

METACLUSTER^{plus} - an R package for probabilistic inference and visualization of context-specific transcriptional regulation of biosynthetic gene clusters

Michael Banf

EducatedGuess.ai GbR, 57290 Neunkirchen, Germany

ARTICLE INFO

Keywords:

transcriptional regulation
biosynthetic gene cluster
probabilistic inference
bioinformatics visualization

ABSTRACT

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. To enhance our understanding of context-specific transcriptional gene cluster regulation, we redefine and augment this probabilistic framework, labelled METACLUSTER^{plus}, integrating gene expression compendia, context-specific annotations, biosynthetic gene cluster definitions, as well as gene regulatory network architectures. Further, it provides a set of appealing and intuitive visualizations of inferred results for analysis and publication. METACLUSTER^{plus} is available at <https://github.com/mbanf/MetaclusterPlus>.

1. Introduction

Plants as well as microbial organisms produce a variety of compounds denoted as specialized metabolites to cope with environmental challenges but the biosynthetic pathways for many of these compounds have not yet been elucidated [15]. Recent studies in plants [6, 13, 10, 14, 16] 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 clusters exhibit high degrees of co-expression among their enzymes [13, 10, 17]. Moreover, the expression patterns of experimentally characterized gene clusters indicate spatial and condition specificity, such as enzymatic genes of clusters synthesizing terpenes in *A. thaliana* and *L. japonicus* [11, 7, 8, 17].


To facilitate a convenient and context-specific activity analysis of metabolic gene clusters, we recently proposed a probabilistic framework [1], denoted METACLUSTER, which automatically identifies conditions and tissues associated with inferred gene clusters within a given differential gene expression compendium. However, of equal importance, in particular with the emergence of large-scale transcription factor binding data such as [2], is the elucidation of metabolic gene cluster transcriptional regulation, since it has been argued that one way to facilitate such coordinated gene expression may be co-regulation due to promoter binding by shared transcription factors [3]. Hence, to enhance our understanding of context-specific transcriptional gene cluster regulation, we redefine and augment this probabilistic framework, hence denoting it METACLUSTER^{plus},

integrating gene expression compendia, 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 [12], 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 (see figure 1). METACLUSTER^{plus} may 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 [13], antiSMASH [5], plantiSMASH [10], and PhytoClust [14], with additional layers of automated high-resolution functionality and transcriptional regulation inference. Further, it provides a set of appealing and intuitive visualizations of inferred results for elucidation and publication.

2. Methods

2.1. Inference of context-specific transcriptional activity and regulation

Schlapfer *et al.* [13] proposed a probabilistic ranking framework based on co-expression to identify sets of high confidence metabolic gene clusters and to prioritize clusters for experimental validation. This framework had been extended in our previous work [1] in order to allow for the identification of context-specific gene expression of metabolic gene clusters. For each gene cluster, a rank was introduced based on combining multiple evidence scores regarding co-expression among cluster genes and the cluster's context specific transcriptional activity, all integrated using Fisher's method [12]. While keeping the multiple evidence integration based

 michael@educatedguess.ai (M. Banf)
ORCID(s):

METACLUSTER^{plus} - inference and visualization of context-specific transcriptional regulation of biosynthetic gene clusters

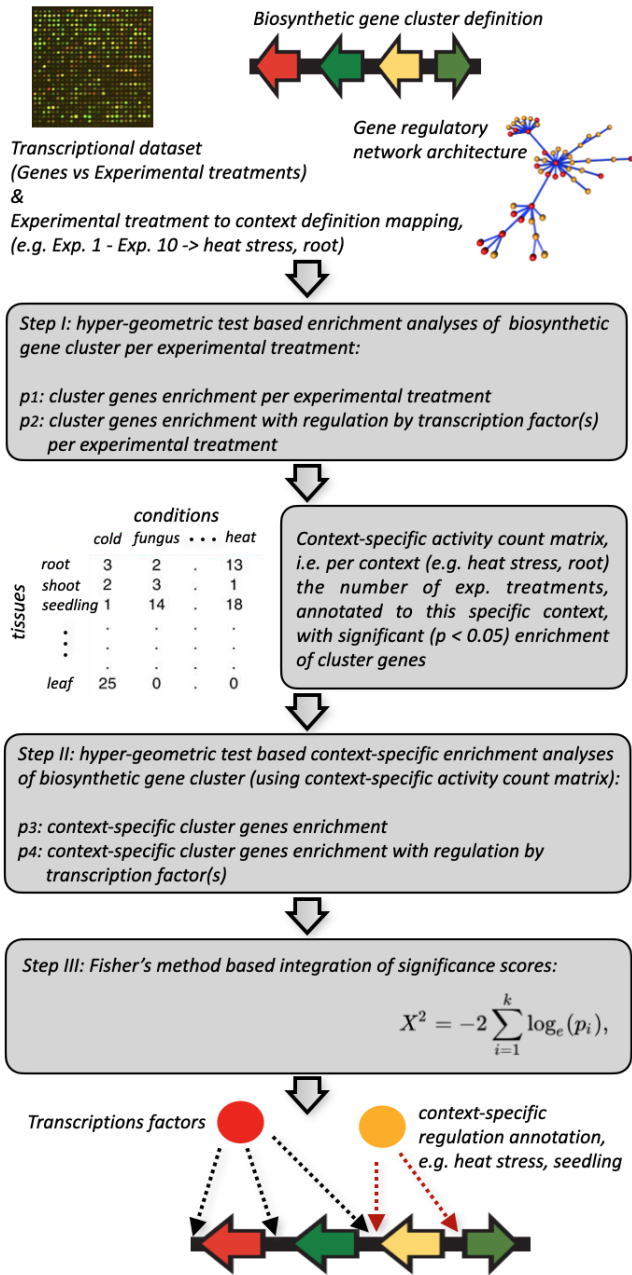


Figure 1: Overview of the probabilistic gen cluster regulation inference framework.

approach, METACLUSTER^{plus} redefines the transcriptional activity inference in order to compensate for a potential weakness in the original framework. It further augments transcriptional activity analysis by another layer, that is the simultaneous inference of context specific transcriptional regulation.

Initially, as in [1], a differential gene expression dataset is constructed by retaining experimental treatments represented by gene expression profiles measuring gene expression responses of wild type to treatment and control conditions and computing the log of fold change difference between the mean of the treatment and control sample replicates. Two sample t-tests are performed 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, an entry in D is assigned 1, -1 or 0 for statistically highly significant ($p < 0.05$) up-, down-, or non-significant differential expression.

Given D , our novel framework first estimates the probability $p_{gc}(e)$ of a gene cluster gc to be transcriptionally active in an experiment e . Since each experimental treatment e is annotated with a specific pair of condition c and tissue t , e may be defined as $e := (c, t)$. We select experimental treatments in the differential expression matrix D that were statistically enriched in gc based on hyper-geometric test following a general hyper-geometric distribution $\binom{G_{gc}}{g_{gc,e}}$.

$\binom{G-G_{gc}}{g_e-g_{gc,e}} / \binom{G}{g_e}$ with G and G_{gc} denoting the total number of genes in the genome and the number of gene cluster genes, within experimental treatment e . g_e and $g_{gc,e}$ represent the number of genes being differentially expressed in experimental treatment e and the subset of differentially expressed cluster genes in e , respectively. Following this hyper-geometric test, all experimental treatments with $p \leq 0.05$ were selected.

Using our manually curated mapping of experimental treatments to conditions and tissues, we then established a conditions vs tissue count matrix C_{gc} for further downstream context analysis (see figure 1). The context count matrix uses the association of individual experimental treatments e to the conditions c and tissues t , i.e. a pair of condition and tissue is incremented, if the corresponding treatment e for the cluster is significantly expressed ($p \leq 0.05$). We then performed hyper-geometric tests to identify enrichment of gc for a specific context, i.e. a pair of (c, t) , defined as probability $p_{gc}(c, t)$ following the distribution $\binom{N_{gc,(c,t)}}{n_{gc,(c,t)}}$.

$\binom{N_{(c,t)}-N_{gc,(c,t)}}{n_{(c,t)}-n_{gc,(c,t)}} / \binom{N_{(c,t)}}{n_{(c,t)}}$ with $N_{(c,t)}$ and $N_{gc,(c,t)}$ denoting the total number of all context pairs (c, t) and the number of all context pairs for a gene cluster gc . Accordingly, $n_{(c,t)}$ and $n_{gc,(c,t)}$ denote the number of a specific context pairs as well as the number of that context pair (c, t) for a specific gene cluster gc .

Defining context specific activity in this manner, we also compensate for potential weakness in the original framework [1] where a rather artificial disentanglement of condition and tissue annotations was introduced due to the sequential approach that separately analyzed enrichment of conditions only first, thereby ignoring putatively conflicting tissues, and subsequently trying to add tissues. Further, as a beneficial side-effect, this further simplifies the whole process of transcriptional activity analysis and allows for a more unambiguous identification of specific cluster genes being as being transcriptional active compared to [1].

Next, we estimate $p_{gc}(c, t, r)$, which represents the probability of cluster gc to be transcriptionally active in a specific context (c, t) with a putative regulator r for all regulators active in (c, t) with a minimum of two putative target genes of gene cluster gc being analyzed. Enrichment analysis is, again, based on a hyper-geometric distribution $\binom{G_{gc}}{g_{gc,(c,t)}} \cdot \binom{G-G_{gc}}{g_{(c,t)}-g_{gc,(c,t)}} / \binom{G}{g_{(c,t)}}$ with G and G_{gc} , with $t_{r,(c,t)}$ and

METACLUSTER^{plus} - inference and visualization of context-specific transcriptional regulation of biosynthetic gene clusters

$t_{gc,r,(c,t)}$ representing the number of targets of regulator r in general and the number of targets genes expressed in experimental treatment (c, t) , respectively. Next, we define context count matrix $C_{gc,r}$ per gene cluster gc and putative regulator r . Again associated individual gene clusters and putative regulation to condition and tissue labels., we performed hyper-geometric test $\binom{N_{gc,(c,t),r}}{n_{gc,(c,t),r}} \cdot \binom{N_{(c,t)} - N_{gc,(c,t),r}}{n_{(c,t)} - n_{gc,(c,t),r}} / \binom{N_{(c,t)}}{n_{(c,t)}}$, with $N_{(c,t)}$ and $N_{gc,(c,t),r}$ denoting the total number of contexts (condition and tissue) pairs and the total number context pairs for a gene cluster and associated regulator r , and a specific context (c, t) . Accordingly, $n_{(c,t)}$ and $n_{gc,(c,t),r}$ denote the number of a specific context as well as the number of that context (c, t) for a specific gene cluster gc and regulator r .

Our method further allows for the integration of additional evidences, here for example the enrichment $p_{gc}(sig)$ of signature enzymes per cluster [13]. Integration of all individual evidence probability scores per gene cluster gc follows our previously proposed approach in [1] based on Fisher's method [12] to estimate a combined p-value $p_{gc}(r \in (c, t))$ to define a final score of involvement of regulator r with gene cluster gc , given some condition c and tissue t .

2.2. Visualization of context-specific transcriptional activity and regulation

Aside a textual representation of the inferred gene cluster regulation (as illustrated in table 1), we equip our framework with chord graph based visualization per gene cluster that allows for i) transcription factor families vs conditions and treatments on a gene cluster level (see figures 2 and 4), as well as ii) transcription factor families vs conditions and treatments vs individual cluster genes on a cluster specific gene level (see figures 3 and 5). Numbers represent an actual count of associations between members of a transcription factor family and the gene cluster or cluster genes, respectively, given a specific condition and treatment.

3. Results and Discussion

To demonstrate the utility of METACLUSTER^{plus} as well as its visualization capabilities, we run our pipeline for metabolic gene cluster predictions in *Arabidopsis thaliana*, acquired from [13]. Here, we highlight prediction and visualization of transcriptional activity and regulation for two examples, the experimentally characterized terpene biosynthetic clusters in *Arabidopsis*, i.e., the thalianol [7] and the marneral [8] cluster. These were clusters C641 and C628 in [13]. We use a recently compiled large-scale gene expression dataset by He *et al.* [9] with 6057 expression profiles, covering 79.7% of the *A. thaliana* ecotype Columbia genome. We retain 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 are assigned to 27 manually curated conditions and 9 tissues (see [1], supplementary methods). As for a gene regulatory network architecture, we harness a recently released, large scale DNA affinity purification sequencing (DAP-seq) based dataset [2], providing a gene reg-

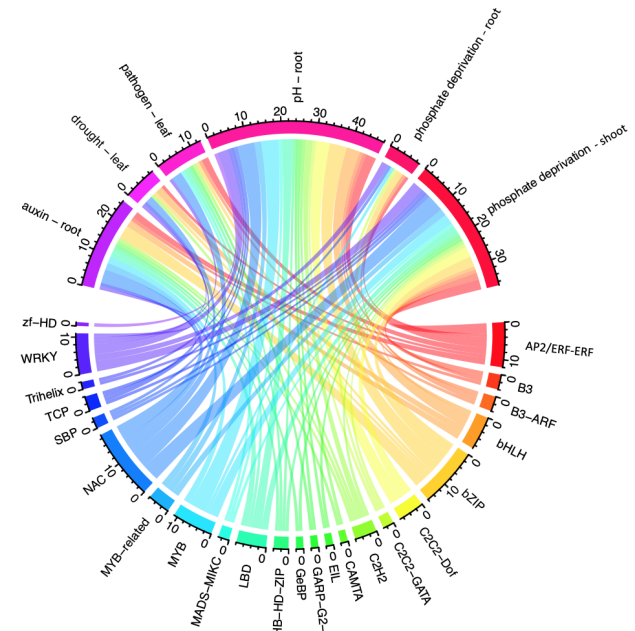


Figure 2: Gene cluster level visualization, i.e. transcription factor families vs conditions and treatments, of transcriptional activity and regulation of the marneral gene cluster.

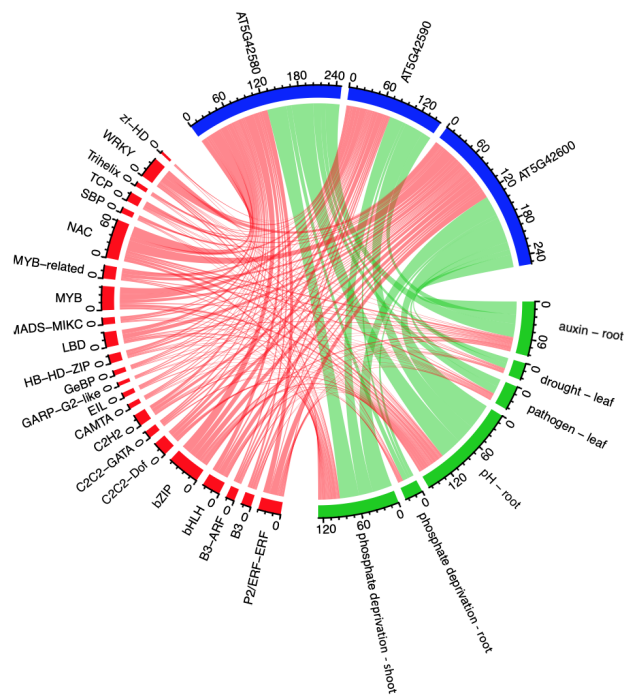


Figure 3: Cluster gene level visualization, i.e. transcription factor families (red) vs conditions and treatments (green) vs cluster genes (blue), of transcriptional activity and regulation of the marneral gene cluster.

ulatory network consists of 349 transcription factors, 26921 target genes and 1791998 connections.

Table 1 illustrates an excerpt of the textual representation of the predicted gene cluster regulations. Figures 2 and

METACLUSTER^{plus} - inference and visualization of context-specific transcriptional regulation of biosynthetic gene clusters

Table 1

Selected example results of transcriptional activity and regulation inference for the thalianol cluster

Transcription factor (family)	Treatment	Tissue	Regulated Genes
AT3G22760 (Trihelix)	salt	leaf	AT5G47950, AT5G48000, AT5G48010
AT4G00730 (C2C2-Dof)	salt	root	AT5G47970, AT5G47990, AT5G48000, AT5G48010
AT3G61150 (WRKY)	phosphate deprivation	shoot	AT5G47980, AT5G47990, AT5G48000, AT5G48010
AT1G76890 (LBD)	phosphate deprivation	shoot	AT5G47980, AT5G47990, AT5G48000, AT5G48010
AT4G14770 (NAC)	phosphate deprivation	shoot	AT5G47950, AT5G48000, AT5G48010
AT5G47370 (NAC)	phosphate deprivation	shoot	AT5G48000, AT5G48010
AT2G22430 (AP2/ERF-ERF)	phosphate deprivation	shoot	AT5G47980, AT5G47990, AT5G48000, AT5G48010
AT2G30590 (C2C2-Dof)	pathogen	seedling	AT5G47950, AT5G47980
AT3G01970 (WRKY)	pathogen	seedling	AT5G47950, AT5G47980, AT5G47990, AT5G48000
AT5G47370 (MADS-MIKC)	pathogen	seedling	AT5G47950, AT5G47980, AT5G47990, AT5G48000
AT3G21890 (HB-HD-ZIP)	auxin	root	AT5G47950, AT5G47980, AT5G47990, AT5G48000 AT5G48010

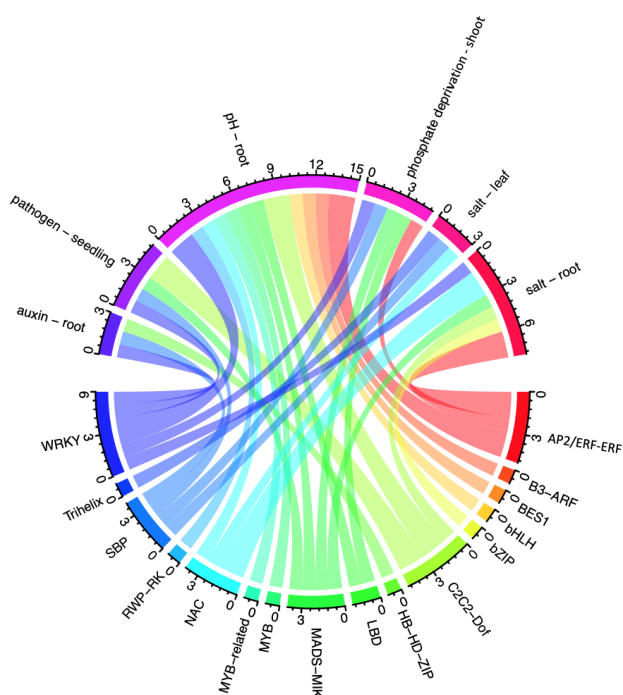


Figure 4: Gene cluster level visualization, i.e. transcription factor families vs conditions and treatments, of transcriptional activity and regulation of the thalianol gene cluster.

4 as well as figures 3 and 5 highlight the corresponding chord graph based visualizations on a gene cluster as well as cluster specific gene levels, respectively. In particular, these graphical representations may serve to provide an immediate visual and high level summary of the given condition-specific regulatory relationships for prioritization and further investigations. For instance, both clusters show an enrichment for regulation by members of the APETALA2/Ethylene Response Factor (AP2/ERF) family, or the WRKY transcription factor family across a variety of stress conditions and tissues, which is corroborated by research on these transcription factor families' influence on specialized metabolism control in plants [20, 19, 18].

Given its utility, we anticipate METACLUSTER^{plus} to be a valuable tool for the efficient integration of heterogeneous datasets in order plan experiments and guide validation of context-specific metabolic gene cluster transcriptional regulation.

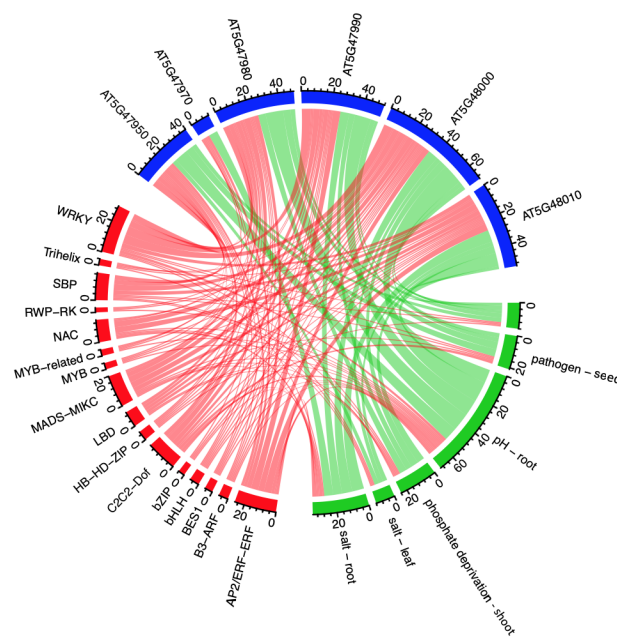


Figure 5: Cluster gene level visualization, i.e. transcription factor families (red) vs conditions and treatments (green) vs cluster genes (blue), of transcriptional activity and regulation of the thalianol gene cluster.

References

- [1] Banf M, Zhao K, Rhee SY. METACLUSTER-an R package for context-specific expression analysis of metabolic gene clusters. *Bioinformatics*. 2019 Sep 1;35(17):3178-3180.
- [2] O'Malley RC, Huang SC, Song L, et al. Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape *Cell*. 2016;165(5):1280-1292.
- [3] Nützmann HW, Huang A, Osbourn A. Plant metabolic clusters - from genetics to genomics. *New Phytol*. 2016 Aug;211(3):771-89.

METACLUSTER^{plus} - inference and visualization of context-specific transcriptional regulation of biosynthetic gene clusters

- [4] Aoki *et al.* (2016) ATTED-II in 2016: a plant coexpression database towards lineage-specific coexpression. *Plant Cell Physiology*, 57, e5
- [5] Blin K., *et al.* (2017) antiSMASH 4.0—improvements in chemistry prediction and gene cluster boundary identification. *Nucleic acids research*. 45 (W1), W36-W41.
- [6] Chae, L. *et al.* (2014) Genomic signatures of specialized metabolism in plants. *Science*, 344(6183):510-3.
- [7] Field B, Osbourn AE. (2008) Metabolic diversification—-independent assembly of operon-like gene clusters in different plants. *Science*. 320(5875):543-7.
- [8] Field B, *et al.* (2011) Formation of plant metabolic gene clusters within dynamic chromosomal regions. *PNAS* 108 (38) 16116-16121.
- [9] He, H., *et al.* (2016) Large-scale atlas of microarray data reveals the distinct expression landscape of different tissues in Arabidopsis. *Plant Journal*. 86(6):472-80.
- [10] Kautsar S.A., *et al.* (2017) plantiSMASH: automated identification, annotation and expression analysis of plant biosynthetic gene clusters. *Nucleic acids research*. 45 (W1), W55-W63.
- [11] Krokida A., *et al.* (2013) A metabolic gene cluster in *Lotus japonicus* discloses novel enzyme functions and products in triterpene biosynthesis. *New Phytologist*. 200(3):675-90
- [12] 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(2):284–295.
- [13] Schlapfer, P. *et al.* (2017), Genome-wide prediction of metabolic enzymes, pathways and gene clusters in plants, *Plant Physiology*, 176 (3), 2583-2583.
- [14] Toepfer N., *et al.* (2017) The PhytoClust tool for metabolic gene clusters discovery in plant genomes, *Nucleic Acids Research*. 45, 12, 7049–7063
- [15] Wink, M. (2010) *Biochemistry of plant sec. metabolism* Wiley-Blackwell.
- [16] Wisecaver J.H, *et al.* (2017) A Global Coexpression Network Approach for Connecting Genes to Specialized Metabolic Pathways in Plants. *Plant Cell*. 29(5):944-959.
- [17] Yu N, *et al.* (2016) Delineation of metabolic gene clusters in plant genomes by chromatin signatures. *Nucleic Acids Research*;44(5):2255-65.
- [18] Schluttenhofer C., Yuan L. (2015) Regulation of specialized metabolism by WRKY transcription factors. *Plant Physiol*. 167(2):295-306.
- [19] Paul P, *et al.* (2019) Mutually Regulated AP2/ERF Gene Clusters Modulate Biosynthesis of Specialized Metabolites in Plants. *Plant Physiology*, 182(2): 840–856.
- [20] Shoji T, Yuan L. (2021) ERF Gene Clusters: Working Together to Regulate Metabolism. *Trends Plant Sci*. 26(1):23-32.