bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Variation-preserving normalization unveils blind spots in gene expression profiling

Carlos P. Roca[1], Susana I. L. Gomes[2], Mónica J. B. Amorim[2], and Janeck J. Scott-Fordsmand[3]

[1]Department of Chemical Engineering, Universitat Rovira i Virgili, Spain

[2]Department of Biology & CESAM, University of Aveiro, Portugal

[3]Department of Bioscience, Aarhus University, Denmark

July 22, 2015

Abstract

RNA-Seq and gene expression microarrays provide comprehensive profiles of gene activity, but lack of reproducibility has hindered their application. A key challenge in the data analysis is the normalization of gene expression levels, which is currently performed following an implicit assumption that most genes are not differentially expressed. Here, we present a mathematical approach to normalization that makes no assumption of this sort. We have found that variation in gene expression is much greater than currently believed, and it can be measured with available technologies. Our results also explain, at least partially, the problems encountered in transcriptomics studies. We expect this improvement in detection to help efforts to realize the full potential of gene expression profiling, especially in analyses of cellular processes involving complex modulations of gene expression.

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Since the discovery of DNA structure by Watson and Crick, molecular biology has progressed increasingly quickly, with rapid advances in sequencing and related genomic technologies. Among these, microarrays and RNA-Seq have been widely adopted to obtain gene expression profiles [1, 2, 3, 4, 5], by measuring the concentration of tens of thousands of mRNA molecules in single assays. Despite their enormous potential [6, 7, 8, 9], problems of reproducibility and reliability [10, 11, 12] have discouraged their use in some areas, e.g. biomedicine [13, 14, 15]. In more mature microarray technologies, issues such as probe design, cross-hybridization, non-linearities and batch effects [16, 17] have been identified as possible culprits, but the problems persist [18, 19].

The normalization of gene expression, which is required to set a common reference level among samples [20, 21, 22, 23, 24], is also reportedly problematic, affecting the reproducibility of both microarray [18, 25] and RNA-Seq [19, 22, 24] results. An underlying assumption of the most widely used normalization methods (such as median and quantile normalization [26] for microarrays, or RPKM [4] and TMM [27] for RNA-Seq), is that most genes are not differentially expressed [24, 28]. This *lack-of-variation assumption* may seem reasonable for many applications, but it has not been confirmed. Furthermore, results obtained with other technologies, particularly qRT-PCR, suggest that it may not be valid [18, 22]. Thus, in attempts to clarify and overcome limitations imposed by this assumption, we have developed an approach to normalization that does not assume lack-of-variation. The analysis of a large gene expression dataset [29] using this approach shows that the assumption can severely undermine the detection of variation in gene expression. We have found that large numbers of differentially expressed genes with substantial expression changes are missed when data are normalized with methods that assume lack-of-variation.

The dataset was obtained from biological triplicates of *Enchytraeus crypticus* (a globally distributed soil organism used in standard ecotoxicity tests), sampled under 51 experimental conditions (42 treatments and 9 controls), involving exposure to several substances, at several concentrations and durations according to a factorial design (Table S1). Gene expression was measured using a customized high-density oligonucleotide microarray [29],

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

and the resulting dataset was normalized with four methods. Two of these methods are the most widely used procedures for microarrays, median (or scale) normalization and quantile normalization [26], whereas the other two, designated *median condition-decomposition normalization* and *standard-vector condition-decomposition normalization*, have been developed for this study.

With the exception of quantile normalization, all used methods apply a multiplicative factor to the expression levels in each sample, equivalent to the addition of a number in the usual $\log_2$-scale for gene expression levels. Solving the *normalization problem* consists of finding these correction factors. The problem can be exactly and linearly decomposed into several sub-problems: one within-condition normalization for each experimental condition and one final between-condition normalization for the condition averages [29]. In the within-condition normalizations, the samples (replicates) subjected to each experimental condition are normalized separately, whereas in the final between-condition normalization average levels for all conditions are normalized together. Because there are no genes with differential expression in any of the within-condition normalizations, the lack-of-variation assumption only affects the final between-condition normalization. The assumption is avoided by using, in this normalization, expression levels only from *no-variation genes*, i.e. genes that show no evidence of differential expression under a statistical test [29]. Both methods of normalization proposed here follow this condition-decomposition approach.

With median condition-decomposition normalization, all normalizations are performed with median values, as in conventional median normalization, but only no-variation genes are included in the between-condition step. Otherwise, if all genes were used in this final step, the resulting total normalization factors would be exactly the same as those obtained with conventional median normalization.

For standard-vector condition-decomposition normalization, a vectorial procedure was developed to carry out each normalization step. The samples of any experimental condition, in a properly normalized dataset, must be *exchangeable*. In mathematical terms, the expression levels of each gene can be considered as an s-dimensional vector, where s is the

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

number of samples for the experimental condition. After standardization (mean subtraction and variance scaling), these standard vectors are located in a $(s-2)$-dimensional hypersphere. The exchangeability mentioned above implies that, when properly normalized, the distribution of standard vectors must be invariant with respect to permutations of the sample labels and must have zero expected value. These properties allow to obtain, under fairly general assumptions, a robust estimator of the normalization factors [29].

An important feature of the approaches to normalization proposed here (linear decomposition into normalization sub-problems per condition, and standard-vector normalization for each sub-problem) is that they do not depend on any particular aspect of the technology of gene expression microarrays or RNA-Seq. The numbers in the input data are interpreted as measured concentrations of mRNA molecules, irrespective of whether they were obtained from fluorescence intensities of hybridized cDNA (microarrays) or from counts of fragments read of mRNA sequences (RNA-Seq). Nevertheless, we consider that specific within-sample corrections for each technology are still necessary and must be applied *before* the between-sample normalizations proposed here. Examples include background correction for microarrays or gene-length normalization (RPKM) for RNA-Seq.

To further explore and compare outcomes of the normalization methods, they were also applied to a synthetic random dataset [29]. This dataset was generated with identical means and variances gene-by-gene to the real dataset, and with the assumption that all genes were no-variation genes. In addition, normalization factors were applied, equal to those obtained from the real dataset. Thus, the synthetic dataset was very similar to the real one, while complying by construction with the lack-of-variation assumption.

Figure 1 displays the results of applying the four normalization methods to the real and synthetic datasets. Each panel shows the interquartile ranges of expression levels for the 153 samples, grouped in triplicates exposed to each experimental condition. Both median (second row) and quantile normalization (third row) yielded similar outputs, for both datasets. In contrast, the condition-decomposition normalizations (fourth and fifth rows) identified marked differences, detecting much greater variation between conditions in the

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

real dataset. Conventional median normalization makes, by design, the median of each sample the same, while quantile normalization makes the full distribution of each sample the same. Hence, if there were differences in medians or distributions of gene expression between experimental conditions, both methods would have removed them. Figures 1G,I show that such variation between conditions was present in the real dataset.

The variation between medians displayed in Figs. 1G,I may seem surprising, given routine expectations based on current methods (Figs. 1C,E). Nevertheless, this variation inevitably results from the imbalance between over- and under-expressed genes. As an illustration, let us consider a case with two experimental conditions, in which the average expression of a given gene is less than the distribution median under one condition, but greater than the median under the other. The variation of this gene alone will change the value of the median to the expression level of the next ranked gene. Therefore, if the number of over-expressed genes is different from the number of under-expressed genes, and enough changes cross the median boundary, then the median will substantially differ between conditions. Only when the differential expression is balanced or small enough, will the median stay the same. This argument applies equally to any other quantile in the distribution of gene expression.

To clarify how the condition-decomposition normalizations preserved the variation between conditions, we studied the influence of the choice of no-variation genes in the final between-condition normalization. To this end, we obtained the between-condition variation with both methods in two families of cases. In one family, no-variation genes were chosen in decreasing order of p-values from an ANOVA test. In the other family, genes were chosen at random. The first option was similar to the approach implemented to obtain the results presented in Fig. 1G–J, with the difference that there the number of genes was chosen automatically by a statistical test. As shown in Fig. 2A, for the real dataset the random choice of genes resulted in $n^{-1/2}$ decays (n being the number of chosen genes), followed by a plateau. The $n^{-1/2}$ decays reflect the standard errors of the estimators of the normalization factors. Selecting the genes by decreasing p-values, however, yielded a completely different result. Up to a certain number of genes, the variance remained

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

similar, but for larger numbers of genes the variance dropped rapidly. Figure 2A shows, therefore, that between-condition variation was removed as soon as the between-condition normalizations used genes that varied in expression level across experimental conditions. The big circles in Fig. 2A indicate the working points of the normalizations used to generate the results displayed in Figs. 1G,I. In fact, these points slightly underestimated the variation between conditions. Although the statistical test for identifying no-variation genes ensured that there was no evidence of variation, inevitably the expression of some selected genes varied across conditions.

Figure 2B displays the results obtained with the synthetic dataset. There were no plateaus when no-variation genes were chosen randomly, only $n^{-1/2}$ decays, and small differences when no-variation genes were selected by decreasing p-values. Big circles show that working points were selected with much larger numbers of genes in the synthetic dataset (Figs. 1H,J) than in the real dataset (Figs. 1G,I). The residual variation, produced by errors in the estimation of the normalization factors, was much smaller than the variation detected in the real dataset, especially for standard-vector condition-decomposition normalization. Overall, Fig. 2 shows that the between-condition variation pictured in Figs. 1G,I is not an artifact caused by using an exceedingly small or extremely particular set of genes in the final between-condition normalization, but that this variation originated from the real dataset.

Finally, Fig. 3A shows the numbers of differentially expressed gene probes (DEGP), identified after normalizing with the four methods, for each of the 42 experimental treatments versus the corresponding control (Table S2). Compared to conventional methods, the number of DEGP detected with the condition-decomposition normalizations was much larger under most treatments, including some whose number of DEGP was larger by more than one order of magnitude. These are statistically significant changes of gene expression, i.e. changes that cannot be explained by chance. More important is the scale of the detected variation, as illustrated by the boxplots in Fig. 3C showing absolute fold changes of DEGP detected after standard-vector condition-decomposition normalization. For all treatments, the entire interquartile range of absolute fold change is above 1.5-fold,

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

and for more than two thirds of the treatments the median absolute fold change is greater than 2. This amount of gene expression variation cannot be neglected, and warrants further research to explore its biological significance.

The lack-of-variation assumption underlying the current methods of normalization was self-fulfilling, removing variation in gene expression that was present in the real dataset. Moreover, it had negative consequences for downstream analyses, as it both removed potentially important biological information and introduced errors in the detection of gene expression. A removal of variation can be understood as errors in the estimation of normalization factors. Considering data and errors vectorially, the length of each vector equals, after centering and up to a constant factor, the standard deviation of the data or error [29]. The addition of an error of small magnitude, compared to the data variance, would have only a minor effect. However, errors of similar or greater magnitude than the data variance may, depending on the lengths and relative angles of the vectors, severely distort the observed data variance. This will in turn cause spurious results in the statistical analyses. Furthermore, the angles between the data and the correct normalization factors (considered as vectors) are random. Data reflect biological variation, while normalization factors respond to technical variation. If the experiment is repeated, even with exactly the same experimental settings, the errors in the normalization factors will vary randomly, causing random spurious results in the downstream analyses. This explains, at least partially, the lack of reproducibility found in transcriptomics studies, especially for the detection of small changes of gene expression, because small variations are most likely to be distorted by errors in the estimates of normalization factors. Accordingly, the largest differences in numbers of DEGP detected by conventional compared to condition-decomposition methods (Fig. 3A) occurred consistently in the treatments with the smallest magnitudes of gene expression changes, e.g. treatments 28, 29 and 33 (Figs. 3B,C).

In summary, this study proves that large numbers of genes change in expression level (often strongly) across experimental conditions, and too extensively to ignore in the normalization of gene expression data. Further, our approach, which avoids the prevailing

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

lack-of-variation assumption, demonstrates that current normalization methods likely remove and distort important variation in gene expression. It also offers a means to investigate broad changes in gene expression that have remained hidden to date. We expect this to provide revealing insights about diverse biomolecular processes, particularly those involving substantial numbers of genes, such as cell differentiation, toxic responses, diseases with non-Mendelian inheritance patterns and cancer. After years of lagging behind the advances in genome sequencing, we believe that the procedures presented here will assist efforts to realize the full potential of gene expression profiling.

References

[1] M. Schena, D. Shalon, R. W. Davis, P. O. Brown, Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470 (1995).

[2] D. J. Lockhart *et al*, Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.* **14**, 1675–1680 (1996).

[3] D. J. Duggan, M. Bittner, Y. Chen, P. Meltzer, J. M. Trent, Expression profiling using cDNA microarrays. *Nat. Genet.* **21**, 10–14 (1999).

[4] A. Mortazavi, B. A. Williams, K. McCue, L. Schaeffer, B. Wold, Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods* **5**, 621–628 (2008).

[5] Z. Wang, M. Gerstein, M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).

[6] T. R. Golub *et al*, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537 (1999).

[7] L. J. van 't Veer *et al*, Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–536 (2002).

[8] N. B. Ivanova *et al*, A stem cell molecular signature. *Science* **298**, 601–604 (2002).

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[9] J.-T. Chi *et al*, Endothelial cell diversity revealed by global expression profiling. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 10623–10628 (2003).

[10] P. K. Tan *et al*, Evaluation of gene expression measurements from commercial microarray platforms. *Nucleic Acids Res.* **31**, 5676–5684 (2003).

[11] S. Frantz, An array of problems. *Nat. Rev. Drug Discov.* **4**, 362–363 (2005).

[12] J. Couzin, Genomics. Microarray data reproduced, but some concerns remain. *Science* **313**, 1559 (2006).

[13] S. Michiels, S. Koscielny, C. Hill, Prediction of cancer outcome with microarrays: A multiple random validation strategy. *Lancet* **365**, 488–492 (2005).

[14] B. Weigelt, J. S. Reis-Filho, Molecular profiling currently offers no more than tumor morphology and basic immunohistochemistry. *Breast Cancer Res.* **12** Suppl 4, S5 (2010).

[15] K. H. Brettingham-Moore, C. P. Duong, A. G. Heriot, R. J. S. Thomas, W. A. Phillips, Using gene expression profiling to predict response and prognosis in gastrointestinal cancers—the promise and the perils. *Ann. Surg. Oncol.* **18**, 1484–1491 (2011).

[16] S. Draghici, P. Khatri, A. C. Eklund, Z. Szallasi, Reliability and reproducibility issues in DNA microarray measurements. *Trends Genet.* **22**, 101–109 (2006).

[17] J. T. Leek *et al*, Tackling the widespread and critical impact of batch effects in high-throughput data. *Nat. Rev. Genet.* **11**, 733–739 (2010).

[18] L. Shi *et al*, The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nat. Biotechnol.* **24**, 1151–1161 (2006).

[19] Z. Su *et al*, A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.* **32**, 903–914 (2014).

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[20] G. K. Smyth, T. P. Speed, Normalization of cDNA microarray data. *Methods* **31**, 265–273 (2003).

[21] R. A. Irizarry *et al*, Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264 (2003).

[22] J. H. Bullard, E. Purdom, K. D. Hansen, S. Dudoit, Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* **11**, 94 (2010).

[23] M. Garber, M. G. Grabherr, M. Guttman, C. Trapnell, Computational methods for transcriptome annotation and quantification using RNA-seq. *Nat. Methods* **8**, 469–477 (2011).

[24] M.-A. Dillies *et al*, A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief. Bioinform.* **14**, 671–683 (2013).

[25] R. Shippy *et al*, Using RNA sample titrations to assess microarray platform performance and normalization techniques. *Nat. Biotechnol.* **24**, 1123–1131 (2006).

[26] B. M. Bolstad, R. A. Irizarry, M. Astrand, T. P. Speed, A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193 (2003).

[27] M. D. Robinson, A. Oshlack, A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).

[28] S. C. Hicks, R. A. Irizarry, quantro: a data-driven approach to guide the choice of an appropriate normalization method. *Genome Biol.* **16**, 117 (2015).

[29] Materials and methods and supplementary text are available online.

[30] S. I. L. Gomes, J. J. Scott-Fordsmand, M. J. B. Amorim, Cellular Energy Allocation to assess the impact of nanomaterials on soil invertebrates (Enchytraeids): Effect of Cu and Ag. *Int. J. Environ. Res. Public Health* **12**, 6858–6878 (2015).

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[31] J. J. Scott-Fordsmand, P. H. Krogh, J. M. Weeks, Responses of *Folsomia fimetaria* (Collembola: Isotomidae) to copper under different soil copper contamination histories in relation to risk assessment. *Environ. Toxicol. Chem.* **19**, 1297–1303 (2000).

[32] *Guidelines for the Testing of chemicals No 202. Daphnia sp. Acute Immobilization Test.* (Organization for Economic Cooperation and Development, Paris, 2004).

[33] Y. Chang, M. L. Lye, H. C. Zeng, Large-scale synthesis of high-quality ultralong copper nanowires. *Langmuir* **21**, 3746–3748 (2005).

[34] *Guidelines for the Testing of chemicals No. 220. Enchytraeid Reproduction Test.* (Organization for Economic Cooperation and Development, Paris, 2004).

[35] S. I. L. Gomes, G. Caputo, N. Pinna, J. J. Scott-Fordsmand, M. J. B. Amorim, Effect of 10 different TiO_2 and ZrO_2 (nano)materials on the soil invertebrate *Enchytraeus crypticus*. *Environ. Toxicol. Chem.* (2015). DOI 10.1002/etc.3080.

[36] M. P. Castro-Ferreira *et al*, Transcriptome assembly and microarray construction for Enchytraeus crypticus, a model oligochaete to assess stress response mechanisms derived from soil conditions. *BMC Genomics* **15**, 302 (2014).

[37] R Core Team. *R: A language and environment for statistical computing.* (R Foundation for Statistical Computing, Vienna, 2015). URL `http://www.R-project.org/`.

[38] W. Huber *et al*, Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods* **12**, 115–121 (2015).

[39] M. E. Ritchie *et al*, Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

[40] J. Durbin, *Distribution Theory for Tests Based on the Sample Distribution Function* (Society for Industrial and Applied Mathematics, Philadelphia, 1973).

[41] M. L. Eaton, *Multivariate Statistics: A Vector Space Approach* (Institute of Mathematical Statistics, Beachwood, Ohio, 2007).

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

[42] K. Fang, S. Kotz, K. W. Ng, *Symmetric Multivariate and Related Distributions* (Chapman and Hall, New York, 1990).

[43] A. K. Gupta, T. Varga, T. Bodnar, *Elliptically Contoured Models in Statistics and Portfolio Theory.* (Springer, New York, 2013).

[44] O. Kallenberg, *Probabilistic Symmetries and Invariance Principles.* (Springer, New York, 2005).

[45] J. D. Storey, R. Tibshirani, Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 9440–9445 (2003).

[46] J. D. Storey, The positive false discovery rate: a Bayesian interpretation and the q-value. *Ann. Stat.* **31**, 2013–2035 (2003).

[47] W. Feller, *An Introduction to Probability Theory and Its Applications*, (Wiley, New York, ed. 2, 1971), vol. 2.

Acknowledgements

This work was funded by the European Union FP7 projects MODERN (Ref. 309314-2) (C.P.R., J.J.S.-F.) and MARINA (Ref. 263215) (J.J.S.-F.), by FEDER through COMPETE (Programa Operacional Factores de Competitividade) and FCT (Fundação para a Ciência e Tecnologia) through project bio-CHIP (Ref. FCT EXPL/AAG-MAA/0180/2013) (S.I.L.G., M.J.B.A.), and by a post-doctoral grant (Ref. SFRH/BPD/95775/2013) (S.I.L.G).

MIAME-compliant microarray data from the experiment were submitted to the Gene Expression Omnibus (GEO) at the NCBI website (platform: GPL20310; series: GSE69746, GSE69792, GSE69793 and GSE69794). Custom code that reproduces all the reported results starting from the raw microarray data is available at the GitHub repository `https://github/carlosproca/gene-expr-norm-paper`.

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

S.I.L.G., M.J.B.A. and J.J.S.-F. designed the toxicity experiment. S.I.L.G. carried out the experimental work and collected the microarray data. C.P.R. designed and implemented the novel normalization methods. C.P.R. performed the statistical analyses. All the authors jointly discussed the results. C.P.R. drafted the paper, with input from all the authors. All the authors edited the final version of the paper.

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figures

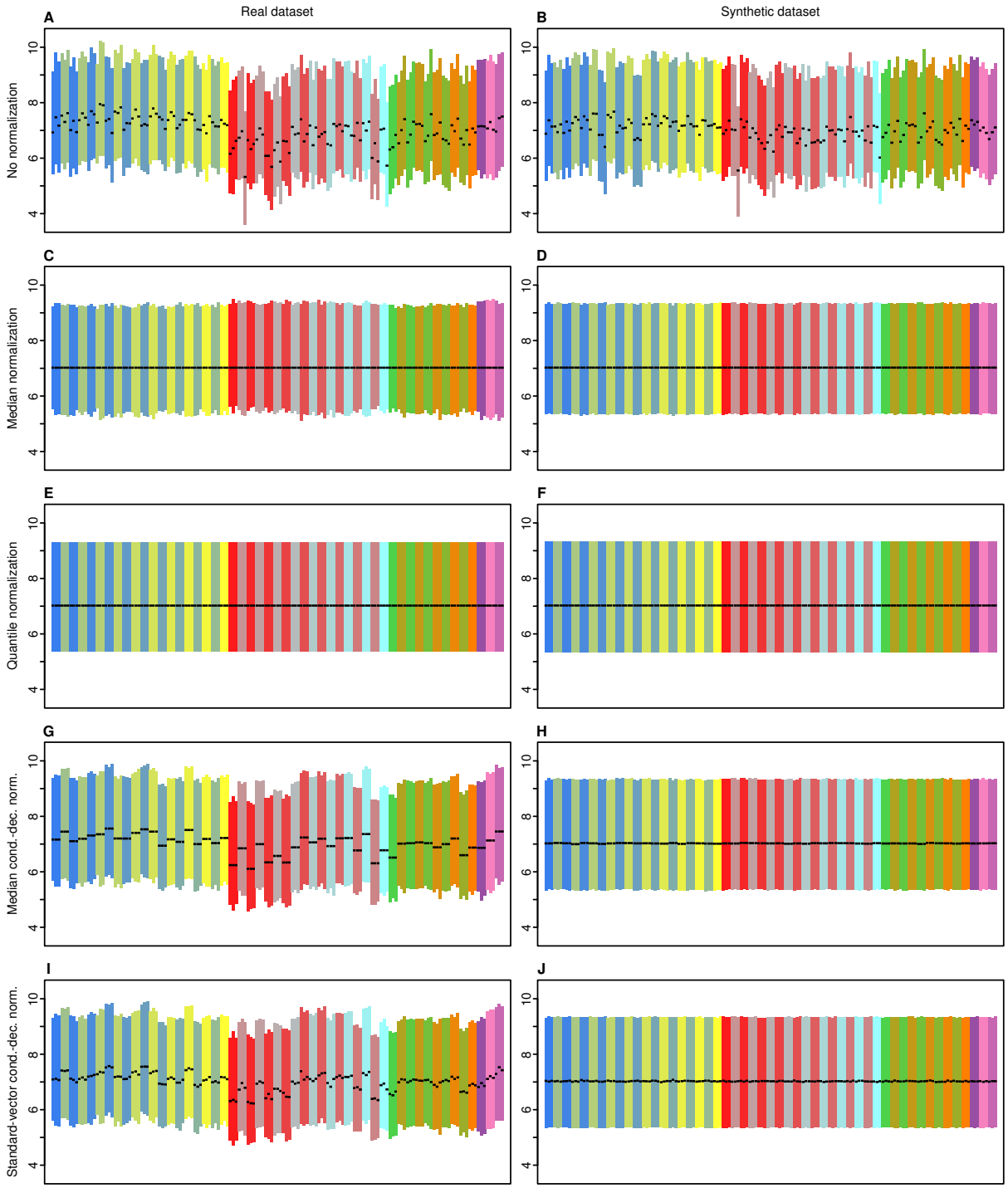

Figure 1

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 1: The condition-decomposition normalizations detected a large amount of between-condition variation in the real expression data, in contrast with conventional methods. All 10 panels show interquartile ranges of expression levels of the 153 samples, grouped by the 51 experimental conditions (Ag, blue-yellow; Cu, red-cyan; Ni, green-orange; UV, purple; see Table S1). Black lines indicate medians. Rows and columns correspond to normalization methods and datasets (as labeled), respectively. In the synthetic dataset no gene was differentially expressed between any two conditions.

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

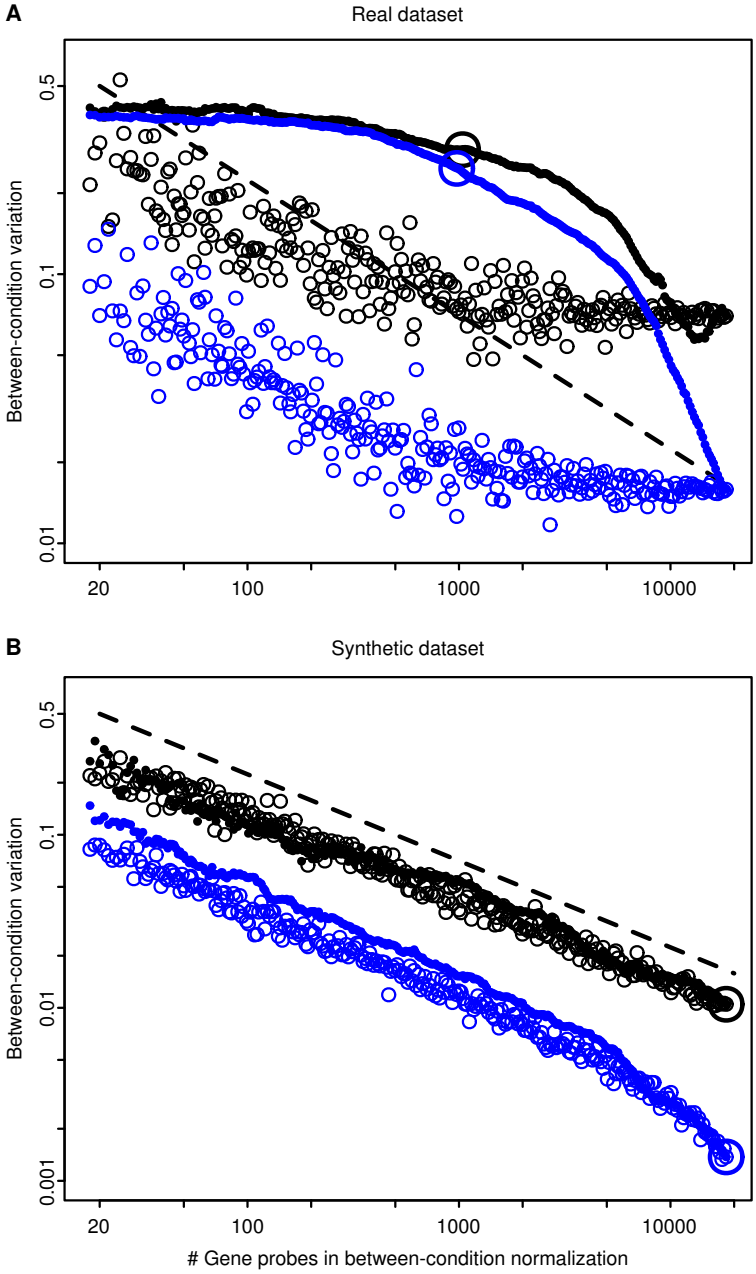

Figure 2

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 2: The selection of genes in the final between-condition normalization was crucial to preserve variation between conditions. The panels show the detected variation as a function of the number of gene probes used in the between-condition normalization of the real dataset (**A**) and synthetic dataset (**B**). Between-condition variation is represented as the standard deviation of the within-condition mean averages (averages of sample mean expression levels, for all samples under the condition). See Fig. S1 for within-condition median averages, with similar results. Each point in either of the panels indicates the variation obtained with one complete normalization (black circles, median condition-decomposition normalization; blue circles, standard-vector condition-decomposition normalization). Gene probes were selected in two ways: randomly (empty circles) or in decreasing order of p-values (filled circles). Big circles show the working points of the algorithms whose results are depicted in Fig. 1G–J. Black dashed lines show references for $n^{-1/2}$ decays, with the same values in both panels.

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

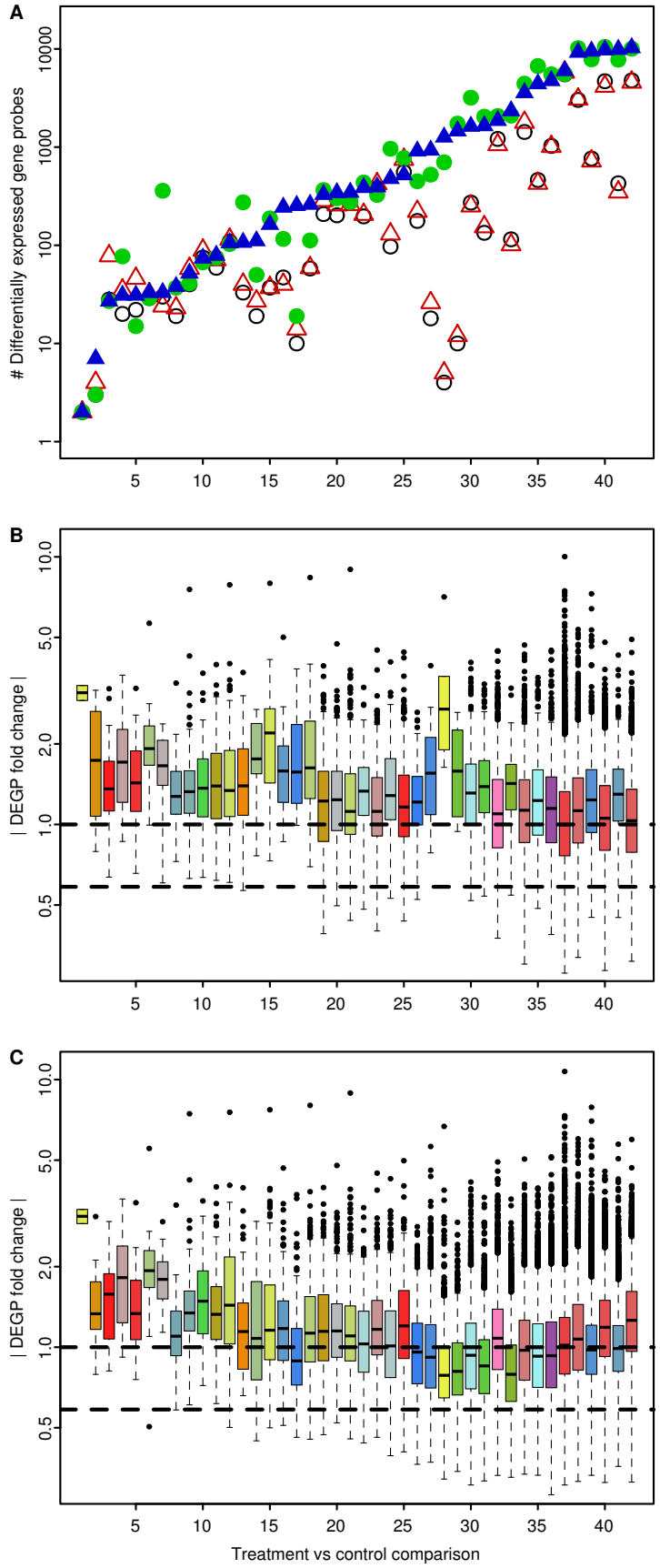

Figure 3

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 3: The condition-decomposition normalizations detected much larger numbers of differentially expressed gene probes (DEGP), with substantial fold changes. (**A**) Number of DEGP for each treatment vs control comparison, obtained after applying the four normalization methods (empty black circles, median normalization; empty red triangles, quantile normalization; filled green circles, median condition-decomposition normalization; filled blue triangles, standard-vector condition decomposition normalization). Significant differential expression was identified with R/Bioconductor package limma. (see Fig. S2 for results with t-tests). Lower panel shows boxplots of absolute values of DEGP fold changes (absolute differences of $\log_2$ expression levels), also per treatment vs control comparison, obtained with quantile normalization (**B**) and standard-vector condition-decomposition normalization (**C**). Boxplots are colored by treatment, with the same color code as in Fig. 1. In both panels comparisons are ordered according to the number of DEGP identified with standard-vector condition-decomposition normalization, increasing from left to right (Table S2). Dashed horizontal lines in the lower panel indicate references of 1.5-fold and 2-fold changes.

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary Tables

Table S1: Experimental conditions of the toxicity experiment on *E. crypticus*, listed in the same order as they appear in each panel of Fig. 1, from left to right.

Condition number	Condition ID	Condition description
1	Ag.AgNO3.EC20.3d	AgNO3 EC20 3 days
2	Ag.AgNO3.EC20.7d	AgNO3 EC20 7 days
3	Ag.AgNO3.EC50.3d	AgNO3 EC50 3 days
4	Ag.AgNO3.EC50.7d	AgNO3 EC50 7 days
5	Ag.Coated.EC20.3d	Ag-NPs PVP-Coated EC20 3 days
6	Ag.Coated.EC20.7d	Ag-NPs PVP-Coated EC20 7 days
7	Ag.Coated.EC50.3d	Ag-NPs PVP-Coated EC50 3 days
8	Ag.Coated.EC50.7d	Ag-NPs PVP-Coated EC50 7 days
9	Ag.NC.EC20.3d	Ag-NPs Non-Coated EC20 3 days
10	Ag.NC.EC20.7d	Ag-NPs Non-Coated EC20 7 days
11	Ag.NC.EC50.3d	Ag-NPs Non-Coated EC50 3 days
12	Ag.NC.EC50.7d	Ag-NPs Non-Coated EC50 7 days
13	Ag.NM300K.EC20.3d	Ag NM300K EC20 3 days
14	Ag.NM300K.EC20.7d	Ag NM300K EC20 7 days
15	Ag.NM300K.EC50.3d	Ag NM300K EC50 3 days
16	Ag.NM300K.EC50.7d	Ag NM300K EC50 7 days
17	Ag.CT.3d	Ag Control 3 days
18	Ag.CT.7d	Ag Control 7 days
19	Ag.CTD.3d	Ag Control Dispersant 3 days
20	Ag.CTD.7d	Ag Control Dispersant 7 days
21	Cu.CuNO3.EC20.3d	CuNO3 EC20 3 days
22	Cu.CuNO3.EC20.7d	CuNO3 EC20 7 days
23	Cu.CuNO3.EC50.3d	CuNO3 EC50 3 days
24	Cu.CuNO3.EC50.7d	CuNO3 EC50 7 days
25	Cu.Cu.NPs.EC20.3d	Cu-NPs EC20 3 days
26	Cu.Cu.NPs.EC20.7d	Cu-NPs EC20 7 days
27	Cu.Cu.NPs.EC50.3d	Cu-NPs EC50 3 days
28	Cu.Cu.NPs.EC50.7d	Cu-NPs EC50 7 days
29	Cu.Cu.Nwires.EC20.3d	Cu-NWires EC20 3 days
30	Cu.Cu.Nwires.EC20.7d	Cu-NWires EC20 7 days
31	Cu.Cu.Nwires.EC50.3d	Cu-NWires EC50 3 days
32	Cu.Cu.Nwires.EC50.7d	Cu-NWires EC50 7 days
33	Cu.Cu.field.EC20.3d	Cu-Field EC20 3 days
34	Cu.Cu.field.EC20.7d	Cu-Field EC20 7 days
35	Cu.Cu.field.EC50.3d	Cu-Field EC50 3 days
36	Cu.Cu.field.EC50.7d	Cu-Field EC50 7 days
37	Cu.CT.3d	Cu Control 3 days
38	Cu.CT.7d	Cu Control 7 days
39	Ni.NiNO3.EC20.3d	NiNO3 EC20 3 days
40	Ni.NiNO3.EC20.7d	NiNO3 EC20 7 days
41	Ni.NiNO3.EC50.3d	NiNO3 EC50 3 days
42	Ni.NiNO3.EC50.7d	NiNO3 EC50 7 days
43	Ni.Ni.NPs.EC20.3d	Ni-NPs EC20 3 days
44	Ni.Ni.NPs.EC20.7d	Ni-NPs EC20 7 days
45	Ni.Ni.NPs.EC50.3d	Ni-NPs EC50 3 days
46	Ni.Ni.NPs.EC50.7d	Ni-NPs EC50 7 days
47	Ni.CT.3d	Ni Control 3 days
48	Ni.CT.7d	Ni Control 7 days
49	Uv.UV.D1.5d	UV Dose 1
50	Uv.UV.D2.5d	UV Dose 2
51	Uv.CT.5d	UV Control

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Table S2: Treatment vs control comparisons, listed in increasing number of differentially expressed gene probes (DEGP) obtained with standard-vector condition-decomposition normalization and limma statistical analysis. This is the same order as in Fig. 3, from left to right.

Comparison number	Treatment ID	Control ID	Treatment description	Number of DEGP
1	Ag.NM300K.EC20.7d	Ag.CTD.7d	Ag NM300K EC20 7 days	2
2	Ni.Ni.NPs.EC20.7d	Ni.CT.7d	Ni-NPs EC20 7 days	7
3	Cu.CuNO3.EC20.3d	Cu.CT.3d	CuNO3 EC20 3 days	27
4	Cu.CuNO3.EC50.7d	Cu.CT.7d	CuNO3 EC50 7 days	31
5	Cu.Cu.NPs.EC20.3d	Cu.CT.3d	Cu-NPs EC20 3 days	31
6	Ag.AgNO3.EC50.7d	Ag.CT.7d	AgNO3 EC50 7 days	33
7	Cu.Cu.NPs.EC20.7d	Cu.CT.7d	Cu-NPs EC20 7 days	33
8	Ag.NM300K.EC20.3d	Ag.CTD.3d	Ag NM300K EC20 3 days	38
9	Ag.NM300K.EC50.3d	Ag.CTD.3d	Ag NM300K EC50 3 days	52
10	Ni.NiNO3.EC20.3d	Ni.CT.3d	NiNO3 EC20 3 days	74
11	Ni.NiNO3.EC20.7d	Ni.CT.7d	NiNO3 EC20 7 days	79
12	Ag.NC.EC20.7d	Ag.CT.7d	Ag-NPs Non-Coated EC20 7 days	106
13	Ni.Ni.NPs.EC50.7d	Ni.CT.7d	Ni-NPs EC50 7 days	107
14	Ag.AgNO3.EC20.7d	Ag.CT.7d	AgNO3 EC20 7 days	111
15	Ag.NC.EC50.7d	Ag.CT.7d	Ag-NPs Non-Coated EC50 7 days	163
16	Ag.NC.EC20.3d	Ag.CT.3d	Ag-NPs Non-Coated EC20 3 days	244
17	Ag.AgNO3.EC50.3d	Ag.CT.3d	AgNO3 EC50 3 days	254
18	Ag.Coated.EC20.7d	Ag.CT.7d	Ag-NPs PVP-Coated EC20 7 days	261
19	Ni.NiNO3.EC50.7d	Ni.CT.7d	NiNO3 EC50 7 days	329
20	Cu.Cu.NPs.EC50.7d	Cu.CT.7d	Cu-NPs EC50 7 days	343
21	Ag.Coated.EC50.7d	Ag.CT.7d	Ag-NPs PVP-Coated EC50 7 days	346
22	Cu.Cu.Nwires.EC50.7d	Cu.CT.7d	Cu-NWires EC50 7 days	387
23	Cu.CuNO3.EC20.7d	Cu.CT.7d	CuNO3 EC20 7 days	393
24	Cu.Cu.Nwires.EC20.7d	Cu.CT.7d	Cu-NWires EC20 7 days	478
25	Cu.CuNO3.EC50.3d	Cu.CT.3d	CuNO3 EC50 3 days	522
26	Ag.AgNO3.EC20.3d	Ag.CT.3d	AgNO3 EC20 3 days	911
27	Ag.Coated.EC20.3d	Ag.CT.3d	Ag-NPs PVP-Coated EC20 3 days	930
28	Ag.NM300K.EC50.7d	Ag.CTD.7d	Ag NM300K EC50 7 days	1,264
29	Ni.Ni.NPs.EC20.3d	Ni.CT.3d	Ni-NPs EC20 3 days	1,460
30	Cu.Cu.field.EC20.7d	Cu.CT.7d	Cu-Field EC20 7 days	1,627
31	Ni.NiNO3.EC50.3d	Ni.CT.3d	NiNO3 EC50 3 days	1,649
32	Uv.UV.D2.5d	Uv.CT.5d	UV Dose 2	1,864
33	Ni.Ni.NPs.EC50.3d	Ni.CT.3d	Ni-NPs EC50 3 days	2,341
34	Cu.Cu.field.EC50.3d	Cu.CT.3d	Cu-Field EC50 3 days	3,578
35	Cu.Cu.field.EC50.7d	Cu.CT.7d	Cu-Field EC50 7 days	4,412
36	Uv.UV.D1.5d	Uv.CT.5d	UV Dose 1	4,746
37	Cu.Cu.NPs.EC50.3d	Cu.CT.3d	Cu-NPs EC50 3 days	5,993
38	Cu.Cu.field.EC20.3d	Cu.CT.3d	Cu-Field EC20 3 days	9,225
39	Ag.Coated.EC50.3d	Ag.CT.3d	Ag-NPs PVP-Coated EC50 3 days	9,474
40	Cu.Cu.Nwires.EC20.3d	Cu.CT.3d	Cu-NWires EC20 3 days	9,753
41	Ag.NC.EC50.3d	Ag.CT.3d	Ag-NPs Non-Coated EC50 3 days	9,876
42	Cu.Cu.Nwires.EC50.3d	Cu.CT.3d	Cu-NWires EC50 3 days	10,287

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary Figures

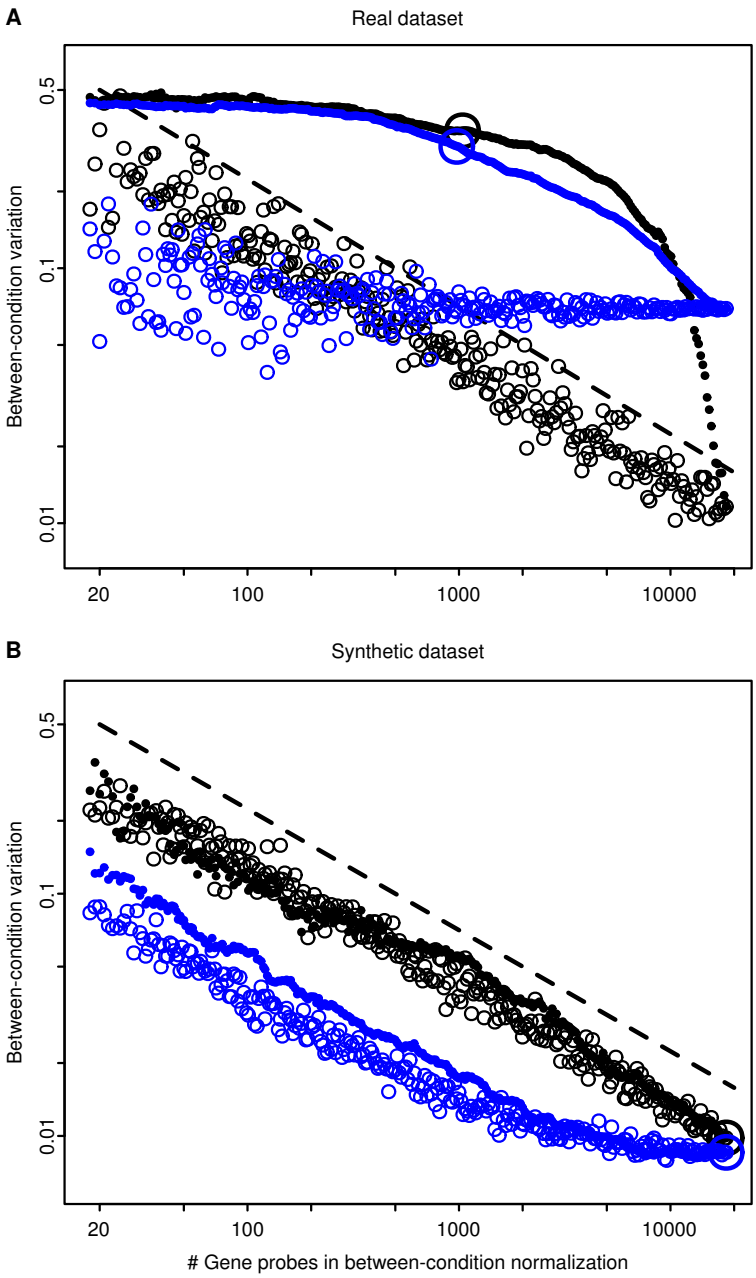

Figure S1

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure S1: Representing between-condition variation as the standard deviation of the within-condition median averages (averages of sample median expression levels, for all samples under the condition) yields similar results to those obtained with within-condition mean averages (Fig. 2). The panels show the detected variation as a function of the number of gene probes used in the between-condition normalization of the real dataset (**A**) and synthetic dataset (**B**). Labeling is the same as in Fig. 2. Each point in either of the panels indicates the variation obtained with one complete normalization (black circles, median condition-decomposition normalization; blue circles, standard-vector condition-decomposition normalization). Gene probes were selected in two ways: randomly (empty circles) or in decreasing order of p-values (filled circles). Big circles show the working points of the algorithms whose results are depicted in Fig. 1G–J. Black dashed lines show references for $n^{-1/2}$ decays, with the same values in both panels.

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

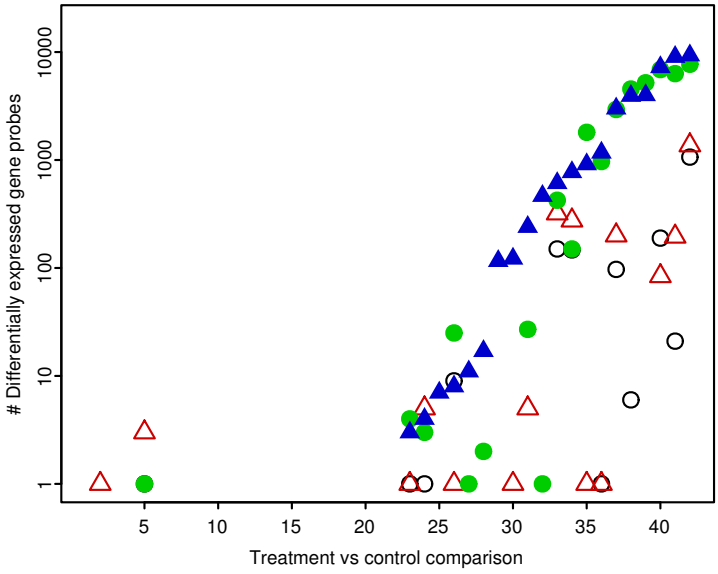

Figure S2: With t-tests, the condition-decomposition normalizations also detected much larger numbers of differentially expressed gene probes (DEGP). The figure shows the number of DEGP obtained with a statistical analysis based on t-tests instead of limma (Fig. 3A). Labeling is the same as in Fig. 3A (empty black circles, median normalization; empty red triangles, quantile normalization; filled green circles, median condition-decomposition normalization; filled blue triangles, standard-vector condition decomposition normalization). Treatment vs control comparisons are ordered according to the number of DEGP identified with standard-vector condition-decomposition normalization, increasing from left to right. This order (not shown) was similar but not exactly the same as in Fig. 3A.

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Materials and Methods

Test organism

The test species was *Enchytraeus crypticus*. Individuals were cultured in Petri dishes containing agar medium, in controlled conditions [30].

Exposure media

For copper (Cu) exposure, a natural soil collected at Hygum, Jutland, Denmark was used [30, 31]. For silver (Ag) and nickel (Ni) exposure, the natural standard soil LUFA 2.2 (LUFA Speyer, Germany) was used [30]. The exposure to ultra-violet (UV) radiation was done in ISO reconstituted water [32].

Test chemicals

The tested Cu forms [30] included copper nitrate ($Cu(NO_3)_2 \cdot 3H_2O > 99\%$, Sigma Aldrich), Cu nanoparticles (Cu-NPs, 20–30 nm, American Elements) and Cu nanowires (Cu-Nwires, synthesized by reduction of copper (II) nitrate with hydrazine in alkaline medium [33]).

The tested Ag forms [30] included silver nitratre $AgNO_3 > 99\%$, Sigma Aldrich), non-coated Ag nanoparticles (Ag-NPs Non-Coated, 20–30 nm, American Elements), Polyvinylpyrrolidone (PVP)-coated Ag nanoparticles (Ag-NPs PVP-Coated, 20–30 nm, American Elements), and Ag NM300K nanoparticles (Ag NM300K, 15 nm, JRC Repository). The Ag NM300K was dispersed in 4% Polyoxyethylene Glycerol Triolaete and Polyoxyethylene (20) orbitan mono-Laurat (Tween 20), thus the dispersant was tested alone as control (CTdisp).

The tested Ni forms included nickel nitrate ($Ni(NO_3)_2 \cdot 6H_2O \geq 98.5\%$, Fluka) and Ni nanoparticles (Ni-NPs, 20 nm, American Elements).

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Spiking procedure

Spiking for the Cu and Ag materials was done as in previous work [30]. For the Ni materials, the Ni-NPs were added to the soil as powder, following the same procedure as for the Cu materials. $NiNO_3$, being soluble, was added to the pre-moistened soil as aqueous dispersions.

The concentrations tested were selected based on the reproduction effect concentrations EC_{20} and EC_{50}, for *E. crypticus*, within 95% of confidence intervals, being: $CuNO_3$ $EC_{20/50}$ = 290/360 mgCu/kg, Cu-NPs $EC_{20/50}$ = 980/1760 mgCu/kg, Cu-Nwires $EC_{20/50}$ = 850/1610 mgCu/kg, Cu-Field $EC_{20/50}$ = 500/1400 mgCu/kg, $AgNO_3$ $EC_{20/50}$ = 45/60 mgAg/kg, Ag-NP PVP-coated $EC_{20/50}$ = 380/550 mgAg/kg, Ag-NP Non-coated $EC_{20/50}$ = 380/430 mgAg/kg, Ag NM300K $EC_{20/50}$ = 60/170 mgAg/kg, CTdisp = 4% w/w Tween 20, $NiNO_3$ $EC_{20/50}$ = 40/60 mgNi/kg, Ni-NPs $EC_{20/50}$ = 980/1760 mgNi/kg.

Four biological replicates were performed per test condition, including controls. For Cu exposure, the control condition for all the treatments consisted of soil from a control area at Hygum site, which has a Cu background concentration of 15 mg/kg [31]. For Ag exposure, two control sets were performed: CT (un-spiked LUFA soil, to be the control condition for $AgNO_3$, Ag-NPs PVP-Coated and Ag-NPs Non-Coated treatments) and CTdisp (LUFA soil spiked with the dispersant Tween 20, to be the control condition for the Ag NM300K treatments). For Ni exposure, the control consisted of un-spiked LUFA soil.

Exposure details

In soil (i.e. for Cu, Ag and Ni) exposure followed the standard ERT [34] with adaptations as follows: twenty adults with well-developed clitellum were introduced in each test vessel, containing 20 g of moist soil (control or spiked). The organisms were exposed for three and seven days under controlled conditions of photoperiod (16:8 h light:dark) and temperature $20 \pm 1\,^{\circ}C$ without food. After the exposure period, the organisms were carefully removed

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

from the soil, rinsed in deionized water and frozen in liquid nitrogen. The samples were stored at −80 °C, until analysis.

For UV exposure, the test conditions [32] were adapted for *E. crypticus* [35]. The exposure was performed in 24-well plates, where each well correspond to a replicate and contain 1 ml of ISO water and five adult organisms with clitellum. The test duration was five days, at 20 ± 1 °C. The organisms were exposed to UV on a daily basis, during 15 minutes per day to two UV intensities (280–400nm) of 1669.25 ± 50.83 and 1804.08 ± 43.10 mW/m^2, corresponding to total UV doses of 7511.6 and 8118.35 J/m^2, respectively. The remaining time was spent under standard laboratory illumination (16:8 h photoperiod). UV radiation was provided by an UV lamp (Spectroline XX15F/B, Spectronics Corporation, NY, USA, peak emission at 312 nm) and a cellulose acetate sheet was coupled to the lamp to cut-off UVC-range wavelengths [35]. Thirty two replicates per test condition (including control without UV radiation) were performed to obtain 4 biological replicates with 40 organisms each for RNA extraction. After the exposure period, the organisms were carefully removed from the water and frozen in liquid nitrogen. The samples were stored at −80 °C, until analysis.

RNA extraction, labeling and hybridization

RNA was extracted from each replicate, which contained a pool of 20 and 40 organisms, for soil and water exposure, respectively. Three biological replicates per test treatment (including controls) were used. Total RNA was extracted using SV Total RNA Isolation System (Promega). The quantity and purity were measured spectrophotometrically with a nanodrop (NanoDrop ND-1000 Spectrophotometer) and its quality checked by denaturing formaldehyde agarose gel electrophoresis.

500 ng of total RNA were amplified and labeled with Agilent Low Input Quick Amp Labeling Kit (Agilent Technologies, Palo Alto, CA, USA). Positive controls were added with the Agilent one-color RNA Spike-In Kit. Purification of the amplified and labeled cRNA was performed with RNeasy columns (Qiagen, Valencia, CA, USA).

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The cRNA samples were hybridized on custom Gene Expression Agilent Microarrays (4 x 44k format), with a single-color design [36]. Hybridizations were performed using the Agilent Gene Expression Hybridization Kit and each biological replicate was individually hybridized on one array. The arrays were hybridized at 65 °C with a rotation of 10 rpm, during 17 h. Afterwards, microarrays were washed using Agilent Gene Expression Wash Buffer Kit and scanned with the Agilent DNA microarray scanner G2505B.

Data acquisition and analysis

Fluorescence intensity data was obtained with Agilent Feature Extraction Software v. 10.7.3.1, using recommended protocol GE1_107_Sep09. Quality control was done by inspecting the reports on the Agilent Spike-in control probes. Background correction was provided by Agilent Feature Extraction software. To ensure an optimal comparison between the different normalization methods, only gene probes with good signal quality (flag IsPosAndSignif = True) in all samples were employed in the analyses. This implied the selection of 18,339 gene probes from a total of 43,750. Analyses were performed with R [37] v. 3.2.0, using R packages plotrix and RColorBrewer, and with Bioconductor [38] v. 3.1 packages genefilter and limma [39].

The synthetic data was generated gene by gene as normal variates with mean and variance equal, respectively, to the sample mean and sample variance of the real data. The applied normalization factors were those detected from the real data with standard-vector condition-decomposition normalization.

Median normalization was performed by subtracting the median of each sample distribution, and then adding the overall median to preserve the global expression level. Quantile normalization was performed as implemented in the limma package.

The two condition-decomposition normalizations proceeded in the same way: first, 51 independent within-condition normalization using all genes; then, final between-condition normalization, iteratively detecting no-variation genes and normalizing until convergence.

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

No-variation genes were identified with one-sided Kolmogorov-Smirnov tests, as goodness-of-fit tests against the uniform distribution, carried out on the greatest p-values obtained from an ANOVA test on the complete dataset (see below). The ANOVA test benefited from the already corrected within-condition variances, provided by the within-condition normalizations. The KS test was rejected at $\alpha = 0.001$.

The criterion for convergence for the median condition-decomposition (CD) normalizations was to require that the relative changes in the standard deviation of the normalization factors were less than 1%, or less than 10% for 10 steps in a row. In the case of standard-vector CD normalizations, convergence required that numerical errors were, compared to the estimated statistical errors (see below), less than 1%, or less than 10% for 10 steps in a row. For Figs. 2 and S1, due to the very low number of gene probes in some cases, the thresholds for convergence for 10 steps in a row were increased to 80% and 50%, respectively, for median CD and standard-vector CD normalization.

In standard-vector CD normalization, the distribution of standard vectors was trimmed in each step to remove the 1% more extreme values of variance.

Differentially expressed gene probes were identified with limma (Fig. 3) or t-tests (Fig. S2), using in all cases a FDR threshold of 5%.

The reference distribution with permutation symmetry shown in the polar plots of the probability density function in Movies S1–S3 was calculated with the 6 permutations of the empirical standard vectors. The Watson U^2 statistic was calculated with the two-sample test [40]. An equal number of samples for comparison was obtained by sampling with replacement the permuted standard vectors.

Mathematical Methods

In a gene expression dataset with g genes, c experimental conditions and n samples per condition, the *observed* expression levels of gene j in condition k, $\mathbf{y}_j^{(k)} = (y_{1j}^{(k)}, \ldots, y_{nj}^{(k)})'$,

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

can be expressed in $\log_2$-scale as

$$\mathbf{y}_j^{(k)} \;=\; \mathbf{x}_j^{(k)} + \mathbf{a}^{(k)}, \tag{1}$$

where $\mathbf{x}_j^{(k)}$ is the vector of *true* gene expression levels and $\mathbf{a}^{(k)}$ is the vector of normalization factors.

Given a sample vector $\mathbf{x}$, the mean vector is $\overline{\mathbf{x}} = \bar{x}\mathbf{1}$, and the residual vector is $\widetilde{\mathbf{x}} = \mathbf{x} - \overline{\mathbf{x}}$. Then, (1) can be linearly decomposed into

$$\overline{\mathbf{y}}_j^{(k)} \;=\; \overline{\mathbf{x}}_j^{(k)} + \overline{\mathbf{a}}^{(k)}, \tag{2}$$

$$\widetilde{\mathbf{y}}_j^{(k)} \;=\; \widetilde{\mathbf{x}}_j^{(k)} + \widetilde{\mathbf{a}}^{(k)}. \tag{3}$$

Equations (3) define the within-condition normalizations for each condition k. The scalar values in (2) are used to obtain the equations on condition means,

$$\overline{\mathbf{y}}_j^{*} \;=\; \overline{\mathbf{x}}_j^{*} + \overline{\mathbf{a}}^{*}, \tag{4}$$

$$\widetilde{\mathbf{y}}_j^{*} \;=\; \widetilde{\mathbf{x}}_j^{*} + \widetilde{\mathbf{a}}^{*}. \tag{5}$$

The between-condition normalization is defined by (5). Equations (4) reduce to a single number, which is irrelevant to the normalization. The complete solution for each condition is obtained with $\mathbf{a}^{(k)} \;=\; \overline{\mathbf{a}}^{(k)} + \widetilde{\mathbf{a}}^{(k)}$.

The n samples of gene j in a given condition can be modeled with the random vectors $\mathbf{X}_j, \mathbf{Y}_j \in \mathbb{R}^n$. Again, $\mathbf{Y}_j = \mathbf{X}_j + \mathbf{a}$, where $\mathbf{a}$ is a fixed vector of normalization factors. It can be proved, under fairly general assumptions, that the true standard vectors have zero expected value

$$\mathrm{E}\left(\sqrt{n-1}\,\frac{\widetilde{\mathbf{X}}_j}{\|\widetilde{\mathbf{X}}_j\|}\right) \;=\; \mathbf{0}, \tag{6}$$

whereas the observed standard vectors verify, as long as $\mathbf{a} \neq \mathbf{0}$,

$$0 \;<\; \mathrm{E}\left(\sqrt{n-1}\,\frac{\widetilde{\mathbf{Y}}_j}{\|\widetilde{\mathbf{Y}}_j\|}\right)' \frac{\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{a}}\|} \;<\; \mathrm{E}\left(\sqrt{n-1}\,\frac{1}{\|\widetilde{\mathbf{Y}}_j\|}\right) \|\widetilde{\mathbf{a}}\|. \tag{7}$$

This motivates the following iterative procedure to solve (3) and (5) (*standard-vector*

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

normalization):

$$\widehat{\mathbf{y}}_j^{(0)} = \widetilde{\mathbf{y}}_j, \tag{8}$$

$$\widehat{\mathbf{y}}_j^{(t)} = \widehat{\mathbf{y}}_j^{(t-1)} - \widehat{\mathbf{b}}^{(t-1)}, \quad \text{for } t \geq 1, \tag{9}$$

$$\widehat{\mathbf{b}}^{(t)} = \frac{\displaystyle\sum_{j=1}^{g} \frac{\widehat{\mathbf{y}}_j^{(t)}}{\|\widehat{\mathbf{y}}_j^{(t)}\|}}{\displaystyle\sum_{j=1}^{g} \frac{1}{\|\widehat{\mathbf{y}}_j^{(t)}\|}}, \quad \text{for } t \geq 0. \tag{10}$$

At convergence, $\lim_{t\to\infty} \widehat{\mathbf{b}}^{(t)} = \mathbf{0}$, which implies $\lim_{t\to\infty} \widehat{\mathbf{y}}_j^{(t)} = \widetilde{\mathbf{x}}_j$ and $\sum_{t=0}^{\infty} \widehat{\mathbf{b}}^{(t)} = \widetilde{\mathbf{a}}$. Convergence is faster the more symmetric the empirical distribution of $\widetilde{\mathbf{x}}_j/\|\widetilde{\mathbf{x}}_j\|$ is on the unit $(n-2)$-sphere. Convergence is optimal with spherically symmetric distributions, such as the Gaussian distribution, because in that case

$$\mathrm{E}\left(\frac{\widetilde{\mathbf{Y}}_j}{\|\widetilde{\mathbf{Y}}_j\|}\right) = \lambda\widetilde{\mathbf{a}}, \quad \text{with} \quad 0 < \lambda < \mathrm{E}\left(\frac{1}{\|\widetilde{\mathbf{Y}}_j\|}\right). \tag{11}$$

Assuming no correlation between genes, an approximation of the statistical error at step t can be obtained with

$$\mathrm{E}\left(\|\widehat{\mathbf{b}}^{(t)}\|\right) \approx \frac{\sqrt{g}}{\displaystyle\sum_{j=1}^{g} \frac{1}{\|\widehat{\mathbf{y}}_j^{(t)}\|}}. \tag{12}$$

This statistical error is compared with the numerical error to assess convergence.

See Supplementary Text below for a detailed exposition of the mathematical methods, and Movies S1–S5 for an illustration.

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Supplementary Methods

Contents

SM1 Vectorial representation of sample data 32

SM2 Linear decomposition of the normalization problem 34

SM3 Permutation invariance of multivariate data 38

SM4 Standard-vector normalization 44

SM5 Identification of non-differentially expressed genes 46

SM1 Vectorial representation of sample data

Let $x_1, \ldots, x_n$ be the samples of n independent and identically distributed random variables $X_1, \ldots, X_n$. Let us represent the samples $x_1, \ldots, x_n$ with the $\mathbb{R}^n$ column vector $\mathbf{x} = (x_1, \ldots, x_n)'$, and let us denote the sample mean by $\bar{x} = \sum_{i=1}^{n} x_i/n$.

Let us define the $\mathbb{R}^n \to \mathbb{R}^n$ vectorial operators mean $(\overline{\cdot})$ and residual $(\widetilde{\cdot})$, respectively, as

$$\overline{\mathbf{x}} \;=\; (\,\bar{x},\, \ldots\, ,\bar{x}\,)' \;=\; \bar{x}\mathbf{1}, \tag{13}$$

$$\widetilde{\mathbf{x}} \;=\; \mathbf{x} - \overline{\mathbf{x}} \;=\; \mathbf{x} - \bar{x}\mathbf{1}, \tag{14}$$

$\mathbf{1}$ being the all-ones column vector of dimension n.

Thus, any sample vector $\mathbf{x} \in \mathbb{R}^n$ can be decomposed as

$$\mathbf{x} \;=\; \overline{\mathbf{x}} + \widetilde{\mathbf{x}}. \tag{15}$$

The mean vector $\overline{\mathbf{x}}$ contains the sample mean, while the residual vector $\widetilde{\mathbf{x}}$ carries the sample variation around the mean.

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The vectorial operators mean (13) and residual (14) are linear.

Proposition. For any two sample vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and any two numbers $\alpha, \beta \in \mathbb{R}$,

$$\overline{\alpha\mathbf{x} + \beta\mathbf{y}} \;=\; \alpha\overline{\mathbf{x}} + \beta\overline{\mathbf{y}}, \tag{16}$$

$$\widetilde{\alpha\mathbf{x} + \beta\mathbf{y}} \;=\; \alpha\widetilde{\mathbf{x}} + \beta\widetilde{\mathbf{y}}. \tag{17}$$

Proof. Let us denote $\mathbf{x} = (x_1, \ldots, x_n)'$ and $\mathbf{y} = (y_1, \ldots, y_n)'$.

$$\begin{aligned}
\overline{\alpha\mathbf{x} + \beta\mathbf{y}} \;&=\; \frac{\sum_{i=1}^n (\alpha x_i + \beta y_i)}{n}\mathbf{1} \;=\; \alpha\frac{\sum_{i=1}^n x_i}{n}\mathbf{1} + \beta\frac{\sum_{i=1}^n y_i}{n}\mathbf{1} \;=\; \alpha\overline{\mathbf{x}} + \beta\overline{\mathbf{y}}, \\
\widetilde{\alpha\mathbf{x} + \beta\mathbf{y}} \;&=\; \alpha\mathbf{x} + \beta\mathbf{y} - \overline{\alpha\mathbf{x} + \beta\mathbf{y}} \;=\; \alpha\mathbf{x} + \beta\mathbf{y} - (\alpha\overline{\mathbf{x}} + \beta\overline{\mathbf{y}}), \\
&=\; \alpha(\mathbf{x} - \overline{\mathbf{x}}) + \beta(\mathbf{y} - \overline{\mathbf{y}}) \;=\; \alpha\widetilde{\mathbf{x}} + \beta\widetilde{\mathbf{y}}. \quad \square
\end{aligned}$$

An essential property of the mean and residual vectors is that they belong to subspaces that are orthogonal complements [41]. Hence, for any sample vector $\mathbf{x} \in \mathbb{R}^n$, the mean vector $\overline{\mathbf{x}}$ belongs to the subspace of dimension 1 spanned by the unit vector $\widehat{\mathbf{1}} = \mathbf{1}/\sqrt{n}$, while the residual vector $\widetilde{\mathbf{x}}$ belongs to the $(n-1)$-dimensional hyperplane orthogonal to $\widehat{\mathbf{1}}$.

The lengths of the mean vector and residual vector are equal, up to a scaling factor, to the sample mean and sample standard deviation, respectively. For a set of samples $x_1, \ldots, x_n$, where $n \geq 2$, let us denote the sample mean as before by $\bar{x} = \sum_{i=1}^n x_i/n$, and the sample variance as $s_x^2 = \sum_{i=1}^n (x_i - \bar{x})^2/(n-1)$. Then, the lengths of the mean and residual vectors obtained from the sample vector $\mathbf{x} = (x_1, \ldots, x_n)'$ are

$$\|\overline{\mathbf{x}}\| \;=\; \sqrt{n\,\bar{x}^2} \;=\; \sqrt{n}\,|\bar{x}|, \tag{18}$$

$$\|\widetilde{\mathbf{x}}\| \;=\; \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \;=\; \sqrt{n-1}\,s_x. \tag{19}$$

Finally, let us define the standard vector of the sample vector $\mathbf{x} = (x_1, \ldots, x_n)'$ $(n \geq 2)$, as

$$\mathrm{stdvec}(\mathbf{x}) \;=\; \sqrt{n-1}\,\frac{\widetilde{\mathbf{x}}}{\|\widetilde{\mathbf{x}}\|}, \tag{20}$$

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

whenever $\widetilde{\mathbf{x}} \neq \mathbf{0}$, or otherwise as $\mathrm{stdvec}(\mathbf{x}) = \mathbf{0}$. $\mathbf{0}$ is the all-zeros column vector of dimension n.

For a given number of samples n, all the non-zero standard vectors belong to the $(n-2)$-sphere of radius $\sqrt{n-1}$, embedded in the $(n-1)$-dimensional hyperplane perpendicular to $\widehat{\mathbf{1}}$. Besides, all the components of a standard vector are equal to the corresponding standardized samples,

$$\sqrt{n-1}\frac{\widetilde{x}_i}{\|\widetilde{\mathbf{x}}\|} = \frac{x_i - \bar{x}}{s_x}. \tag{21}$$

For the degenerate case of having only two samples ($n = 2$), the only possible values of a non-zero standard vector are $\pm(\,1/\sqrt{2},\ -1/\sqrt{2}\,)'$.

SM2 Linear decomposition of the normalization problem

Let us consider a gene expression dataset, with g genes and c experimental conditions. Each condition k has s_k samples. The total number of samples is $s = \sum_{k=1}^{c} s_k$.

Let us denote the *observed* expression level of gene j in the sample i of condition k by $y_{ij}^{(k)}$. We assume that the observed level $y_{ij}^{(k)}$ is equal, in the usual $\log_2$-scale, to the addition of the normalization factor $a_i^{(k)}$ to the *true* gene expression level $x_{ij}^{(k)}$,

$$y_{ij}^{(k)} = x_{ij}^{(k)} + a_i^{(k)}. \tag{22}$$

Solving the *normalization problem* amounts to finding the normalization factors $a_i^{(k)}$ from the observed values $y_{ij}^{(k)}$. The normalization factors can be understood as sample-wide changes in the concentration of mRNA molecules by multiplicative factors equal to $2^{a_i^{(k)}}$. These changes are caused by technical reasons in the assay and are independent of the biological variation in the true levels $x_{ij}^{(k)}$.

Let us represent the true and observed expression levels, $x_{ij}^{(k)}$ and $y_{ij}^{(k)}$, of gene j in the

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

samples $i = 1 \ldots s_k$ of condition k, by the s_k-dimensional vectors

$$\mathbf{x}_j^{(k)} = (x_{1j}^{(k)}, \ldots, x_{s_k j}^{(k)})', \tag{23}$$

$$\mathbf{y}_j^{(k)} = (y_{1j}^{(k)}, \ldots, y_{s_k j}^{(k)})'. \tag{24}$$

Let us also represent the unknown normalization factors of condition k by the s_k-dimensional vector

$$\mathbf{a}^{(k)} = (a_1^{(k)}, \ldots, a_{s_k}^{(k)})'. \tag{25}$$

From (22)–(25), the normalization problem can be written in vectorial form as

$$\mathbf{y}_j^{(k)} = \mathbf{x}_j^{(k)} + \mathbf{a}^{(k)}. \tag{26}$$

Applying the vectorial operators mean (13) and residual (14), we obtain

$$\overline{\mathbf{y}}_j^{(k)} = \overline{\mathbf{x}}_j^{(k)} + \overline{\mathbf{a}}^{(k)}, \tag{27}$$

$$\widetilde{\mathbf{y}}_j^{(k)} = \widetilde{\mathbf{x}}_j^{(k)} + \widetilde{\mathbf{a}}^{(k)}. \tag{28}$$

The residual-vector equations (28) correspond to the c within-condition normalizations. Each within-condition normalization uses the equations (28) particular to a condition k, for the subset of genes $\mathcal{G}_k \subseteq \{1, \ldots, g\}$ that have expression level available and of enough quality in that experimental condition.

Let us denote the condition means for each gene as

$$\overline{x}_j^{(k)} = \frac{\sum_{i=1}^{s_k} x_{ij}^{(k)}}{s_k}, \tag{29}$$

$$\overline{y}_j^{(k)} = \frac{\sum_{i=1}^{s_k} y_{ij}^{(k)}}{s_k}, \tag{30}$$

$$\overline{a}^{(k)} = \frac{\sum_{i=1}^{s_k} a_i^{(k)}}{s_k}, \tag{31}$$

so that

$$\overline{\mathbf{x}}_j^{(k)} = \overline{x}_j^{(k)} \mathbf{1}_{s_k}, \tag{32}$$

$$\overline{\mathbf{y}}_j^{(k)} = \overline{y}_j^{(k)} \mathbf{1}_{s_k}, \tag{33}$$

$$\overline{\mathbf{a}}^{(k)} = \overline{a}^{(k)} \mathbf{1}_{s_k}, \tag{34}$$

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

$\mathbf{1}_{s_k}$ being the all-ones column vector of dimension s_k.

Then, the mean-vector equations (27) can be written as

$$\overline{y}_j^{(k)}\mathbf{1}_{s_k} = \overline{x}_j^{(k)}\mathbf{1}_{s_k} + \overline{a}^{(k)}\mathbf{1}_{s_k}, \tag{35}$$

so they reduce to the scalar equations

$$\overline{y}_j^{(k)} = \overline{x}_j^{(k)} + \overline{a}^{(k)}. \tag{36}$$

Let us define the vectors of conditions means as

$$\mathbf{x}_j^* = (\,\overline{x}_j^{(1)}, \ldots, \overline{x}_j^{(c)}\,)', \tag{37}$$

$$\mathbf{y}_j^* = (\,\overline{y}_j^{(1)}, \ldots, \overline{y}_j^{(c)}\,)', \tag{38}$$

$$\mathbf{a}^* = (\,\overline{a}^{(1)}, \ldots, \overline{a}^{(c)}\,)', \tag{39}$$

and let us express the condition-mean equations in vectorial form as

$$\mathbf{y}_j^* = \mathbf{x}_j^* + \mathbf{a}^*. \tag{40}$$

Applying again the mean and variance operators, we obtain

$$\overline{\mathbf{y}}_j^* = \overline{\mathbf{x}}_j^* + \overline{\mathbf{a}}^*, \tag{41}$$

$$\widetilde{\mathbf{y}}_j^* = \widetilde{\mathbf{x}}_j^* + \widetilde{\mathbf{a}}^*. \tag{42}$$

The residual-vector equations on condition means (42) correspond to the single between-condition normalization, in a similar way as (28) do for the each of the within-condition normalizations. There is one equation (42) per gene. The only equations used in the between-condition normalization are those of the subset of genes $\mathcal{G}^* \subseteq \{1, \ldots, g\}$ that show no evidence of variation across experimental conditions, according to a statistical test.

Given that $\overline{\mathbf{a}}^* = \overline{a}^*\mathbf{1}_c$, (41) has the only unknown $\overline{a}^*$. The meaning of $\overline{a}^*$ is a conversion factor between the scale the true and observed expression levels. This factor depends on

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the technology used to measure the expression levels and finding it is out of the scope of the normalization problem. Therefore, without loss of generality, we assume $\overline{a}^* = 0$, so

$$\overline{\mathbf{a}}^* = \mathbf{0}_c, \tag{43}$$

$$\mathbf{a}^* = \widetilde{\mathbf{a}}^*. \tag{44}$$

The solution of the between-condition normalization, $\widetilde{\mathbf{a}}^*$, allows to find the mean vectors of the normalization factors $\overline{\mathbf{a}}^{(k)}$, via (34), (39) and (44). The within-condition normalizations yield the residual vectors $\widetilde{\mathbf{a}}^{(k)}$. The complete solution to the normalization problem is finally obtained, for each condition k, with

$$\mathbf{a}^{(k)} = \overline{\mathbf{a}}^{(k)} + \widetilde{\mathbf{a}}^{(k)}. \tag{45}$$

Thus, the original normalization problem (26) has been divided in $c+1$ normalization subproblems on residual vectors, stated by (28) and (42). In fact, this linear decomposition is possible for any partition of the set of s samples. The choice of the partition as the one defined by the experimental conditions is motivated by the need to control the biological variation among the genes used in each normalization. All the $c+1$ normalizations face the same kind of *normalization of residuals problem*, which we define in general as follows.

Normalization of Residuals Problem. Let y_{ij} be the i-th observed value of feature j, in a dataset with $n \geq 2$ observations for each of the m features. The observed values y_{ij} are equal to the true values x_{ij} plus the normalization factors a_i, which are constant across features. In vectorial form, there are m equations

$$\mathbf{y}_j = \mathbf{x}_j + \mathbf{a}, \tag{46}$$

where the vectors belong to $\mathbb{R}^n$. As a consequence

$$\widetilde{\mathbf{y}}_j = \widetilde{\mathbf{x}}_j + \widetilde{\mathbf{a}}. \tag{47}$$

Solving the *normalization of residuals problem* amounts to finding the residual vector of normalization factors $\widetilde{\mathbf{a}}$ from the observed residual vectors $\widetilde{\mathbf{y}}_j$. In the within-condition

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

normalizations, the features are gene expression levels, with one observation per sample of the corresponding experimental condition. In the between-condition normalization, the features are means of gene expression levels, with one observation per condition.

There is, however, an additional requirement imposed by the methods with which we propose to solve the between-condition normalization. We would like to consider the condition means $\overline{x}_j^{(k)}$ in (36) as sample data across conditions. This only holds when all the conditions have the same number of samples. Otherwise, we balance the condition means so that they result from the same number of samples in all conditions, according to the procedure described in the following.

Let s^* be the minimum number of samples across conditions, $s^* = \min\{s_1, \ldots, s_c\}$. Let $\mathcal{S}_j^{(k)}$ be independent random samples (without replacement) of size s^* from the set of indexes $\{1, \ldots, s_k\}$, with one sample per gene j and condition k. Then, the balanced condition means are defined as

$$\overline{x}_j^{(k)*} = \frac{\sum_{i \in \mathcal{S}_j^{(k)}} x_{ij}^{(k)}}{s^*}, \tag{48}$$

$$\overline{y}_j^{(k)*} = \frac{\sum_{i \in \mathcal{S}_j^{(k)}} y_{ij}^{(k)}}{s^*}, \tag{49}$$

$$\overline{a}_j^{(k)*} = \frac{\sum_{i \in \mathcal{S}_j^{(k)}} a_i^{(k)}}{s^*}. \tag{50}$$

From (22), the balanced condition means verify a relationship similar to (36),

$$\overline{y}_j^{(k)*} = \overline{x}_j^{(k)*} + \overline{a}_j^{(k)*}. \tag{51}$$

Moreover, the average of $\overline{a}_j^{(k)*}$ across the sampling subsets $\mathcal{S}_j^{(k)}$ is equal to the unknown $\overline{a}^{(k)}$. This implies that (51) are, on average, equivalent to (36). Hence, we use the following vectors of balanced conditions means

$$\mathbf{x}_j^* = (\overline{x}_j^{(1)*}, \ldots, \overline{x}_j^{(c)*}), \tag{52}$$

$$\mathbf{y}_j^* = (\overline{y}_j^{(1)*}, \ldots, \overline{y}_j^{(c)*}), \tag{53}$$

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

instead of (37), (38), in order to build the condition-mean equations (40). This balancing of the condition means is only required when the experimental conditions have different number of samples.

SM3 Permutation invariance of multivariate data

Let x_{ij} and y_{ij} be, respectively, the true and observed values of a dataset with n observations of m features, as defined in the *normalization of residuals problem* above.

We have assumed that the n true values $x_{1j}, \ldots, x_{nj}$ of feature j are samples of independent and identically distributed random variables $X_{1j}, \ldots, X_{nj}$. These random variables can be represented with the random vector $\mathbf{X}_j = (X_{1j}, \ldots, X_{nj})'$, carried by the probability space $(\Omega, \mathcal{F}, \mathrm{P})$ and with induced space $(\mathbb{R}^n, \mathbb{B}^n, \mathbb{P})$. Let us define the random vectors $\overline{\mathbf{X}}_j$ and $\widetilde{\mathbf{X}}_j$ with the vectorial operators mean (13) and residual (14), respectively,

$$\overline{X}_j = \sum_{i=1}^{n} \frac{X_{ij}}{n}, \tag{54}$$

$$\overline{\mathbf{X}}_j = (\overline{X}_j, \ldots, \overline{X}_j)' = \overline{X}_j \mathbf{1}, \tag{55}$$

$$\widetilde{\mathbf{X}}_j = \mathbf{X}_j - \overline{\mathbf{X}}_j = \mathbf{X}_j - \overline{X}_j \mathbf{1}. \tag{56}$$

$\mathbf{X}_j = \overline{\mathbf{X}}_j + \widetilde{\mathbf{X}}_j$ holds for any random vector $\mathbf{X}_j$, as well as the other properties presented above. Let us assume that $\mathrm{E}(\,\|\mathbf{X}_j\|\,) < \infty$ and that $\mathrm{P}(\,\|\widetilde{\mathbf{X}}_j\| = 0\,) = 0$, which imply that $\widetilde{\mathbf{X}}_j / \|\widetilde{\mathbf{X}}_j\|$ has length 1 almost surely.

The standard random vector $\sqrt{n-1}\,\widetilde{\mathbf{X}}_j / \|\widetilde{\mathbf{X}}_j\|$ is a pivotal quantity, where the location (mean) and scale (standard deviation) of feature j have been removed. The probability distribution of $\widetilde{\mathbf{X}}_j / \|\widetilde{\mathbf{X}}_j\|$ across the remaining degrees of freedom over the unit $(n-2)$-sphere is governed by the parametric family of the random variables $X_{1j}, \ldots, X_{nj}$. Moreover, the independence and identity of distribution across the n observations implies that the distribution of $\mathbf{X}_j$ is *exchangeable*, i.e. invariant with respect to permutations of the observation labels. As a result, $\widetilde{\mathbf{X}}_j / \|\widetilde{\mathbf{X}}_j\|$ is also permutation invariant, which geometrically corresponds to symmetries with respect to the $n!$ permutations of the axes

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

in the n-dimensional space of random vectors, projected onto the $(n-1)$-dimensional hyperplane of residual vectors.

Residual vectors and standard vectors have been widely studied, especially in relation to elliptically symmetric distributions and linear models [42, 43], and to the invariances of probability distributions [44]. Here, we consider these vectors from the viewpoint of the problem of normalizing multivariate data, and its relationship with permutation invariance.

It is well know that, for a multivariate distribution with independent and identically distributed components, the expected value of the standard vector is zero [41], given that it is so for each component. We prove this here for completeness, and to show that it is also a necessary consequence of the permutation invariance of the distribution.

Proposition. The expected value of any true (i.e. without normalization issues) standard vector is zero. If the $n \geq 2$ samples of feature j are independent and identically distributed, then

$$\mathrm{E}\left(\sqrt{n-1}\,\frac{\widetilde{\mathbf{X}}_j}{\|\widetilde{\mathbf{X}}_j\|}\right) \;=\; \mathbf{0}. \tag{57}$$

Proof. Let $\mathcal{P}_n$ be the set of all the permutation matrices in $\mathbb{R}^{n\times n}$. Then, for any $P \in \mathcal{P}_n$, $\widetilde{\mathbf{X}}_j/\|\widetilde{\mathbf{X}}_j\|$ is equal in distribution to $P\,\widetilde{\mathbf{X}}_j/\|\widetilde{\mathbf{X}}_j\|$. This implies that

$$\mathrm{E}\left(\frac{\widetilde{\mathbf{X}}_j}{\|\widetilde{\mathbf{X}}_j\|}\right) \;=\; \mathrm{E}\left(P\frac{\widetilde{\mathbf{X}}_j}{\|\widetilde{\mathbf{X}}_j\|}\right) \;=\; P\,\mathrm{E}\left(\frac{\widetilde{\mathbf{X}}_j}{\|\widetilde{\mathbf{X}}_j\|}\right).$$

The only vectors that are invariant with respect to all possible permutations are those that have all components identical. Therefore, $\mathrm{E}(\,\widetilde{\mathbf{X}}_j/\|\widetilde{\mathbf{X}}_j\|\,) = \alpha\widehat{\mathbf{1}}$, with $\alpha \in \mathbb{R}$. However, $\widetilde{\mathbf{X}}_j'\widehat{\mathbf{1}} = 0$, so that $\alpha = \mathrm{E}(\,\widetilde{\mathbf{X}}_j/\|\widetilde{\mathbf{X}}_j\|\,)'\,\widehat{\mathbf{1}} = 0$. Hence $\mathrm{E}(\,\widetilde{\mathbf{X}}_j/\|\widetilde{\mathbf{X}}_j\|\,) = \mathbf{0}$. $\square$

For each true random vector $\mathbf{X}_j$, there is an observed random vector $\mathbf{Y}_j = \mathbf{X}_j + \mathbf{A}$, where $\mathbf{A}$ is the random vector of normalization factors. The random vectors $\mathbf{X}_j$ and $\mathbf{A}$ are independent, representing biological and technical variation, respectively. Therefore, and without loss of generality, we assume in what follows a fixed vector of normalization

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

factors $\mathbf{a}$, i.e. we condition on the event $\{\mathbf{A} = \mathbf{a}\}$. We also assume that $\mathrm{P}(\,\|\widetilde{\mathbf{Y}}_j\| = 0\,) = 0$, which implies that $\widetilde{\mathbf{Y}}_j/\|\widetilde{\mathbf{Y}}_j\|$ has length 1 almost surely.

In contrast to the true standard vector $\sqrt{n-1}\,\widetilde{\mathbf{X}}_j/\|\widetilde{\mathbf{X}}_j\|$, the observed standard vector $\sqrt{n-1}\,\widetilde{\mathbf{Y}}_j/\|\widetilde{\mathbf{Y}}_j\|$ is biased toward the direction of $\widetilde{\mathbf{a}}$, with the result that the expected value is not zero.

Proposition. If the $n \geq 2$ samples of feature j are independent and identically distributed, whenever $\widetilde{\mathbf{a}} \neq \mathbf{0}$,

$$\mathrm{E}\left(\sqrt{n-1}\,\frac{\widetilde{\mathbf{Y}}_j}{\|\widetilde{\mathbf{Y}}_j\|}\right) \neq \mathbf{0}. \tag{58}$$

When $n = 2$, there is the additional requirement that $\mathrm{P}(\,\|\widetilde{\mathbf{X}}_i\| < \|\widetilde{\mathbf{a}}\|\,) > 0$. This threshold of detection only occurs for the degenerate case of $n = 2$.

Proof. Let us consider the projection of $\widetilde{\mathbf{Y}}_j/\|\widetilde{\mathbf{Y}}_j\|$ on $\widetilde{\mathbf{a}}$, compared to the projection of $\widetilde{\mathbf{X}}_j/\|\widetilde{\mathbf{X}}_j\|$.

When the vectors $\widetilde{\mathbf{X}}_j$ and $\widetilde{\mathbf{a}}$ are collinear,

$$\frac{\widetilde{\mathbf{X}}_j'\,\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{X}}_j\|\,\|\widetilde{\mathbf{a}}\|} = \pm 1, \quad \text{and} \quad \frac{\widetilde{\mathbf{Y}}_j'\,\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{Y}}_j\|\,\|\widetilde{\mathbf{a}}\|} = \pm 1,$$

with

$$\frac{\widetilde{\mathbf{Y}}_j'\,\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{Y}}_j\|\,\|\widetilde{\mathbf{a}}\|} \geq \frac{\widetilde{\mathbf{X}}_j'\,\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{X}}_j\|\,\|\widetilde{\mathbf{a}}\|}.$$

This is the only case when $n = 2$. The additional requirement ensures that, for $n = 2$,

$$\mathrm{P}\left(\frac{\widetilde{\mathbf{Y}}_j'\,\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{Y}}_j\|\,\|\widetilde{\mathbf{a}}\|} > \frac{\widetilde{\mathbf{X}}_j'\,\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{X}}_j\|\,\|\widetilde{\mathbf{a}}\|}\right) > 0,$$

which implies

$$\mathrm{E}\left(\frac{\widetilde{\mathbf{Y}}_j'\,\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{Y}}_j\|\,\|\widetilde{\mathbf{a}}\|}\right) > \mathrm{E}\left(\frac{\widetilde{\mathbf{X}}_j'\,\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{X}}_j\|\,\|\widetilde{\mathbf{a}}\|}\right).$$

Otherwise, when $n > 2$ and the vectors $\widetilde{\mathbf{X}}_j$ and $\widetilde{\mathbf{a}}$ are not collinear, they lie on a plane. The vector $\widetilde{\mathbf{Y}}_j = \widetilde{\mathbf{X}}_j + \widetilde{\mathbf{a}}$ is the diagonal of the parallelogram defined by $\widetilde{\mathbf{X}}_j$ and $\widetilde{\mathbf{a}}$. Hence

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

the angle between $\widetilde{\mathbf{Y}}_j$ and $\widetilde{\mathbf{a}}$ is strictly less than the angle between $\widetilde{\mathbf{X}}_j$ and $\widetilde{\mathbf{a}}$, so the cosine of the angle is strictly greater. Thus,

$$\frac{\widetilde{\mathbf{Y}}_j'\,\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{Y}}_j\|\,\|\widetilde{\mathbf{a}}\|} > \frac{\widetilde{\mathbf{X}}_j'\,\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{X}}_j\|\,\|\widetilde{\mathbf{a}}\|}.$$

Due to the permutation symmetries in the distribution of $\widetilde{\mathbf{X}}_j/\|\widetilde{\mathbf{X}}_j\|$, when $n > 2$ the vector $\widetilde{\mathbf{X}}_j$ has non-zero probability of being not collinear with $\widetilde{\mathbf{a}}$, i.e. $\mathrm{P}(\,|\widetilde{\mathbf{X}}_j'\,\widetilde{\mathbf{a}}| < 1\,) > 0$.

Therefore,

$$\mathrm{P}\left(\frac{\widetilde{\mathbf{Y}}_j'\,\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{Y}}_j\|\,\|\widetilde{\mathbf{a}}\|} > \frac{\widetilde{\mathbf{X}}_j'\,\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{X}}_j\|\,\|\widetilde{\mathbf{a}}\|}\right) > 0,$$

which again implies

$$\mathrm{E}\left(\frac{\widetilde{\mathbf{Y}}_j'\,\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{Y}}_j\|\,\|\widetilde{\mathbf{a}}\|}\right) > \mathrm{E}\left(\frac{\widetilde{\mathbf{X}}_j'\,\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{X}}_j\|\,\|\widetilde{\mathbf{a}}\|}\right).$$

Finally,

$$\left\|\mathrm{E}\left(\frac{\widetilde{\mathbf{Y}}_j}{\|\widetilde{\mathbf{Y}}_j\|}\right)\right\| \geq \mathrm{E}\left(\frac{\widetilde{\mathbf{Y}}_j}{\|\widetilde{\mathbf{Y}}_j\|}\right)'\frac{\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{a}}\|} > \mathrm{E}\left(\frac{\widetilde{\mathbf{X}}_j}{\|\widetilde{\mathbf{X}}_j\|}\right)'\frac{\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{a}}\|} = 0. \quad \square$$

As a consequence, the *normalization of residuals problem* may be restated as the problem of finding the normalization factors $\widetilde{\mathbf{a}}$ from the observed vectors $\widetilde{\mathbf{y}}_j$, such that the standard vectors $\sqrt{n-1}\,(\widetilde{\mathbf{y}}_j - \widetilde{\mathbf{a}})/\|\widetilde{\mathbf{y}}_j - \widetilde{\mathbf{a}}\|$ are invariant against permutations of the observation labels. Or equivalently, such that the standard vectors $\sqrt{n-1}\,(\widetilde{\mathbf{y}}_j - \widetilde{\mathbf{a}})/\|\widetilde{\mathbf{y}}_j - \widetilde{\mathbf{a}}\|$ have zero mean. The following property provides an approach to the solution.

Proposition. Whenever $\widetilde{\mathbf{a}} \neq \mathbf{0}$, the component of the expected value of $\widetilde{\mathbf{Y}}_j/\|\widetilde{\mathbf{Y}}_j\|$ parallel to $\widetilde{\mathbf{a}}$ verifies

$$0 < \mathrm{E}\left(\frac{\widetilde{\mathbf{Y}}_j}{\|\widetilde{\mathbf{Y}}_j\|}\right)'\frac{\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{a}}\|} < \mathrm{E}\left(\frac{1}{\|\widetilde{\mathbf{Y}}_j\|}\right)\|\widetilde{\mathbf{a}}\|. \tag{59}$$

As in (58), when $n = 2$ we also assume that $\mathrm{P}(\,\|\widetilde{\mathbf{X}}_j\| < \|\widetilde{\mathbf{a}}\|\,) > 0$.

Proof. The first inequality holds from the previous proof. Concerning the second inequality, let us consider

$$\frac{\widetilde{\mathbf{Y}}_j'\,\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{Y}}_j\|\,\|\widetilde{\mathbf{a}}\|} = \frac{(\widetilde{\mathbf{X}}_j + \widetilde{\mathbf{a}})'\,\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{X}}_j + \widetilde{\mathbf{a}}\|\,\|\widetilde{\mathbf{a}}\|} = \frac{\|\widetilde{\mathbf{X}}_j\|}{\|\widetilde{\mathbf{X}}_j + \widetilde{\mathbf{a}}\|}\,\frac{\widetilde{\mathbf{X}}_j'\,\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{X}}_j\|\,\|\widetilde{\mathbf{a}}\|} + \frac{\|\widetilde{\mathbf{a}}\|}{\|\widetilde{\mathbf{Y}}_j\|}.$$

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

We need to prove that the first term on the RHS has negative expected value. Let us decompose this term into the positive and negative parts,

$$\frac{\|\widetilde{\mathbf{X}}_j\|}{\|\widetilde{\mathbf{X}}_j+\widetilde{\mathbf{a}}\|}\,\frac{\widetilde{\mathbf{X}}_j'\,\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{X}}_j\|\,\|\widetilde{\mathbf{a}}\|} = \left(\frac{\|\widetilde{\mathbf{X}}_j\|}{\|\widetilde{\mathbf{X}}_j+\widetilde{\mathbf{a}}\|}\,\frac{\widetilde{\mathbf{X}}_j'\,\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{X}}_j\|\,\|\widetilde{\mathbf{a}}\|}\right)^+ - \left(\frac{\|\widetilde{\mathbf{X}}_j\|}{\|\widetilde{\mathbf{X}}_j+\widetilde{\mathbf{a}}\|}\,\frac{\widetilde{\mathbf{X}}_j'\,\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{X}}_j\|\,\|\widetilde{\mathbf{a}}\|}\right)^-,$$

where $X^+ = \max(X,0)$ and $X^- = -\min(X,0)$.

Because $\|\widetilde{\mathbf{X}}_j+\widetilde{\mathbf{a}}\|^2 = \|\widetilde{\mathbf{X}}_j\|^2 + \|\widetilde{\mathbf{a}}\|^2 + 2\widetilde{\mathbf{X}}_j'\widetilde{\mathbf{a}}$,

$$\left(\frac{\|\widetilde{\mathbf{X}}_j\|}{\|\widetilde{\mathbf{X}}_j+\widetilde{\mathbf{a}}\|}\,\frac{\widetilde{\mathbf{X}}_j'\,\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{X}}_j\|\,\|\widetilde{\mathbf{a}}\|}\right)^+ \le \left(\frac{\|\widetilde{\mathbf{X}}_j\|}{\sqrt{\|\widetilde{\mathbf{X}}_j\|^2+\|\widetilde{\mathbf{a}}\|^2}}\,\frac{\widetilde{\mathbf{X}}_j'\,\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{X}}_j\|\,\|\widetilde{\mathbf{a}}\|}\right)^+,$$

$$\left(\frac{\|\widetilde{\mathbf{X}}_j\|}{\|\widetilde{\mathbf{X}}_j+\widetilde{\mathbf{a}}\|}\,\frac{\widetilde{\mathbf{X}}_j'\,\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{X}}_j\|\,\|\widetilde{\mathbf{a}}\|}\right)^- \ge \left(\frac{\|\widetilde{\mathbf{X}}_j\|}{\sqrt{\|\widetilde{\mathbf{X}}_j\|^2+\|\widetilde{\mathbf{a}}\|^2}}\,\frac{\widetilde{\mathbf{X}}_j'\,\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{X}}_j\|\,\|\widetilde{\mathbf{a}}\|}\right)^-.$$

These inequalities are identities when $\widetilde{\mathbf{X}}_j'\,\widetilde{\mathbf{a}}$ is of opposite sign to $(\,\cdot\,)^\pm$, or when $\widetilde{\mathbf{X}}_j'\,\widetilde{\mathbf{a}} = 0$. Because of the permutation symmetries of $\widetilde{\mathbf{X}}_j/\|\widetilde{\mathbf{X}}_j\|$, it follows that $\mathrm{P}(\,\widetilde{\mathbf{X}}_j'\,\widetilde{\mathbf{a}} \neq 0\,) > 0$, which implies

$$\mathrm{P}\left(\left(\frac{\|\widetilde{\mathbf{X}}_j\|}{\|\widetilde{\mathbf{X}}_j+\widetilde{\mathbf{a}}\|}\,\frac{\widetilde{\mathbf{X}}_j'\,\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{X}}_j\|\,\|\widetilde{\mathbf{a}}\|}\right)^+ < \left(\frac{\|\widetilde{\mathbf{X}}_j\|}{\sqrt{\|\widetilde{\mathbf{X}}_j\|^2+\|\widetilde{\mathbf{a}}\|^2}}\,\frac{\widetilde{\mathbf{X}}_j'\,\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{X}}_j\|\,\|\widetilde{\mathbf{a}}\|}\right)^+\right) > 0,$$

$$\mathrm{P}\left(\left(\frac{\|\widetilde{\mathbf{X}}_j\|}{\|\widetilde{\mathbf{X}}_j+\widetilde{\mathbf{a}}\|}\,\frac{\widetilde{\mathbf{X}}_j'\,\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{X}}_j\|\,\|\widetilde{\mathbf{a}}\|}\right)^- > \left(\frac{\|\widetilde{\mathbf{X}}_j\|}{\sqrt{\|\widetilde{\mathbf{X}}_j\|^2+\|\widetilde{\mathbf{a}}\|^2}}\,\frac{\widetilde{\mathbf{X}}_j'\,\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{X}}_j\|\,\|\widetilde{\mathbf{a}}\|}\right)^-\right) > 0,$$

and hence

$$\mathrm{E}\left(\left(\frac{\|\widetilde{\mathbf{X}}_j\|}{\|\widetilde{\mathbf{X}}_j+\widetilde{\mathbf{a}}\|}\,\frac{\widetilde{\mathbf{X}}_j'\,\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{X}}_j\|\,\|\widetilde{\mathbf{a}}\|}\right)^+\right) < \mathrm{E}\left(\left(\frac{\|\widetilde{\mathbf{X}}_j\|}{\sqrt{\|\widetilde{\mathbf{X}}_j\|^2+\|\widetilde{\mathbf{a}}\|^2}}\,\frac{\widetilde{\mathbf{X}}_j'\,\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{X}}_j\|\,\|\widetilde{\mathbf{a}}\|}\right)^+\right),$$

$$\mathrm{E}\left(\left(\frac{\|\widetilde{\mathbf{X}}_j\|}{\|\widetilde{\mathbf{X}}_j+\widetilde{\mathbf{a}}\|}\,\frac{\widetilde{\mathbf{X}}_j'\,\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{X}}_j\|\,\|\widetilde{\mathbf{a}}\|}\right)^-\right) > \mathrm{E}\left(\left(\frac{\|\widetilde{\mathbf{X}}_j\|}{\sqrt{\|\widetilde{\mathbf{X}}_j\|^2+\|\widetilde{\mathbf{a}}\|^2}}\,\frac{\widetilde{\mathbf{X}}_j'\,\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{X}}_j\|\,\|\widetilde{\mathbf{a}}\|}\right)^-\right).$$

For any permutation matrix $P \in \mathcal{P}_n$,

$$\frac{\|\widetilde{\mathbf{X}}_j\|}{\sqrt{\|\widetilde{\mathbf{X}}_j\|^2+\|\widetilde{\mathbf{a}}\|^2}} = \frac{\|P\,\widetilde{\mathbf{X}}_j\|}{\sqrt{\|P\,\widetilde{\mathbf{X}}_j\|^2+\|\widetilde{\mathbf{a}}\|^2}} \quad \text{surely,}$$

$$\frac{\widetilde{\mathbf{X}}_j}{\|\widetilde{\mathbf{X}}_j\|} = P\,\frac{\widetilde{\mathbf{X}}_j}{\|\widetilde{\mathbf{X}}_j\|} \quad \text{in distribution,}$$

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

so that

$$\frac{\|\widetilde{\mathbf{X}}_j\|}{\sqrt{\|\widetilde{\mathbf{X}}_j\|^2+\|\widetilde{\mathbf{a}}\|^2}}\,\frac{\widetilde{\mathbf{X}}_j}{\|\widetilde{\mathbf{X}}_j\|} \;=\; P\,\frac{\|\widetilde{\mathbf{X}}_j\|}{\sqrt{\|\widetilde{\mathbf{X}}_j\|^2+\|\widetilde{\mathbf{a}}\|^2}}\,\frac{\widetilde{\mathbf{X}}_j}{\|\widetilde{\mathbf{X}}_j\|} \quad \text{in distribution,}$$

which together with

$$\left(\frac{\|\widetilde{\mathbf{X}}_j\|}{\sqrt{\|\widetilde{\mathbf{X}}_j\|^2+\|\widetilde{\mathbf{a}}\|^2}}\,\frac{\widetilde{\mathbf{X}}_j}{\|\widetilde{\mathbf{X}}_j\|}\right)'\widehat{\mathbf{1}} \;=\; 0 \quad \text{surely,}$$

implies, as in (57), that

$$\mathrm{E}\left(\frac{\|\widetilde{\mathbf{X}}_j\|}{\sqrt{\|\widetilde{\mathbf{X}}_j\|^2+\|\widetilde{\mathbf{a}}\|^2}}\,\frac{\widetilde{\mathbf{X}}_j}{\|\widetilde{\mathbf{X}}_j\|}\right) \;=\; \mathbf{0}.$$

Therefore,

$$\mathrm{E}\left(\left(\frac{\|\widetilde{\mathbf{X}}_j\|}{\sqrt{\|\widetilde{\mathbf{X}}_j\|^2+\|\widetilde{\mathbf{a}}\|^2}}\,\frac{\widetilde{\mathbf{X}}_j'\,\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{X}}_j\|\,\|\widetilde{\mathbf{a}}\|}\right)^{+}\right) \;=\; \mathrm{E}\left(\left(\frac{\|\widetilde{\mathbf{X}}_j\|}{\sqrt{\|\widetilde{\mathbf{X}}_j\|^2+\|\widetilde{\mathbf{a}}\|^2}}\,\frac{\widetilde{\mathbf{X}}_j'\,\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{X}}_j\|\,\|\widetilde{\mathbf{a}}\|}\right)^{-}\right).$$

Back to the initial expected values, it follows that

$$\mathrm{E}\left(\left(\frac{\|\widetilde{\mathbf{X}}_j\|}{\|\widetilde{\mathbf{X}}_j+\widetilde{\mathbf{a}}\|}\,\frac{\widetilde{\mathbf{X}}_j'\,\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{X}}_j\|\,\|\widetilde{\mathbf{a}}\|}\right)^{+}\right) \;<\; \mathrm{E}\left(\left(\frac{\|\widetilde{\mathbf{X}}_j\|}{\|\widetilde{\mathbf{X}}_j+\widetilde{\mathbf{a}}\|}\,\frac{\widetilde{\mathbf{X}}_j'\,\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{X}}_j\|\,\|\widetilde{\mathbf{a}}\|}\right)^{-}\right),$$

which implies

$$\mathrm{E}\left(\frac{\|\widetilde{\mathbf{X}}_j\|}{\|\widetilde{\mathbf{X}}_j+\widetilde{\mathbf{a}}\|}\,\frac{\widetilde{\mathbf{X}}_j'\,\widetilde{\mathbf{a}}}{\|\widetilde{\mathbf{X}}_j\|\,\|\widetilde{\mathbf{a}}\|}\right) \;<\; 0. \quad \square$$

The Gaussian multivariate distribution, among others, has spherical symmetry besides permutation symmetry. For parametric families with spherical symmetry, the true standard vector $\sqrt{n-1}\,\widetilde{\mathbf{X}}_j/\|\widetilde{\mathbf{X}}_j\|$ has uniform distribution over the $(n-2)$-sphere. As a result, the components of $\widetilde{\mathbf{Y}}_j/\|\widetilde{\mathbf{Y}}_j\|$ perpendicular to $\widetilde{\mathbf{a}}$ are antisymmetric with respect to the direction of $\widetilde{\mathbf{a}}$, so that they cancel out in expectation. That is, for parametric families with spherical symmetry, and as long as $\widetilde{\mathbf{a}} \neq \mathbf{0}$,

$$\mathrm{E}\left(\frac{\widetilde{\mathbf{Y}}_j}{\|\widetilde{\mathbf{Y}}_j\|}\right) \;=\; \lambda\widetilde{\mathbf{a}}, \quad \text{with} \quad 0<\lambda<\mathrm{E}\left(\frac{1}{\|\widetilde{\mathbf{Y}}_j\|}\right). \tag{60}$$

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

SM4 Standard-vector normalization

The properties (59), (60) suggest the use of

$$\widehat{\mathbf{b}} = \frac{\displaystyle\sum_{j=1}^{m} \frac{\widetilde{\mathbf{y}}_j}{\|\widetilde{\mathbf{y}}_j\|}}{\displaystyle\sum_{j=1}^{m} \frac{1}{\|\widetilde{\mathbf{y}}_j\|}} \tag{61}$$

to approximate the unknown residual vector of normalization factors $\widetilde{\mathbf{a}}$. The following iterative method implements this approach to solve the *normalization of residuals problem*.

Let us define the following recursive sequence, where each step t comprises m vectors $\widehat{\mathbf{y}}_j^{(t)}$ ($j \in \{1, \ldots, m\}$) and one vector $\widehat{\mathbf{b}}^{(t)}$,

$$\widehat{\mathbf{y}}_j^{(0)} = \widetilde{\mathbf{y}}_j, \tag{62}$$

$$\widehat{\mathbf{y}}_j^{(t)} = \widehat{\mathbf{y}}_j^{(t-1)} - \widehat{\mathbf{b}}^{(t-1)}, \quad \text{for } t \geq 1, \tag{63}$$

$$\widehat{\mathbf{b}}^{(t)} = \frac{\displaystyle\sum_{i=1}^{m} \frac{\widehat{\mathbf{y}}_j^{(t)}}{\|\widehat{\mathbf{y}}_j^{(t)}\|}}{\displaystyle\sum_{i=1}^{m} \frac{1}{\|\widehat{\mathbf{y}}_j^{(t)}\|}}, \quad \text{for } t \geq 0. \tag{64}$$

We assume that $\widehat{\mathbf{y}}_j^{(t)} \neq \mathbf{0}_n$, for all $j \in \{1, \ldots, m\}$ and all $t \geq 0$. Nonetheless, an implementation of this algorithm benefits from trimming out a small fraction (e.g. 1%) of the features with lesser $\|\widehat{\mathbf{y}}_j^{(t)}\|$ in (64), in order to avoid numerical singularities.

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Let us write $\widehat{\mathbf{y}}_j^{(t)}$ as a function of the unknowns $\widetilde{\mathbf{x}}_j$ and $\widetilde{\mathbf{a}}$. For any $t \geq 1$,

$$\begin{aligned}
\widehat{\mathbf{y}}_j^{(t)} &= \widehat{\mathbf{y}}_j^{(t-1)} - \widehat{\mathbf{b}}^{(t-1)}, && (65)\\
&= \widehat{\mathbf{y}}_j^{(t-2)} - \widehat{\mathbf{b}}^{(t-2)} - \widehat{\mathbf{b}}^{(t-1)}, && (66)\\
&\vdots && (67)\\
&= \widehat{\mathbf{y}}_j^{(0)} - \sum_{r=0}^{t-1} \widehat{\mathbf{b}}^{(r)}, && (68)\\
&= \widetilde{\mathbf{y}}_j - \sum_{r=0}^{t-1} \widehat{\mathbf{b}}^{(r)}, && (69)\\
&= \widetilde{\mathbf{x}}_j + \widetilde{\mathbf{a}} - \sum_{r=0}^{t-1} \widehat{\mathbf{b}}^{(r)}. && (70)
\end{aligned}$$

Note that (70) is also valid for $t = 0$.

Let us also define the vectors $\widehat{\mathbf{a}}^{(t)}$, for $t \geq 0$, which describe the vector of normalization factors still to be removed at step t,

$$\widehat{\mathbf{a}}^{(t)} = \widetilde{\mathbf{a}} - \sum_{r=0}^{t-1} \widehat{\mathbf{b}}^{(r)}, \tag{71}$$

so that, by (70), for $t \geq 0$,

$$\widehat{\mathbf{y}}_j^{(t)} = \widetilde{\mathbf{x}}_j + \widehat{\mathbf{a}}^{(t)}. \tag{72}$$

Therefore, the recursive sequence (62)–(64) faces a new, weaker *normalization of residuals problem* at each step t, with true residual vectors $\widetilde{\mathbf{x}}_j$, observed residual vectors $\widehat{\mathbf{y}}_j^{(t)}$ and unknown normalization factors $\widehat{\mathbf{a}}^{(t)}$. The step t results in the estimation of normalization factors $\widehat{\mathbf{b}}^{(t)}$, which are removed from $\widehat{\mathbf{y}}_j^{(t)}$, generating the next step. At the beginning, $\widehat{\mathbf{y}}_j^{(0)} = \widetilde{\mathbf{y}}_j$ and $\widehat{\mathbf{a}}^{(0)} = \widetilde{\mathbf{a}}$.

At convergence, $\lim_{t\to\infty} \widehat{\mathbf{b}}^{(t)} = \mathbf{0}$. Equations (57), (58), (64) imply that, in such a case, $\lim_{t\to\infty} \widehat{\mathbf{y}}_j^{(t)} = \widetilde{\mathbf{x}}_j$ and $\sum_{t=0}^{\infty} \widehat{\mathbf{b}}^{(t)} = \widetilde{\mathbf{a}}$. Convergence is optimal when the parametric family of the m features has spherical symmetry, Gaussian being the most prominent case. Otherwise, the more uniform the distribution of standard vectors $\sqrt{n-1}\,\widetilde{\mathbf{x}}_j/\|\widetilde{\mathbf{x}}_j\|$ is on the $(n-2)$-sphere, the faster the sequence (62)–(64) converges. See examples of convergence in Movies S1–S3.

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

SM5 Identification of non-differentially expressed genes

Let us consider a gene expression dataset, with g genes and c experimental conditions. Each condition k has s_k samples. The total number of samples is $s = \sum_{k=1}^{c} s_k$. Let us assume that $c \geq 2$ and that $s_k \geq 2$, for all conditions $k \in \{1, \ldots, c\}$. Let us also assume that, among the g genes, there is a fraction π_0 of non-differentially expressed (non-DE) genes, with $0 \leq \pi_0 \leq 1$, while the remaining fraction $1 - \pi_0$ comprises the differentially expressed (DE) genes [45].

Let us consider the usual ANOVA test comparing average expression levels across conditions, gene-by-gene. Under the null hypothesis of a non-differentially expressed gene, the corresponding F-statistic follows the F-distribution with $c - 1$ and $s - c$ degrees of freedom. The test of this hypothesis yields a p-value p_j for each gene $j \in \{1, \ldots, g\}$. The obtained p-values p_j follow a probability distribution that can be considered as the mixture of two probability distributions, F_0 and F_1, for the non-DE genes and the DE genes, respectively [46]. The fraction π_0 of non-DE genes follows the uniform distribution on the interval $[0, 1]$,

$$F_0(p) \;=\; p, \tag{73}$$

while the fraction $1 - \pi_0$ of DE genes follows a distribution that verifies, for any $p \in (0, 1)$,

$$F_1(p) \;>\; p, \tag{74}$$

and the mixture distribution is

$$F(p) \;=\; \pi_0 F_0(p) + (1 - \pi_0) F_1(p). \tag{75}$$

Let us further assume that there exists a p^*, with $0 < p^* < 1$, such that $F_1(p) = 1$ for every $p \geq p^*$. In other words, all DE genes have p-value p_j from the ANOVA test such that $p_j \leq p^*$, while only some genes among the non-DE genes have p-value with $p_j > p^*$. This implies that the mixture distribution of p-values is uniform on the interval $[p^*, 1]$,

$$F(p) \;=\; \pi_0 p + 1 - \pi_0, \quad \text{for } p^* \leq p \leq 1, \tag{76}$$

$$f(p) \;=\; \pi_0, \quad \text{for } p^* < p < 1. \tag{77}$$

bioRxiv preprint doi: https://doi.org/10.1101/021212; this version posted July 22, 2015. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

On the other hand, for any set of n samples $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$ obtained from n independent and identically distributed uniform random variables on the interval $[a, b]$, all the distances between consecutive ordered samples (including boundaries), $x_{(1)} - a, x_{(2)} - x_{(1)}, \ldots, x_{(n)} - x_{(n-1)}, b - x_{(n)}$, obey the same distribution [47]. Then, it can be realized that, for any j such that $2 \leq j \leq n-1$, the two subsets of samples $x_{(1)}, \ldots, x_{(j-1)}$ and $x_{(j+1)}, \ldots, x_{(n)}$ follow uniform distributions on the intervals $[a, x_{(j)}]$ and $[x_{(j)}, b]$, respectively.

Based on these facts, to identify non-DE genes we propose finding the minimum $p_{(j)}$, from the ordered sequence of p-values $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(g)}$, such that a goodness-of-fit test for the uniform distribution on the interval $[p_{(j)}, 1]$, performed on $p_{(j+1)}, \ldots, p_{(g)}$, is not rejected. As a result, the genes corresponding to the p-values $p_{(j)}, p_{(j+1)}, \ldots, p_{(g)}$ are considered as non-DE genes.

Given the concavity of $F(p)$, the goodness-of-fit test used is the one-sided Kolmogorov-Smirnov test on positive deviations of the empirical distribution function.

See Movies S4–S5 for examples of this approach to identifying non-differentially expressed genes.