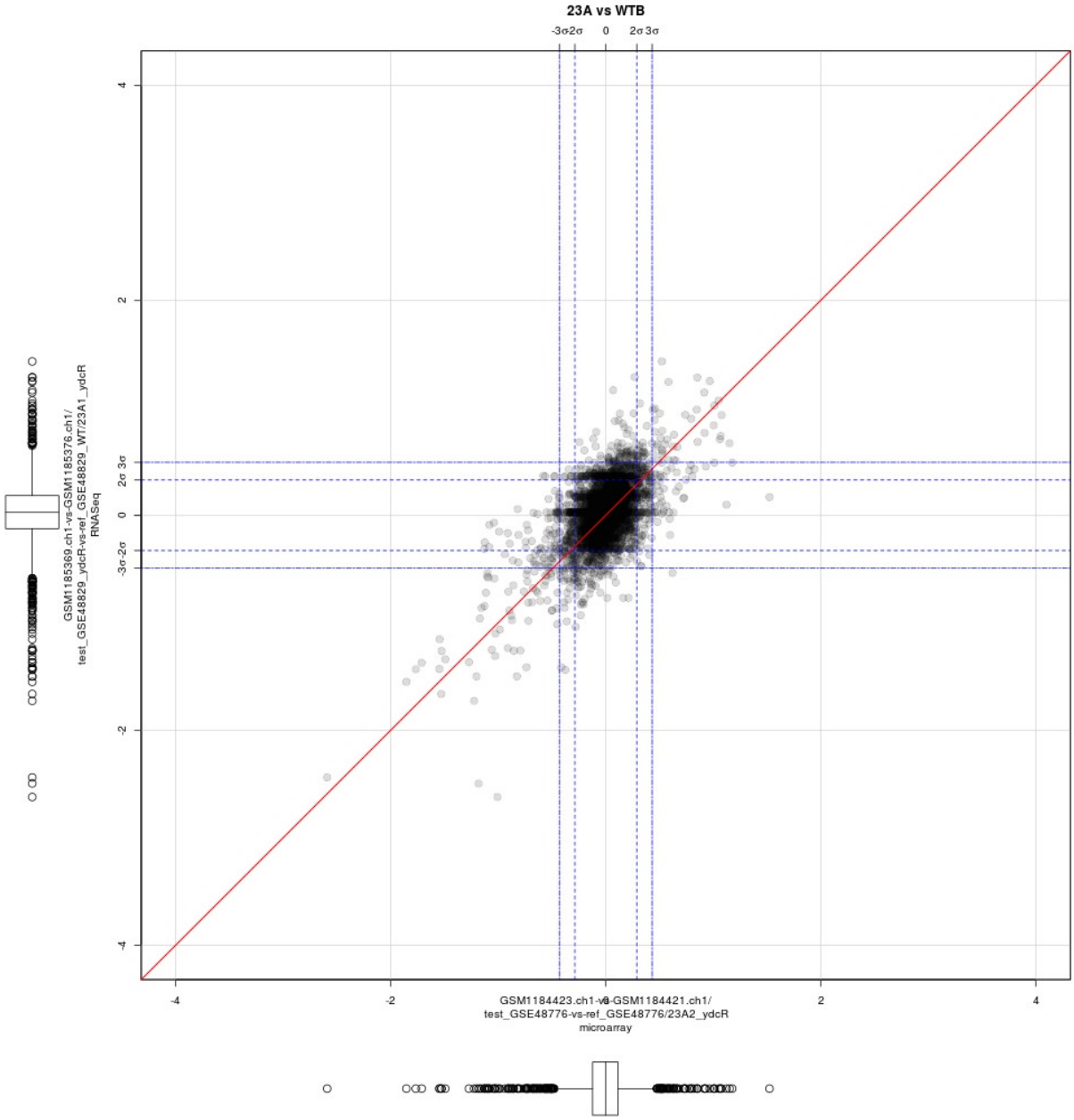

Figure S2: Comparison of the logratios between both technologies RNA-seq (Y-axis) and microarrays (X-axis) for the condition contrast comparing Δ*ydcR* to the wild-type (*23A vs WTB*). Red line: bisector; dashed blue line: 95% confidence interval of the estimated robust noise distribution; dashed-dotted blue line: 99% confidence interval of the estimated robust noise distribution; the box plots alongside the X and Y axis give further information on the shape of the logratio distributions for the microarrays and RNA-seq respectively.

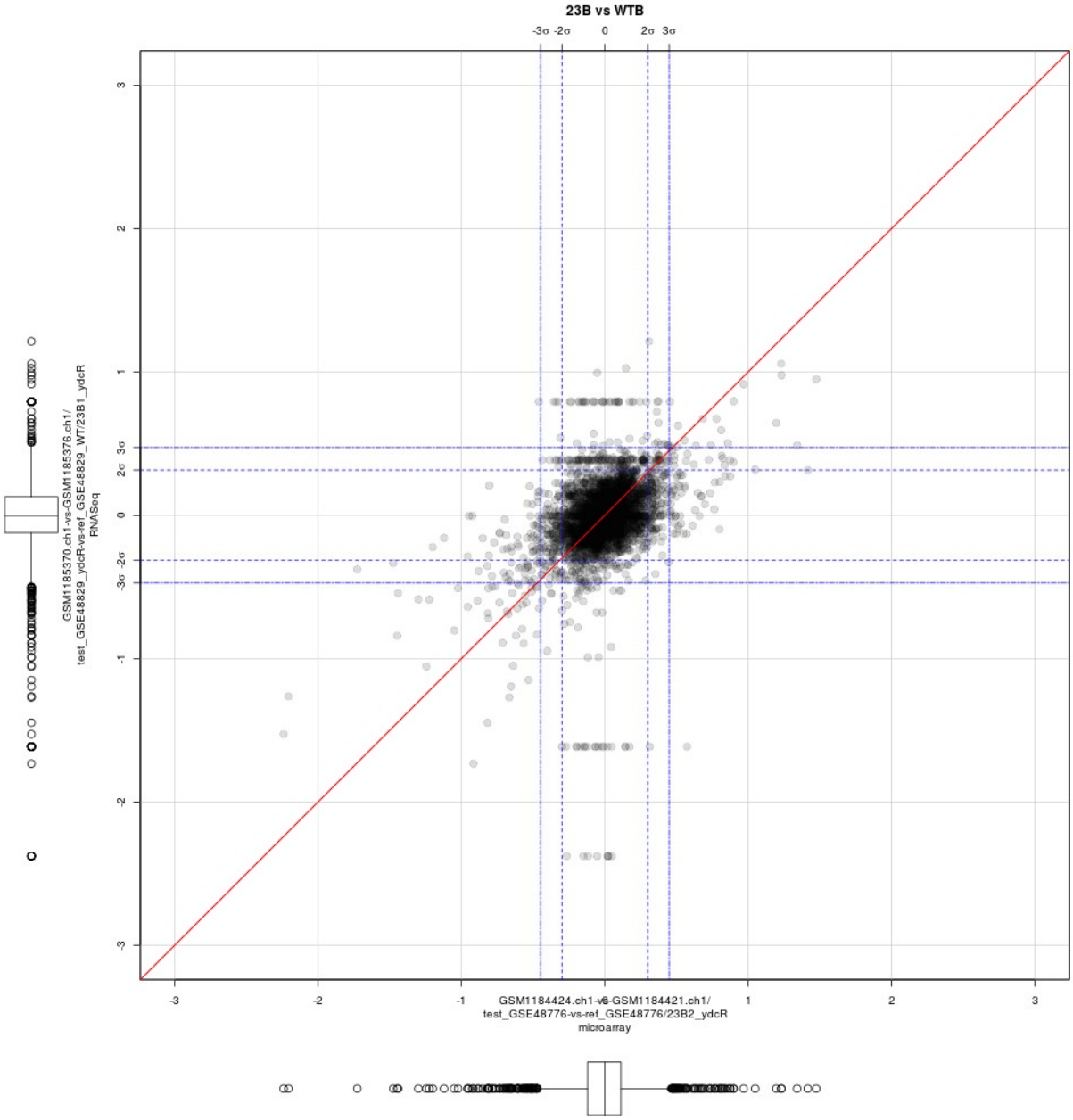

Figure S3: Comparison of the logratios between both technologies RNA-seq (Y-axis) and microarrays (X-axis) for the condition contrast comparing Δ*ydcR* to the wild-type (*23B vs WTB*). Red line: bisector; dashed blue line: 95% confidence interval of the estimated robust noise distribution; dashed-dotted blue line: 99% confidence interval of the estimated robust noise distribution; the box plots alongside the X and Y axis give further information on the shape of the logratio distributions for the microarrays and RNA-seq respectively.

Materials and Methods

Bacterial strains and growth conditions

Escherichia coli wild-type, Δ*ydcR* and Δ*yjiR* strains (27) were grown at 37°C with agitation in M9 salts (Sigma-Aldrich) supplemented with 0.2 % glucose, 2 mM MgSO4 and 100 µM CaCl2.

RNA isolation

Overnight cultures were diluted 100-fold into fresh medium and grown until mid-exponential phase (OD595 = 0.15). The RNA content of 40 ml bacterial culture was stabilized by adding 1/5 volume of ice-cold phenol:ethanol (5:95) after which cells were harvested by centrifugation. The cell pellet was frozen in liquid nitrogen and stored at -80°C. Total RNA was isolated as described in (26). Briefly, cell pellets were resuspended in 1 ml of TRIzol after which the cells were lysed by mechanical disruption using a Precellys 24 (Bertin Technologies) at 6500 rpm for 45 seconds with 0.25 ml of 0.1 mm glass beads. The organic and aqueous phase of the lysed cells were separated by Phase Lock Gel tubes (heavy type) after which total RNA was isolated using the Purelink RNA MiniKit (Ambion). Two treatments of 2 µl TURBO DNase (Ambion) were carried out to remove DNA contamination, which was checked afterwards by PCR (30 cycles). RNA was precipitated in 3 volumes of isopropanol and 1/10 volume of sodium acetate, washed twice in ethanol and dissolved in nuclease-free ultrapure water. RNA integrity was evaluated using Experion RNA StdSens Chips (Bio-Rad). RNA quantity and purity was assessed by measuring the A260/A280 and A260/A230 ratio of all samples using the NanoDrop ND-1000. The ratios of all samples were ≥ 1.8.

RNA-seq – Illumina MiSeq

RNA concentration and purity were determined spectrophotometrically using the Nanodrop ND-1000 (Nanodrop Technologies) and RNA integrity was assessed using a Bioanalyser 2100 (Agilent). Ribosomal RNA was removed using the Ribo-Zero Magnetic kit for Gram-negative bacteria (MRZGN126) (Epicentre) starting from 2 µg total RNA and using the protocol from the manufacturer. rRNA depleted RNA was finally purified by ethanol precipitation and resuspended in 7 µl water. 1 µl was used for Nanodrop quantification and 0.5 µl was diluted to the appropriate concentration for analysis on the BioAnalyer pico RNA chip. 2.5 µl purified rRNA depleted RNA was used as the input for the Illumina TruSeq RNA sample prep kit (RS-122-2001) using the instructions provided in the Illumina TruSeq RNA sample preparation version 2 guide (part# 15026495 revision C – May 2012) starting on page 47. 2.5 µl rRNA depleted RNA was mixed with 15.5 µl EPF and incubated in the thermocycler for the "Elution 2 – Frag – Prime" program. With the adaptor ligation, care was taken that each sample was ligated with a differently barcoded adaptor and that the combination of barcodes resulted in sufficient diversity when reading each barcode position. Upon adaptor ligation and purification, half of the volume rather than the total amount was used in the subsequent enrichment step. Amplification was for 15 cycles and upon magnetic bead purification, the final libraries were eluted in 3 µl water. Each library was quantified by a Qubit (Life Technologies) measurement and the average length of the library molecules was

determined by running each library on a BioAnalyzer High Sensitivity chip. Molarity was calculated from both the concentration and average length. Each library was diluted to 10 nM and an equal volume was taken from each library to yield an equimolar 10 nM library pool. Following the instructions provided in the MiSeq System User Guide, the library pool was denatured and diluted to 6 pM. 5% of PhiX was added and the total mixture was run on a MiSeq 50 cycles kit.

Microarrays – Affymetrix *E. coli* Genome 2.0 arrays

RNA concentration and purity were determined spectrophotometrically using the Nanodrop ND-1000 (Nanodrop Technologies) and RNA integrity was assessed using a Bioanalyser 2100 (Agilent). Per sample, an amount of 50 ng of total RNA was amplified and converted to cDNA using the NuGEN Ovation Pico WTA System v2. cDNA was fragmented and biotin labeled using the NuGEN Encore Biotin Module. All steps were carried out according to the manufacturers protocol (NuGEN). A mixture of purified and fragmented biotinylated cDNA and hybridisation controls (Affymetrix) was hybridised on Affymetrix *E. coli* Genome 2.0 arrays followed by staining and washing in a GeneChip® fluidics station 450 (Affymetrix) according to the manufacturer's procedures. To assess the raw probe signal intensities, chips were scanned using a GeneChip® scanner 3000 (Affymetrix).

RNA-seq: Read Mapping and Quantification of transcript level

Reads in FASTQ format were mapped to the Escherichia coli K12 MG1655 genome sequence (GenBank accession no. NC_000913) using Bowtie version 2.0.5 (20) with default settings. The level of transcription for each gene was estimated using HTSeq-count v0.5.4p2 with the union resolution mode to deal with reads that overlap more than one gene (http://www-huber.embl.de/users/anders/HTSeq/doc/count.html). The obtained read counts are further processed with dedicated analysis pipelines to account for the specifics of RNA-seq, such as the commonly observed heteroscedasticity (40) due to dispersion dependent on the mean count level (low counts show a much higher dispersion than high counts) for which we compensate for by taking advantage of variance-stabilizing data transformation techniques (37-39).

REFERENCES

1. Engelen,K., Fu,Q., Meysman,P., Sánchez-Rodríguez,A., De Smet,R., Lemmens,K., Fierro,A.C. and Marchal,K. (2011) COLOMBOS: Access Port for Cross-Platform Bacterial Expression Compendia. *PLoS ONE*, **6**, e20938.

20. Dimmer,E.C., Huntley,R.P., Alam-Faruque,Y., Sawford,T., O'Donovan,C., Martin,M.J., Bely,B., Browne,P., Mun Chan,W., Eberhardt,R., et al. (2011) The UniProt-GO Annotation database in 2011. *Nucleic acids research*, **40**, D565–570.

26. Vercruysse,M., Fauvart,M., Cloots,L., Engelen,K., Thijs,I.M., Marchal,K. and Michiels,J. (2010) Genome-wide detection of predicted non-coding RNAs in Rhizobium etli expressed during free-living and host-associated growth using a high-resolution tiling array. *BMC genomics*, **11**, 53.

27. Baba,T., Ara,T., Hasegawa,M., Takai,Y., Okumura,Y., Baba,M., Datsenko,K.A., Tomita,M., Wanner,B.L. and Mori,H. (2006) Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Molecular systems biology*, **2**, 2006.0008.

28. Bradford,J.R., Hey,Y., Yates,T., Li,Y., Pepper,S.D. and Miller,C.J. (2010) A comparison of massively parallel nucleotide sequencing with oligonucleotide microarrays for global transcription profiling. *BMC genomics*, **11**, 282.

29. Guida,A., Lindstädt,C., Maguire,S.L., Ding,C., Higgins,D.G., Corton,N.J., Berriman,M. and Butler,G. (2011) Using RNA-seq to determine the transcriptional landscape and the hypoxic response of the pathogenic yeast Candida parapsilosis. *BMC genomics*, **12**, 628.

30. Liu,F., Jenssen,T.-K., Trimarchi,J., Punzo,C., Cepko,C.L., Ohno-Machado,L., Hovig,E. and Kuo,W.P. (2007) Comparison of hybridization-based and sequencing-based gene expression technologies on biological replicates. *BMC genomics*, **8**, 153.

31. Malone,J.H. and Oliver,B. (2011) Microarrays, deep sequencing and the true measure of the transcriptome. *BMC biology*, **9**, 34.

32. Marioni,J.C., Mason,C.E., Mane,S.M., Stephens,M. and Gilad,Y. (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, **18**, 1509–17.

33. Nookaew,I., Papini,M., Pornputtpong,N., Scalcinati,G., Fagerberg,L., Uhlén,M. and Nielsen,J. (2012) A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. *Nucleic acids research*, 10.1093/nar/gks804.

34. Robles,J.A., Qureshi,S.E., Stephen,S.J., Wilson,S.R., Burden,C.J. and Taylor,J.M. (2012) Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. *BMC genomics*, **13**, 484.

35. 't Hoen,P. a C., Ariyurek,Y., Thygesen,H.H., Vreugdenhil,E., Vossen,R.H. a M., de Menezes,R.X., Boer,J.M., van Ommen,G.-J.B. and den Dunnen,J.T. (2008) Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic acids research*, **36**, e141.

36. Sîrbu,A., Kerr,G., Crane,M. and Ruskin,H.J. (2012) RNA-Seq vs dual- and single-channel microarray data: sensitivity analysis for differential expression and clustering. *PloS one*, **7**, e50986.

37. Durbin,B.P., Hardin,J.S., Hawkins,D.M. and Rocke,D.M. (2002) A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, **18**, S105–S110.

38. Lin,S.M., Du,P., Huber,W. and Kibbe,W.A. (2008) Model-based variance-stabilizing transformation for Illumina microarray data. *Nucleic acids research*, **36**, e11.

39. Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome biology*, **11**, R106.

40. Sun,Z. and Zhu,Y. (2012) Systematic comparison of RNA-Seq normalization methods using measurement error models. *Bioinformatics (Oxford, England)*, **28**, 2584–91.