


RESEARCH

Open Access



RNA-sequence data normalization through in silico prediction of reference genes: the bacterial response to DNA damage as case study

Bork A. Berghoff¹, Torgny Karlsson², Thomas Källman^{3,4}, E. Gerhart H. Wagner⁵ and Manfred G. Grabherr^{3,4*} 

* Correspondence:

manfred.grabherr@imbim.uu.se

³Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden

⁴Bioinformatics Infrastructure for Life Sciences (BILS), Science for Life Laboratories, Uppsala University, Uppsala, Sweden

Full list of author information is available at the end of the article

Abstract

Background: Measuring how gene expression changes in the course of an experiment assesses how an organism responds on a molecular level. Sequencing of RNA molecules, and their subsequent quantification, aims to assess global gene expression changes on the RNA level (transcriptome). While advances in high-throughput RNA-sequencing (RNA-seq) technologies allow for inexpensive data generation, accurate post-processing and normalization across samples is required to eliminate any systematic noise introduced by the biochemical and/or technical processes. Existing methods thus either normalize on selected known reference genes that are invariant in expression across the experiment, assume that the majority of genes are invariant, or that the effects of up- and down-regulated genes cancel each other out during the normalization.

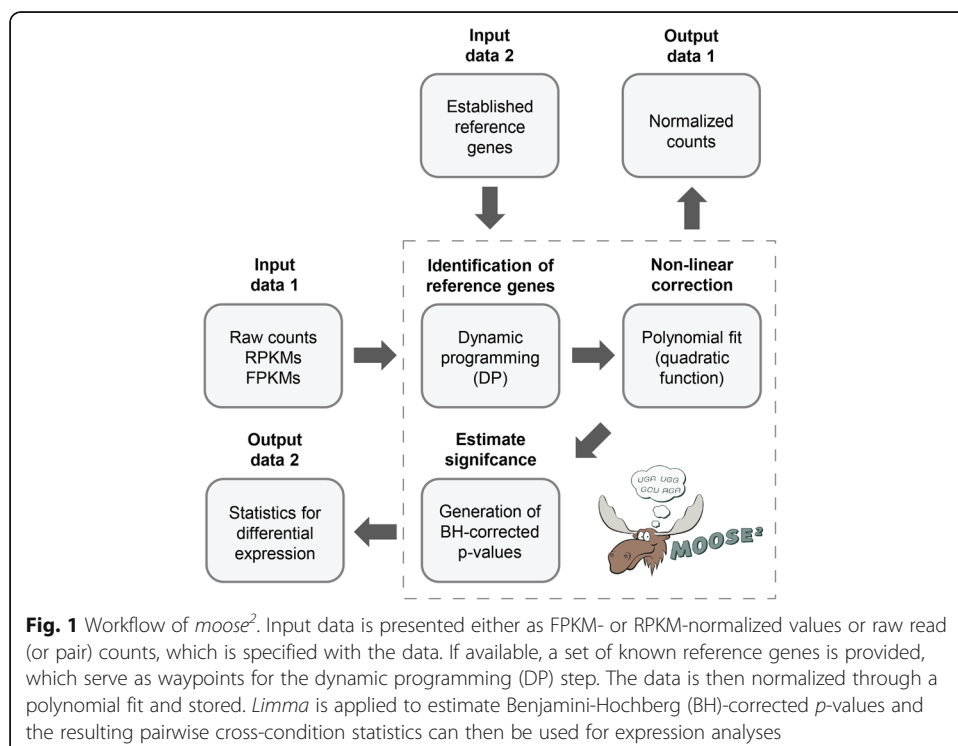
Results: Here, we present a novel method, *moose*², which predicts invariant genes in silico through a dynamic programming (DP) scheme and applies a quadratic normalization based on this subset. The method allows for specifying a set of known or experimentally validated invariant genes, which guides the DP. We experimentally verified the predictions of this method in the bacterium *Escherichia coli*, and show how *moose*² is able to (i) estimate the expression value distances between RNA-seq samples, (ii) reduce the variation of expression values across all samples, and (iii) to subsequently reveal new functional groups of genes during the late stages of DNA damage. We further applied the method to three eukaryotic data sets, on which its performance compares favourably to other methods. The software is implemented in C++ and is publicly available from <http://grabherr.github.io/moose2/>.

Conclusions: The proposed RNA-seq normalization method, *moose*², is a valuable alternative to existing methods, with two major advantages: (i) in silico prediction of invariant genes provides a list of potential reference genes for downstream analyses, and (ii) non-linear artefacts in RNA-seq data are handled adequately to minimize variations between replicates.

Keywords: RNA-seq, Transcriptomics, Normalization, Gene expression, DNA damage, Stress response

Background

RNA-sequencing (RNA-seq) has revolutionized transcriptomics by means of sensitivity, accuracy, and resolution. Additionally, RNA-seq does not rely on prior knowledge of whether any particular RNA is present, and therefore represents a powerful tool for the identification of unknown RNAs. In a typical RNA-seq experiment, total RNA or a particular RNA fraction is isolated from samples that either represent different biological conditions, or replicates from the same condition. After validation of RNA quality, the RNA is subjected to cDNA synthesis via primers that specifically match adapter sequences, or through random priming. If random priming is used, adapter sequences are introduced during subsequent steps. The cDNA is finally amplified by PCR to yield ready-to-use libraries that can be sequenced using different technologies. RNA-seq indirectly measures the abundance of transcripts by the number of reads or fragments generated from a particular transcript. Since the total amount of RNA present in a cell or sample is unknown, data for each sample are either normalized individually by the total read counts per sample and transcript length into RPKM or FPKM values [1], or over all samples by methods such as Upper Quartile (UQ) normalization [2], DESeq2 [3], or Trimmed-Mean of M-values (TMM) normalization [4] (for a review, see ref. [5]). Assumptions underlying the latter methods are that: (i) the mean expression of genes, on which the normalization is computed, does not change across experiments; and (ii) that a single global scaling factor is valid over the entire dynamic range of expression. Importantly, normalization methods may perform poorly if the assumptions do not match the biological experiment [6]. As an alternative to global scaling factors, the use of reference genes, i.e. genes that are invariable in expression regardless of condition or sample, has been suggested [7, 8]. Here, we present a novel method, *moose*² (Fig. 1), which uses known reference genes if available, and additionally predicts



reference genes in silico by a dynamic programming (DP) scheme. Application of a polynomial model then allows for normalizing of the entire data set in a non-linear fashion depending on transcript abundance. Hence, our approach specifically aims at satisfying the assumptions that: (i) there is a small, identifiable subset of genes that is not differentially expressed in the given experiment; and (ii) that a quadratic function approximates any non-linear characteristics of expression measurements across the dynamic range.

To validate this method, we examined the bacterial response (in *E. coli*) to the chemotherapeutic drug mitomycin C (MMC), which we investigated at early and late time-points by RNA-seq. MMC is a potent DNA crosslinker that will ultimately generate double-stranded breaks (DSB) in DNA and thereby activate the so-called SOS response. The SOS response is initiated whenever DNA damage occurs. This generates single-stranded DNA (ssDNA) which is bound by the RecA protein. RecA-nucleofilaments subsequently trigger autocleavage of the LexA repressor that controls a regulon of >50 genes in *E. coli*, many of which have functions in DNA repair [9–12]. The response to DNA damage has been intensively studied by microarray analysis of *E. coli* cells that have been treated with UV light, MMC, or quinolone antibiotics [13–17]. However, none of these studies followed the response to high levels of DNA damage over an extended period of time, nor did they capture possible, more subtle changes in gene expression. Here, relative changes in transcript levels were calculated from an RNA-seq study of *E. coli* cells treated with a high dose of MMC for up to 90 min. *moose*²-performed better than the other tested normalization methods in terms of (i) the Euclidean distances, which are lower in within-replicate comparisons than in cross-condition comparisons, as is to be expected; and (ii) minimizing the variation of expression values across samples. Thus, the *moose*² results could be used to predict the expression profiles of functional groups, such as the LexA regulon, which gave exciting new insights into the bacterial response after prolonged DNA damage. In addition to this bacterial system, we also applied the approach to three eukaryotic data sets and compared the results to other methods.

Methods

Cultivation of bacteria and sampling

Escherichia coli MG1655 cells, obtained from CGSC (Coli Genetic Stock Center) at Yale University (<http://cgsc.biology.yale.edu>) were grown aerobically in Luria broth (LB) at 37 °C. Triplicate overnight cultures were diluted 1:100 into fresh LB and grown for 2 h to reach an OD₆₀₀ of ~0.35. Mitomycin C (MMC) was added at a final concentration of 2.5 µg/ml to induce DNA damage. Samples were withdrawn at 0, 30, and 90 min, and immediately mixed with 0.25 vol of RNA stop solution (95% ethanol, 5% phenol) on ice. Cells were pelleted by centrifugation and frozen in liquid nitrogen. Pellets were thawed on ice and immediately processed for RNA extraction.

RNA extraction and quality control

Total RNA was prepared by the hot acid-phenol method [18]. Cells were resuspended in lysis buffer (100 mM Tris pH 7.5, 40 mM EDTA, 200 mM NaCl, 0.5% SDS) and incubated at 65 °C for 5 min. After adding acidic phenol (pH 4.0) to the cell lysate,

extraction mixtures were incubated at 65 °C for 3 min, frozen in liquid nitrogen, and centrifuged for phase separation. RNA was precipitated from supernatants using isopropanol and pelleted by centrifugation. RNA was washed in 75% ethanol and resuspended in RNase-free water. Samples were treated with DNase I (Thermo Scientific) and extracted with phenol/chloroform, followed by precipitation with isopropanol as before. PCR using primers BB1 (GCT TTA CAG GGG AGA CAA) and BB2 (AAC CCG CAC GCT AAA TAT) was applied to test for DNA contamination. Absorbance ratios of A260/A280 and A260/A230 were determined using a NanoDrop ND-1000 spectrophotometer to assure purity of RNA. RNA integrity was assessed on 1% agarose gels containing 25 mM guanidinium thiocyanate and by analysis with an Agilent 2100 Bioanalyzer (Agilent Technologies) using the Agilent RNA 6000 Pico Kit for total prokaryotic RNA.

Library preparation, RNA-sequencing and read mapping

Preparation of cDNA libraries

Sequencing libraries were prepared with the Encore Complete Prokaryotic RNA-Seq DR Multiplex System (NuGEN) using 200 ng total RNA as input. After cDNA synthesis, cDNA was fragmented by ultra-sonication using the Covaris S-Series System according to the recommendations of the Encore protocol. Adapters containing unique barcode sequences and target sites for Illumina sequencing primers were ligated to cDNA fragments. After strand selection and adapter cleavage, cDNA was amplified to yield strand-specific ready-to-use sequencing libraries. Length distribution of amplified cDNA fragments was validated by Agilent 2100 Bioanalyzer (Agilent Technologies) using the Agilent High Sensitivity DNA Kit. Ultra-sonication and length distribution analysis were performed by the SNP&SEQ Technology Platform in Uppsala, Sweden (www.sequencing.se).

Quality control of sequencing libraries

The quality of the libraries was evaluated using the Advanced Analytical Technologies Fragment Analyzer and a DNA-kit (DNF910). The adapter-ligated fragments were quantified by qPCR using the Library quantification kit for Illumina (KAPA Biosystems) on a StepOnePlus instrument (Applied Biosystems/Life Technologies) prior to cluster generation and sequencing.

Cluster generation and sequencing

An 11 pM solution of sequencing library was subjected to cluster generation and paired-end sequencing with 100 bp read length on the HiSeq 2500 system (Illumina Inc.) using v3 chemistry according to the manufacturer's protocols. Base calling was done by RTA 1.17.21.3 and the resulting.bcl files were demultiplexed and converted to fastq format with tools provided by CASAVA 1.8.4 (Illumina Inc.), allowing for one mismatch in the index sequence. Additional statistics on sequence quality were compiled with an in-house script from the fastq-files, RTA and CASAVA output files. Sequencing was performed by the SNP&SEQ Technology Platform in Uppsala, Sweden (www.sequencing.se).

Read mapping

Reads were mapped against the *E. coli* K-12 MG1655 genome sequence using Papaya (<http://sourceforge.net/projects/satsuma/>) from the Satsuma [19] package, and counted against the NCBI GenBank annotations (NC_000913), requiring a minimum alignment length of 60 nt and identity >0.98.

qRT-PCR

RNA concentrations were determined with a Qubit 2.0 Fluorometer (Invitrogen) using the Qubit RNA HS Assay Kit (Molecular Probes). Primers for qRT-PCR were optimized using primer design software (Additional file 1). The Brilliant III Ultra-Fast SYBR Green QRT-PCR Master Mix (Agilent Technologies) was applied to perform an ultra-fast one-step protocol on a StepOnePlus Real-Time PCR System (Applied Biosystems/Life Technologies) using the following settings for amplification: 50 °C - 10 min, 95 °C - 3 min, 45× (95 °C - 5 s, 60 °C - 10 s). Initial RNA concentrations were set to 1 ng/μl, except for *ssrA* and *rrsA* (set to 10 pg/μl). Melting curves were recorded to monitor amplification specificity. All samples were measured as technical triplicates. Cycle threshold (Ct) values were automatically determined in the linear amplification phase as implemented in the StepOne Software v2.3. Relative fold changes of gene expression were calculated according to the $2^{-\Delta\Delta C_t}$ method [20]. The average Ct values of the six reference genes *cysG*, *idnT*, *hcaT*, *ihfB*, *ssrA*, and *rrsA* were used for normalization.

Normalization methods

Identifying putative invariant genes

In order to increase the number of invariant reference genes, we developed a numerical method to identify such genes in silico. The identification can be regarded as an optimization problem, or more specifically a particular type of linear programming called minimum cost network flow problem [21]. Particularly, the idea is to find the cheapest path in a directed, acyclic n_s -dimensional graph (n_s denotes the total number of samples), from the most lowly, non-zero expressed gene (source) to the most highly expressed gene (target), given a number of constraints. Any edge (i, j) that connects two nodes (genes) in the graph points in the direction from the lower ranked gene to the higher ranked gene, where the ranking, $i = 1, \dots, n$, is determined by sorting the values of the gene expression averaged over all samples. The genes with identically zero expression are omitted from the analysis. Note that in the graph, each gene is connected to every other gene. The linear program may be written in the form (to indicate the direction in the graph, the first index denotes the lower ranked gene which it is always smaller than the second index, denoting the higher ranked gene)

$$\begin{aligned}
 &\text{minimize } z = \sum_{i=1}^{n-1} \sum_{j=i+1}^n c_{ij} x_{ij}, \\
 &\text{subject to } \sum_{j=2}^n x_{ij} = 1, i = 1, \\
 &\quad \sum_{j=i+1}^n x_{ij} - \sum_{j=1}^{i-1} x_{ji} = b_i, i = 2, \dots, n-1, \\
 &\quad \sum_{j=1}^{n-1} x_{ji} = 1, i = n,
 \end{aligned} \tag{1}$$

where $x_{ij} \in \{0, 1\}$ is the indicator variable that signals whether the edge between the two genes i and j belongs to the cheapest path ($x_{ij} = 1$) or not ($x_{ij} = 0$), c_{ij} is the cost

associated with the path connecting gene i and j , b_i denotes the source or sink at node i , while n denotes the number of nodes. The cost c_{ij} is given by

$$c_{ij} = d_{ij} + (j-i-1)m + k_{ij}h, \quad j > i, \quad (2)$$

where d_{ij} is the normalised Euclidean distance between gene i and j , $m = 4.0$ is a flat score introduced as a penalty for taking a path between two genes which are not immediately adjacent in ranking, and $k_{ij}h$ ($h = 5.0$ is constant) denotes the penalty given to those edges for which a number of k_{ij} samples of the higher ranked gene j have a lower expression (RPKM) than the corresponding samples of the lower ranked gene i . Finally, the sources and sinks of the interior nodes $i = 2, \dots, n - 1$ are given by.

$$b_i = \begin{cases} -r, & i = \text{ranking of known reference gene} \\ 0, & \text{otherwise.} \end{cases}$$

In order to speed up the computation, the linear program in Eq. (1) is solved by applying a dynamic programming algorithm. Also, in this way, the costs c_{ij} , given by the expression in Eq. (2), do not have to be pre-computed. The specific values of the reward sinks $b_i = -r = -800$ ($r \gg 1$ in order to ensure that known reference genes are included in the path), which is roughly the number of genes divided by the number of reference genes. The penalties $m = 4.0$ and $h = 5.0$, which control the number of identified in silico reference genes, are chosen to produce about 30 reference genes that are roughly evenly spaced in $\log(\text{expression})$ space (for parameter choice, see Additional file 2 and the *moose*² manual on the web site at <http://grabherr.github.io/moose2/>). In case no reference genes are provided, the algorithm selects the statistically best estimate from the experimental data.

Non-linear (polynomial) correction

For each individual sample, we fit a linear model of second order (i.e., a parabola), such that

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i, \quad i = 1, \dots, n, \quad (3)$$

based on the n in silico invariant genes (including any pre-defined house-keeping genes). In (3), x_i denotes the logarithm of the RPKM-value of invariant gene i for the specific sample in question, while y_i denotes the mean of the $\log(\text{RPKM})$ of gene i , taken over all samples, and ε_i denotes the error term. In matrix notation, (3) may be written as $Y = X\beta + \varepsilon$ and the ordinary least-squares (OLS) estimates, $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)^T$, of the model parameters are then simply given by $\hat{\beta} = (X^T X)^{-1} X^T Y$, where the sum of squared residuals

$$Q = \sum_{i=1}^n (y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2))^2$$

have been minimized s.t. $\partial Q / \partial \beta_0 = \partial Q / \partial \beta_1 = \partial Q / \partial \beta_2 = 0$. This fit is then used to assign a corrected \log -RPKM value,

$$\hat{y}_k = \hat{\beta}_0 + \hat{\beta}_1 x_k + \hat{\beta}_2 x_k^2,$$

to each gene k , based on the gene's observed \log -RPKM value (x_k) for the sample in question. Note that the estimates $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)^T$ will be different for each sample.

Implementation

The algorithms are implemented in C++ and are publicly available from <https://github.com/grabherr/moose2> under the General Public License (GPL).

Linear (global) normalization

For linear normalization we used four different algorithms: RPKM [1], UQ [2], DESeq2 [3], and TMM [4]. DESeq2 and TMM normalization were performed using the R statistical language (<http://www.r-project.org/>) and Bioconductor (<http://www.bioconductor.org/>) packages 'DESeq2' [3] and 'edgeR' [22]. In the 'DESeq2' package, functions *estimateSizeFactors* and *sizeFactors* were called to receive sample-specific normalization factors. In the 'edgeR' package, function *calcNormFactors* was called to output scaling factors together with the original library size. The product of the original library size and the scaling factor, the so-called effective library size, was used as a sample-specific normalization factor. All fold-changes were calculated as direct \log_2 ratios from the normalized read counts.

Estimating p-values

For identification of differentially expressed genes the *moose*²-corrected counts were first transformed using the *voom* function then subjected to the actual gene expression analysis using *Limma* [23]. In short, *voom* estimates mean-variance relationships of log-counts for the samples and *Limma* identifies differentially expressed genes in a linear modelling framework using an empirical Bayesian method to moderate standard errors and estimate fold changes between conditions. From these values, moderated t-statistic and the corresponding *p*-values are estimated. Finally, Benjamini-Hochberg correction of *p*-values was used to control the false discovery rate.

Cluster analyses

Hierarchical cluster analysis of RNA-Seq samples was performed using the R statistical language (<http://www.r-project.org/>). Function *dist* (method = euclidean) was called to generate distance matrices based on \log_2 -transformed read counts (with a pseudocount of 1). Alternatively, the 'DESeq2' package provides log-transformation schemes that are based on per-gene dispersion estimates. Expression data, normalized by DESeq's *sizeFactors*, were applied to the regularized log transformation (function *rlog*) and variance stabilizing transformation (VST; function *varianceStabilizingTransformation*), using dispersion estimates calculated with function *estimateDispersions*. Distance matrices were subsequently generated. All distance matrices were used as input for function *hclust* (method = ward.D2) to generate cluster trees.

Expression cluster analysis of time-series data based on fuzzy *c*-means was performed using the R statistical language (<http://www.r-project.org/>) and Bioconductor (<http://www.bioconductor.org/>) package 'Mfuzz' [24]. For soft clustering of the Top-1000 list (see main text), the fuzzification parameter was set to $m = 2$ and the number of clusters to $c = 6$. Results are displayed in Additional file 3.

Functional annotation clustering of genes was performed using the DAVID bioinformatics database [25] (<http://david.abcc.ncifcrf.gov/home.jsp>). Medium classification stringency with default settings was used to generate clusters of gene ontology (GO) terms based on biological process (BP), cellular component (CC), and molecular function (MF). Pathway clustering was performed in a similar way using the KEGG (<http://www.genome.jp/kegg/>) pathway option. Results are displayed in Additional file 4.

RNA-seq data

The original RNA-seq datasets for *E. coli* are distributed with *moose*² (<https://github.com/grabherr/moose2>). *Moose*²-normalized data including significance analysis can be found as Additional file 5.

Results

Distortion of RNA-seq read counts depends on transcript abundance

We performed an experiment in which we exposed *E. coli* cells (strain MG1655) to high doses of MMC and measured gene expression at 0 (control), 30, and 90 min in three biological replicates each by Illumina sequencing (Fig. 2a). Earlier microarray-based studies on the SOS response, using UV light or MMC, showed a global change in gene expression, but might have missed effects after prolonged time of severe DNA damage [13, 14]. To establish a set of reference genes, we monitored mRNA levels of the widely used housekeeping genes *ihfB*, *ssrA*, and *rrsA*, as well as expression of recently suggested reference genes *cysG*, *idnT*, and *hcaT* [26] by qRT-PCR. Figure 2b indicates that the reference genes were expressed at stable levels and not subject to systematic bias (Additional file 6). However, calculating the pairwise correlations of the six reference genes over all non-normalized RNA-seq samples suggested that even though the moderately expressed *cysG*, *hcaT*, and *idnT* levels were positively correlated

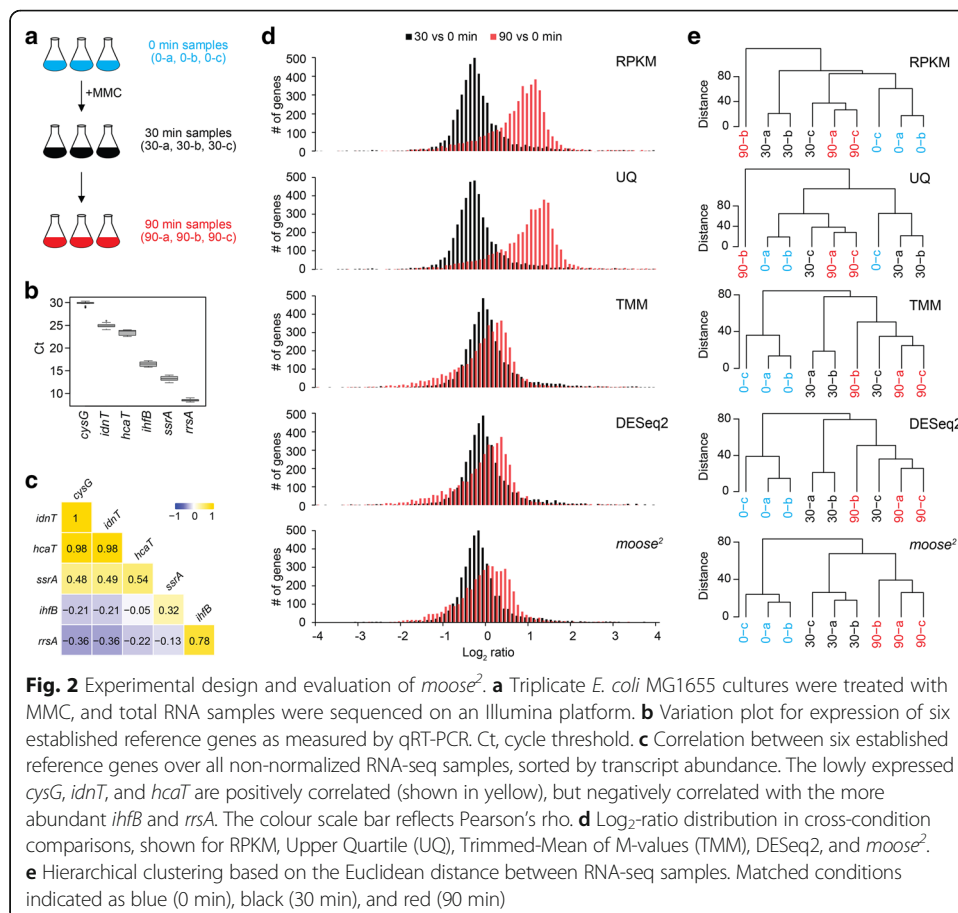


Fig. 2 Experimental design and evaluation of *moose*². **a** Triplicate *E. coli* MG1655 cultures were treated with MMC, and total RNA samples were sequenced on an Illumina platform. **b** Variation plot for expression of six established reference genes as measured by qRT-PCR. Ct, cycle threshold. **c** Correlation between six established reference genes over all non-normalized RNA-seq samples, sorted by transcript abundance. The lowly expressed *cysG*, *idnT*, and *hcaT* are positively correlated (shown in yellow), but negatively correlated with the more abundant *ihfB* and *rrsA*. The colour scale bar reflects Pearson's rho. **d** Log₂-ratio distribution in cross-condition comparisons, shown for RPKM, Upper Quartile (UQ), Trimmed-Mean of M-values (TMM), DESeq2, and *moose*². **e** Hierarchical clustering based on the Euclidean distance between RNA-seq samples. Matched conditions indicated as blue (0 min), black (30 min), and red (90 min)

(Pearson's $\rho \geq 0.98$), the estimated correlation with the highly expressed *rrsA* and *ihfB* was negative (Pearson's $\rho \leq -0.05$, Fig. 2c). Thus, the presence of a systematic bias likely distorts the expression measurements by RNA-seq in dependence of transcript abundance.

In silico reference genes allow for non-linear transformation of expression values

We first applied the RPKM, UQ, TMM and DESeq2 normalization methods, and plotted the \log_2 -ratio distributions comparing the combined replicates across time-points (Fig. 2d). We next computed a distance matrix based on the respective Euclidean distances of all genes (as \log_2 -transformed expression values) across samples, and performed hierarchical clustering (Fig. 2e). Notably, none of the normalization schemes correctly grouped all samples by experiment, indicating that the distances are not consistently lower in within-replicate comparisons. Moreover, reducing the variance between samples by using per-gene dispersion estimates for the log-transformation (e.g., *rlog* and *VST* in the 'DESeq2' package) did not reproduce the correct grouping of biological replicates (Additional file 7). Processing the data with *moose*², guided by the six established reference genes (Fig. 2b), predicted 27 additional in silico reference genes (Table 1. For a more detailed analysis on how selecting in silico genes depends on the choice of conditions, see Additional file 8). While TMM, DESeq2 and *moose*² estimate the peaks of the cross-experiment comparisons around the same position (Fig. 2d), *moose*² reduces the tails on both sides in the 90-to-0-min distribution. Moreover, Euclidean distances based on *moose*²-normalized expression values correctly resolve the grouping of biological replicates (Fig. 2e). Notably, the coefficients for the quadratic correction term, ranging from -0.044 to 0.029 , were weakly inversely correlated (Pearson's $\rho = -0.72$, $p < 0.018$) with the total number of raw reads of each sample, possibly indicating that nonlinearity could have been introduced during library construction, e.g. during random priming, or in the sequencing process. Finally, accurate data normalization is expected to reduce the variation of expression values across all samples. Relative log expression (RLE) boxplots represent the distribution of \log_2 ratios for all genes between one particular sample and the median across all samples. The RLE boxplots should be ideally centered on zero and exhibit a similar dispersion. In contrast to RPKM and UQ normalization, TMM and DESeq2 shifted the mean values close to zero, but without major effect on the variation, while *moose*² clearly reduced the variation (Fig. 3).

We next investigated the individual contributions stemming from (a) predicted in silico reference genes; (b) the quadratic correction term; and (c) guiding the in silico prediction by established reference genes. We thus eliminated each feature individually, and found that the sample grouping was only correct when including both in silico prediction and non-linear correction based on the quadratic term (Additional file 9). To verify, we normalized the data with DESeq2 based on (a) the six established reference genes, and (b) the 33 predicted invariant genes, confirming that a non-linear correction term is required for correct sample grouping (Additional file 10).

To assess the role of accurate reference genes, we ran *moose*² with (a) six genes that were randomly selected over the expression range; and (b) the six most differentially expressed genes (Additional file 9). Interestingly, the resulting sample groupings are correct even in the second case, due to *moose*² rejecting five out of the six genes as too costly to use in the dynamic programming step, reverting to a different set of

Table 1 Expression-invariant genes in *E. coli* during DNA damage as identified by the DP scheme and used for *moose*²

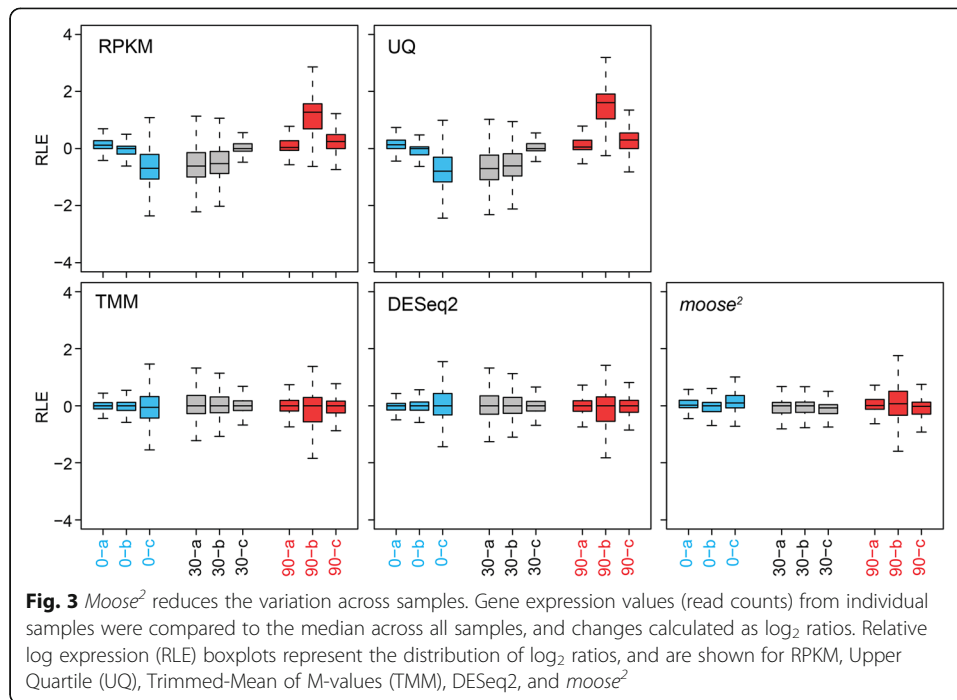
Gene	Product	Process/Function
<i>cysG</i> ^a	uroporphyrin III C-methyltransferase	Other
<i>dnaG</i>	DNA primase	DNA replication
<i>dtpB</i>	dipeptide/tripeptide:H ⁺ symporter DtpB	Transport
<i>ftsX</i>	Cell division protein ftsX	Cell division
<i>ftsY</i>	Cell division protein ftsY	Cell division
<i>glyY</i>	tRNA _{gly}	Translation
<i>gyrB</i>	DNA gyrase, subunit B	DNA replication
<i>hcaT</i> ^a	putative transport protein, major facilitator superfamily (MFS)	Transport
<i>idnT</i> ^a	L-idonate / 5-ketogluconate / gluconate transporter IdnT	Transport
<i>ihfB</i> ^a	integration host factor (IHF), beta subunit	Transcription
<i>lhr</i>	member of ATP-dependent helicase superfamily II	DNA replication
<i>mutM</i>	formamidopyrimidine DNA glycosylase	DNA repair
<i>mutY</i>	A/G-specific adenine glycosylase	DNA repair
<i>ndk</i>	nucleoside diphosphate kinase	Other
<i>nfuA</i>	iron-sulfur cluster scaffold protein	Other
<i>pnp</i>	polynucleotide phosphorylase monomer	RNA processing
<i>rbbA</i>	ribosome-associated ATPase	Translation
<i>rhsB</i>	RhsB protein in <i>rhs</i> element	Other
<i>rpsU</i>	30S ribosomal subunit protein S21	Translation
<i>rrsA</i> ^a	16S ribosomal RNA (<i>rrsA</i>)	Translation
<i>rrsE</i>	16S ribosomal RNA (<i>rrsE</i>)	Translation
<i>rrsG</i>	16S ribosomal RNA (<i>rrsG</i>)	Translation
<i>secB</i>	SecB chaperone	Protein localization
<i>spoT</i>	Guanosine-3',5'-bis(diphosphate) 3'-pyrophosphohydrolase	Other
<i>ssrA</i> ^a	tmRNA	Trans-translation
<i>tfaR</i>	Rac prophage; predicted tail fiber assembly protein	Prophage
<i>thrW</i>	tRNA _{thrW}	Translation
<i>valS</i>	Valyl-tRNA synthetase	Translation
<i>yedJ</i>	predicted phosphohydrolase	Other
<i>ynaE</i>	Rac prophage; cold shock protein, function unknown	Prophage
<i>yphG</i>	conserved protein	Unknown
<i>zntA</i>	zinc, cadmium and lead efflux system	Transport
<i>zupT</i>	heavy metal divalent cation transporter ZupT	Transport

^aEstablished reference genes in *E. coli* (see main text and Fig. 2b) were used as an input for *moose*²

predictions. In either experiment, however, the RLE boxplots are not as closely centered on zero and/or the variance is larger than when incorporating the six established reference genes used in the full analysis (Additional file 11).

In-silico predicted expression-invariant transcripts contain housekeeping genes

Genes with stable expression patterns across a variety of conditions often serve house-keeping functions, such as transcription, translation, or replication. The DP scheme in the *moose*² pipeline predicted 27 expression-invariant genes, many of which have



indeed housekeeping functions (Table 1). The most dominant group comprises genes with a function in translation, including ribosomal RNAs, one ribosomal protein, transfer RNAs, and one aminoacyl tRNA synthetase. Furthermore, the predicted invariant genes were enriched for functions in DNA replication and repair. These findings support the accuracy of the DP scheme for the identification of reference genes.

We tested the 33 reference genes (six established and 27 predicted) for their correlation to each other across all RNA-seq samples. Since normalization is expected to reduce systematic errors in read counts, such as library size effects, one would expect more unstructured correlation patterns for expression-invariant genes after normalization. While global normalization (e.g. by DESeq2; Additional file 12) produced a structured correlation pattern with many estimated coefficients (Pearson's Rho) clearly divergent from zero, *moose*² produced a less structured pattern, which would be expected, and more coefficients close to zero (Additional file 13). This analysis suggested that systematic biases are efficiently reduced by *moose*².

Accurate prediction of expression changes reveals new features of the bacterial response to DNA damage

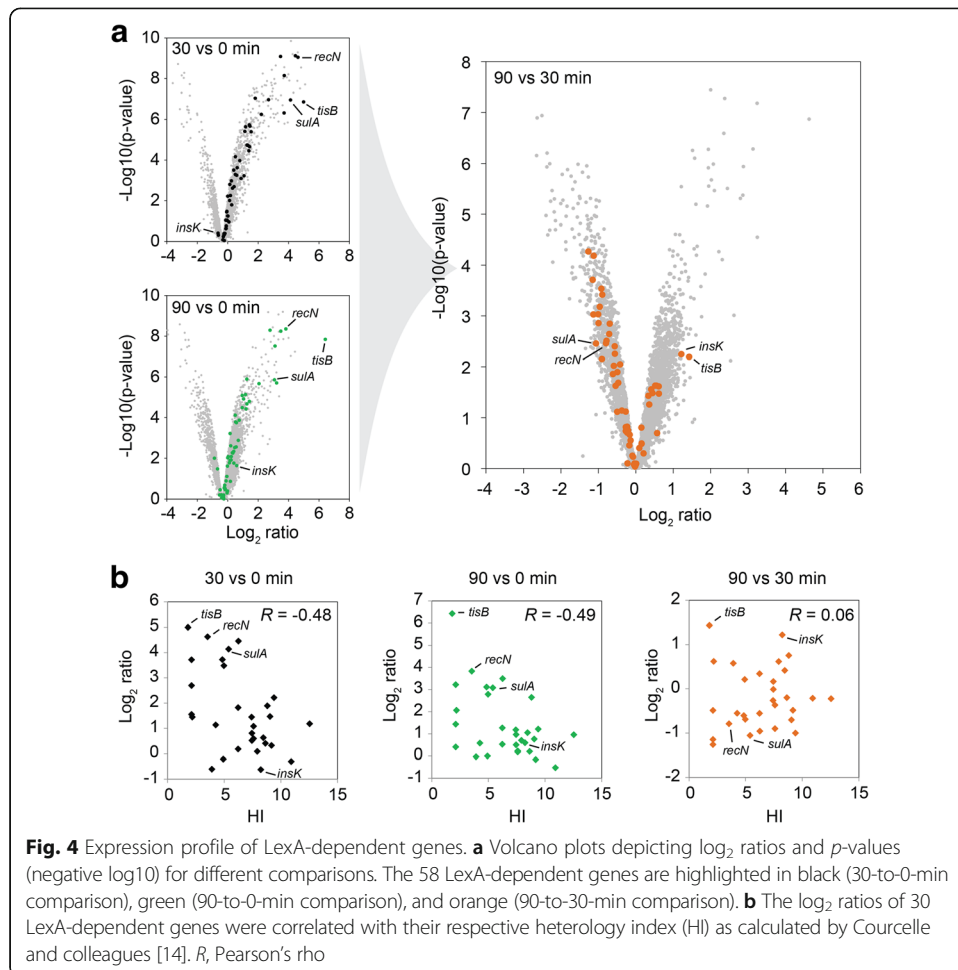
For further validation, we selected 19 *E. coli* genes, eight being members of the SOS response, for qRT-PCR, widely accepted as the gold-standard for assessing gene expression changes [2, 27]. While the Pearson correlation coefficients between qRT-PCR and *moose*² were comparable to the TMM and DESeq2 methods (Additional file 14), linear regression showed that the constant term, describing the global shift of the data, was closest to zero for *moose*² (−0.072 and −0.021 respectively), compared to TMM (0.264 and −0.381) and DESeq2 (0.204 and −0.413). Even though expression ratios predicted by *moose*² might be slightly underestimated in some cases (Additional file 14), the overall qRT-PCR results were reliably reproduced by *moose*².

We subsequently used *moose*² to globally predict changes in gene expression. From microarray analyses, it is known that ~30 LexA-dependent genes are induced upon UV irradiation [14], and in total more than 1000 genes might be affected by DNA damage [13]. We sorted our expression data according to *p*-values computed by *Limma* to visualize coverage of LexA-dependent genes. According to the RegulonDB database [28], 58 genes might be LexA-dependent. A *p*-value cutoff of $p < 9.563 \cdot 10^{-4}$ generated a list of 1000 genes (Top-1000), representing ~24% of the whole data set, including 31 LexA-dependent genes (Additional file 15). We therefore considered the Top-1000 list as a reliable resource for functionally relevant features and examined the directions in which expression changes between time-points, starting with LexA-dependent genes. Several genes were clearly up-regulated upon MMC treatment as exemplified by *recN*, *sulA*, and *tisB* (Fig. 4a). There were however genes (e.g. *insK*) that responded to MMC only after 90 min, which motivated us to calculate expression changes for the 90-to-30-min comparison. Interestingly, most LexA-dependent genes exhibited reduced transcript levels in this comparison, as observed for *recN* and *sulA*. By contrast, only two LexA-dependent genes were found in the same comparison to be clearly increased at the transcript level (\log_2 ratio > 1). This applied to the toxin gene *tisB* and the putative transposase gene, *insK* (Fig. 4a). Most LexA-dependent genes are preceded by a LexA-box sequence. The heterology index (HI) defines the similarity of a particular LexA-box to the consensus of all LexA-box sequences [10]. Low HI values represent high similarity to the consensus. We compared the \log_2 ratios of 30 genes to their corresponding HI values [14], and found an inverse correlation for the 30-to-0-min and 90-to-0-min comparisons as expected (Pearson's Rho of -0.48 and -0.49 respectively; Fig. 4b). In contrast, there was no correlation between \log_2 ratios and HI values for the 90-to-30-min comparison (Pearson's Rho of 0.06), suggesting that changes on transcript level between 30 and 90 min of MMC treatment do not depend on LexA.

The Top-1000 list was applied to soft clustering to generate six expression clusters (Fig. 5a and Additional file 3). LexA-dependent genes were mainly found in expression clusters that exhibited induction at time-point 30 min (clusters 1, 2, and 6). Functional annotation clustering of gene ontology (GO) terms was applied to identify cellular functions that are enriched in distinct expression clusters (Additional file 4), with a focus on the 90-to-30-min comparison (Fig. 5b). Genes with a function in the cell envelope, as e.g. cell wall biosynthesis [*mltB* (murein transglycosylase B) and *oppB* (subunit of murein tripeptide ABC transporter)] or sugar import [*ptsG* (glucose PTS permease) and *manY* (mannose PTS permease)], were decreased in expression after prolonged MMC treatment. The same applied to several genes encoding ribosomal proteins, tRNAs, and aminoacyl tRNA synthetases (Fig. 5b). By contrast, genes with a function intrinsic to the inner membrane or in nitrogen compound biosynthetic processes were only induced at time-point 90 min, as already observed for e.g. *insK*. Among those, several genes encode transporters, and pathway analysis further highlighted genes with a role in purine, amino acid, and sulfur metabolism (Additional file 4).

***Moose*² can be applied to complex eukaryotic samples**

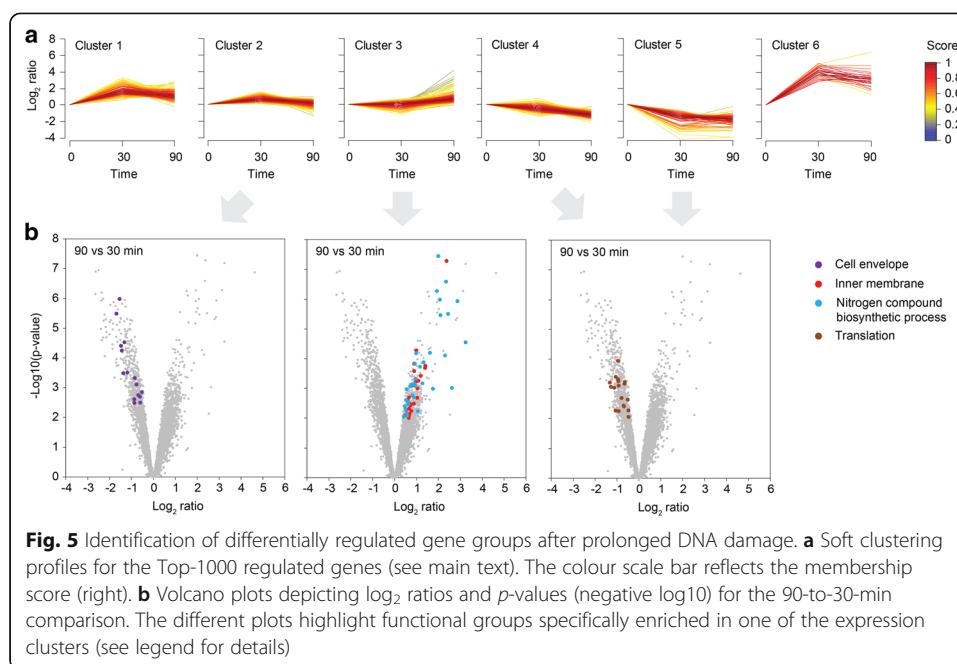
We next explored whether *moose*² outperforms other methods, such as TMM and DESeq2, on data sets beyond cultured bacteria, by applying it to three eukaryotic data



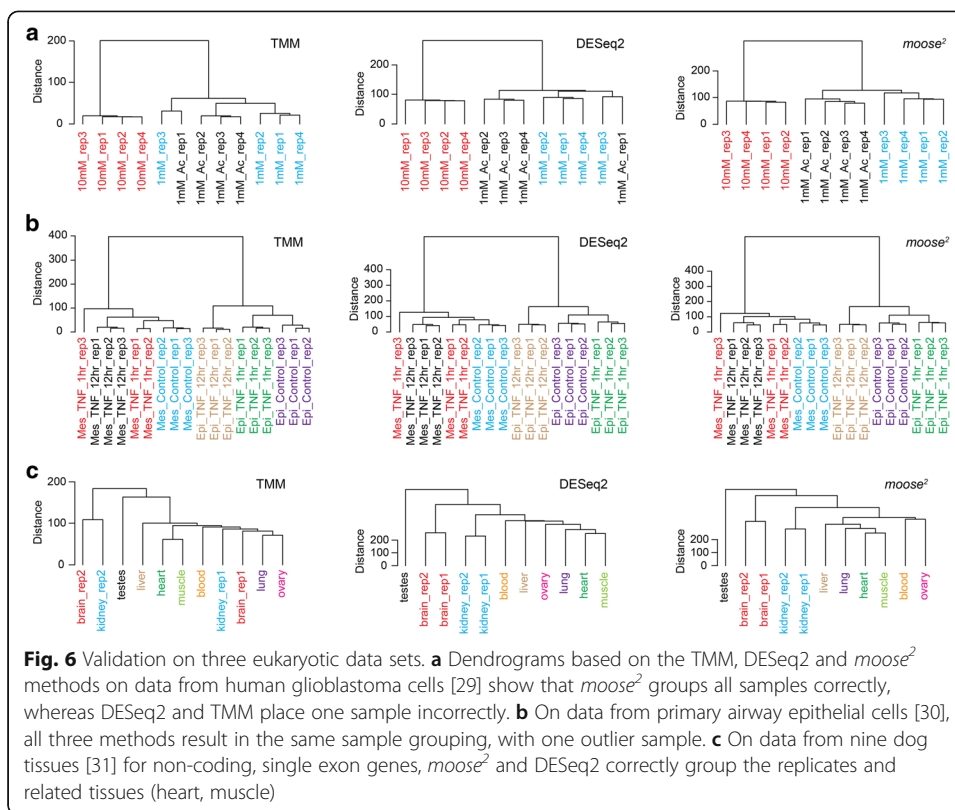
sets. On expression data generated to investigate the glucose- and acetate-regulated transcripts in human glioblastoma cells [29], *moose*² groups all samples correctly, whereas DESeq2 and TMM incorrectly place one sample each (Fig. 6a). On data examining the TGF β -induced program in human primary airway epithelial cells [30], all three methods result in the same sample grouping, with one sample, Mes_TNF_1hr_rep3, placed outside its clade (Fig. 6b). However, comparing scatter plots between two replicate pairs indicate that sample Mes_TNF_1hr_rep3 is an outlier with a higher variance in expression (data not shown), and that this sample is likely not suitable for a completely accurate analysis. Lastly, on data from nine dog tissues [31] with two sample replicates, *moose*² and DESeq2 correctly group the subset of single-exon, non-coding transcripts, which have been reported to exhibit sample-specific expression patterns, by replicates and related tissues (heart, muscle), with the exception of liver, which appears related to kidney in the expression of protein-coding genes [31] (Fig. 6c).

Discussion

Adequate normalization of RNA-seq data is an essential step required to reliably predict differentially expressed genes [2]. The correct choice of a normalization method depends on the assumptions that are valid for the particular biological system under investigation. For example, the RPKM method (normalization by library size) assumes



that the RNA amount per cell is not changed between conditions, an assumption that is easily violated by differential expression of highly expressed genes [2, 5]. Methods such as DESeq2 and TMM assume that the number of up- and down-regulated genes are balanced between conditions, i.e. expression changes are symmetric. These methods might, therefore, perform poorly when the symmetry of expression changes is skewed towards one direction [6]. The genomes of bacteria comprise relatively few genes (e.g. *E. coli*: ~4500) compared to other organisms, thus any response to outside stimuli might involve a large fraction of genes. This can be exemplified by bacteria exposed to extreme stresses [13, 32] or environments such as macrophages [33, 34], where hundreds of genes are differentially expressed, representing up to one fourth of the whole genome. Even though expression changes are not necessarily asymmetric, many experiments show a clear trend towards either side of regulation. As an alternative to the aforementioned methods, expression data can be normalized by applying control genes, which are either external controls (spike-ins) [7, 35, 36], or invariant genes [2, 37]. The underlying assumption for the latter is that at least a small number of genes exist that are not subject to changes in expression in the given experiment. Here, we applied a new method, *moose*², which first identifies such a small subset of invariant genes in silico, and further applies different normalization factors based on the expression strength of each individual gene. In an experiment exposing *E. coli* to high doses of the DNA damaging agent MMC for an extended period of time, global changes in gene expression can be expected [13, 15], and were validated here (Additional file 16). Importantly, gene expression changes might not be symmetric, since the \log_2 ratio distributions (Fig. 2d) are skewed towards up- and down-regulation for the 30-to-0-min and 90-to-0-min comparisons, respectively. The assumption of symmetric gene expression changes is therefore unjustified, thus necessitating an approach that relies on a different assumption, as the existence of invariant genes. The dynamic programming step in the *moose*² pipeline, guided by six established reference genes, predicted 27



additional genes to be expressed at stable levels across experiments. The majority of these invariant genes are known to perform housekeeping functions. We do note, however, that the term “invariant” only applies to genes that do not change expression in any given experiment, so that the selection of these genes depends on the conditions. While the prediction of in silico genes appears stable in the experiments presented here, we caution that there are cases in which this scheme might perform poorly, notably when analyzing large numbers of conditions, or expression across species. Furthermore, in case invariant genes do not exist, the main assumption of *moose*² is violated and alternative methods are preferable. For example, global changes in gene expression, where most of the genes are up-regulated, have been observed in tumor cells and termed transcriptional amplification [38]. In this special case, invariant genes do not exist and external controls (spike-ins) are needed for adequate normalization of RNA-seq data: in the study of Lovén et al. [35] cyclic loess normalization on the spike-ins was successfully applied. Hence, there are limitations to the usage of *moose*², even though it is expected to perform well for most experimental settings. However, the experimentalist should in every case carefully check the justification of the assumptions before deciding on a normalization method.

Reduction of in-between replicate variation by non-linear correction schemes has already been suggested for microarray experiments [39–41], and our data indicate the general strength of such methods in removing technical bias in expression data as well. Interestingly, for our *E. coli* data set, we found that the quadratic correction term used in the *moose*² pipeline performs best, when based on in silico prediction of invariant genes (Additional file 9). The in silico prediction step is clearly a reasonable basis for a

non-linear transformation. The accuracy of subsequently calculated \log_2 ratios was experimentally verified through qRT-PCR for selected genes (Additional file 14). *Moose*², therefore, allows for precisely identifying differentially expressed genes, e.g. using *Limma* [23], or other methods that are continuously being developed and refined in parallel to advances in RNA-seq technologies [42].

The costs for high-throughput sequencing have rapidly decreased over the years, and sequencing of any prokaryotic or eukaryotic organism has become achievable. However, for many organisms, well-annotated reference genomes are unavailable. De novo transcriptome assembly represents an attractive strategy to assess non-sequenced organisms, despite being a bigger informatics challenge than reference-based transcriptome assembly [43, 44]. Furthermore, for downstream analyses such as qRT-PCR, robust reference genes are often needed, but generally not known for non-sequenced organisms. Identification of invariant genes by *moose*² might help to establish reference genes for accurate normalization of qRT-PCR [45]. Different methods have been described for data-driven identification of reference genes [37], and ideally, these methods should be combined with *moose*² to define a reliable set of reference genes. We predict that applying de novo transcriptome assembler together with reference gene identification will benefit the establishment of new model organisms.

As a showcase, we used the *moose*² approach to investigate the response of *E. coli* to prolonged DNA damage caused by MMC. Since there are many treatments that can evoke DNA damage, like ionizing radiation, UV light, DNA gyrase inhibitors, and DNA crosslinkers, the gene expression changes presented here are considered as the MMC-specific response to DNA damage. Also, in a comparative study, aiming to define a global network scheme based on compilations of microarrays, it was found that the SOS response is the only transcriptional response that is consistently triggered upon DNA damage regardless of the toxic agent [16]. As expected and observed here, the degree of induction of several SOS response genes relies on the HI value of the corresponding LexA-box: the lower the HI value, the higher the induction (Fig. 4b). The 90-to-30-min comparison however revealed that most LexA-dependent genes clearly decrease in expression level at the late time-point, which cannot be attributed to their HI values. Since most of the LexA-dependent genes solely depend on LexA and Sigma70 for transcription [28], it is likely that transcript stability and other post-transcriptional mechanisms are pivotal. The strong expression increase of the toxin gene *tisB* (Fig. 4a) is of particular interest, since TisB targets the inner membrane to impair the proton motive force, which then contributes to persister cell formation under DNA-damaging conditions [46–49]. Persisters are transiently drug-tolerant cells that are arrested in their growth due to the action of toxins. The TisB-dependent growth arrest might be accompanied by downstream expression changes, relevant to the persister phenomenon. Gene expression at an early time-point of DNA damage (here 30 min) generally represents the effort to counteract the stressful condition, i.e. inhibiting cell division and repairing DNA damages. The situation changes at late stages (here 90 min), when a fraction of cells has experienced a high level of DNA damage and consequently died, while the surviving subpopulation (i.e. persisters) have only faced moderate DNA damage [50]. So, the SOS response is expected to decline, and this is exactly what we observe (Fig. 4a). Since our RNA-seq data are based on bulk experiments, conclusions have to be drawn cautiously, and some of the gene expression changes at 90 min may reflect the surviving

subpopulation. The clear up-regulation of transporters and enzymes involved in purine and amino acid metabolism (Fig. 5b) might factor into long-term survival strategies of the bacteria. Interestingly, the same GO terms have been found to be under-represented during short periods of DNA damage [16], and might therefore be highly specific to late adaptation processes.

Conclusions

In summary, we present a novel method, *moose*², and show that it corrects for systematic bias in RNA-seq expression data from a bacterial data set by normalizing expression values against a set of genes that were predicted as invariant *in silico*. Moreover, when applied to more complex eukaryotic data sets, the method performs consistently as well as, or better than, other RNA-seq normalization methods, indicating that its algorithm is also applicable to a wider set of organisms. The software is modular and can easily be integrated with other methods that require a set of invariant genes for normalization. *Moose*² is written in C++ and freely available as source code under the General Public License from <http://grabherr.github.io/moose2/>.

Additional files

Additional file 1: Primers used for qRT-PCR. (PDF 18 kb)

Additional file 2: Number of *in silico* invariant genes depending on the choice of parameters. The number of genes depends on h and m , with several settings resulting in the same set of 33 genes used in this analysis. (DOCX 8 kb)

Additional file 3: Expression cluster analysis of time-series data. The data sheet contains the results for the expression cluster analysis of time-series data of the Top-1000 list. Bioconductor package 'Mfuzz' was applied for soft clustering. (XLSX 47 kb)

Additional file 4: Functional annotation cluster analysis. Expression clusters as determined by soft clustering were applied to functional annotation clustering using the DAVID bioinformatics database. The data sheet contains the results for gene ontology (GO) terms (BP: biological process; CC: cellular component; MF: molecular function) and pathway analyses (KEGG). (XLSX 288 kb)

Additional file 5: *Moose*²-normalized RNA-seq data. The data sheet contains the RNA-seq read counts after *moose*² normalization and p -values for cross-condition comparisons. (XLSX 723 kb)

Additional file 6: Box plots of reference genes broken down by condition. While there are minor differences in expression levels, there is no systematic trend in either direction. (TIFF 238 kb)

Additional file 7: Per-gene dispersion estimates for the log-transformation. Using *rlog* and *VST* in the 'DESeq2' package did not reproduce the correct grouping of biological replicates. (TIFF 694 kb)

Additional file 8: *In silico* reference genes predicted on including two conditions. To examine how the choice of *in silico* invariant genes depends on the choice of conditions, we applied *moose*² to subsets consisting of two conditions each. While the number of predictions increases when including only two conditions (0/30, 0/90, and 30/90), there are a number of predicted genes shared among the data sets. Grey shading in the table indicates the six established reference genes. The Venn diagram visualizes overlaps between predictions. (PDF 78 kb)

Additional file 9: Contribution of *in silico* predictions and quadratic correction. Shown are the sample groupings for default parameters, as used in our experiment (a); grouping using a linear fit (b); and no predictions and quadratic fit (c). Out of these, only the combination of predictions and quadratic fit achieve the correct grouping. Also shown are the results when supplying a list of six differently expressed (DE) genes as reference genes (d), in which case five genes are rejected; six randomly selected genes also resolve the grouping correctly (e). (TIFF 1271 kb)

Additional file 10: DESeq2 analyses in different modes. Supplying DESeq2 with the *moose*² predictions does not accurately resolve the sample grouping. (TIFF 392 kb)

Additional file 11: Relative log expression (RLE) boxplots for *moose*²-normalized data. Centering on zero and/or variance are improved when *in silico* predictions are based on a set of established reference genes. (TIFF 494 kb)

Additional file 12: Correlation plots for invariant genes. Genes, that were predicted by *moose*² to be expression-invariant, were correlated with each other according to their transcript counts across all RNA-seq samples. (A) Raw read counts (no normalization) and (B) DESeq2-normalized read counts. The scale bar depicts Pearson's Rho. (TIFF 6253 kb)

Additional file 13: Correlation plots for invariant genes. Genes, that were predicted by *moose*² to be expression-invariant, were correlated with each other according to their transcript counts across all RNA-seq samples. (A) Raw read counts (no normalization) and (B) *moose*²-normalized read counts. The scale bar depicts Pearson's Rho. (TIFF 6341 kb)

Additional file 14: Correlations of gene expression changes for 19 selected genes. Log₂ ratios derived from five normalization methods (RPKM, UQ, TMM, DESeq2, and *moose*²) are compared to qRT-PCR measurements for 30-to-0-min (upper panel) and 90-to-0-min comparisons (lower panel). *R*, Pearson's rho. (TIFF 655 kb)

Additional file 15: Coverage curve for LexA-dependent genes. *E. coli* MG1655 genes were sorted according to their *p*-values, starting with the lowest value. The Top-1000 list (grey box, *p*-value cutoff: $p < 9.563 \times 10^{-4}$) includes 31 out of 58 LexA-dependent genes. (TIFF 426 kb)

Additional file 16: MA-plots for *moose*²-normalized RNA-seq data. The average of expression values (log₂-transformed read counts) is plotted on the x-axis, and the y-axis shows log₂ ratios between conditions. Every dot represents one gene. Differentially expressed genes (DEGs) were determined using *Limma* ($p < 9.563 \times 10^{-4}$) and are shown in red. (TIFF 688 kb)

Acknowledgements

We thank Mirthe Hoekzema for comments on the manuscript and Klev Diamanti for support with R statistical language. Sequencing was performed by the SNP&SEQ Technology Platform in Uppsala. The facility is part of the National Genomics Infrastructure (NGI) Sweden and Science for Life Laboratory. The SNP&SEQ Platform is also supported by the Swedish Research Council and the Knut and Alice Wallenberg Foundation.

Funding

This work was in part supported by a Swedish Research Council Formas grant to M.G.G. We further acknowledge the European Molecular Biology Organization (EMBO) for a long-term fellowship [ALTF 62–2012 to B.A.B.] and the German Research Foundation (DFG) for a research fellowship [BE 5210/1–1 and BE 5210/2–1 to B.A.B.]. Lab work was funded through the Swedish Research Council [VR 621–2010-5233 to E.G.H.W.].

Availability of data and materials

Project name: *moose*².

Project home page: <http://grabherr.github.io/moose2/>

Operating system: Linux.

Programming language: C++.

License: GNU GPL.

The datasets generated and analyzed during the current study, including the read counts, are publicly available from <https://github.com/grabherr/moose2> under the General Public License (GPL), or are included in this published article and its additional information files.

Authors' contributions

BAB, EGHW, and MGG designed the work. BAB performed the experiments. BAB, TK, TKäl, and MGG analyzed the data. All authors contributed to writing the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹Institut für Mikrobiologie und Molekularbiologie, Justus-Liebig-Universität, Giessen, Germany. ²Department of Immunology, Genetics and Pathology, Uppsala University, Uppsala, Sweden. ³Department of Medical Biochemistry and Microbiology, Uppsala University, Uppsala, Sweden. ⁴Bioinformatics Infrastructure for Life Sciences (BILS), Science for Life Laboratories, Uppsala University, Uppsala, Sweden. ⁵Department of Cell and Molecular Biology, Uppsala University, Uppsala, Sweden.

Received: 28 May 2017 Accepted: 22 August 2017

Published online: 05 September 2017

References

1. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*. 2008;5:621–8.
2. Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*. 2010;11:94.
3. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15:550.
4. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*. 2010;11:R25.

5. Dillies MA, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform.* 2012;14:671–83.
6. Evans C, Hardin J, Stoebel DM. Selecting between-sample RNA-Seq normalization methods from the perspective of their assumptions. *Brief Bioinform.* 2017. doi:10.1093/bib/bbx008. [Epub ahead of print]
7. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat Biotech.* 2014;32:896–902.
8. Zhuo B, Emerson S, Chang JH, Di Y. Identifying stably expressed genes from multiple RNA-Seq data sets. *PeerJ.* 2016;4:e2791.
9. Little JW. Mechanism of specific LexA cleavage: autodigestion and the role of RecA coprotease. *Biochimie.* 1991;73:411–22.
10. Lewis L, Harlow GR, Gregg-Jolly LA, Mount DW. Identification of high affinity binding sites for LexA which define new DNA damage-inducible genes in *Escherichia Coli*. *J Mol Biol.* 1994;241:507–23.
11. Fernandez De Henestrosa AR, Ogi T, Aoyagi S, Chafin D, Hayes JJ, Ohmori H, et al. identification of additional genes belonging to the LexA regulon in *Escherichia Coli*. *Mol Microbiol.* 2000;35:1560–72.
12. Wade JT, Reppas NB, Church GM, Struhl K. Genomic analysis of LexA binding reveals the permissive nature of the *Escherichia Coli* genome and identifies unconventional target sites. *Genes Dev.* 2005;19:2619–30.
13. Khil PP, Camerini-Otero RD. Over 1000 genes are involved in the DNA damage response of *Escherichia Coli*. *Mol Microbiol.* 2002;44:89–105.
14. Courcelle J, Khodursky A, Peter B, Brown PO, Hanawalt PC. Comparative gene expression profiles following UV exposure in wild-type and SOS-deficient *Escherichia Coli*. *Genetics.* 2001;158:41–64.
15. Kyeong SJ, Xie Y, Hiasa H, Khodursky AB. Analysis of pleiotropic transcriptional profiles: a case study of DNA gyrase inhibition. *PLoS Genet.* 2006;2:1464–76.
16. Hong J, Ahn JM, Kim BC, Gu MB. Construction of a functional network for common DNA damage responses in *Escherichia Coli*. *Genomics.* 2009;93:514–24.
17. Sangurdekar DP, Srien F, Khodursky AB. A classification based framework for quantitative description of large-scale microarray data. *Genome Biol.* 2006;7:R32.
18. Blomberg P, Wagner EG, Nordström K. Control of replication of plasmid R1: the duplex between the antisense RNA, CopA, and its target, CopT, is processed specifically in vivo and in vitro by RNase III. *EMBO J.* 1990;9:2331–40.
19. Grabherr MG, Russell P, Meyer M, Mauceci E, Alföldi J, Palma F Di, et al. Genome-wide synteny through highly sensitive sequence alignment: Satsuma. *Bioinformatics* 2010;26:1145–1151.
20. Livak KJ, Schmittgen TD. Analysis of relative gene expression data using real-time quantitative PCR and the 2^{−(Delta Delta C(T))} method. *Methods.* 2001;25:402–8.
21. Griva I, Nash SG, Sofer A. Linear and nonlinear optimization: second edition. SIAM 2009. ISBN: 978-0-898716-61-0
22. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.* 2010;26:139–40.
23. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 2015;43:e47.
24. Kumar L. E Futschik M. Mfuzz: a software package for soft clustering of microarray data. *Bioinformatics.* 2007;23:5–7.
25. Huang DW, Lempicki RA, Sherman BT. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc.* 2009;4:44–57.
26. Zhou K, Zhou L, Lim Q, Zou R, Stephanopoulos G, Too H-P. Novel reference genes for quantifying transcriptional responses of *Escherichia Coli* to protein overexpression by quantitative PCR. *BMC Mol Biol.* 2011;12:18.
27. Canales RD, Luo Y, Willey JC, Austerhammer B, Barbacioru CC, Boysen C, et al. Evaluation of DNA microarray results with quantitative gene expression platforms. *Nat Biotechnol.* 2006;24:1115–22.
28. Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muñoz-Rascado L, García-Sotelo JS, et al. RegulonDB v8.0: Omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Res.* 2013;41(Database issue):D203–D213.
29. Lee JV, Carrer A, Shah S, Snyder NW, Wei S, Venneti S, et al. Akt-dependent metabolic reprogramming regulates tumor cell histone acetylation. *Cell Metab.* 2014;20:306–19.
30. Tian B, Li X, Kalita M, Widen SG, Yang J, Bhavnani SK, et al. Analysis of the TGFβ-induced program in primary airway epithelial cells shows essential role of NF-κB/RelA signaling network in type II epithelial mesenchymal transition. *BMC Genomics.* 2015;16:529.
31. Hoepfner MP, Lundquist A, Pirun M, Meadows JRS, Zamani N, Johnson J, et al. An improved canine genome and a comprehensive catalogue of coding genes and non-coding transcripts. *PLoS One.* 2014;9:e91172.
32. Kröger C, Colgan A, Srikumar S, Händler K, Sivasankaran SK, Hammarlöf DL, et al. An infection-relevant transcriptomic compendium for salmonella enterica serovar typhimurium. *Cell Host Microbe.* 2013;14:683–95.
33. Tolman JS, Valvano MA. Global changes in gene expression by the opportunistic pathogen *Burkholderia cenocepacia* in response to internalization by murine macrophages. *BMC Genomics.* 2012;13:63.
34. Srikumar S, Kröger C, Hébrard M, Colgan A, Owen SV, Sivasankaran SK, et al. RNA-seq Brings New Insights to the Intra-Macrophage Transcriptome of *Salmonella Typhimurium*. *PLoS Pathog.* 2015;11(11).
35. Lovén J, Orlando DA, Sigova AA, Lin CY, Rahl PB, Burge CB, et al. Revisiting global gene expression analysis. *Cell.* 2012;151:476–82.
36. Jiang L, Schlesinger F, Davis CA, Zhang Y, Li R, Salit M, et al. Synthetic spike-in standards for RNA-seq experiments. *Genome Res.* 2011;21:1543–51.
37. Mar JC, Kimura Y, Schroder K, Irvine KM, Hayashizaki Y, Suzuki H, et al. Data-driven normalization strategies for high-throughput quantitative RT-PCR. *BMC Bioinformatics.* 2009;10:110.
38. Lin CY, Lovén J, Rahl PB, Paranal RM, Burge CB, Bradner JE, et al. Transcriptional amplification in tumor cells with elevated c-Myc. *Cell.* 2012;151:56–67.
39. Workman C, Jensen LJ, Jarmer H, Berka R, Gautier L, Nielsen HB, et al. A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biol.* 2002;3:research0048.
40. Edwards D. Non-linear normalization and background correction in one-channel cDNA microarray studies. *Bioinformatics.* 2003;19:825–33.

41. Faller D, Voss HU, Timmer J, Hobohm U. Normalization of DNA-microarray data by nonlinear correlation maximization. *J Comput Biol.* 2003;10:751–62.
42. Huang HC, Niu Y, Qin LX. Differential expression analysis for RNA-Seq: an overview of statistical methods and computational software. *Cancer Inform.* 2015;14:57–67.
43. Martin JA, Wang Z. Next-generation transcriptome assembly. *Nat Rev Genet.* 2011;12:671–82.
44. Moreton J, Izquierdo A, Emes RD. Assembly, assessment and availability of de novo generated eukaryotic transcriptomes. *Front Genet.* 2015;6:1–9.
45. Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, et al. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol.* 2002;3:RESEARCH0034.
46. Unoson C, Wagner EGH. A small SOS-induced toxin is targeted against the inner membrane in *Escherichia Coli*. *Mol Microbiol.* 2008;70:258–70.
47. Dörr T, Vulic M, Lewis K. Ciprofloxacin causes persister formation by inducing the TisB toxin in *Escherichia Coli*. *PLoS Biol.* 2010;8:e1000317.
48. Gurnev PA, Ortenberg R, Dörr T, Lewis K, Bezrukov SM. Persister-promoting bacterial toxin TisB produces anion-selective pores in planar lipid bilayers. *FEBS Lett.* 2012;586:2529–34.
49. Berghoff BA, Hoekzema M, Aulbach L, Wagner EGH. Two regulatory RNA elements affect TisB-dependent depolarization and persister formation. *Mol Microbiol.* 2017;103:1020–33.
50. Dörr T, Lewis K, Vulic M. SOS response induces persistence to fluoroquinolones in *Escherichia coli*. *PLoS Genet.* 2009;5:e1000760.

Submit your next manuscript to BioMed Central
and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

