BMC Bioinformatics

RESEARCH Open Access

Optimal dimensionality selection for independent component analysis of transcriptomic data

John Luke McConn, Cameron R. Lamoureux, Saugat Poudel, Bernhard O. Palsson and Anand V. Sastry*

*Correspondence:
avsastry@eng.ucsd.edu
University of California San Diego, San Diego, USA

Abstract

Background: Independent component analysis is an unsupervised machine learning algorithm that separates a set of mixed signals into a set of statistically independent source signals. Applied to high-quality gene expression datasets, independent component analysis effectively reveals both the source signals of the transcriptome as co-regulated gene sets, and the activity levels of the underlying regulators across diverse experimental conditions. Two major variables that affect the final gene sets are the diversity of the expression profiles contained in the underlying data, and the user-defined number of independent components, or dimensionality, to compute. Availability of high-quality transcriptomic datasets has grown exponentially as high-throughput technologies have advanced; however, optimal dimensionality selection remains an open question.

Methods: We computed independent components across a range of dimensionalities for four gene expression datasets with varying dimensions (both in terms of number of genes and number of samples). We computed the correlation between independent components across different dimensionalities to understand how the overall structure evolves as the number of user-defined components increases. We then measured how well the resulting gene clusters reflected known regulatory mechanisms, and developed a set of metrics to assess the accuracy of the decomposition at a given dimension.

Results: We found that over-decomposition results in many independent components dominated by a single gene, whereas under-decomposition results in independent components that poorly capture the known regulatory structure. From these results, we developed a new method, called OptICA, for finding the optimal dimensionality that controls for both over- and under-decomposition. Specifically, OptICA selects the highest dimension that produces a low number of components that are dominated by a single gene. We show that OptICA outperforms two previously proposed methods for selecting the number of independent components across four transcriptomic databases of varying sizes.

Conclusions: OptICA avoids both over-decomposition and under-decomposition of transcriptomic datasets resulting in the best representation of the organism's underlying transcriptional regulatory network.

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 [1]. For example, ICA may be used to solve the 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 can effectively recover the original independent audio signals, and their relative volumes for each microphone.

Beyond deconvoluting audio signals, ICA is widely applicable to several other fields involving signal separation or feature extraction [2, 3]. With the advancement of high-throughput gene expression profiling, ICA has proven to be useful in analyzing highly multivariate microarray and RNA sequencing (RNA-seq) gene expression datasets [4–8]. The resulting ICA decomposition of such datasets yields sets of independent components (composed of weights for all genes), and the components relative strength across all samples or experimental conditions (i.e. activities). Significant genes from these components can then be determined and paired to a reference gene set through statistical tests.

Applied to large microbial transcriptomic datasets, this kind of analysis can result in highly accurate reconstructions of their underlying transcriptional regulatory networks (TRNs). A TRN is a collection of regulatory interactions between transcriptional regulators (such as transcription factors) and their regulated genes. For example, ICA decomposition of an *Escherichia coli* expression compendium containing 278 expression profiles (named PRECISE) revealed 92 independently modulated gene sets, termed iModulons. Most iModulons had significant overlap with known regulons (i.e., groups of co-regulated genes), which allowed many iModulons to be directly associated with distinct transcription factors [4]. Regulons are groups of co-regulated genes derived from biomolecular binding experiments that form a bottom-up representation of the TRN. On the other hand, iModulons are co-expressed gene sets that are associated with specific transcriptional regulators to form a top-down, data-driven TRN that is amenable to rapid scaling.

Similar analysis has since been carried out on a *Bacillus subtilis* microarray dataset with 269 expression profiles [9] and on a compendium of 108 RNA-seq profiles of *Staphylococcus aureus*, named *Staph*PRECISE [10], which revealed 83 and 29 similarly informative iModulons, respectively. Although independent components extracted from human datasets also capture biologically relevant gene clusters [11, 12], the similarity between these gene clusters and the known TRN is obscured by the inherent complexity of eukaryotic transcriptional regulation.

As gene expression datasets incorporate more unique expression profiles across varied growth conditions, it becomes possible to develop a comprehensive reconstruction of an organism's TRN. To achieve this, the output of ICA decompositions depends primarily on two inputs—the number of high-quality data sources across diverse growth conditions; and the number of independent components to compute. While the former has become more accessible through public data repositories [13], determining the optimal dimensions for the decomposition remains an open question [14]. Previous investigation into this problem showed that searching for too many components can result in many components driven by small gene sets, whereas searching for too few components can obscure the biological interpretation [15].

Several methods have been suggested and employed to answer this question. One such method entails setting the number of dimensions equal to the number of principal components, determined through principal component analysis (PCA), which account for a certain level of variance in the data [4, 8]. Alternatively, the Maximally Stable Transcriptome Dimension (MSTD) has been suggested, defined as the maximum dimension before ICA begins to produce a large proportion of unstable components [15]. Although these are the primary two methods used to identify the optimal dimensionality when applying ICA to gene expression compendia, these methods have not yet been rigorously tested against well-characterized microbial TRNs.

In this study, we investigated how different dimensions for ICA affect the accuracy of the inferred TRN and evaluated the performance of existing dimensionality selection methods on four high-quality transcriptomic datasets that contained expression profiles across diverse growth conditions. We found that these previously proposed selection methods were inconsistent, resulting in either over-decomposition or under-decomposition in several cases. From these results, we developed a new method that identifies the optimal dimension that maximizes independent components that represent known regulation, while minimizing the presence of biologically meaningless over-decomposed components. The new method, named OptICA, ensures that future studies will select the ideal dimensionality to optimize the reconstructed TRNs from new transcriptomic datasets. Furthermore, OptICA relies on the finding that independent components are well-conserved across dimensions, suggesting that this method could be used in other applications of ICA.

Results

Independent components form a "tree" across dimensions

To develop an understanding of how independent components evolve across different dimensions, we decomposed four transcriptomic datasets (Table 1) using FastICA [16] across the full range of possible dimensions. Since FastICA is inherently stochastic, we applied a clustering approach that only retains independent components that persist across multiple runs (see Methods). Therefore, the number of robust components identified at a particular dimension may be lower than expected, as unstable components are discarded (Additional file 1: Figure S1). Application of a threshold on the gene weights

Table 1 Overview of datasets

Dataset Name	References	Organism	Technology	Number of samples
PRECISE 1.0	[4]	*Escherichia coli*	RNA-seq	278
PRECISE 2.0	[17]	*Escherichia coli*	RNA-seq	815
*Staph*PRECISE	[10]	*Staphylococcus aureus*	RNA-seq	108
B. subtilis Microarray	[9, 18]	*Bacillus subtilis*	Microarray	269

of each of these robust components resulted in discrete groups of genes for each component, named iModulons.

Several RNA-seq and microarray datasets were utilized for this analysis, including the original version of PRECISE (PRECISE 1.0) [4], an expanded version (PRECISE 2.0) [17], a compendium of *S. aureus* RNA-seq expression profiles (StaphPRECISE) [10] and a *B. subtilis* microarray dataset [9, 18] (Table 1). Each dataset was decomposed across a range of dimensions, from low to nearly fully decomposed (i.e., one dimension for each sample), and independent components were compared between adjacent dimensions to form a "dimensionality tree".

Dimensionality trees convey the evolution of the independent component structure across dimensions, as shown by the dimensionality tree for PRECISE 1.0 (Fig. 1). At low dimensions, only a few independent components are identified. The iModulons derived

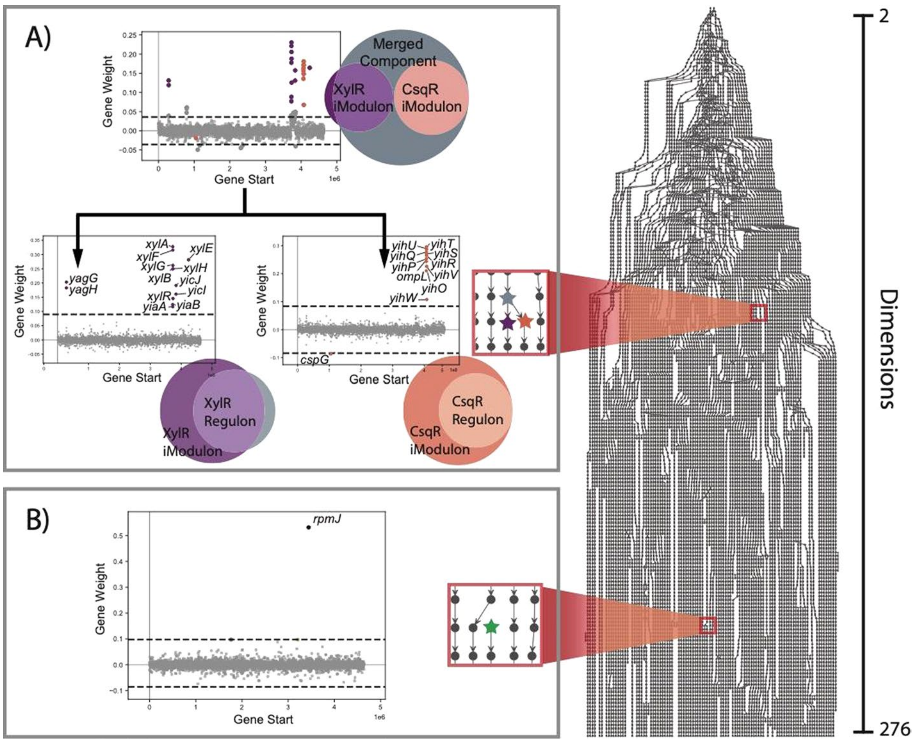

Fig. 1 The dimensionality tree of PRECISE 1.0 reveals how its ICA decomposition evolves across a range of dimensions. Each point represents a computed component, and a row of points represents all components calculated at a particular dimensionality. Connections between components of adjacent dimensions were established where their correlation was greater than 0.3. **A** The red box highlights a section of the dimensionality tree where component splitting occurs. At low dimensions, components undergo reorganization whereby disparate gene sets, initially contained in a single component (gray star node), split into multiple components which more accurately reflect underlying transcriptional regulatory mechanisms (purple and orange star nodes). The "Gene Start" (position of gene along the genome) vs "Gene Weight" (independent component gene weight) graphs illustrate a specific example of component splitting. The gene sets contained in the XylR and CsqR iModulons are contained in the same independent component at low dimensionality but are split into separate components at higher dimensions. These split components more accurately map to a single regulon. **B** The red box highlights a point of the dimensionality tree at which a single gene component appears. At high dimensions new components appear which are uncorrelated to those of the preceding dimension (green star node). Often, these components contain a single highly weighted gene and signify the commencement of over-decomposition

from these independent components often contain multiple related regulons (Fig. 1A). As the number of dimensions increases, these iModulons tend to split such that each regulon is contained within its own iModulon. As existing components split and new components appear, there is a net increase in total independent components until a relatively stable decomposition structure is reached. The appearance of additional components beyond this stable region suggests the commencement of over-decomposition evidenced by the appearance of components with a single highly weighted gene (Fig. 1B).

Dimensionality trees were computed for all four transcriptomic datasets, and demonstrated how the overall structure of the ICA decomposition evolves as more components are computed. With few exceptions, once a component was initially discovered at a particular dimension, it proved to be conserved in higher order decompositions (Additional file 1: Figure S2). This realization suggests that across the stable decomposition region, the independent component structure does not materially change. Dimensionality selection techniques that produce useful decompositions of transcriptomic data would likely target points within this range, following the initial linear increase in computed components but before the commencement of over-decomposition.

Dimension selection methods often result in over or under decomposition

To identify the optimal dimension, the information from the dimensionality tree was summarized into various categories of iModulons. An iModulon was classified as “regulatory” if it was significantly enriched with a specific regulon (Fisher’s exact test, FDR < 1e−5). Since iModulons found at high dimensions are often highly similar to known regulons, we also tracked the number of “conserved” iModulons, or iModulons at lower dimensions that were similar to an iModulon detected at the largest dimension. Finally, to capture over-decomposition, we tracked the number of iModulons that contained a single highly weighted gene, and conversely, contained more than one gene. The categories were not exclusive; an iModulon at a particular dimension could be a regulatory, non-single gene, conserved iModulon.

These classifications were plotted across the full range of possible dimensionalities for the four datasets (Fig. 2). These charts clearly showed that the number of regulatory iModulons and non-single gene iModulons sharply increased at lower dimensions. For the *E.coli* datasets (PRECISE 1.0 and 2.0), the number of regulatory iModulons and non-single gene iModulons plateaued and did not substantially increase even at the full decomposition. However, the number of conserved iModulons increased during plateau, indicating that the iModulons were experiencing internal reorganization. On the other hand, we observed large numbers of single-gene iModulons at high dimensions, signifying over-decomposition. For PRECISE 2.0, the largest dataset in this study, nearly all new components above a dimension of 500 were single-gene components.

Previously published ICA decompositions of three of the datasets (PRECISE 1.0, *Staph*PRECISE, and *B. subtilis*) utilized the same method for establishing dimensionality. Referred to as the PC-VA method herein, this technique sets the ICA dimension based on the number of principal components which explain a certain level of variance (e.g., 95%) in the data. However, the PC-VA method selected sub-optimal dimensions for two datasets. For the *B. subtilis* dataset, the selected dimension resulted in nearly half of the possible regulatory iModulons. Selecting a higher dimension in this case would have

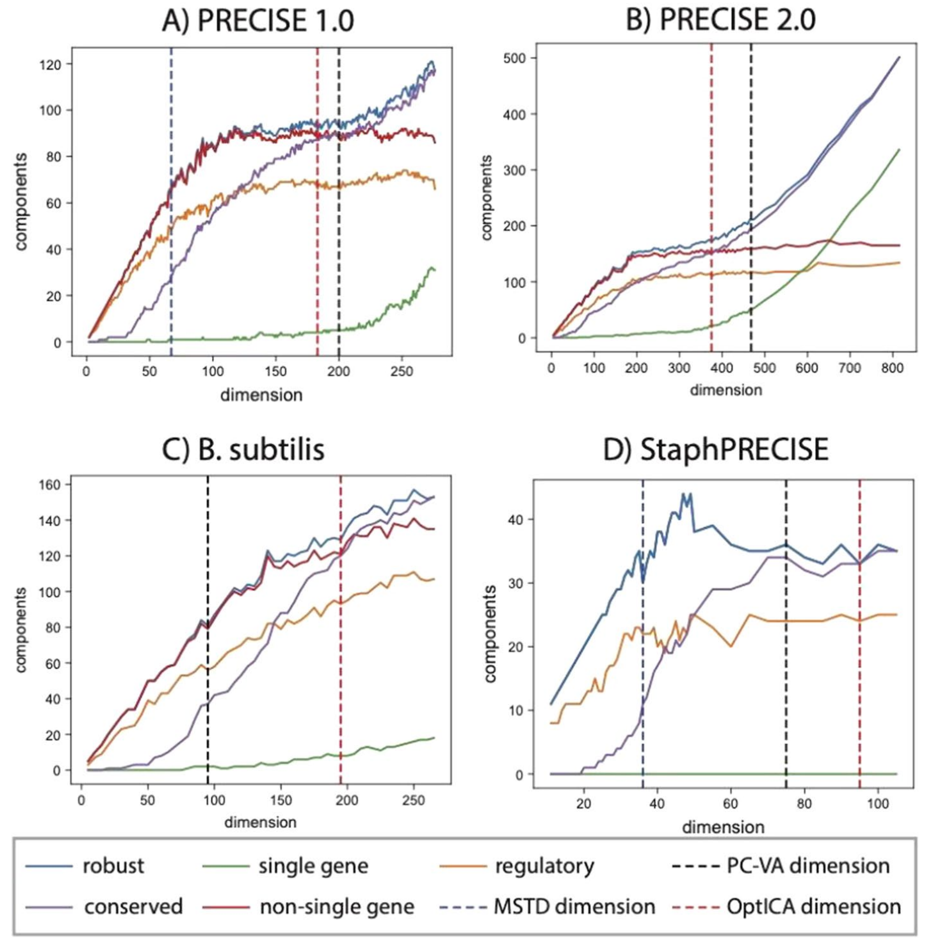

Fig. 2 Classification of independent components across dimensions for each dataset. The PC-VA, OptICA, and MSTD dimensions are shown where applicable. The four datasets are **A** PRECISE 1.0, **B** PRECISE 2.0, **C** *B. subtilis*, and **D** *Staph*PRECISE. Note that since no single gene components are found for the StaphPRECISE dataset, the non-single gene and robust component lines overlap

better captured the true TRN of the organism. On the other hand, the PC-VA method resulted in 51 single-gene iModulons from the PRECISE 2.0 dataset out of 179 total iModulons (nearly 30%), indicating that the dataset was over-decomposed.

We also attempted to compute the MSTD for each dataset. However, the algorithm did not converge on a stable dimension for either the PRECISE 2.0 dataset or the *B. subtilis* dataset. The MSTD identified for the PRECISE 1.0 and *Staph*PRECISE datasets seemed to under-decompose the datasets, as the number of robust iModulons, non-single gene iModulons, and regulatory iModulons were still rising at this dimensionality (Fig. 2A, D).

OptICA, a novel ICA dimensionality selection technique, controls over- and under-decomposition

Based on the observation that components are conserved across dimensions, we proposed a new method to identify the optimal dimension of the dataset. An informative decomposition would maximize the discovery of these conserved components, while minimizing the number of components with single genes. Therefore, OptICA selects the dimension at which the number of conserved components equals the number of non-single gene components.

We used three criteria to evaluate the performance of the MSTD, PC-VA dimension and the OptICA dimension: (1) the number of iModulons enriched with a transcriptional regulator (i.e., regulatory iModulons); (2) the number of single-gene iModulons; and (3) the F1-score of the regulatory iModulons (Fig. 3). The F1-score is the harmonic average of precision and recall, which measure the number of false positives and false negatives, respectively, in iModulons as compared to published regulons in the literature. Over-decomposition would result in a high number of single-gene iModulons, whereas under-decomposition would result in a low number of regulatory iModulons and/or a low average F1-score.

Across datasets, the OptICA method resulted in the most consistent results, selecting fully decomposed dimensions prior to the occurrence of rampant over-decomposition. The PC-VA and OptICA dimensions resulted in similar decompositions for PRECISE 1.0, while the MSTD occurred at a point that captured fewer regulatory components with less congruence with associated regulons (i.e., lower F1-score) (Fig. 3A). The OptICA dimension of PRECISE 2.0 resulted in a substantial reduction in the number of single gene components compared to the PC-VA dimension, with a minimal reduction in the number of regulatory components and mean F1-score (Fig. 3B). Alternatively, the OptICA dimension of the *B. subtilis* dataset resulted in capturing substantially more

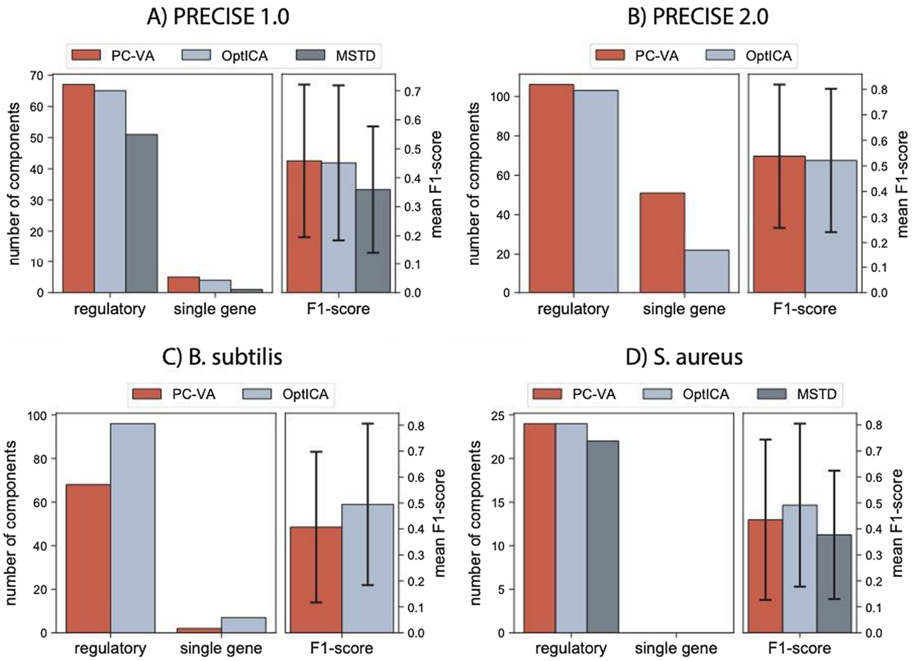

Fig. 3 Bar charts indicating three parameters used to compare decompositions conducted at different dimensions (PC-VA, OptICA, and MSTD, when available). The left bar chart shows the number of regulatory and single gene components computed at each dimensionality, whereas the right bar chart shows the mean F1-score. Error bars represent standard deviations. Data was plotted for **A** PRECISE 1.0, **B** PRECISE 2.0, **C** the *B. subtilis* dataset, and **D** *Staph*PRECISE

regulatory components than PC-VA, which better modeled underlying regulatory mechanisms (i.e., higher F1-score) with a slight increase in single gene components computed (Fig. 3C). The OptICA and PC-VA dimensions captured the same number of regulatory components in *Staph*PRECISE; however, the OptICA dimension resulted in slightly higher congruence between the regulatory components and their associated regulons, evidenced by a higher mean F1-score (Fig. 3D).

From this analysis, we found that the OptICA dimension controlled for both over- and under-decomposition better than the PC-VA dimension and MSTD across the four datasets.

OptICA results in more accurate TRN representations

To gain a deeper understanding of how different dimensions affect how well iModulons mirror the known TRN, we tracked the average F1-score across all dimensions for the four datasets (Fig. 4A, B). The F1-scores seem to initially rapidly increase, and then stabilize, similar to the trajectory of the number of regulatory iModulons. Overall, the average F1-scores did not significantly differ between the PC-VA dimensions and the OptICA dimensions. This was likely because the F1-scores had neared their maximum values.

However, the OptICA dimension resulted in a meaningful increase in average F1-scores for the *B. subtilis* dataset, and a substantial increase in the number of iModulons which were perfectly aligned with a known regulon (F1-score = 1.0) (Fig. 4C, D). This improvement partially resulted from the resolution of merged components present at the PC-VA dimension (Fig. 4E). Several iModulons from the originally published decomposition contained gene sets known to be regulated by different mechanisms; these components were effectively split at the OptICA dimension. For example, the original NadR/BirA iModulon was split into two components at the OptICA dimension which were highly congruent with their associated regulon.

Discussion

Two important factors strongly 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 described above. These methods were tested on several transcriptomic datasets and in some cases, were found to select dimensions which under- or over-decompose the datasets evaluated, necessitating an alternative method for setting the dimensionality of ICA.

The results presented herein reveal several insights to more optimally select this specification for transcriptomic datasets. ICA was conducted on four transcriptomic datasets across a range of dimensions, showing that the overall structure of the decomposition 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.

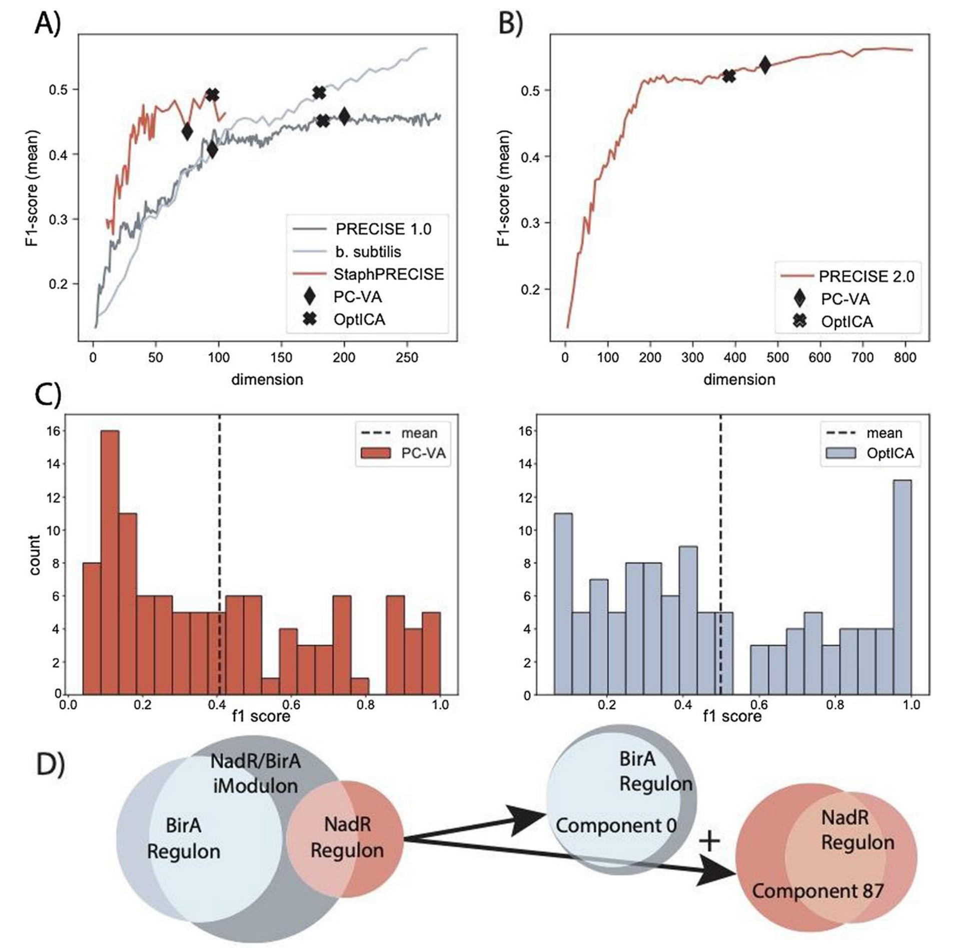

Fig. 4 Comparing regulatory components to their associated regulon suggests the dimension selected by OptICA more accurately models the TRN. **A** The average F1-scores for PRECISE 1.0, *Staph*PRECISE, and the *B. subtilis* dataset at each dimensionality. **B** The average F1-scores for the PRECISE 2.0 dataset at each dimensionality. **C** The higher dimension selected by this new method applied to the *B. subtilis* dataset resulted in higher mean F1-scores and a substantial increase in components with perfect scores (exact precision and recall between the component and associated regulon). **D** OptICA improved F1-scores and resolved under decomposition by splitting merged components computed at the PC-VA dimension. For example, the originally published decomposition reported the NadR/BirA iModulon which contained genes belonging to both the NadR and BirA regulons. The dimension selected by this new method computed separate components which more accurately model the underlying regulatory mechanisms

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 computed. As shown by PRECISE 2.0, if a dataset is large enough, it is conceivable that each gene could be decomposed into its own iModulon, obfuscating the true structure of the dataset. The dimensionality selection method presented here, OptICA, 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 resulting components at the OptICA dimension are likely to capture primarily the conserved, biologically relevant components.

OptICA has two additional advantages. First, the algorithm allows for incremental, on-line learning, where the optimality of the decomposition can be assessed at each new dimension. This avoids the need to perform ICA at high dimensions, which is computationally expensive. Second, it does not require prior knowledge of the true transcriptional regulatory network since it does not rely on assessing regulatory iModulons. This enables the use of OptICA on transcriptional datasets for organisms with uncharacterized TRNs.

Overall, OptICA results in improved transcriptomic decompositions for both small and large RNA-seq datasets, avoiding both over- and under-decomposition. We validated OptICA against known transcriptional regulatory networks and found that it outperformed previously published algorithms for identifying the optimal dimensionality. OptICA is organism-invariant, and we foresee that it will assist in developing many models of transcriptional regulatory networks.

Methods

Conducting independent component analysis on gene expression data

The Scikit-learn (v0.23.2) [19] implementation of FastICA [20] was executed 100 times with random seeds and a convergence tolerance of 10^{-7}. The resulting independent components (ICs) were clustered using DBSCAN [21] to identify robust ICs, using an epsilon of 0.1 and minimum cluster seed size of 50. To account for identical components with opposite signs, the following distance metric was used for computing the distance matrix:

$$d_{x,y} = 1 - \left|\left|\rho_{x,y}\right|\right|$$

where $\rho_{x,y}$ is the Pearson correlation between components x and y. The final robust ICs were defined as the centroids of the cluster. Again, to account for identical components with opposite signs, we choose one component as the canonical direction, and flip all other components in the cluster to ensure that the Pearson correlation is positive between all members of the cluster before computing the centroid.

For PRECISE 1.0, which contains 278 expression profiles, the multi-start ICA process was run computing every dimension from 2 to 276 components. For the *Staph*PRECISE dataset, every 5th dimension was analyzed from 11 components through 105 components. For the *B. subtilis* dataset, every 5th dimension was analyzed from 5 components through 265 components. For PRECISE 2.0, the process was run computing every 5th dimension from 5 to 815 components.

Building dimensionality trees and classifying conserved components

Cosine distance was computed using the full set of gene weights to compare components across dimensionalities; this metric is referred to as correlation herein. The correlation between components of each subset and those of 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 (final components) 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 heat maps of conserved component occurrence in each dimension. Where the correlation was greater than an established threshold a final component was said to exist in a preceding decomposition and those components were said to be conserved. To establish these thresholds the components in the highest dimension decomposition were compared pairwise to all components computed at lower dimensions. Correlation was calculated for each pair and histograms of the highest values associated with a particular component in the final decomposition were plotted, resulting in a distribution of highly correlated components (Additional file 1: Figure S3). The elbow point of this distribution, determined by the Kneedle algorithm [22], was used to establish a threshold correlation to classify a component as a conserved component.

Identifying significant genes in an independent component

To perform regulator enrichments on components, genes with significantly high weightings must be identified. To keep this method agnostic to the prior regulatory structure, we applied the Scikit-learn [19] implementation of K-means clustering to the absolute values of the gene weights in each independent component. All genes in the top two clusters were deemed significant, and the set of significant genes in each independent component was called the iModulon.

Associating iModulons to regulators

The set of significant genes in each component, or the iModulon, was compared to each regulon by the two-sided Fisher's exact test ($\text{FDR} < 10^{-5}$) to determine regulator enrichment. F1-scores were calculated to evaluate these associations. The F1-score is the harmonic average of precision and recall between a component and its linked regulon. Precision is the proportion of genes in the component that are present in the associated regulon and recall is the proportion of genes in the regulon that are present in the associated component. Prior information about regulator binding sites was borrowed from previous studies [4, 9, 10].

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

All components computed from a multi-start ICA decomposition, as described above, were counted as "robust components". 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. Components, which contained significant gene sets, that were associated with a previously published regulon as described above were classified as regulatory.

Determining the PC-VA, MSTD and OptICA 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 [15]. Briefly, ICA is carried out K times across a range of dimensions, from Mmin to Mmax. At each value of M, MxK components are computed, clustered, and ranked by stability. The MSTD point corresponds to the inflection point in the distribution in these stability profiles. The OptICA dimension was defined as the point where the number of non-single gene components was equal to the number of conserved components in that decomposition.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-021-04497-7.

Additional file 1: Figure S1. Due to density-based clustering of ICA run with randomized restarts, the number of robust components do not directly correlate with the selected dimensionality. **Figure S2.** Across all dimensions components revealed in the final decomposition were found to be well conserved at the established threshold. All components in the final decomposition were correlated pairwise with those of all preceding decompositions. Once a component present in the final decomposition was calculated at a lower dimension that component continued to persist within decompositions of higher dimensionality. A component in the final decomposition was said to be present in a particular decomposition where its correlation with a component in that decomposition was above the established threshold. Across all datasets, PRECISE 1.0, PRECISE 2.0, *B. subtilis*, and *Staph*PRECISE, components were found to be well conserved, rarely dropping below their established threshold once computed at a lower dimension. PRECISE 1.0 shown here for example. **Figure S3.** Components from the final, fully decomposed dataset were correlated pairwise with components of all preceding decompositions. Histograms of the highest correlations for each component across all dimensions were plotted for (A) PRECISE 1.0, (B) PRECISE 2.0, (C) *B. subtilis*, and (D) *Staph*PRECISE. The elbow point of these highly correlated values served as the threshold to classify a particular component as conserved.

Acknowledgements
The authors would like to thank Kevin Rychel and Katherine Decker for helpful discussions.

Authors' contributions
JLM, AVS and CRL designed the study. JLM performed most of the analysis, with contributions from AVS, CRL, and SP. JLM and AVS wrote the manuscript. All authors read and approved the final manuscript.

Funding
This work was funded by the Novo Nordisk Foundation Center for Biosustainability and the Technical University of Denmark (Grant number NNF10CC1016517) and by the NIH NIAID (Grant number U01AI124316). This research used resources of the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

Availability of data and materials
All data generated or analyzed during this study are available in iModulonDB (https://imodulondb.org/). Code to perform ***optICA*** is available at https://github.com/avsastry/modulome-workflow/tree/main/4_optICA.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 26 May 2021 Accepted: 16 November 2021
Published online: 08 December 2021

References

1. Hyvärinen A, Oja E. Independent component analysis: algorithms and applications. Neural Netw. 2000;13:411–30.
2. Delorme A, Makeig S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. J Neurosci Methods. 2004;134:9–21.
3. Scholz M, Gatzek S, Sterling A, Fiehn O, Selbig J. Metabolite fingerprinting: detecting biological features by independent component analysis. Bioinformatics. 2004;20:2447–54.

4. Sastry AV, Gao Y, Szubin R, Hefner Y, Xu S, Kim D, et al. The *Escherichia coli* transcriptome mostly consists of independently regulated modules. Nat Commun. 2019;10:5536.
5. Kong W, Vanderburg CR, Gunshin H, Rogers JT, Huang X. A review of independent component analysis application to microarray gene expression data. Biotechniques. 2008;45:501–20.
6. Engreitz JM, Daigle BJ Jr, Marshall JJ, Altman RB. Independent component analysis: mining microarray data for fundamental human gene expression modules. J Biomed Inform. 2010;43:932–44.
7. Biton A, Bernard-Pierrot I, Lou Y, Krucker C, Chapeaublanc E, Rubio-Pérez C, et al. Independent component analysis uncovers the landscape of the bladder tumor transcriptome and reveals insights into luminal and basal subtypes. Cell Rep. 2014;9:1235–45.
8. Teschendorff AE, Journée M, Absil PA, Sepulchre R, Caldas C. Elucidating the altered transcriptional programs in breast cancer using independent component analysis. PLoS Comput Biol. 2007;3:e161.
9. Rychel K, Sastry AV, Palsson BO. Machine learning uncovers independently regulated modules in the *Bacillus subtilis* transcriptome. Nat Commun. 2020;11:6338.
10. Poudel S, Tsunemoto H, Seif Y, Sastry AV, Szubin R, Xu S, et al. Revealing 29 sets of independently modulated genes in *Staphylococcus aureus*, their regulators, and role in key physiological response. Proc Natl Acad Sci USA. 2020;117:17228–39.
11. Karczewski KJ, Snyder M, Altman RB, Tatonetti NP. Coherent functional modules improve transcription factor target identification, cooperativity prediction, and disease association. PLoS Genet. 2014;10:e1004122.
12. Saelens W, Cannoodt R, Saeys Y. A comprehensive evaluation of module detection methods for gene expression data. Nat Commun. 2018;9:1090.
13. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res. 2013;41:D991–5.
14. Way GP, Zietz M, Rubinetti V, Himmelstein DS, Greene CS. Compressing gene expression data using multiple latent space dimensionalities learns complementary biological representations. Genome Biol. 2020;21:109.
15. Kairov U, Cantini L, Greco A, Molkenov A, Czerwinska U, Barillot E, et al. Determining the optimal number of independent components for reproducible transcriptomic data analysis. BMC Genomics. 2017;18:712.
16. Hyvarinen A. Fast ICA for noisy data using Gaussian moments. In: 1999 IEEE international symposium on circuits and systems (ISCAS). vol 5. 1999. p. 57–61.
17. Lamoureux CR, Decker KT, Sastry AV, McConn JL. PRECISE 2.0-an expanded high-quality RNA-seq compendium for *Escherichia coli* K-12 reveals high-resolution transcriptional regulatory structure. bioRxiv. 2021. https://doi.org/10.1101/2021.04.08.439047v1.abstract.
18. Nicolas P, Mäder U, Dervyn E, Rochat T, Leduc A, Pigeonneau N, et al. Condition-dependent transcriptome reveals high-level regulatory architecture in Bacillus subtilis. Science. 2012;335:1103–6.
19. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
20. Hyvärinen A. Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans Neural Netw. 1999;10:626–34.
21. Ester M, Kriegel H-P, Sander J, Xu X, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Kdd 1996; p. 226–31.
22. Satopaa V, Albrecht J, Irwin D, Raghavan B. Finding a“ kneedle” in a haystack: Detecting knee points in system behavior. In: 2011 31st international conference on distributed computing systems workshops. IEEE; 2011. p. 166–71.

Publisher's Note

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