Bioinformatics, 35(17), 2019, 3178–3180
doi: 10.1093/bioinformatics/btz021
Advance Access Publication Date: 18 January 2019
Applications Note

OXFORD

Gene expression

METACLUSTER—an R package for context-specific expression analysis of metabolic gene clusters

Michael Banf[1,2,*], Kangmei Zhao[1] and Seung Y. Rhee[1,*]

[1]Department of Plant Biology, Carnegie Institution for Science, Stanford, CA 93405, USA and [2]EducatedGuess.ai, Siegen, Germany

*To whom correspondence should be addressed.
Associate Editor: Oliver Stegle

Received on June 19, 2018; revised on November 22, 2018; editorial decision on December 30, 2018; accepted on January 14, 2019

Abstract

Summary: Plants and microbes produce numerous compounds to cope with their environments but the biosynthetic pathways for most of these compounds have yet to be elucidated. Some biosynthetic pathways are encoded by enzymes collocated in the chromosome. To facilitate a more comprehensive condition and tissue-specific expression analysis of metabolic gene clusters, we developed METACLUSTER, a probabilistic framework for characterizing metabolic gene clusters using context-specific gene expression information.

Availability and implementation: METACLUSTER is freely available at https://github.com/mbanf/METACLUSTER.

Contact: michael@educatedguess.ai or srhee@carnegiescience.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Plants and microbes produce a vast array of compounds called specialized or secondary metabolites to cope with their environments but the biosynthetic pathways for many of these compounds have not yet been elucidated (Wink, 2010). Recent studies in plants (Chae *et al.*, 2014; Kautsar *et al.*, 2017; Schlapfer *et al.*, 2017; Töpfer *et al.*, 2017; Wisecaver *et al.*, 2017) revealed a widespread occurrence of metabolic enzymes that collocate in the chromosome. This offers an intriguing possibility for uncovering new biosynthetic pathways encoded by these metabolic gene clusters. To this end, co-expression analysis can provide valuable insights as characterized specialized metabolic pathways and gene clusters show a high degree of co-expression among their enzymes (Kautsar *et al.*, 2017; Schlapfer *et al.*, 2017; Yu *et al.*, 2016). Moreover, the expression patterns of experimentally characterized gene clusters indicate spatial and condition specificity, such as enzymatic genes of clusters synthesizing terpenes in *Arabidopsis thaliana* and *Lotus japonicus* (Field and Osbourn, 2008; Field *et al.*, 2011; Krokida *et al.*, 2013; Yu *et al.*, 2016). Thus, general co-expression analyses integrating a diverse range of experimental treatments and tissue types may mask condition- and tissue-specific co-expression among enzymatic genes within a cluster (Obayashi *et al.*, 2011). Some cluster prediction algorithms, such as plantiSMASH, can compute the co-expression for genes within predicted gene clusters, given user-provided gene expression data (Kautsar *et al.*, 2017). However, these algorithms do not autonomously distinguish tissue types or experimental conditions if integrated gene expression datasets are used. To facilitate a convenient and context-specific analysis of metabolic gene clusters, we present a probabilistic framework, called METACLUSTER, which automatically identifies conditions and tissues associated with inferred gene clusters within a given differential gene expression compendium. METACLUSTER can be applied to any organism, gene cluster descriptions and differential gene expression datasets, thereby providing a valuable complementary framework to augment gene cluster inference approaches, such as PlantClusterFinder (Schlapfer *et al.*, 2017), antiSMASH (Blin *et al.*, 2017), plantiSMASH (Kautsar *et al.*, 2017) and PhytoClust (Töpfer *et al.*, 2017), with additional layers of automated high-resolution functionality inference.

2 Materials and methods

2.1 Pre-processing expression data and conditions

We constructed a differential gene expression dataset by retaining only the data from transcriptome experiments with various treatments, and computing the log of fold change difference between the mean of the treatment and control sample replicates. We performed two sample *t*-tests per gene on each of the experiments to evaluate the significance of a gene's differential expression between treatment and control, producing a ternary matrix D over all genes. For each gene and experimental treatment, we defined an entry in D, and assigned 1, −1 or 0 for significant ($P < 0.05$) up-, down- or non-significant differential expression. Furthermore, we assigned all experimental treatments to manually defined condition c and tissue t group combinations.

2.2 Co-differential expression and co-expression analyses between pairs of genes using Monte Carlo simulation

To generate a context-specific gene pair co-expression dataset, we first built upon an idea originally proposed by Less *et al.* (2011) to identify gene pairs with a statistically significant number of shared experimental treatments, within which both genes are differentially expressed with similar directionality using the ternary matrix D. To determine a significance threshold for this co-differential expression between a cluster gene pairs, we calculated a distribution of co-differential expression of gene pairs by chance via shuffling all entries in D independently per gene. We defined the significance threshold as the 95th percentile of this distribution, corresponding to an empirical P-value ≤ 0.05. This way, we obtained: (i) gene pairs to be considered for further metabolic gene cluster analysis, and (ii) the corresponding experimental treatments for further analysis and annotation. Next, we computed the Pearson's correlation coefficient (pcc) between all significantly co-differentially expressed metabolic gene pairs in a gene cluster, using only shared experimental treatments that were identified per enzyme pair in the previous step. To define appropriate significance thresholds, we established a by chance co-expression distribution, considering the numbers of shared condition-tissue sets between enzymatic gene pairs. To this end, we identified the number of shared condition-tissue sets per enzyme pair and computed Pearson's correlation coefficients of 100 randomly selected gene pairs from the same selected experimental subsets. Based on these by chance correlation values, we constructed a distribution $P_{\text{random}}(pcc)$. Again, we selected the 95th percentile of this distribution, corresponding to an empirical P-value ≤ 0.05, to infer significant condition-specific co-expression between pairs of genes.

2.3 Context-specific transcriptional activity analysis of gene clusters

We proposed a probabilistic framework for each gene cluster gc to score its likelihood of being transcriptionally active in a given condition and tissue pair (c, t). Here, context-specific transcriptional activity of a cluster gc is defined based on four types of transcriptional behavior of the gc (see Supplementary Material): (i) co-differential expression among gc's genes, (ii) co-expression among gc's genes, (iii) the probability of gc to be transcriptionally active in condition c and (iv) the probability of gc to be transcriptionally active in tissue t given condition c. These four P-values were combined as $p_{gc\in(c,t)}$ using Fisher's method (Li *et al.*, 2014). In case a tissue could not be identified for condition c, we kept the cluster's condition annotation but referred to the tissue as non-specific, disregarding $p_{gc}(t|c)$. Finally, only the gene clusters gc with $p_{gc\in(c,t)} \leq 0.05$ were considered transcriptionally active.

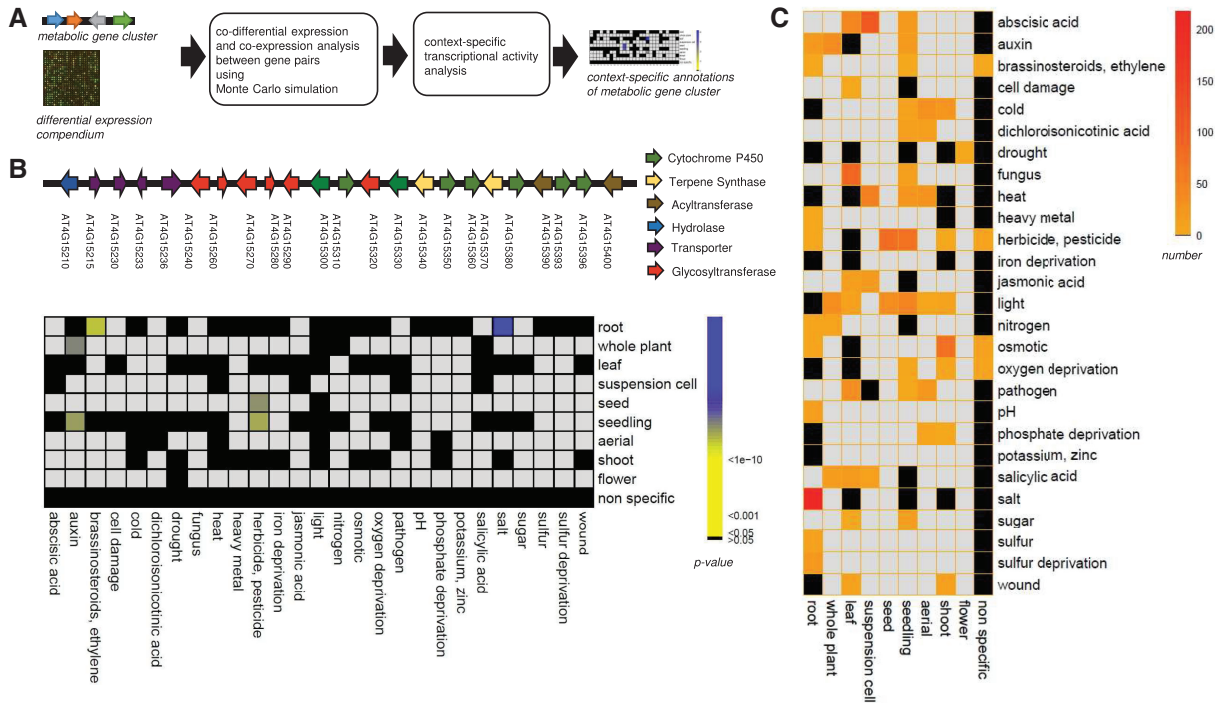

Fig. 1. (A) The METACLUSTER framework. (**B**) Cluster diagram and transcriptional activity map of the arabidiol/baruol cluster (Yu *et al.*, 2016) (C463 based on the prediction in Schlapfer *et al.*, 2017). Colors indicate the inferred P-value of the cluster to be transcriptionally active per condition and tissue. Gray tiles indicate condition-tissue combinations that are missing in the differential expression dataset. (**C**) Transcriptional activity map of the 317 inferred context-specific gene clusters. Color values denote the number of the transcriptionally active gene clusters per condition-tissue. Black tiles indicate condition-tissue combinations with no inferred transcriptionally active clusters

3 Results and discussion

To demonstrate the utility of METACLUSTER, we performed a context-specific expression analysis of metabolic gene cluster predictions acquired from (Schlapfer *et al.*, 2017). We used a recently compiled large-scale gene expression dataset by He *et al.* (2016) with 6057 expression profiles, covering 79.7% of the *A.thaliana* ecotype Columbia genome. We retained 435 experimental treatments represented by 1825 expression profiles measuring gene expression responses of wild-type plants to treatment and control conditions. All 435 experimental treatments were assigned to 27 manually curated conditions and 9 tissues (Fig. 1, see Supplementary Material). Two sample *t*-tests between treated and untreated samples produced the differential expression matrix *D* containing 435 values per gene. We predicted 317 (out of 674) metabolic gene clusters with at least 3 co-differentially expressed enzymes to be transcriptionally active (Fig. 1). In total, 1380 metabolic enzymes in 317 metabolic gene clusters were predicted to be transcriptionally active in specific conditions and tissues. We observed a significant overlap with the 'high-confidence' gene clusters, which were previously supported by the integrated co-expression analysis in Schlapfer *et al.* (2017) (fold change: 1.4, *P*-value: 0.009, hyper-geometric test). In addition, our set of enzymes included 371 signature and tailoring enzymes, which was significantly higher compared with the gene set in Schlapfer *et al.* (2017) (fold change 1.28, *P*-value: 0.0006, fisher's exact test). Furthermore, we recovered all experimentally characterized terpene biosynthetic clusters in Arabidopsis, i.e. the thalianol (Field and Osbourn, 2008), marneral (Field *et al.*, 2011), tirucalla-7, 24-dien-3beta-ol (Boutanaev *et al.*, 2015) and the arabidiol cluster (Yu *et al.*, 2016). These were clusters C641, C628, C615 and C463 in (Schlapfer *et al.*, 2017). Our analysis also revealed that these clusters are highly transcriptionally active in roots, which is consistent with the expression patterns revealed by functional characterization of these clusters and compounds (Field and Osbourn, 2008; Go *et al.*, 2012). Furthermore, they are predicted to be transcriptionally active in abiotic stress, hormone related conditions or pesticide treatments corroborating the suggestions made by (Chu *et al.*, 2011; de Silva *et al.*, 2011; Smith *et al.*, 2018). The inferred transcriptional activity in response to pesticide treatments is of particular interest, since biopesticides based on plant extracts are considered as promising, natural alternatives to conventional synthetic pesticides (Smith *et al.*, 2018). Given its utility, we anticipate METACLUSTER to help guide the experimental validation of gene cluster predictions in order to further our understanding of the chemical diversity of Nature's pharmacopeia.

Acknowledgements

We thank Pascal Schlapfer and Jan Nasemann for insightful discussions and software tests.

Funding

M.B. was supported by the Alexander von Humboldt Foundation. This work was supported by the Carnegie Institution for Science endowment, grants from the National Science Foundation [IOS-1546838, IOS-1026003], Department of Energy [DE-SC0008769, DE-SC0018277] and National Institutes of Health [1U01GM110699–01A1].

Conflict of Interest: none declared.

References

Blin,K. *et al.* (2017) antiSMASH 4.0–improvements in chemistry prediction and gene cluster boundary identification. *Nucleic Acids Res.*, **45**, W36–W41.

Boutanaev,A.M. *et al.* (2015) Investigation of terpene diversification across multiple sequenced plant genomes. *Proc. Natl. Acad. Sci. USA*, **112**, E81–E88.

Chae,L. *et al.* (2014) Genomic signatures of specialized metabolism in plants. *Science*, **344**, 510–513.

Chu,H.Y. *et al.* (2011) From hormones to secondary metabolism: the emergence of metabolic gene clusters in plants. *Plant J.*, **66**, 66–79.

de Silva,K. *et al.* (2011) Arabidopsis thaliana calcium-dependent lipid-binding protein (AtCLB): a novel repressor of abiotic stress response. *J. Exp. Bot.*, **62**, 2679–2689.

Field,B. and Osbourn,A.E. (2008) Metabolic diversification–independent assembly of operon-like gene clusters in different plants. *Science*, **320**, 543–547.

Field,B. *et al.* (2011) Formation of plant metabolic gene clusters within dynamic chromosomal regions. *Proc. Natl. Acad. Sci. USA*, **108**, 16116–16121.

Go,Y.S. *et al.* (2012) Identification of marneral synthase, which is critical for growth and development in Arabidopsis. *Plant J.*, **72**, 791–804.

He,H. *et al.* (2016) Large-scale atlas of microarray data reveals the distinct expression landscape of different tissues in Arabidopsis. *Plant J.*, **86**, 472–480.

Kautsar,S.A. *et al.* (2017) plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic Acids Res.*, **45**, W55–W63.

Krokida,A. *et al.* (2013) A metabolic gene cluster in Lotus japonicus discloses novel enzyme functions and products in triterpene biosynthesis. *New Phytol.*, **200**, 675–690.

Less,H. *et al.* (2011) Coordinated gene networks regulating Arabidopsis plant metabolism in response to various stresses and nutritional cues. *Plant Cell*, **23**, 1264–1271.

Li,Q. *et al.* (2014) Fisher's method of combining dependent statistics using generalizations of the gamma distribution with applications to genetic pleiotropic associations. *Biostatistics*, **15**, 284–295.

Obayashi,T. *et al.* (2011) ATTED-II updates: condition-specific gene coexpression to extend coexpression analyses and applications to a broad range of flowering plants. *Plant Cell Physiol.*, **52**, 213–219.

Schlapfer,P. *et al.* (2017) Genome-wide prediction of metabolic enzymes, pathways and gene clusters in plants. *Plant Physiol.*, **176**, 2583–2583.

Smith,G.H. *et al.* (2018) Terpene based biopesticides as potential alternatives to synthetic insecticides for control of aphid pests on protected ornamentals. *Crop Protect.*, **110**, 125–130.

Töpfer,N. *et al.* (2017) The PhytoClust tool for metabolic gene clusters discovery in plant genomes. *Nucleic Acids Res.*, **45**, 12, 7049–7063.

Wink,M. (2010) *Biochemistry of Plant Secondary Metabolism*. Wiley-Blackwell, New Jersey.

Wisecaver,J.H. *et al.* (2017) A global coexpression network approach for connecting genes to specialized metabolic pathways in plants. *Plant Cell*, **29**, 944–959.

Yu,N. *et al.* (2016) Delineation of metabolic gene clusters in plant genomes by chromatin signatures. *Nucleic Acids Res.*, **44**, 2255–2265.