EducatedGuess.ai
Machine Learning for Life Sciences

A probabilistic framework for context-specific transcriptional regulation analysis of biosynthetic gene clusters

Michael Banf, EducatedGuess.ai, Siegen/Germany
michael@educatedguess.ai

German Conference on Bioinformatics 2021
September 6 - 8 2021

Introduction

Fungi and plants reveal widespread occurrences of metabolic enzymes co-located on the chromosome, some already characterized as being biosynthetic pathways for specialized metabolites, such as terpenes synthesizing enzyme clusters in *Lotus japonicus* and *Arabidopsis thaliana.* These clusters display context-specific co-expression of clustered enzymes, indicating a shared transcriptional response in a spatial and condition specific manner, and co-regulation due to promoter binding by shared transcription factors may be one way to facilitate coordinated expression.

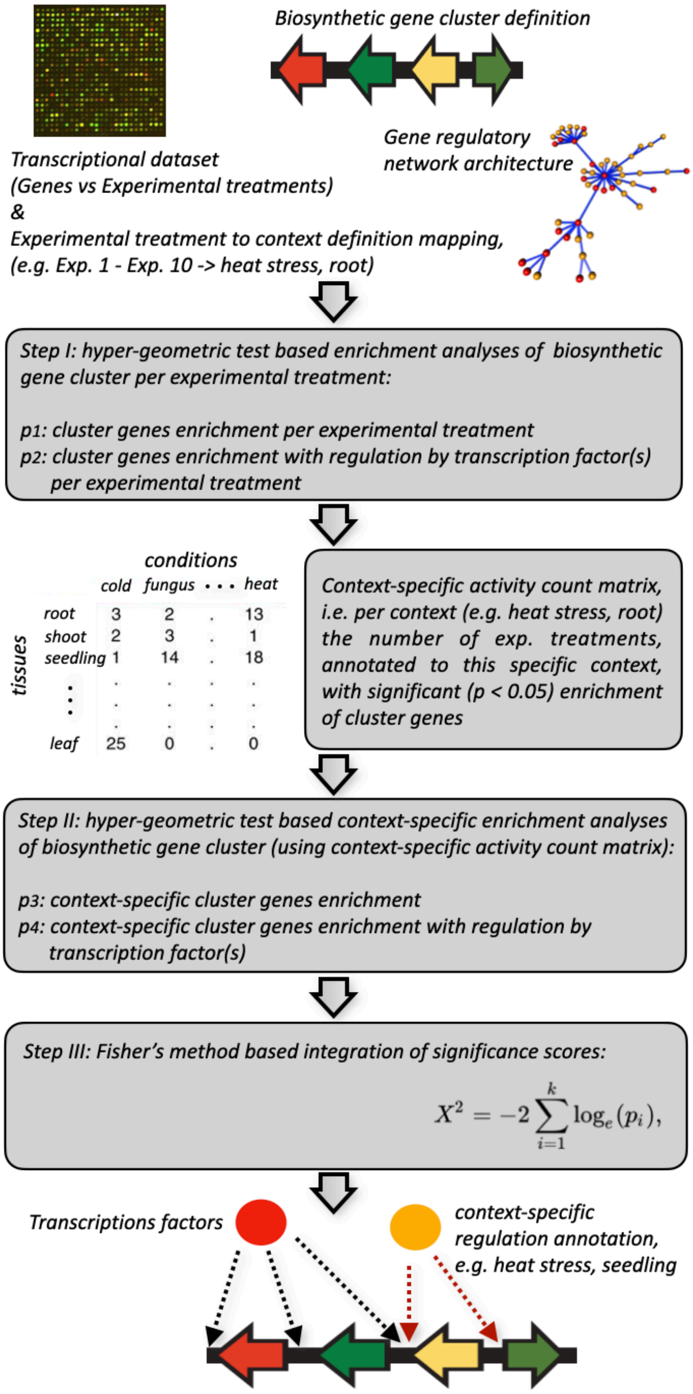
Results

To enhance our understanding of context-specific transcriptional gene cluster regulation, we propose a probabilistic framework integrating gene expression data, context-specific annotations, biosynthetic gene cluster definitions, as well as gene regulatory network architectures. Cluster regulation is then inferred based on a series of statistical analyses, integrated via Fisher's method, including statistical significance scores of metabolic cluster activity in a specific context, metabolic cluster enzyme co-regulation by a transcription factor within that context, as well as optional cluster evidence scores, such as enrichment of signature enzymes per cluster.

An initial analysis of a set of 674 metabolic gene cluster predictions in *Arabidopsis thaliana*, gene expression data of 435 individual experiments, assigned to 27 manually curated conditions and 9 tissues, as well as a DNA-binding prediction based gene regulatory network including 880 transcription factors, inferred 75% of the clusters to be context-specifically active and regulated by, in total, 57% of the candidate regulators. Furthermore, we recovered all characterized terpene clusters with condition and tissue specificity predictions being consistent with expression patterns revealed by functional characterization of these clusters.

In particular, these clusters were statistically enriched for regulation by members of the APETALA2/Ethylene Response Factor (AP2/ERF) family, which is corroborated by research on this transcription factor family's influence on specialized metabolism control in plants.

Methods



Biosynthetic gene cluster definition

Transcriptional dataset (Genes vs Experimental treatments)
&
Experimental treatment to context definition mapping, (e.g. Exp. 1 - Exp. 10 -> heat stress, root)

Gene regulatory network architecture

Step I: hyper-geometric test based enrichment analyses of biosynthetic gene cluster per experimental treatment:

p1: cluster genes enrichment per experimental treatment
p2: cluster genes enrichment with regulation by transcription factor(s) per experimental treatment

		conditions		
tissues	*cold*	*fungus*	• • •	*heat*
root	3	2	.	13
shoot	2	3	.	1
seedling	1	14	.	18
⋮	.	.	.	.
leaf	25	0	.	0

Context-specific activity count matrix, i.e. per context (e.g. heat stress, root) the number of exp. treatments, annotated to this specific context, with significant ($p < 0.05$) enrichment of cluster genes

Step II: hyper-geometric test based context-specific enrichment analyses of biosynthetic gene cluster (using context-specific activity count matrix):

p3: context-specific cluster genes enrichment
p4: context-specific cluster genes enrichment with regulation by transcription factor(s)

Step III: Fisher's method based integration of significance scores:

$$X^2 = -2\sum_{i=1}^{k} \log_e(p_i),$$

Transcriptions factors

context-specific regulation annotation, e.g. heat stress, seedling

Conclusion

Given its utility, we anticipate our approach to efficiently integrate heterogeneous datasets in order to help guide the experimental validation of context-specific metabolic gene cluster transcriptional regulation, thereby allowing for a better understanding and usage of the chemical diversity of Nature's pharmacopoeia.