FunGeneClusterS

Predicting fungal gene clusters from genome and transcriptome data

Vesth, Tammi Camilla; Brandl, Julian; Andersen, Mikael Rørdam

Published in:
Synthetic and Systems Biotechnology

Link to article, DOI:
10.1016/j.synbio.2016.01.002

Publication date:
2016

Document Version
Publisher's PDF, also known as Version of record

Synthetic and Systems Biotechnology 1 (2016) 122–129

Contents lists available at ScienceDirect

Synthetic and Systems Biotechnology

journal homepage: keaipublishing.com/synbio

FunGeneClusterS: Predicting fungal gene clusters from genome and transcriptome data

Tammi C. Vesth, Julian Brandl, Mikael Rørdam Andersen *

Department of Systems Biology, Technical University of Denmark, Søltofts Plads 223, Denmark

ARTICLE INFO

Article history:
Received 18 October 2015
Received in revised form 14 December 2015
Accepted 5 January 2016
Available online

Keywords:
Secondary metabolism
Gene clusters
Transcriptomics
Genomics
Bioinformatics
Aspergillus niger
Aspergillus nidulans

ABSTRACT

Introduction: Secondary metabolites of fungi are receiving an increasing amount of interest due to their prolific bioactivities and the fact that fungal biosynthesis of secondary metabolites often occurs from co-regulated and co-located gene clusters. This makes the gene clusters attractive for synthetic biology and industrial biotechnology applications. We have previously published a method for accurate prediction of clusters from genome and transcriptome data, which could also suggest cross-chemistry, however, this method was limited both in the number of parameters which could be adjusted as well as in user-friendliness. Furthermore, sensitivity to the transcriptome data required manual curation of the predictions. In the present work, we have aimed at improving these features.

Results: FunGeneClusterS is an improved implementation of our previous method with a graphical user interface for off- and on-line use. The new method adds options to adjust the size of the gene cluster(s) being sought as well as an option for the algorithm to be flexible with genes in the cluster which may not seem to be co-regulated with the remainder of the cluster. We have benchmarked the method using data from the well-studied *Aspergillus nidulans* and found that the method is an improvement over the previous one. In particular, it makes it possible to predict clusters with more than 10 genes more accurately, and allows identification of co-regulated gene clusters irrespective of the function of the genes. It also greatly reduces the need for manual curation of the prediction results. We furthermore applied the method to transcriptome data from *A. niger*. Using the identified best set of parameters, we were able to identify clusters for 31 out of 76 previously predicted secondary metabolite synthases/synthetases. Furthermore, we identified additional putative secondary metabolite gene clusters. In total, we predicted 432 co-transcribed gene clusters in *A. niger* (spanning 1.323 genes, 12% of the genome). Some of these had functions related to primary metabolism, e.g. we have identified a cluster for biosynthesis of biotin, as well as several for degradation of aromatic compounds. The data identifies that suggests that larger parts of the fungal genome than previously anticipated operates as gene clusters. This includes both primary and secondary metabolism as well as other cellular maintenance functions.

Conclusion: We have developed FunGeneClusterS in a graphical implementation and made the method capable of adjustments to different datasets and target clusters. The method is versatile in that it can predict co-regulated clusters not limited to secondary metabolism. Our analysis of data has shown not only the validity of the method, but also strongly suggests that large parts of fungal primary metabolism and cellular functions are both co-regulated and co-located.

1. Introduction

The identification of genes involved in production of medically or industrially relevant chemical compounds is of great financial and public health interest. Anti-cancer and anti-infective agents are to a large extent comprised of or derived from natural products.[1] A group of compounds which are thought to be specifically promising within these fields are the secondary (non growth associated) metabolites (SMs). These compounds are often synthesized in a modular process where an initial polymer backbone is modified by

* Corresponding author. Department of Systems Biology, Technical University of Denmark, Søltofts Plads 223, Denmark. Tel.: +4545252675; fax: +4545884148.
E-mail address: mr@bio.dtu.dk (M.R. Andersen).
Peer review under responsibility of KeAi Communications Co., Ltd.

http://dx.doi.org/10.1016/j.synbio.2016.01.002

a number of tailoring enzymes.[2] The genes involved in this process have been found to cluster together on the fungal chromosome and are frequently located at the telomeric regions.[3] In connection with the increased interest in these compounds, there is also a high demand for fast, cheap and automated ways of detecting compounds and genes associated with them. Computational methods can be utilized to speed up the identification of these target genes. The number of known and well-studied secondary metabolite clusters from fungi is relatively small compared to the number of species and isolated to a few well-studied laboratory strains from a limited number of organisms.

Here we present a new implementation of a method for predicting secondary metabolism genes from the combination of genome sequences and transcription data. The present method is an update of our previous algorithm.[4] The first version of the method was developed to predict co-regulated gene clusters in fungal genomes based on a combination of transcriptome data and genome coordinates of the genes. The goal was primarily to be able to identify secondary metabolite (SM) gene clusters. The method was – as a proof of concept – shown to accurately predict known SM gene clusters in *Aspergillus nidulans* using a combination of automated analysis and manual curation. In some cases, we observed variations between the genes predicted in known clusters, and the experimentally verified cluster members. In these cases, we identified the correct position, but not the right boundaries of the clusters. These deviations could however be explained by errors in older gene calling and erroneous genes included in older publications. As such, the method was shown to correctly identify all known gene clusters from the genome sequence and transcription data.

The measure developed in the first paper was described as the Cluster Score (CS) and calculates a numeric value describing the similarity of gene expression profiles of genes located next to each other. A high CS indicates that neighboring genes have very similar expression profiles. It is hypothesized that genes involved in the production of a specific compound will have similar expression profiles. The CS was calculated using three genes up- and downstream from the position being evaluated and used the Pearson correlation coefficient. The calculation sets inverse correlations to 0 and genes located less than 4 genes away from the contig ends are assigned CS scores of 0.[4]

The method for computing the CS was fixed to use the Pearson correlation coefficient, which is a parametric correlation. Here we investigate the effect of using two non-parametric methods, the Kendall and Spearman rank-sum correlations in addition to the Pearson correlation. The graphical interface presented here allows the user to choose between the three methods for the optimal solution of the given dataset.

The method described above is here named FunGenClusterS (Fungal Gene Clusters with R and Shiny) and has been updated in this version to include a number of minor modifications adding flexibility in the prediction method as well as a graphical interface for easier access to the method and the output. In particular we have improved the algorithm, so that the manual curation from the previous method has been greatly reduced. The graphical interface can be run from the web (from https://fungiminions.shinyapps.io/FunGeneClusterS) or on a local installation (using R).

Revisions to the calculation options and methods include:

- User-defined window size for cluster score calculation. The window size determines the number of up- and down-stream genes to include in the running calculation as well as the plotting window around identified clusters
- User-defined correlation type, choose between Spearman, Pearson, and Kendall-based cluster scores. Genes at contig start and ends are dependent on the genes available up- or downstream depending on the availability and selected window size
- Automatic plots of CS score for each area where a possible cluster is found for improved data analysis
- List of genes in clusters with flanking genes according to selected window size
- Calculation speed increased

The generalized equation for calculation of CS can then be written as:

$$CS_{\pm w} = \sum_{i=-w}^{-1}\left(\frac{s_{0,i} + \|s_{0,i}\|}{2}\right)^2 + \sum_{i=1}^{w}\left(\frac{s_{0,i} + \|s_{0,i}\|}{2}\right)^2 \quad (1)$$

where w is the window size (values can be integers from 1 to 6) and s is the correlation coefficient as calculated using Spearman, Kendall or Pearson correlation coefficients. The interface was constructed using the R package Shiny and allows for local data upload and download of results both as table and graphs.

The original publication listed predicted clusters that were very similar to already identified and published gene clusters.[4] These predictions were however a product of automated prediction in combination with manual curation of predictions. The original CS method did not predict larger clusters (15 genes) without manual curation because a single gene with a CS below the threshold would break the cluster, thereby creating a cluster on each side of the low CS gene. This could of course be an actual biological scenario, but as it appeared multiple times, it seems more likely to be an error derived from variations in the quality of the probes on the DNA microarray. We therefore propose adding an option to the method which allows the user to ignore a number of genes within the analysis to allow for flexibility in automated prediction of these larger clusters. The option covers skipping 0–5 genes (default is 0). The different parameter values were evaluated based on the number of predicted clusters as well as the sizes of the predicted clusters (number of genes that are co-expressed). Finally the predictions were compared to secondary metabolism clusters identified using the SMURF algorithm.[5] This specific type of gene clusters represents groups of genes that might be co-regulated and represent different steps in a biosynthetic pathway producing a secondary metabolite. As FunGeneClusterS predicts all types of co-regulated genes, we expect to predict more clusters than SMURF, among others, primary metabolism genes. As some secondary metabolism clusters are silent under most laboratory conditions, we expect not to find these based on expression data. This is not a larger disadvantage, as such cluster would normally not produce compounds for which genes would then be needed. Using secondary metabolism clusters for benchmarking FunGenClusterS will therefore provide some information on the predictive power for this type of clusters but not all co-regulated gene clusters.

2. Material and methods

2.1. Computation

All calculations were performed in R version 3.2.2 (2015-08-14)[6] using the following packages: data.table (version 1.9.6), gplots (version 2.17.0), reshape2 (version 1.4.1), ggplot2 (version 1.0.1), and shiny (version 0.12.2).[7–11]

The application is available from shinyapps.io under the name FunGeneClusterS and source code is available for download from the same resource. All code is published under the LGPL 3.0 license (http://opensource.org/licenses/LGPL-3.0). The identified clusters were compared to the findings of the SMURF algorithm using standard settings.[5] Detection of secondary metabolite synthases and synthetases was performed using the *A. nidulans* FGSC A4[12] and *A. niger* CBS 513.88 genome,[13] which was converted into gene names for version 3.0 of the *A. niger* ATCC 1015[14] using bidirectional best BLASTp hits.[15]

2.2. Transcriptome data

Analysis was performed using gene expression data from *A. nidulans* as previously described.[4] In addition, gene expression data was obtained from a number of experiments of *A. niger*[14,16–18] adding up to a total of 13 different growth conditions with replicates for *A. niger* (complete list of information in Suppl. Mat. S1).

3. Results

FunGeneClusterS is implemented as a graphical interface to predict co-regulated genes. The functionality of this interface has been developed to provide maximum visibility and data availability to the user. The application has five tab panels at the top representing different types of data and information. Fig. 1 shows the data upload page of the application.

3.1. Graphical user interface

3.1.1. Introduction

Here the user will find a brief introduction to the method itself as well as screen-shots of figures, results and literature references. This page also provides detailed descriptions of the input data format. Source code and example data files are available from this page.

3.1.2. Upload and settings

The user must provide a gene annotation file and a file containing transcriptome data. Gene names used in these two data files must match or will otherwise be removed from the data. The files are selected through the file browsing system of the computer. Each setting has a default value, window-size (3), gene skipping (0) and correlation type (Pearson), which are the same parameters as used in our previous work.[4] As the analysis below shows, these are still the best parameters for most co-regulated clusters. The correlation types include two non-parametric measures (Spearman and Kendall) and one parametric (Pearson).

3.1.3. R data

Offers an R-like view of the different data structures created during the analysis. Each button will trigger the construction and display of a new data frame. This tab is intended for illustrative purposes and provides explanation for each step in the calculation.

3.1.4. View datasets

Drop down menu revealing menus which all display a table of values in each data structure. The tables can be searched and are wrapped in individual pages to make them easier to view in a browser. Each page offers a download of the constructed data.

3.1.5. View plots

Drop down menu offering a number of different illustrations of the provided and computed data:

3.1.5.1. Annotation. A histogram view of the annotation data shows how many genes are identified on each contig of the genome sequence.

3.1.5.2. Expression. A boxplot view of the expression data shows the maximum, minimum and quantile values of the expression signals for each experiment.

3.1.5.3. Random quantiles. The plot shows the quantile distributions of your real data versus the pseudo-random data. The 0.95 value of the Random Quantile table is the value with a 5% false discovery rate. These values are used as cutoffs for identifying gene clusters. As the cutoff for the false discovery rate is determined by a random scrambling of the expression data, the cutoff will vary slightly between analysis runs. This might result in slight differences in cluster predictions.

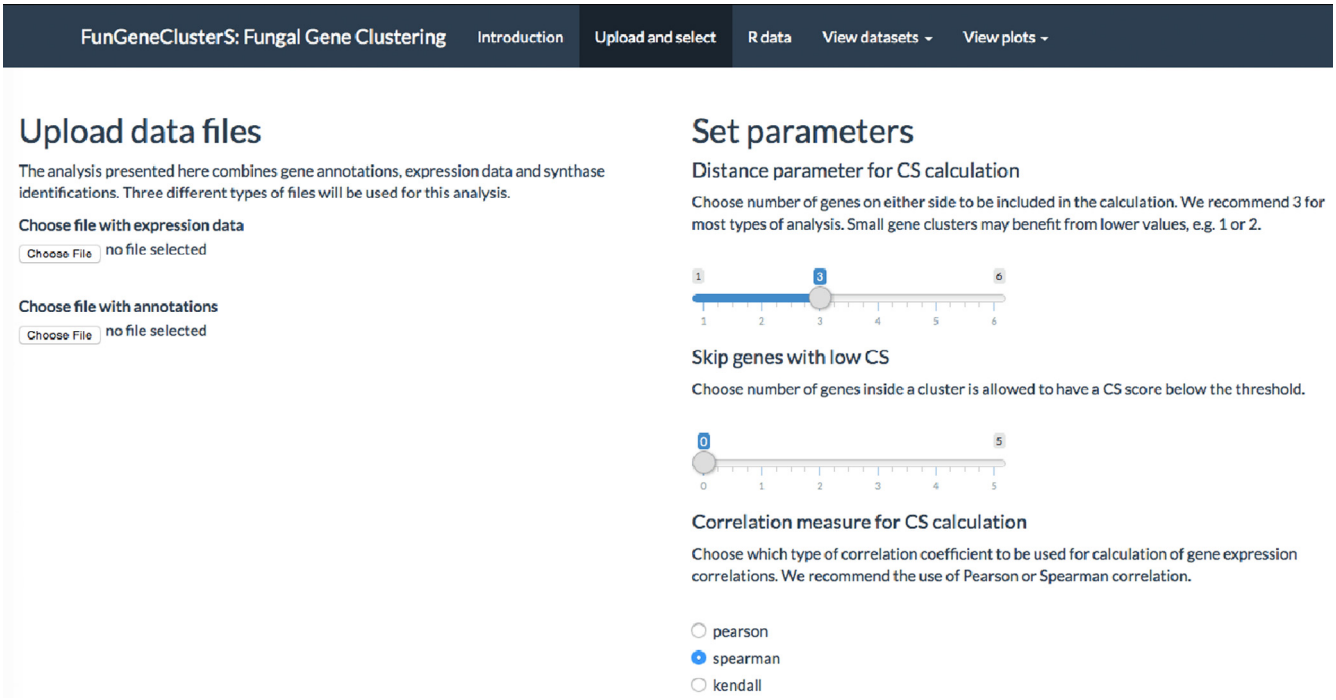

Fig. 1. Data upload page of the graphical interface for FunGeneClusterS. Note that Shiny is currently not fully compatible with Safari. We recommend using Google Chrome or Firefox. The panels for other options can be seen at the top.

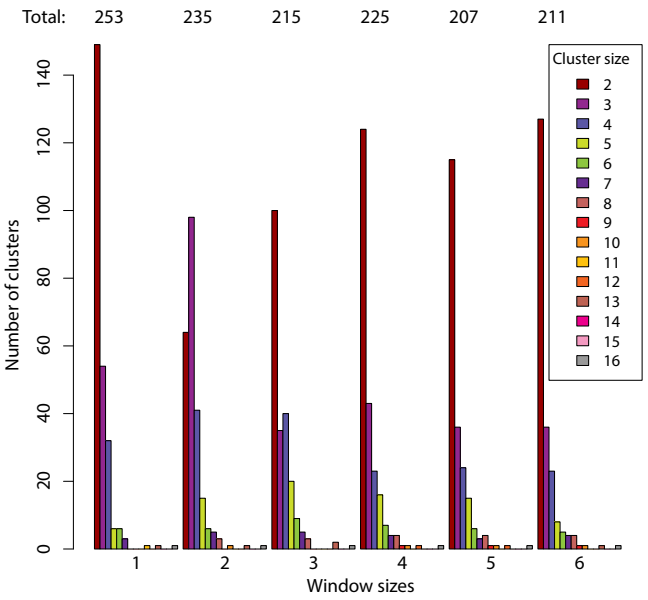

Fig. 2. Prediction of co-regulated gene clusters in *A. nidulans* using FunGenClusterS. Histograms for each window size (*w* values 1–6) show the number of clusters of different lengths (number of genes in a predicted cluster) identified using Pearson correlation and not allowing gene skipping. With a window size of one there are more clusters with size 1. The largest number of genes in a cluster was 16.

3.1.5.4. Cluster scores. The plot represents the CS values for each gene plotted across the individual contig/chromosome. A length distribution shows the size of the predicted clusters. The application displays a condensed view of all the contigs and will not be useful for manual detection of clusters. Instead, a result bundle is made available for download on the same page. This bundle creates a multipage PDF file for each contig with each page representing a high CS region. Each page displays the CS curve around every single possible cluster gene. The bundle also contains a table with gene names and CS scores for each gene above the CS cutoff. These figures and tables represent the predicted clusters. Note that the table contains all genes with a CS above the threshold. This also includes single genes with high scores but these are not regarded as clusters.

The interface will make it possible for users to perform a complete analysis with data files and figures as well as provide the opportunity to create new figures from the supplied data files. The effect of the added options for window size and gene skipping will be investigated in the following.

3.2. Benchmark of FunGeneClusterS with original data from Aspergillus nidulans

The first implementation of this method proved accurate in predicting known gene clusters involved in secondary metabolism by combining automatic cluster prediction and manual curation of CS graphs. A first step was thus to evaluate the performance for this part of the functionality.

The effect of window size was tested on the *A. nidulans* data using values between 1 and 6 and three correlation coefficient types (Fig. 2 uses Pearson correlation and Suppl. Mat. S2 shows results for Kendall and Spearman). It is particularly observed that the two non-parametric methods (Kendall and Spearman) are sensitive to a window size of 1. The predicted number of clusters range from 462/405 to 231/216 clusters as the window increases. The difference between the two is minimal. In terms of size, the largest cluster is predicted to consist of 9 genes and can be identified at window sizes between 3 and 6 (3 and 6 for Kendall, 3–6 for Spearman). For all methods, a window size of 2 is most efficient at predicting clusters of size 3, whereas clusters of size 4 and above are found in the largest numbers using a $w = 3$. Larger clusters can be identified using the Pearson correlation (up to 16 genes) and this method is the most robust to the window size (253 to 211 clusters for window 1 and 6) (Fig. 2).

We have compared the predictions of gene clusters using Pearson and Spearman-based CS to the set of backbone genes predicted by SMURF for *A. nidulans* (Suppl. Mat S3). For $w = 3$, Pearson predicts the presence of 20 out of 58 backbones (34%), whereas Spearman only predicts 14%. Pearson is therefore clearly superior to Spearman, and predicts a large number of the SMURF clusters.

Overall, the improved method performs comparably to the original with the parameter skipping set to 0 and using a Pearson correlation. It is worth noting that the predicted clusters are not limited to secondary metabolism, but other co-expressed clusters in the genome are also found. This makes the method attractive for finding co-regulated biosynthetic or catabolic pathways of any type.

Due to examination of the predicted clusters and prior experience (see above), we were interested in including genes within the clusters, which did not have CS above the threshold (gene skipping). We tested the effect of allowing clusters to skip single genes without correlation values above the threshold for both Pearson and Spearman correlation, with window sizes of 2 or 3 (Suppl. Mat. S3). Here, the overall picture remains the same for the accuracy of detecting clusters with secondary metabolite genes. Pearson identifies 2–3 times as many of the SMURF clusters as Pearson, and includes a lower total number of genes, suggesting a higher accuracy in prediction. However, allowing a skip of one, large gene clusters are more accurately joined. One example is the sterigmatocystein cluster, which consists of at least 25 genes.[19] While the full cluster is not predicted without skipping in our previous method, adding skipping allows the joining of all members of this cluster in analysis with *w* set to 2 or 3, and skipping set to 1. For other known gene clusters, e.g. the cluster coding for the biosynthesis of yanuthone D,[20] the cluster boundaries was most accurately predicted using a window size of 2 and a skip of 1. Furthermore, allowing a skip also correctly predicted a gene cluster around the genes AN2032 and AN2035, which had to be manually curated in the previous version. Thus for detection of large clusters, it seems to be favorable to add the option of skipping 1 gene.

3.3. Application of the method to A. niger *identifies more than 400 clusters, within both primary and secondary metabolism*

FunGeneClusterS was also tested on *A. niger* using a set of 55 transcriptome experiments. The data was scanned using window sizes between 1 and 6 and three correlation coefficient types (Suppl. Mat. S4). We observe the same general trends as for the *A. nidulans*: The two non-parametric methods (Spearman and Kendall) have minimal differences, and are very sensitive at window size of 1 and range from around 605/619 to 335/338 clusters as the window increases. Pearson predicts a slightly larger number of clusters for all window sizes, ranging from 660 to 397 clusters. The effect of different window sizes on the length of individual predicted clusters can be observed in the raw correlation scores (Suppl. Mat. S8). For all correlation coefficients, $w = 2$ is best for predicting clusters of size 3. Performance is very similar for window sizes from 3 to 6, however, $w = 3$ gives the highest number of predicted cluster for sizes of 4 and above. Thus, the method seems to have robust performance between datasets.

The impact of the dataset size on the number of predicted clusters has been evaluated using a window size of 3 and the Pearson correlation coefficient. Replicates have been averaged and the resulting unique experimental conditions have been sampled by randomly removing experiments from the dataset in a sequential

manner (Suppl. Mat. S6). This analysis suggests that there is no added value from having more than 9 experiments for the current dataset.

For a more detailed analysis of clusters in *A. niger*, we evaluated cluster predictions in a data set including Pearson and Spearman correlation, window sizes of 2 and 3, and skipping values of 0 and 1 (Suppl. Mat. S5). The predicted clusters were compared to a set of 80 synthases predicted by SMURF. Here we see a verification of the analysis generated for *A. nidulans*, in that the Pearson-based calculations predict 24–28 of backbone genes, whereas the Spearman-based method only predicts 12–15 of the backbones depending on the method. If the clusters are evaluated based on any gene in a cluster and not the prediction of the cluster backbone, 56% of the SMURF clusters can be identified using FunGeneClusterS (45 clusters, window size 1, no gene skipping, Pearson) (see Suppl. Mat. S7). As the window size gets larger, fewer clusters are predicted and at $w = 6$ only 35% of the clusters are identified. This effect could be attributed to only a few genes in a SMURF cluster being co-expressed. When the window grows, parts of the cluster do not show a high CS score. The total number of predicted clusters is much higher than using SMURF, which is to be expected as other types of gene clusters are expected to be co-expressed.

Again, we see an improvement of predictions with skipping added. Fig. 3A shows the CS graph of a cluster that is predicted to be two clusters of 9 and 2 genes respectively, but these are joined when allowing for one gene skipped. The cluster is predicted to contain a polyketide synthase (Gene ID 44965). In 3B, the expression values are plotted to mark how similar the expression profiles of the two are.

Examining the predicted clusters, we were also able to identify new clusters associated with secondary metabolism, which were not predicted in the original set of predicted synthases. The largest predicted cluster using Pearson, $w = 3$, skipping set to 0, contains 16 members (Gene IDs 202333, 42743, 123782, 211875, 211877, 124807, 211885, 140623, 131352, 188056, 54123, 54124, 54125, 54126, 54127, and 42759). Looking closer at the gene annotation, gene 211885 seems to be a putative PKS, and the remaining cluster members all seem to have activities related to secondary metabolism. Other examples not found in the SMURF predictions include six genes putatively assigned to fumonisin biosynthesis, with gene ID 205913 being the putative polyketide synthase. Thus, this method can, unsupervised, detect new secondary metabolite gene clusters.

However, with 432 predicted co-regulated clusters (spanning 1,323 genes) for the parameters described above, it is clear that not all of these clusters can be associated with secondary metabolism. Scanning the predicted clusters, we also find a number of clusters which seem to be related to primary metabolism. One interesting example is seven genes including the biotin synthase gene (*bioA*, Gene ID 51633), which all seem to be involved in biotin biosynthesis. This feature has been reported to be the case for *A. nidulans*,[21] but to our knowledge, has not been reported for *A. niger*. Fig. 4 shows the synteny of the gene cluster around *bioA* as seen in 17 other *Aspergillus* genomes. Note that the gene cluster seems to be scrambled in Section *Fumigati* (*A. fumigatus* and *A. fischerianus*/*Neosartorya fischeri*), but the genes are still co-located.

Another example of co-transcribed pathways from primary metabolism is a gene cluster of five members (Gene IDs 38849, 199148, 38851, 38852, and 199151), which all have predicted activities involved in degradation of aromatic compounds. This could be a part of the homogentisate pathway, or biosynthesis of a pigment. In addition, we also see the interesting finding that four genes around the oxaloacetate acetylhydrolase gene (*oahA*) are co-regulated. This is interesting as OahA is responsible for the oxalic acid hypersecretion phenotype characteristic of *A. niger* and related black Aspergilli. In summary, we also find that parts of primary metabolism are both clustered and co-expressed and can be detected using this method.

4. Discussion

The data presented here illustrates the unsupervised identification of gene clusters in two Aspergillus species. The implementation offers options for window size, skipping of genes in the cluster, and correlation method to provide the best possible opportunity for the user to adjust the method to the data. The analysis here shows that all of these parameters can have a great impact on the predicted clusters. For *A. nidulans*, there was a great change in cluster number (253 to 402) when changing the correlation method from Pearson to Spearman, but for *A. niger*, the change was negligible for $w = 1$ (660 to 605, Suppl. Mat. S4). However, for $w = 3$, the number of predicted clusters is higher for Pearson (453 versus 376 with Spearman). This shows that the method performance is dependent on the provided data.

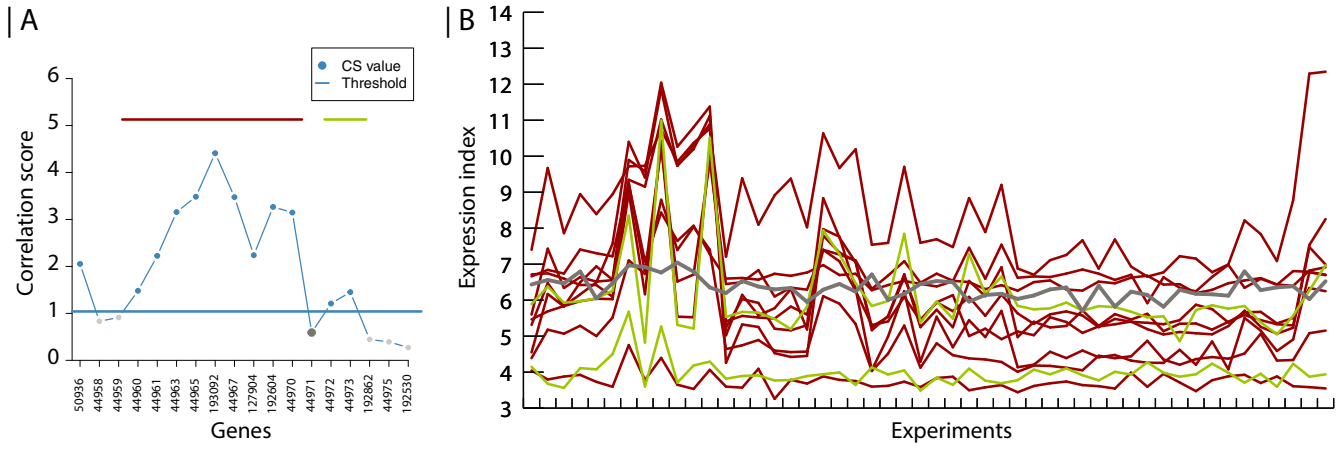

Fig. 3. Co-expressed gene predictions for *A. niger* using FunGeneClusterS. A: Clustering scores for a 12 member gene cluster predicted using window size 3, Pearson correlation and a skipping value of 1. B: Expression values for all members of the joined cluster. Each line represents a gene with the experiments shown horizontally. The 9 member gene cluster are shown as red lines, the 2 member gene cluster is colored in green, and the expression values for the skipped gene are marked in gray. Note how the predicted gene cluster shows very similar expression profiles across the experiments.

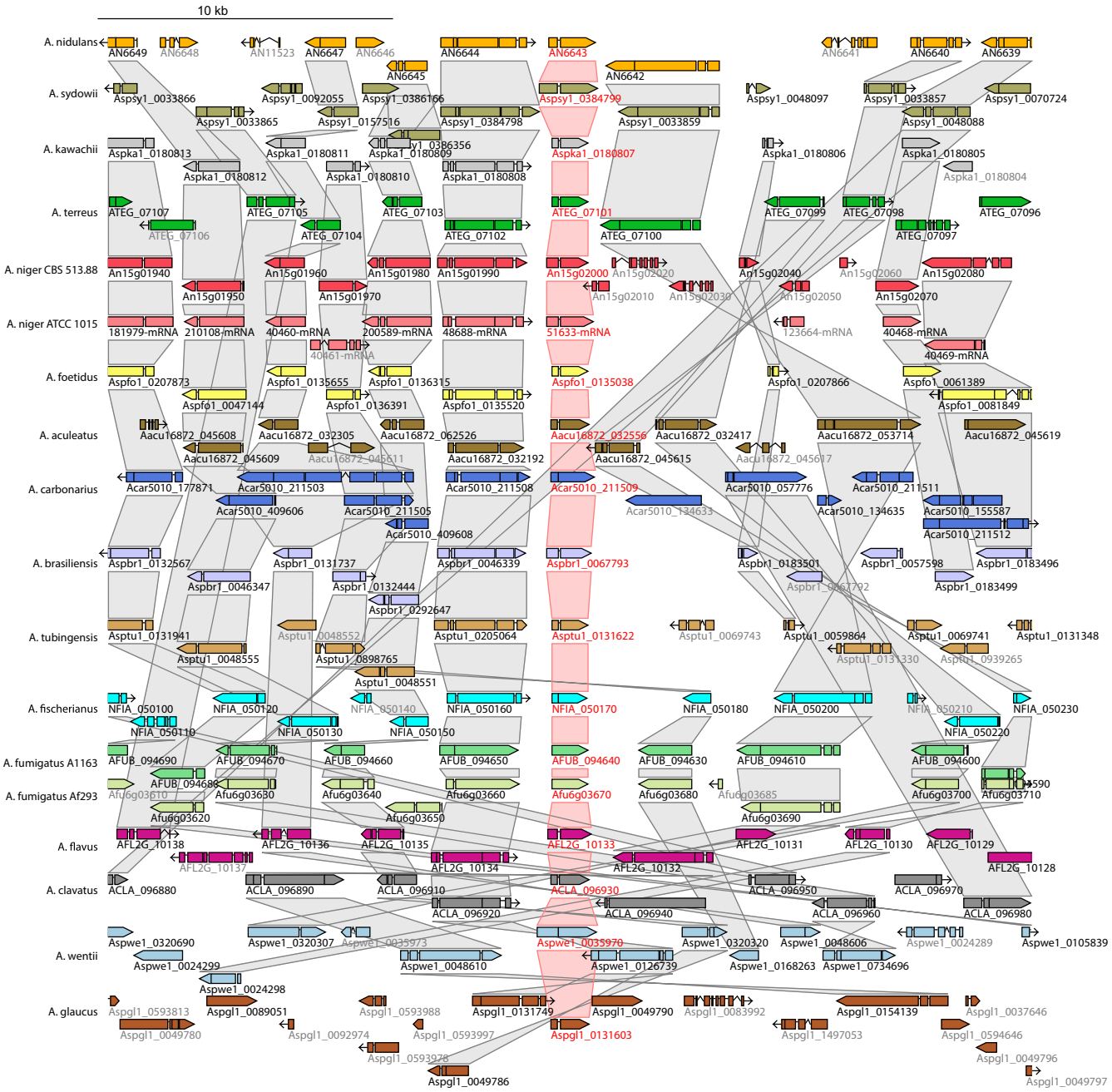

Fig. 4. Synteny map of region around the biotin synthase in 18 *Aspergillus* genomes. The map was generated using AspGD.org[22] including all available genomes, except the distantly related *A. zonatus*, which did not show clustering of biotin biosynthesis.

The window size can be set to predict very large clusters although this might very well result in false positives. Using a smaller window, on the other hand, has the disadvantage of predicting very small clusters. By changing this parameter from 1 to 6, it was possible to reduce the number of clusters by 50% in most cases and in the same process, the number of gene clusters with only two members was also reduced with 50% or more. Clearly, with the two data sets examined here, Pearson correlation performs the best in identification of secondary metabolite gene clusters.

The last parameter to be tested was the option to skip a number of genes with low CS to combine smaller clusters. This proved to merge some single genes with high CS scores into clusters, thereby expanding the cluster boundaries. Combining window sizes and gene skipping will offer the user the option to elongate or shorten clusters depending on the point of interest and data. In general, several analyses should be performed to find the most optimal set of parameters, preferably by comparison to one or more characterized gene clusters in the genome to identify best sets.

In this implementation, we also changed the way the method deals with genes at the end of each contig. Although no examples of this problem were found in the current data, some examples might be found where a cluster spans two contigs. In the previous method, the ends of contigs were simply treated as 0; here they are given a value based on the up or downstream genes. Although this will not ensure the detection of cross-contig clusters, it will show a signal for those genes. In the case where full chromosomes have been assembled, this is of high importance, as it is known that gene clusters are often found at the telomeric ends of chromosomes.

The graphical interface of this implementation also offers a number of output files and plots which makes analysis of the results easier. The clusters are provided as a simple text file as well as a multipage PDF plot. This plot shows each region of the chromosome where a gene with a high CS value has been detected. One file is produced per scaffold in the analysis. These plots offer a great opportunity to visually inspect the identified clusters and refine the parameters of the analysis if the predicted clusters are not as desired.

In the detailed analysis of cluster predictions for *A. niger* ATCC 1015, we have predicted 404 gene clusters in the genome of using Pearson correlation, $w = 3$, and a skipping allowance of 1 (Suppl. Mat. S5). Here, we find the best prediction of secondary metabolite gene clusters based on the number of secondary metabolite synthetases found (28), but some gene clusters are only identified with $w = 2$. In fact, combining the predictions for $w = 2$ and $w = 3$ gives a total number of 31 out of 76 clusters predicted, which is an even better performance than we saw for the *A. nidulans* data.

Generally, we don't predict all backbone genes in either data set; we actually predict less than 50%. While this may seem as a low rate, it is not feasible to expect that all backbones would be predicted here, as several of these genes will be silent under the tested conditions (roughly 1/3). Additionally, for many genes (in particular NRPSs and DMATS), there will not be co-regulated tailoring genes next to the synthases/synthetases, in which case our cluster prediction will not detect them.

Interestingly, our analysis also identified multiple gene clusters which seem to be involved with primary metabolism. This trait has been observed previously at the genome level for e.g. *Saccharomyces cerevisiae*,[23] however, here we also see it at the transcriptional level (co-transcription). As the examples above shows, in particular it seems to be activities involved in metabolism of aromatic compounds. A related observation has been made in another *Aspergillus* species, where the degradative pathway of the aromatic acid tyrosine appears to be clustered.[24] We also see – as in the case of oxalic acid – non-growth related compounds that are secreted in large amounts similar to secondary metabolites. Indeed, a recent review[25] suggested – based on genomics data – that gene clusters of metabolic enzymes to be a general phenomenon in fungi, a trait that we see here to also hold true for *A. niger* at the transcriptional level.

5. Conclusions

Here we have presented a new and optimized version of the cluster finding method published by Andersen et al. in 2013.[4] The implementation includes improvements to the computational setup to increase speed, addition of parameter settings of window size and correlation method, introduction of the option to include genes with low scores to a cluster of high scoring genes, and a graphical interface. The graphical interface both makes the method more accessible to users, and outputs a large number of the key predictions as graphs.

We have shown that the additional options allow for a more fine-tuned analysis and that larger clusters can be identified using the skip option. This allows for errors in expression data to have less effect on the analysis as a whole. Recommended starting points are Pearson correlation, $w = 3$, skip = 1, but for some gene clusters, $w = 2$ and skip = 0 can be more accurate. The application is available through https://fungiminions.shinyapps.io/FunGeneClusterS.

In analyzing new data for *A. niger*, we have not only generated cluster predictions for 31 secondary metabolite clusters based on predictions of the backbone, but for an additional 371 gene clusters. Several of these seem to be additional secondary metabolite gene clusters, which have not been predicted with previous methods. Furthermore, we predict – to the best of our knowledge – for the first time, gene clusters in *A. niger* involved in primary metabolism; in particular, clusters involved in biotin biosynthesis, degradation of aromatics, and biosynthesis of oxalic acid.

In summary, we present a versatile method for gene cluster prediction in fungi based on transcriptome data, which can accurately predict co-regulated clusters not limited to secondary metabolism.

Appendix: Supplementary material

Supplementary data to this article can be found online at doi:10.1016/j.synbio.2016.01.002.

References

1. Newman DJ, Cragg GM. Natural products as sources of new drugs over the 30 years from 1981 to 2010. *J Nat Prod* 2012;**75**(3):311–35. doi:10.1021/np200906s; http://dx.doi.org/10.1021/np200906s.
2. Liu T, Chiang Y, Somoza AD, Oakley BR, Wang CC. Engineering of an "Unnatural" natural product by swapping polyketide synthase domains in Aspergillus nidulans. *J Am Chem Soc* 2011;**133**(34):13314–6. doi:10.1021/ja205780g; http://dx.doi.org/10.1021/ja205780g.
3. Keller NP, Turner G, Bennett JW. Fungal secondary metabolism – from biochemistry to genomics. *Nat Rev Microbiol* 2005;**3**(12):937–47. doi:10.1038/nrmicro1286; http://dx.doi.org/10.1038/nrmicro1286.
4. Andersen MR, Nielsen JB, Klitgaard A, Petersen LM, Zachariasen M, Hansen TJ, et al. Accurate prediction of secondary metabolite gene clusters in filamentous fungi. *Proc Natl Acad Sci U S A* 2013;**110**(1):E99–107. doi:10.1073/pnas.1205532110; http://dx.doi.org/10.1073/pnas.1205532110.
5. Khaldi N, Seifuddin FT, Turner G, Haft D, Nierman WC, Wolfe KH, et al. SMURF: genomic mapping of fungal secondary metabolite clusters. *Fungal Genet Biol* 2010;**47**(9):736–41. doi:10.1016/j.fgb.2010.06.003.SMURF; http://dx.doi.org/10.1016/j.fgb.2010.06.003.SMURF.
6. R. Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2015.
7. Dowle M, Srinivasan A, Short T, Lianoglou S, Saporta R, Antonyan E. with contributions from R. Saporta, E. Antonyan, Data.table: Extension of Data.frame, r package version 1.9.6, 2015. http://CRAN.R-project.org/package=data.table.
8. Warnes GR, Bolker B, Bonebakker L, Gentleman R, Liaw WHA, Lumley T, et al., gplots: Various R Programming Tools for Plotting Data, r package version 2.17.0, 2015. http://CRAN.R-project.org/package=gplots.
9. Wickham H. Reshaping data with the reshape package. *J Stat Softw* 2007;**21**(12):1–20. http://www.jstatsoft.org/v21/i12/.
10. Wickham H. ggplot2: elegant graphics for data analysis. New York: Springer; 2009 http://had.co.nz/ggplot2/book.
11. Chang W, Cheng J, Allaire J, Xie Y, McPherson J. shiny: Web Application Framework for R, r package version 0.12.2, 2015. http://CRAN.R-project.org/package=shiny.
12. Galagan JE, Calvo SE, Cuomo C, Ma L-J, Wortman JR, Batzoglou S, et al. Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae. *Nature* 2005;**438**(7071):1105–15. doi:10.1038/nature04341; http://www.ncbi.nlm.nih.gov/pubmed/16372000.
13. Pel HJ, de Winde JH, Archer DB, Dyer PS, Hofmann G, Schaap PJ, et al. {G}enome sequencing and analysis of the versatile cell factory {A}spergillus niger {CBS} 513.88. *Nat Biotechnol* 2007;**25**(2):221–31.
14. Andersen MR, Salazar MP, Schaap PJ, van de Vondervoort PJ, Culley D, Thykaer J, et al. Comparative genomics of citric-acid-producing Aspergillus niger ATCC 1015 versus enzyme-producing CBS 513.88. *Genome Res* 2011;**21**(6):885–97. doi:10.1101/gr.112169.110; http://dx.doi.org/10.1101/gr.112169.110.
15. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. *BMC Bioinformatics* 2009;**10**:421.
16. Andersen MR, Vongsangnak W, Panagiotou G, Salazar MP, Lehmann L, Nielsen J. A trispecies Aspergillus microarray: comparative transcriptomics of three aspergillus species. *Proc Natl Acad Sci U S A* 2008;**105**(11):4387–92. doi:10.1073/pnas.0709964105; http://dx.doi.org/10.1073/pnas.0709964105.
17. Andersen MR, Giese M, de Vries RP, Nielsen J. Mapping the polysaccharide degradation potential of Aspergillus niger. *BMC Genomics* 2012;**13**:313. doi:10.1186/1471-2164-13-313; http://dx.doi.org/10.1186/1471-2164-13-313.
18. Andersen MR, Lehmann L, Nielsen J. Systemic analysis of the response of Aspergillus niger to ambient pH. *Genome Biol* 2009;**10**(5):R47. doi:10.1186/gb-2009-10-5-r47; http://dx.doi.org/10.1186/gb-2009-10-5-r47.
19. Brown DW, Yu JH, Kelkar HS, Fernandes M, Nesbitt TC, Keller NP, et al. Twenty-five coregulated transcripts define a sterigmatocystin gene cluster in Aspergillus nidulans. *Proc Natl Acad Sci U S A* 1996;**93**(4):1418–22.
20. Holm DK, Petersen LM, Klitgaard A, Knudsen PB, Jarczynska ZD, Nielsen KF, et al. Molecular and chemical characterization of the biosynthesis of the 6-MSA-derived meroterpenoid yanuthone D in Aspergillus niger. *Chem Biol* 2014;**21**(4):519–29. doi:10.1016/j.chembiol.2014.01.013; http://dx.doi.org/10.1016/j.chembiol.2014.01.013.
21. Magliano P, Flipphi M, Sanglard D, Poirier Y. Characterization of the Aspergillus nidulans biotin biosynthetic gene cluster and use of the bioDA gene as a new transformation marker. *Fungal Genet Biol* 2011;**48**(2):208–15. http://www.ncbi.nlm.nih.gov/pubmed/20713166.

22. Arnaud MB, Cerqueira GC, Inglis DO, Skrzypek MS, Binkley J, Chibucos MC, et al. The Aspergillus Genome Database (AspGD): recent developments in comprehensive multispecies curation, comparative genomics and community resources. *Nucleic Acids Res* 2012;**40**(Database issue):D653–9. doi:10.1093/nar/gkr875; http://dx.doi.org/10.1093/nar/gkr875.
23. Lee JM, Sonnhammer ELL. Genomic gene clustering analysis of pathways in eukaryotes. *Genome Res* 2003;**13**(5):875–82. doi:10.1101/gr.737703; http://www.genome.org/cgi/doi/10.1101/gr.737703.
24. Greene GH, McGary KL, Rokas A, Slot JC. Ecology drives the distribution of specialized tyrosine metabolism modules in fungi. *Genome Biol Evol* 2014;**6**(1):121–32. doi:10.1093/gbe/evt208.
25. Wisecaver JH, Rokas A. Fungal metabolic gene clusters-caravans traveling across genomes and environments. *Front Microbiol* 2015;**6**:161. doi:10.3389/fmicb.2015.00161; http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4347624&tool=pmcentrez&rendertype=abstract.