UNIVERSITY OF CALIFORNIA SAN DIEGO

Independent Component Analysis in Alternate Dimensions – Parameter-free dimensionality selection for ICA of transcriptomic data

A Thesis submitted in partial satisfaction of the
requirements for the degree
Master of Science

in

Bioengineering

by

John Luke McConn

Committee in Charge:

Professor Bernhard Ø. Palsson, Chair
Professor Pedro Cabrales
Professor Gert Cauwenberghs

2020

The Thesis of John Luke McConn is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California San Diego

2020

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to thank Professor Bernhard Ø. Palsson for providing the opportunity and wonderfully collaborative lab environment to carry out this work.

I would also like to thank the members of the Systems Biology Research Group for their collaboration and guidance especially Dr. Anand Sastry, Cam Lamoureux, Katherine Decker and Dr. Daniel Zielinski.

ABSTRACT OF THE THESIS

Independent Component Analysis in Alternate Dimensions – Parameter-free dimensionality selection for ICA of transcriptomic data

by

John Luke McConn

Master of Science in Bioengineering

University of California San Diego, 2020

Professor Bernhard Ø. Palsson, Chair

Independent Component Analysis (ICA) is an unsupervised machine learning algorithm which models a complex multivariate dataset as a linear combination of statistically independent hidden factors. Applied to high-quality gene expression data, it effectively reveals these hidden factors of the transcriptional regulatory network as sets of co-regulated genes and their corresponding activities across diverse growth

conditions. The two main variables affecting the output of ICA are the data itself and the user-defined number of components to compute. High quality transcriptomic data has become more accessible as high-throughput technologies have advanced while dimensionality selection remains open to research interest. Several methods for optimally selecting dimensionality have been proposed previously and two are tested herein, but were found to provide inconsistent results, ostensibly under-decomposing a dataset in some cases while over-decomposing in others. A new method for effectively setting dimensionality is proposed in this study which aims to consistently maximize the number of biologically relevant components revealed while minimizing the potential for over-decomposition.

INTRODUCTION

Independent Component Analysis (ICA) is an unsupervised machine learning algorithm which models a multivariate dataset as a linear combination of statistically independent hidden factors or components. Modeling data in this way is useful when the goal is signal deconvolution or separation of mixed signals into their individual sources and corresponding strengths. An illustrative example of ICA's utility is the widely referenced cocktail party problem, in which multiple mixed audio signals (i.e. people speaking simultaneously at a cocktail party) are recorded by microphones dispersed throughout a room. Each device records a unique linear mixture of the original signals depending on its proximity to each speaker. Applying ICA to this set of mixed recordings could effectively recover the independent audio signals which mix to produce the different recordings.

Beyond deconvoluting audio signals, ICA is widely applicable to several other fields involving signal separation or feature extraction and with the advancement of high-throughput gene expression profiling, ICA has proven to be useful in analyzing highly multivariate microarray gene expression and RNA sequencing (RNA-seq) datasets [1][2]. The coalescence of expression profiles across diverse conditions into repositories, such as the Gene Expression Omnibus (GEO) with over 400,000 microarrays, has provided a rich source of data for these efforts [3]. Additionally, high-quality RNA-seq profiles from *E. coli* K-12 MG1655 and BW25113 across hundreds of diverse growth conditions have been compiled into PRECISE, Precision RNA-seq Expression Compendium for Independent Signal Exploration, for the explicit purpose of probing the transcriptional regulatory network through the application of ICA [1].

The diverse gene expression profiles contained in these compendia can be thought of as the mixed recordings and each profile's associated growth condition is equivalent to the placement of a microphone. Application of ICA then reveals groups of co-regulated genes as the independent signals and their associated strengths across each growth condition. Prior ICA studies of human microarray data have often identified groups of co-regulated genes which map readily to known metabolic pathways [3]. The activities of these gene sets can then be evaluated to classify tumor samples or disease states, for example [4]. ICA decomposition of PRECISE revealed 92 independently modulated gene sets, termed iModulons, many of which contained significant overlap with known regulons and could be directly linked to a single transcription factor [1]. Similar analysis has since been carried out on a *B. subtilis* microarray dataset with 269 expression profiles and has revealed 83 similarly informative iModulons [5].

As compendia of such gene expression data incorporate more individual growth conditions it becomes possible to develop a comprehensive model of transcriptional regulation. To achieve this, the output of ICA decompositions depends on two primary inputs – high quality data sources across diverse growth conditions such as those described and the number of independent components to compute. While the former has become more accessible, the latter remains an area of open research interest. Unlike other forms of matrix decomposition, such as PCA, which computes a component for each data dimension, ICA decomposes a dataset into a user-specified number of dimensions. In the context of the cocktail party problem, this can be a trivial specification as the sources are potentially visible and countable. In the context of gene

expression data, where source signals are not visible, how many dimensions should be computed?

Several methods have been suggested and employed to answer this question. One such method entails setting the ICA dimensionality equal to the number of principal components, determined through principal component analysis (PCA), which account for a certain level of variance in the data. Other means devised specifically for transcriptomic data, such as selecting the Maximally Stable Transcriptome Dimension (MSTD), defined as the maximum dimension before ICA begins to produce a large proportion of unstable components, have also been suggested [6].

In this study, these methods are evaluated and the effects of dimensionality variation on ICA decomposition of transcriptomic data are clarified. In several cases, previously proposed dimensionality selection methods were found to be inconsistent, resulting in over-decomposition in one case while under-decomposing a different dataset. Based on this result a new method of ICA dimensionality selection is proposed for transcriptomic data, which aims to maximize the biologically relevant components while minimizing the biologically meaningless single gene components discovered in the resulting ICA decomposition.

METHODS

2.1. Conducting independent component analysis on gene expression data

Independent component analysis decomposes a matrix of gene expression data (X, m genes by n growth conditions) into a matrix of independent components (S, m genes by y components) and each components activity across conditions (A, y components by n growth conditions).

$$\boldsymbol{X} = \boldsymbol{S} * \boldsymbol{A}$$

For each data set, ICA was performed as detailed by Sastry et al. across a range of dimensionalities 100 times with random seeds and with convergence tolerance set to 10^{-5} [1]. For PRECISE 1.0, which contains 278 expression profiles, ICA was run from 2 dimensions up to 276 dimensions. For PRECISE 2.0, it was done from 5 dimensions up to 815, at intervals of 5 dimensions and for the *B. subtilis* data set, from 5 to 265, at intervals of 5 dimensions.

2.2 Building dimensionality trees and comparison of final, conserved components to the components of preceding dimensions

The cosine distance between components of each subset and those at the subsequent dimension was computed. Where this value was greater than 0.3 a connection was established between those components to build the dimensionality tree. Components from the highest dimension from each subset were similarly correlated to components of each preceding dimension. The highest of these values was used to associate a final component with each preceding component to build the heat maps of final component occurrence in each subset. Where the cosine distance was greater

than an established threshold a final component was said to exist in a preceding decomposition.

2.3 Establishing thresholds for final components

The components in the highest dimension decomposition were compared pairwise to all components computed at lower dimensions. Cosine distance was calculated for each pair and histograms of these values were plotted, resulting in a bimodal distribution of highly correlated and uncorrelated components. The elbow point of the higher mode was used to establish a threshold correlation to classify a component as a final component.

2.4 Classification of components as robust, regulatory, single gene or non-single gene

The total number of components computed from a particular ICA decomposition were classified as robust. A component was classified as single gene if the highest gene weight was more than twice the next highest; the number of non-single gene components was determined by subtracting the number of single gene components from the number of robust components. The two-sided Fisher's exact test (FDR < 10^{-5}) was used to compare significant genes in each component to regulon gene sets to classify components as regulatory.

2.5 Determining the PC-VA, MSTD and novel method dimensions

Principal component analysis was conducted on each expression matrix, the principal components were ordered by their associated percentage of explained variance, the point at which cumulative explained variance equaled 99% determined the PC-VA dimensionality. The MSTD, or dimension at which ICA begins to compute a high

proportion of unstable components, was determined as previously described [6]. The new method dimension was defined as the point where the number of non-single gene components was equal to the number of final components in that decomposition.

RESULTS

3.1. The S-matrix structure varies across dimensionalities while individual components are conserved

Several RNA-seq and microarray datasets were utilized for this analysis including the original version of PRECISE (PRECISE 1.0), an expanded version (PRECISE 2.0), and the *B. subtilis* microarray dataset referenced above [7]. These datasets were decomposed by independent component analysis from low dimensionality up to essentially fully decomposed (one dimension for each expression profile) to produce dimensionality trees such as the one shown in Figure 1, which convey the evolution of the S-matrix structure across dimensionalities.

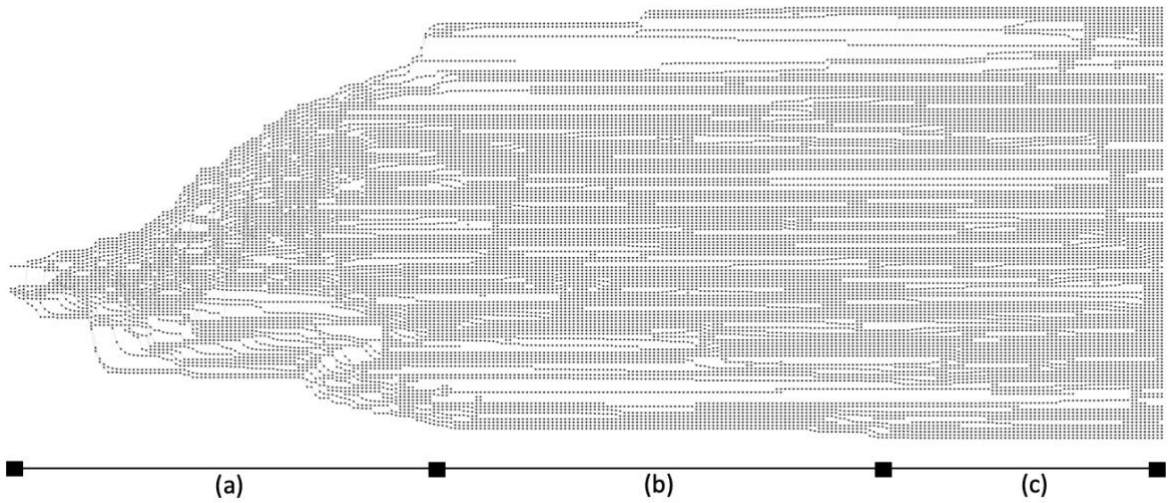

Figure 1. Dimensionality tree of PRECISE 1.0 reveals how the S-matrix evolves as the decomposition is carried out at higher dimensions. (a) Initially, the number of robust components increases roughly linearly until (b) an essentially stable structure is reached. (c) Beyond this stable region over-decomposition appears to occur as additional components are revealed at higher dimensions. Connections between components of subsequent dimensionality subsets where established where the cosine distance was at least 0.3.

In the case of PRECISE 1.0, at low dimensions, as components merge, split and appear there is a net increase in total robust components until a stable decomposition structure is reached. The appearance of additional components beyond this stable region

suggests the commencement of over-decomposition. Connections between components of adjacent subsets were established where the cosine distance between components was greater than 0.3.

These dimensionality trees demonstrate how the overall structure of the ICA decomposition evolves as more components are computed. To gain a clearer understanding of how the individual components themselves evolve across dimensions all components in each decomposition were compared to those in the final, fully decomposed S-matrix.

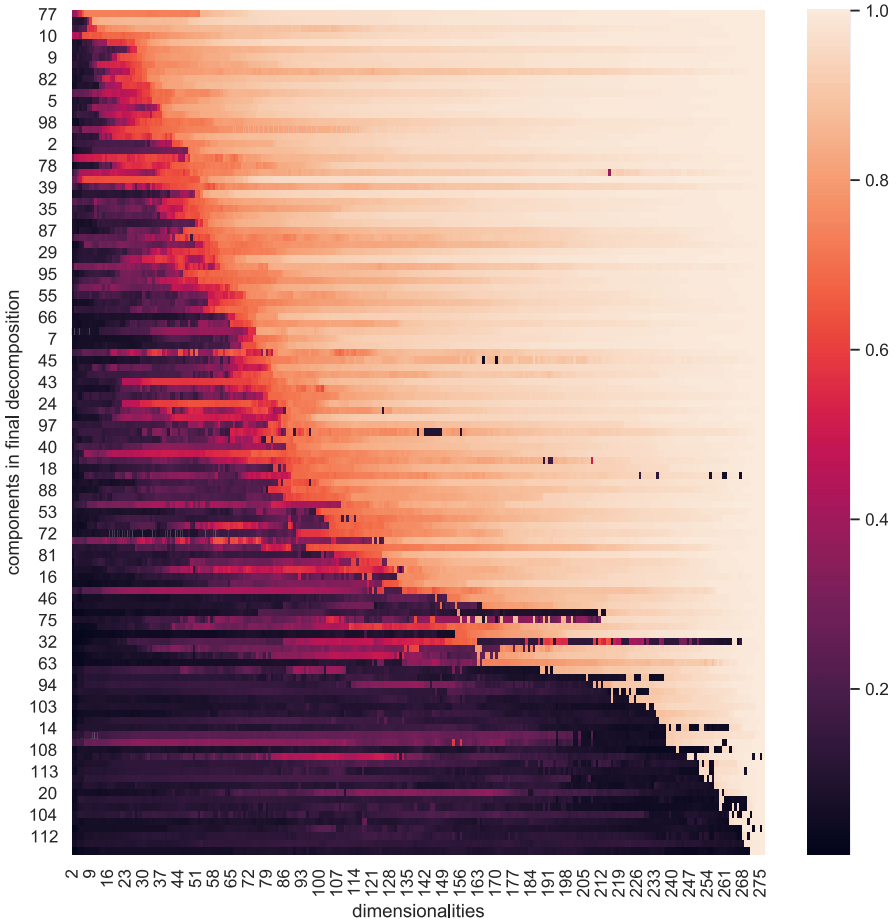

Figure 2. Cosine distance between the components in the final decomposition of PRECISE 1.0 and the components discovered in each preceding decomposition. Once a component is discovered, it is well conserved as it continues to appear in the majority of decompositions at higher dimensionality and does not materially change beyond that threshold.

With few exceptions, once a component was discovered at a particular dimension it proved to be conserved in decompositions at higher dimensions. In the case of PRECISE 1.0, this realization suggests that across the stable decomposition region, both the overall S-matrix structure and the individual components do not materially change.

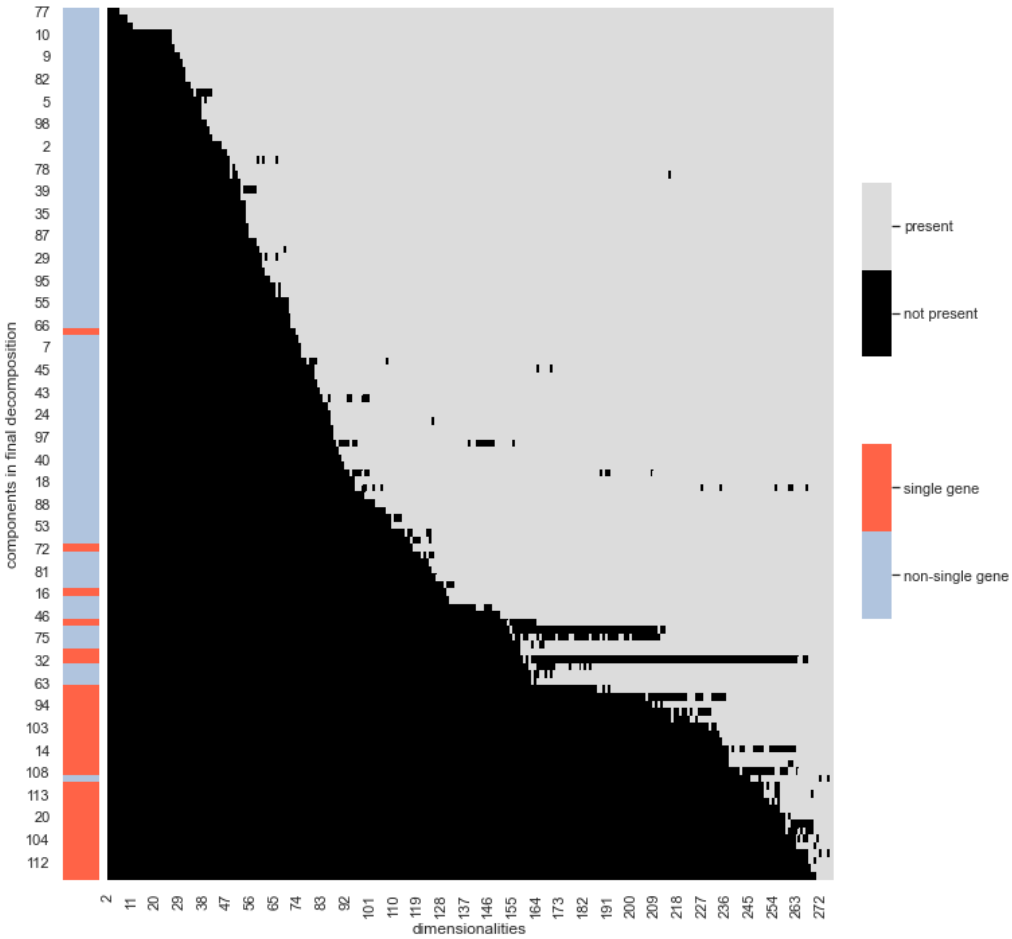

Figure 3. A binarized heatmap revealing the presence of final components in each preceding decomposition and ordered by their first appearance in a particular decomposition. Components in the final decomposition were said to be present in a preceding decomposition if the cosine distance between the two was greater than the established threshold. Components first revealed at lower dimensionality were well conserved while those at higher are primarily single gene suggesting the commencement of over-decomposition.

A component present in the final decomposition was said to be present in earlier decompositions if the cosine distance between the two vectors was greater than the

established threshold. In the majority of cases, once a component was discovered in an early subset it remained stable, rarely dropping below even a cosine distance of 0.9 when compared to the synonymous component in the final decomposition. In addition, the majority of components found at higher dimensionalities contain only a single gene, suggesting the commencement of over-decomposition rather than any biological relevance.

3.2. Dimension selection methods often result in over- or under-decomposition

Two commonly utilized methods of setting ICA dimensionality were evaluated on PRECISE 1.0 initially. The first involves setting the ICA dimensionality based on the number of principal components which explain a certain level of variance in the data (referred to as PC-VA herein); the other, determining the point at which ICA begins to reveal a high proportion of unstable components or the Maximally Stable Transcriptome Dimension (MSTD). The initial ICA studies of PRECISE and the B. subtilis datasets were carried out using the PC-VA method.

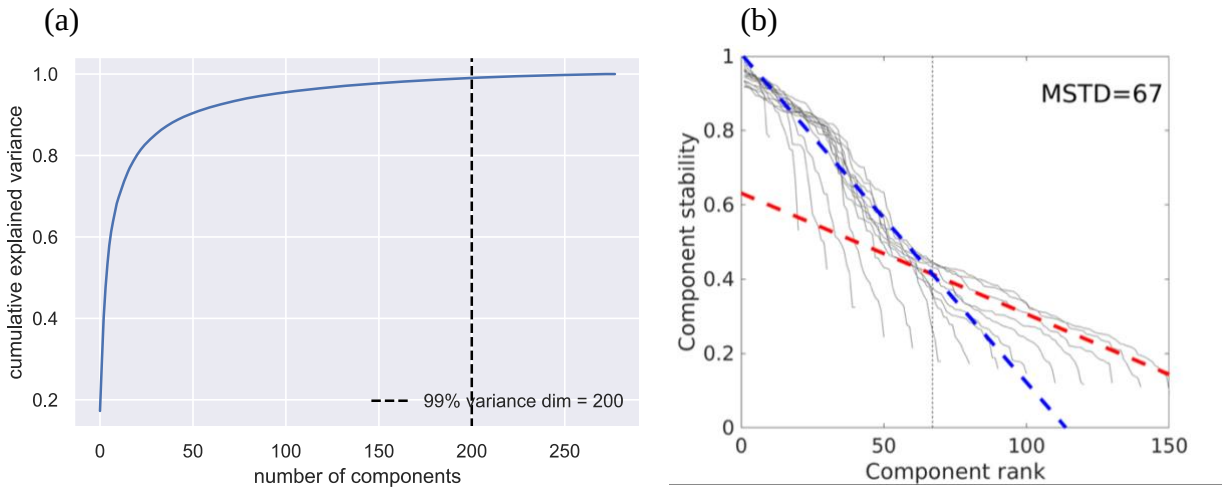

Figure 4. The resulting dimension based on two methods of ICA dimensionality selection. (a) Setting the ICA dimensionality based on the number of principal components which explain 99% of the variance in the data (PC-VA). (b) The Maximally Stable Transcriptome Dimension (MSTD). Notably, the two methods select for drastically different dimensions.

Components were classified across all dimensions to visualize where these points occur in the evolution of the S-matrix structure, shown in Figure 5. Robust components are the total number of components present in a particular decomposition, final components are the ones which also occur in the final decomposition, regulatory components are those that contain groups of genes which are known to be regulated by a common transcription factor, single gene components are those whose highest gene weight is more than twice the next highest, and non-single gene components are those that are robust but not single gene.

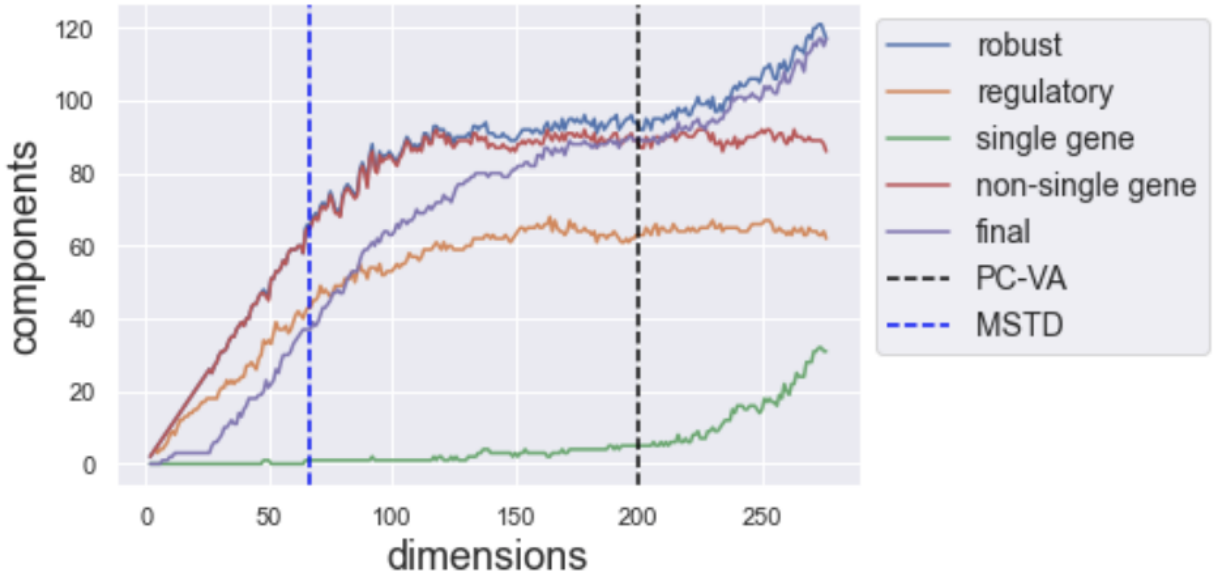

Figure 5. Classification of PRECISE 1.0 components across all dimensions and the points selected by the PC-VA and MSTD methods. MSTD appears to pick an under-decomposed point; while, 99PCA appears to select an appropriate decomposition in the stable S-matrix structure region before over-decomposition begins.

In the case of PRECISE 1.0, these two methods resulted in the selection of drastically different optimal dimensionalities. The MSTD of 67 dimensions appears to be an under-decomposed point before the stable S-matrix structure is reached; while, the PC-VA method appears to select an appropriate dimension in the stable region before over-decomposition begins.

The PC-VA method was then evaluated using the *B. subtilis* and PRECISE 2.0 datasets to ensure efficacy across different datasets; however, this method proved to be unreliable in both cases. Components were classified and tallied across dimensionalities in the same manner as PRECISE 1.0. In the *B. subtilis* dataset, the PC-VA method appeared to select a point of under-decomposition. This finding is consistent with the fact that many of the initially characterized iModulons were noted to contain gene groups regulated by different transcription factors in the same component, suggesting under-decomposition [7].

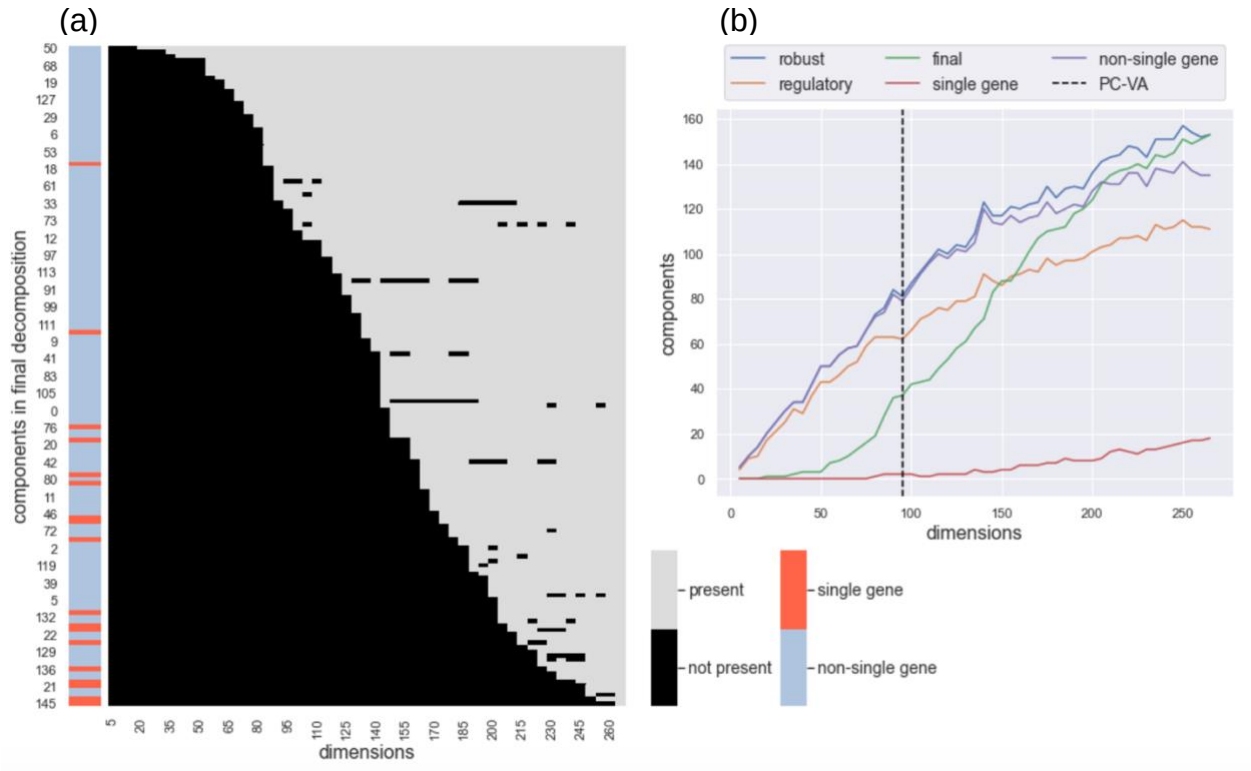

Figure 6. (a) Comparison of the final decomposition components of the *B. subtilis* dataset with those in all preceding decompositions. Again, components are largely conserved over the range of dimensionalities and the emergence of single gene components primarily occurs at the highest dimensions. (b) Component classification across dimensions and the resulting dimensionality selected by the PC-VA method.

Notably, the S-matrix structure of this dataset does not evolve in the same manner as PRECISE 1.0. There is a consistent increase in new components across dimensionalities and there is no point where the S-matrix appears to reach a stable structure before over-decomposition begins. Similar to PRECISE 1.0, comparison of the fully decomposed components with each of those in the preceding decompositions demonstrates that once a component is revealed at low dimensionality it remains fairly consistent at higher level decompositions. Analysis of PRECISE 2.0 revealed similar insights; however, the PC-VA method selected an over-decomposed dimension, marked by a high proportion of single gene (51) to total robust components (211) or around 24%. Components revealed at lower dimensionalities were again conserved at higher levels.

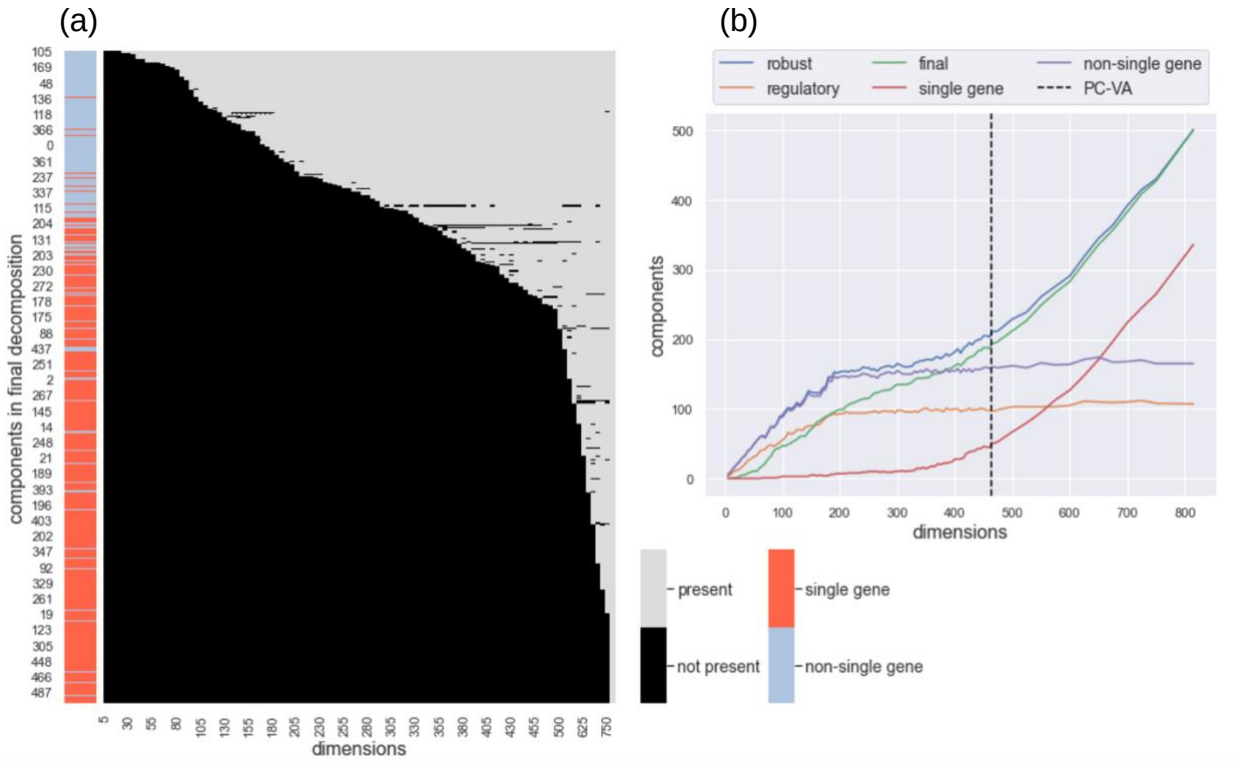

Figure 7. (a) Comparison of the final decomposition components of PRECISE 2.0 with those in all preceding decompositions. Again, components are largely conserved over the range of dimensionalities and the emergence of single gene components primarily occurs at higher dimensions. (b) Component classification across dimensions and the resulting dimensionality selected by the PC-VA method.

Across the studied transcriptomic datasets, the two tested methods for setting dimensionality resulted in inconsistencies. The MSTD method selected an under-decomposed point for PRECISE 1.0; while the PC-VA method appeared to select an appropriate dimension. This method was then applied to the B. subtilis and PRECISE 2.0 datasets to confirm consistent efficacy but was found to select an under-decomposed point in one case and an over-decomposed point in the other.

3.3. A newly proposed method for selecting ICA dimensionality

The method presented herein improves upon these inconsistencies by selecting a point of dimensionality where the number of final components in that decomposition is equal to the number of non-single gene components; thereby filtering out most biologically irrelevant single gene components and maximizing the number of conserved components across dimensions. In all cases, this method results in an improved decomposition or in the case of PRECISE 1.0 selects a similar dimensionality to the PC-VA method yielding a similarly high-quality decomposition.

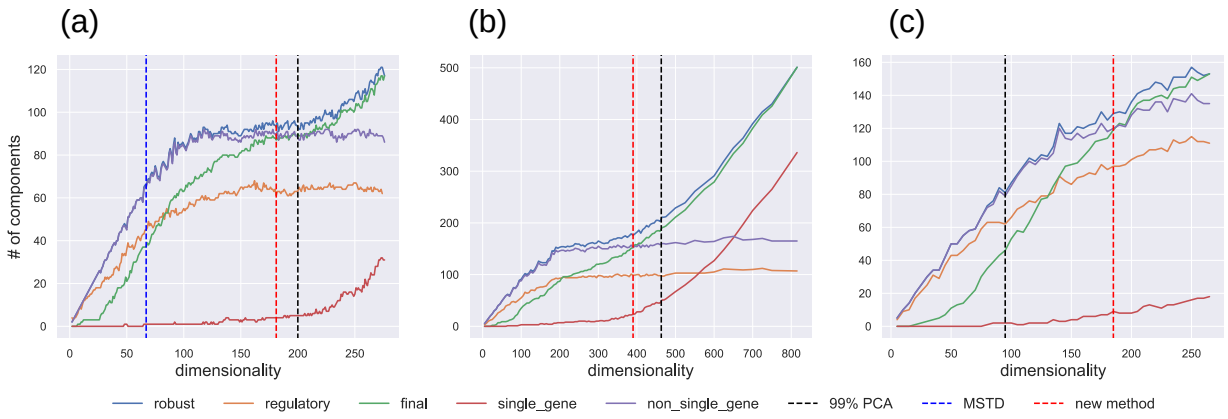

Figure 8. Component classifications across the corresponding dimension range for (a) PRECISE 1.0, (b) PRECISE 2.0 and (c) *B. subtilis*. The new method resulted in selecting a lower dimension for PRECISE 2.0 improving over-decomposition and a higher dimension for *B. subtilis* resolving under-decomposition. It selected a slightly lower but comparable point for PRECISE 1.0 yielding a similar S-matrix structure to that from PC-VA decomposition.

The PRECISE 1.0 components calculated from the PC-VA method were previously characterized and serve as a useful point of comparison for the MSTD and new method decompositions [1]. These components, termed iModulons, were named in most cases by the primary regulator that controls the highly weighted genes within each. The highest correlation between each iModulon and the components of the MSTD and new method decomposition are shown in Figure 9. The MSTD method resulted in selecting a dimension which does not reconstitute many of these well characterized iModulons; whereas, the new method results in a decomposition containing all but two iModulons, SgrT, a single gene component containing only the sgrT gene and FecI.

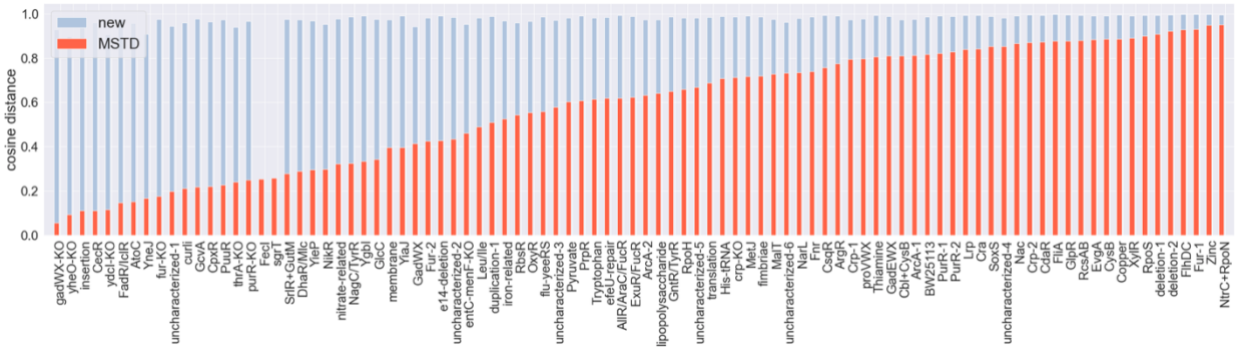

Figure 9. Comparing previously characterized iModulons, named for the transcription factors that regulate their gene enrichments in most cases, from PRECISE 1.0 reveals that the decomposition from the new method reconstituted the vast majority of those components with high fidelity. As expected the under-decomposed MSTD S-matrix did not reproduce many of the well characterized iModulons.

In the case of PRECISE 2.0, over-decomposition was improved as the new method selected a dimensionality which reduces the number of single gene components to 20 or 11% of total components (from 51 or 24% in the PC-VA method decomposition). At the lower dimension selected by this new method, there was no loss in information from the regulatory components as well. Top F1 scores, or the harmonic average of precision and recall between the component and its associated regulon were

slightly improved from 0.55 at the PC-VA dimension to 0.59 at the dimension selected by the new method.

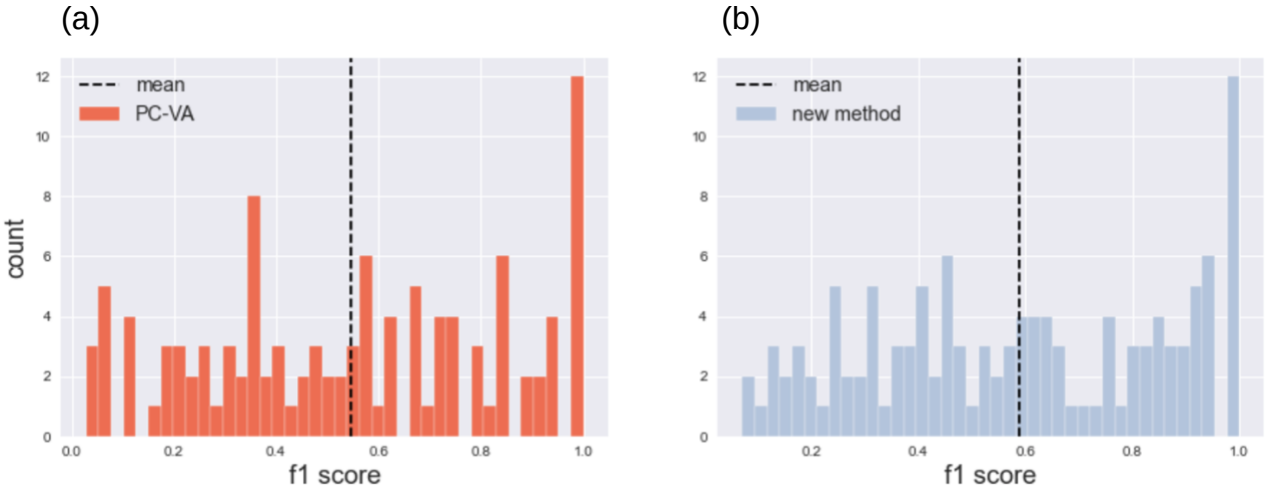

Figure 10. Top F1 scores for regulatory components of PRECISE 2.0 decompositions at the (a) PC-VA dimension and (b) the dimension selected by the new method. Average F1 scores were slightly improved from 0.55 to 0.59 at the lower dimension selected by the new method.

Additionally, the under-decomposition seen in the *B. subtilis* dataset was improved through the new dimensionality selection technique. iModulons for this dataset have been previously characterized as well and again, serve as a useful point of comparison [6].

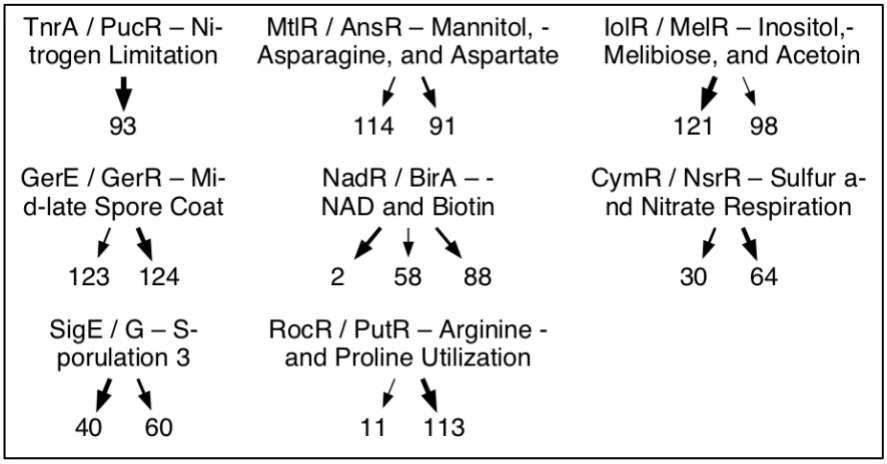

Figure 11. Several previously characterized *B. subtilis* iModulons contained gene groups that are regulated by different transcription factors, suggesting under-decomposition. The new method split most of these iModulons into components containing genes more consistent with a single regulatory mechanism resolving the under-decomposition.

Many of the originally characterized iModulons appeared to be merged, containing multiple gene sets that are known to be regulated by disparate transcription factors. In many cases, the decomposition resulting from the new method did, in fact, split these independently regulated gene sets into different components. For example, the originally characterized CymR/NsrR component was split into components 30 and 64 in the decomposition that resulted from the new method. Component 30 was enriched with genes regulated by NsrR and component 64 contained genes regulated by CymR. In some cases, these splits did not divide the component by regulator but the higher dimension decomposition did improve consistency between the component genes and associated regulon. The average F1 score of the merged iModulons increased from 0.40 using the PC-VA method to 0.55 using the new dimensionality selection method.

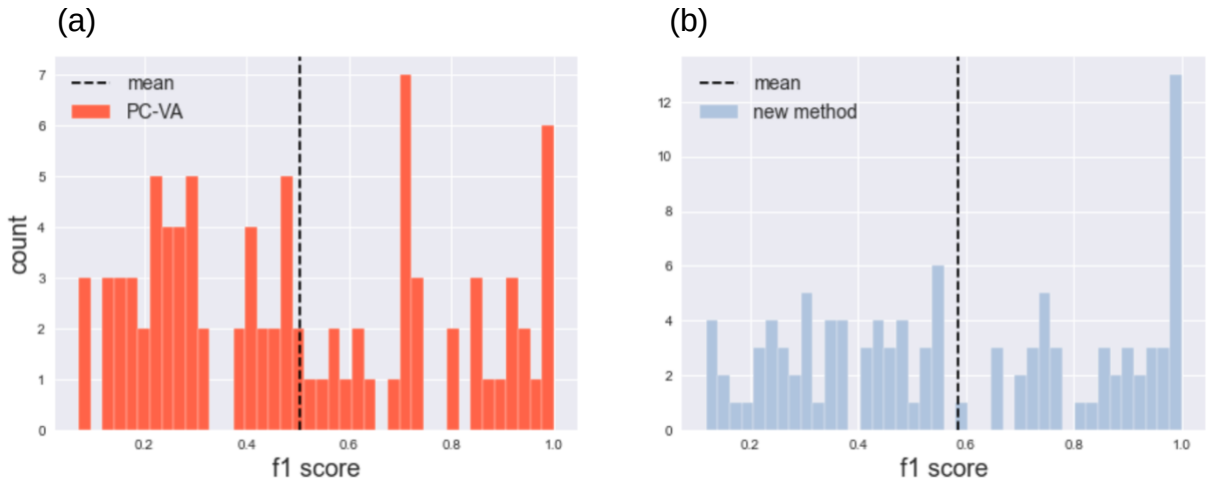

Figure 12. Top F1 scores, the harmonic mean of precision and recall between regulatory components and its associated regulon for (a) the iModulons resulting from PC-VA method of dimensionality selection and (b) the components resulting from the new dimensionality selection technique.

In addition, across all regulatory components the average F1 score increased from 0.50 to 0.58 (p-value=0.049). Again, suggesting increased consistency between

components and their associated regulon and improvement to the problem of under-decomposition.

DISCUSSION

Two important factors greatly influence the output of an ICA decomposition—the dataset of interest and the user-defined number of components to compute. Several methods have been suggested to optimally set this value in a parameter-free manner, including the MSTD and PC-VA methods previously described. These methods were tested on three transcriptomic datasets and, in some cases, were found to select dimensions which under- or over-decompose the datasets evaluated, necessitating alternative methods for setting ICA dimensionality.

The results presented herein reveal several insights to more optimally select this specification for transcriptomic datasets. ICA was conducted on several RNA-seq datasets across a range of dimensions and as expected, the overall structure of the S-matrix evolves as more components are computed. In other words, as the dimensionality is increased new robust components are revealed; additionally, once a component is revealed at lower dimensions, it is well conserved across higher dimensions. This realization essentially sets a lower dimension limit for an informative decomposition which should reveal as many of these conserved components as possible.

Alternatively, an upper limit for an informative decomposition would minimize the chance for over-decomposition, which is signified by an increase in the proportion of single gene components revealed. The dimensionality selection method presented here achieves both by finding the point across the dimensionality range where the number of conserved components is equal to the number of non-single gene components in that decomposition. Because components are well conserved across dimensions and single

gene components are most often revealed at higher dimensions when over-decomposition has set in, the decomposition at this point is likely to capture primarily the conserved, biologically relevant components.

In the case of PRECISE 1.0, the MSTD method resulted in selecting an under-decomposed dimension; while the PC-VA method appeared to select an appropriate dimensionality with limited single gene components and readily characterizable iModulons, designated by the common transcription factor of their enriched genes. The PC-VA method was tested on a *B. subtilis* and expanded *E. coli* transcriptomic dataset, PRECISE 2.0 to ensure efficacy; however, in these cases under- and over-decomposition occurred, respectively.

The method described herein improved these issues in both cases. The *B.subtils* dataset appeared to be under-decomposed by the PC-VA method selecting a relatively low dimensionality. The original characterization of the components into iModulons supported this notion as several contained disparate gene sets known to be regulated by different mechanisms. The new method selected a higher dimension which split many of these merged components thereby improving the decomposition. In the case of PRECISE 2.0 which was over-decomposed by the PC-VA method, the new method improved dimensionality selection evidenced by the lower proportion of single gene components in the decomposition. Lastly, applied to PRECISE 1.0 the method resulted in the reconstitution of all but two of the original iModulons, one of which was a single gene component.

REFERENCES

[1] Sastry, A.V., Gao, Y., Szubin, R., Hefner, Y., Xu, S., Kim, D., Choudhary, K.S., Yang, L., King, Z.A., Palsson, B.O. The *Escherichia coli* transcriptome mostly consists of independently regulated modules. *Nat Commun* 10, 5536 (2019). https://doi.org/10.1038/s41467-019-13483-w

[2] Kong, W., Vanderburg, C.R., Gunshin, H., Rogers, J.T., Huang, X. “A review of independent component analysis application to microarray gene expression data.” *BioTechniques*vol. 45,5 (2008). https://doi.org/10.2144/000112950

[3] Engreitz, J.M., Daigle, J.B., Marshall, J.J., Altman, R.B. Independent component analysis: mining microarray data for fundamental human gene expression modules. *Journal of biomedical informatics* 43, 6 (2010). https://doi.org/10.1016/j.jbi.2010.07.001

[4] Biton, A., Bernard-Pierrot, I., Lou, Y., Krucker, C., Chapeaublanc, E., Rubio-Pérez, C., López-Bigas, N., Kamoun, A., Neuzillet, Y., Gestraud, P., Grieco, L., Rebouissou, S., de Reyniès, A., Benhamou, S. Lebret, T., Southgate, J., Barillot, E., Allory, Y., Zinovyev, A., Radvanyi, F. Independent component analysis uncovers the landscape of the bladder tumor transcriptome and reveals insights into luminal and basal subtypes. *Cell Reports* 9, 4 (2014). https://doi.org/10.1016/j.celrep.2014.10.035

[5] Rychel, K., Sastry, A.V., Palsson, B.O. Machine learning uncovers independently regulated modules in the *Bacillus sbutilis* transcriptome. *bioRxiv*, 2020, Preprint. https://doi.org/10.1101/2020.04.26.062638

[6] Kairov, U., Cantini, L., Greco, A. Molkenov, A., Czerwinka, U., Barillot, E., Zinovyez, A. Determining the optimal number of independent components for reproducible transcriptomic data analysis. *BMC Genomics* 18, 712 (2017). https://doi.org/10.1186/s12864-017-4112-9

[7] Nicolas, P., Mäder, U, Dervyn, E., Rochat, T., Leduc, A., Pigeonneau, N., Bidnenko, E., Marchadier, E., Hoebeke, M., Aymerich, S., Becher, D., Bisicchia, P., Botella, E., Delumeau, O., Doherty, G., Denham, E.L., Fogg, M.J., Fromion, V., Goelzer, A., Hansen, A., Härtig, E., Harwood, C.R., Homuth, G., Jarmer, H., Jules, M., Klipp, E., Le Chat, L., Lecointe, F., Lewis, P., Liebermeister, W., March, A., Mars, R.A.T., Nannapaneni, P., Noone, D., Pohl, S., Rinn, B., Rügheimer, F., Sappa, P.K., Samson, F., Schaffer, M., Schwinowski, B., Steil, L., Stülke, J., Wiegert, T., Devine, K.M., Wilkinson, A.J., Maarten van Dijl, J., Hecker, M., Völker, U., Bassières, P., Noirot, P. Condition-Dependent Transcriptome Reveals High-Level Regulatory Architecture in Bacillus subtilis. Science 335, 1103–1106 (2012). https://doi.org/10.1126/science.1206848