Peer-reviewed version available at *Genes* **2020**, *11*, 634; doi:10.3390/genes11060634

Article

metaRE R package for meta-analysis of transcriptome data to identify the cis-regulatory code behind the transcriptional reprogramming

Daria D. Novikova [1,2], Pavel A. Cherenkov[3], Yana G. Sizentsova[1], Victoria V. Mironova[1,3*]

[1] Institute of Cytology and Genetics, Lavrentyeva avenue 10, Novosibirsk, 630090, Russian Federation;
[2] Laboratory of Biochemistry, Wageningen University, Stippeneng 4, Wageningen, 6708WE, Netherlands;
[3] Novosibirsk State University, 2 Pirogova Street, Novosibirsk, 630090, Russian Federation;
[*] Correspondence: victoria.v.mironova@gmail.com

Abstract: At the molecular level, response to an external factor or an internal condition causes reprogramming of temporal and spatial transcription. When an organism undergoes physiological and/or morphological changes, several signaling pathways are activated simultaneously. Examples of such complex reactions are the response to temperature changes, dehydration, various biologically active substances, and others. Synergistic action of multiple pathways greatly complicates the experimental study of the molecular genetic mechanisms of the organism's reactions. As a result, a significant part of the regulatory ensemble in such complex reactions remains unidentified. We developed *metaRE*, an R package for the systematic search for cis-regulatory elements enriched in the promoters of the genes significantly changed their transcription in a complex reaction. *metaRE* mines multiple expression profiling datasets generated to test the same organism's response and identifies simple and composite cis-regulatory elements systematically associated with differential expression of genes. Here we showed *metaRE* performance for identification of cold stress-responsive cis-regulatory code in *Arabidopsis thaliana*. *MetaRE* identified potential binding sites for known as well as unknown cold response regulators. Software with source files, documentation, and example data files are freely available online at the repository (https://github.com/cheburechko/MetaRE).

Keywords: meta-analysis, transcription factor, binding sites, genomics, transcriptomics, chilling stress, CBF, DREB, CAMTA1

1. Introduction

More than two decades have passed since the establishment of whole-genome expression profiling methods. Nowadays, thousands of transcriptomes are publicly available. Typically, several related experiments studying the same phenomenon can be found, thus, providing a rich set of material for analysis. Meta-analysis is applicable to sets of experiments testing the same hypotheses to extract robust signals and repetitive features that are impossible to derive from the individual experiments.

The typical example of meta-analysis is the defining of robust differentially expressed genes (DEGs) over many transcriptomic datasets. This approach is widely used in medical genomics to identify the gene signatures associated with a condition or disease, e.g., in [1-3]. To account for the most reliable and reproducible gene signatures, different authors applied such meta-analysis procedures as Fisher's methods, Stouffer's method, permutation, or machine-learning procedures. Recently, a ready-to-use framework GSMA has been developed to solve this task for any problem of interest [3].

Alternatively, a meta-analysis of transcriptome datasets can help to understand the cis-regulatory code behind the transcriptional response. The simplest way is to analyze the upstream

regions of the robust DEGs for overrepresented sequences, e.g., as in [4, 5]. However, the way to detect the robust gene sets might be comprehensive. He and co-authors (2016) analyzed DEGs in nine transcriptomic datasets on breast cancer: DEGs were identified by Fisher's method for p-values combination [6]. Subsequent enrichment analysis of motifs in promoters of DEGs was estimated by Fisher's exact test and allowed identifying transcription factors associated with breast cancer.

A better way to identify a full set of cis-elements, or a "cistrome", associated with a transcriptional response, is a meta-analysis of individual transcriptomes and not the robust DEGs. Authors of the cis-Metalysis program performed a meta-analysis of transcriptomics data on bee [7]. They revealed enrichment of transcription factors binding sites in the DEGs and their association with external factors that cause similar changes in the organism. An interesting approach has been applied to study the citrome for iron deficiency response in Arabidopsis (*Arabidopsis thaliana*) roots [8]. Authors searched for the enrichment of k-mers in upstream regulatory regions of Fe-responsive genes taken from several experiments. They applied the machine learning algorithm Random Forest to identify enriched elements in different functional clusters of co-expressed genes revealed. However, on the different steps of their study, authors used separate tools and approaches aiming at a specific goal of identifying clusters of Fe-responsive genes regulated by the same pulls of cis-regulatory elements.

The methods for comprehensive meta-analysis of transcription profiles for cis-elements prediction described above have proven to be powerful in specific studies. However, they were not implemented in a ready-to-use package. Here we developed a powerful but versatile pipeline for cistrome-wide meta-analysis, implemented as a *metaRE* R package. In this study, we show the performance of *metaRE* on cold stress-responsive transcriptome datasets. This analysis identified the binding sites for all known regulators of cold stress, but also predicted some new perspective candidates.

2. Materials and Methods

2.1. metaRE R package structure and functionality

metaRE R package implements a pipeline to search for consensus sequences enriched in the promoters of DEGs. We used *C++* to speed up slow components and the *Rcpp* package to integrate the *C++* code into R [9, 10]. *metaRE* package performs a five-step analysis: *(1)* DEGs identification; *(2)* cis-regulatory consensus element search; *(3)* calculation of association between consensus presence and changes in gene expression; *(4)* meta-analysis over multiple datasets; *(5)* permutation test. The pipeline is detailed below and in Figure 1.

(1) DEGs identification

As an input, *metaRE* uses transcriptome data. For users' convenience, we applied *GEOquery* [11], *limma* [12] and *edgeR* [13, 14] packages to identify DEGs in the datasets from the GEO database [15]. *metaRE* function *prepareGEO* allows loading and adjusting the preprocessed GEO data frames. Functions *processMicroarray* and *processRNAcounts* could be used to identify DEGs in a single dataset using *limma* (microarray and RNA-seq, respectively), functions generate a new table for a particular experiment with user-defined expression classes. The function *preprocessGeneExpressionData* can perform the same analysis for multiple datasets at once, it generates the final data frame *GeneClassificationMatrix*, which combines information about DEGs from all experiments in the meta-

analysis.

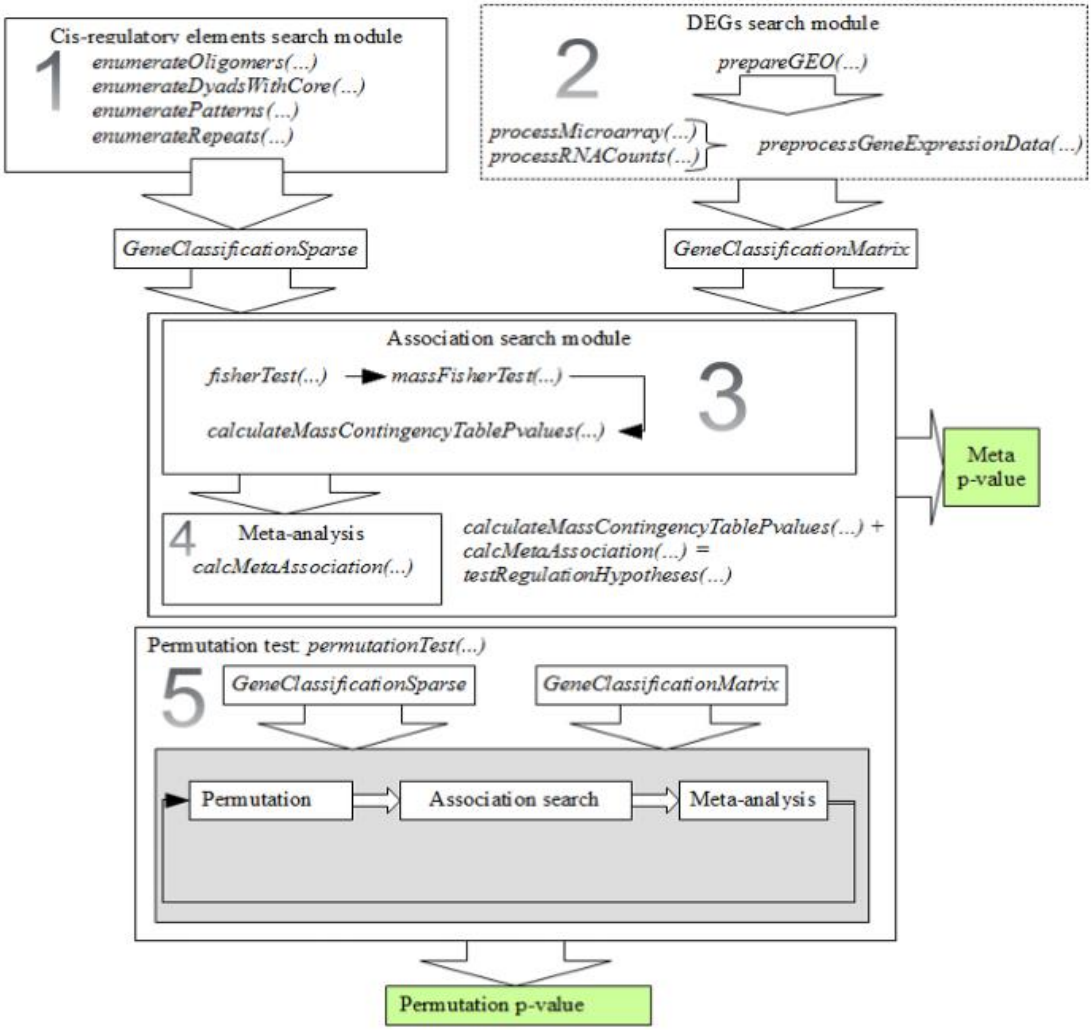

Figure 1. Scheme of *metaRE* modules that implement a five-step pipeline for the search for cis-elements significantly associated with differential gene expression over multiple datasets. DEGs - differentially expressed genes. Different modules are highlighted with squares; final sets of p-values are painted green. Described in the Methods steps are enumerated on the figure.

(2) Cis-regulatory consensus elements search

Another input data for the *metaRE* package are the regulatory region sequences in *fasta* format. *MetaRE* uses the *Biostrings* R package [16] to upload the sequences from BioMart [17]. Next, *metaRE* annotates each sequence for the presence of a potential cis-element in the following format. Function *enumerateOligomers* searches for all possible k-mers without considering complementarity, e.g., in case of hexamers, *metaRE* searches for 2080 non-redundant hexamers comprising 2016 complementary pairs and 64 palindromes instead of 4096 possible combinatorial variants. In addition to k-mers, it is possible to annotate systematically the regulatory regions with the information about all possible spaced repeats with the same k-mer as a core (*enumerateRepeats*), spaced bipartite elements with different k-mers as the cores (*enumerateDyadsWithCore*). It is also possible to search for a predetermined list of motifs described with 15 letters IUPAC ambiguity code (*enumeratePatterns*). For the *enumerateRepeats* and *enumerateDyadsWithCore* functions, it is possible to set maximum and minimum spacer length in both cases. *MetaRE* will search for k-mers' combinations with given spacer length diapason. For all the functions, the logic remains the same: reverse complement k-mers are considered to be the same element. Thus, the number of k-mers/bipartite elements/repeats/predetermined motifs in the analysis is reduced comparing to the number of possible combinatorial variants.

The output of the second step of the procedure is a named list of integer vectors. Names are the consensus sequence; vectors are the indices of genes in which these sequences are present.

(3) Calculation of association between cis-regulatory element presence and changes in gene expression

At this step, for each cis-element and each experiment, an association with differentially expressed genes is estimated, separately for all regulation classes. A p-value for the association is calculated using a 2x2 contingency table by Fisher's exact test [18, 19]. The test estimates the probability of getting such an association between two variables in the contingency table. In this case, the variables are "presence/absence of the cis-element" and "DEG/non-DEG". In *metaRE* the procedure is implemented by a function *calculateMassContingencyTablePvalues*. The result is a float matrix of p-values for the association between the cis-element presence and up/down-regulation, where rows correspond to the cis-elements, columns correspond to the datasets in which cells are calculated p-values.

(4) Meta-analysis

Function *calcMetaAssociation* used to combine the p-values calculated for a particular cis-element over many datasets. *MetaRE* uses Fisher's method to calculate meta-p-values (Figure 1, [18]). Due to multiple testing for many cis-elements, *calcMetaAssociation* also estimates an adjusted p-value, for which the user can choose one of the following multiple correction methods: Bonferroni, Bonferroni-Holm [20, 21], Benjamini-Hochberg [22] and Benjamini-Yakuteli [23]. Users also can set the cutoff threshold for adjusted meta-p-value – the cis-elements which pass the cutoff are to be tested on the step *(5)*.

(5) Permutation test

Finally, *metaRE* applies the permutation test to the cis-elements with significant adjusted meta-p-values. *MetaRE* uses *foreach* package (CRAN project) for parallel permutation testing. *PermutationTest* function shuffles the regulatory regions between the genes and recalculates meta-p-value for each cis-element in the analysis. We optimized the procedure so that every iteration run *permutationTest* stores the preliminary results in 'outfile' and removes the cis-elements that will not pass the cutoff threshold. After performing *M* permutations, *the function* computes the permutation p-value for cis-elements left in the analysis as $p=(m+1)/(M+1)$, where m is a number of recorded p-values not greater than the meta-p-value. It also computes adjusted permutation-p-values to consider the multiple testing (for the amount of cis-elements predetermined on the step (4)).

In the end, the cis-elements with adjusted permutation-p-value below the cutoff threshold are considered to be significantly associated with the differential expression.

2.2. Motifs comparison

To annotate cis-elements, we used the TOMTOM tool from Meme Suit [24] with the reference databases DAPv1, PBM, and Cis-BP. The best match with E-value<0.05 was taken into the annotation.

2.3. Datasets

Transcriptome datasets on low positive temperature treatment were retrieved from the GEO database. 22 out of 40 datasets passed the quality control for well-clustered replicas giving a sufficient number of DEGs (see Table S1). The identification of DEGs was made using the Benjamini–Hochberg method [22] to control the false discovery rate (FDR < 0.05).

3. Results

3.1. MetaRE R package for cistrome-wide association study

We developed *metaRE* R package which identifies the cistrome associated with the case of study via a meta-analysis of multiple transcriptomic experiments. *MetaRE* pipeline includes five steps: (1) DEGs identification in many transcriptomic datasets, (2) search for cis-regulatory elements in upstream gene sequences, (3) assessment of the association between cis-regulatory element presence and the changes in gene expression in each transcriptomic dataset, (4) meta-analysis over multiple datasets and (5) permutation test to study the robustness of the prediction. The first step is performed

in *metaRE* using standard R packages, or the user can upload processed data. At the second step, *metaRE* generates the information about the presence/absence of all combinatorially possible nucleotide sequences of a particular length and structure (encoded in the 15-nucleotide IUPAC alphabet) in a set of nucleotide sequences (for instance, promoter regions, transcription factors binding regions, etc.). The package allows the user to identify cis-regulatory elements of different lengths, which could consist of one element, repeats, or bipartite elements with a variable or fixed spacer and order of elements. In the third step, *metaRE* assesses the association between each cis-elements and differential gene expression in each of the datasets. At the fourth step, *metaRE* combines the p-values taken from the separate datasets and highlights which of the cis-elements are systematically overrepresented. In the last step, *metaRE* tests the independence of obtained results from external factors by permutation test.

The main advantage of the *metaRE* package is that it identifies a reliable and reproducible set of cis-regulatory elements associated with the transcriptional response over many independent datasets, rather than in a single gene set. The R package can be used for the study cases on any organism with a sequenced genome. It is possible to adjust the procedure by changing the statistical tests, thresholds, cis-elements structure in search, promoters' length, etc. Other nucleotide sequences could be used instead of the promoters, e.g., 3'UTRs or ChIP-Seq profiles. Thus, *metaRE* gives the user freedom to adjust the package to the particular study, which is essential considering the differences and quality of raw data, annotation of the genome of different species, and knowledge on the location of cis-regulatory elements.

MetaRE was tested on several independent studies on different organisms, for instance, cold-induced Zebrafish transcriptomes, dioxin-induced human and mouse transcriptomes [25], auxin-induced Arabidopsis transcriptomes [18]. The application of *metaRE* was efficient for all of the cases. Here we discuss *metaRE* performance to identify cold-responsive cistrome in Arabidopsis.

3.2. MetaRE for identification of cold-responsive cistrome

To demonstrate the utility of *metaRE* package we performed analysis on cold stress-induced transcriptome datasets generated with a good-quality on Arabidopsis so far (Table S1, [5, 26-33]). Cold-responsive transcriptome datasets were divided into two groups by time of response: early response, up to six hours of cold exposure, and late response, 12-24 hours of cold exposure.

On the step (1), *metaRE* identified DEGs (FDR<0.05) lists for all of 22 transcriptomic datasets. We varied the threshold for fold-change from none, to 1.5 and 2. As a result, three summary tables were generated summarizing information about the differential transcriptional response.

On the step (2), *metaRE* annotated Arabidopsis' regulatory regions [-1500; -1] taken from TAIR10 by the diversity of non-redundant *k*-mers. In this study, we searched for hexa-, hepta-, and octamers.

On the steps (3-5), *metaRE* identified all *k*-mers associated with the transcriptional cold stress response. We tried two multiple testing corrections (Bonferroni-Hochberg or Bonferroni) and set the stringent threshold for adjusted meta-p-value < 1E-10 and adjusted permutation p-value ≤ 1E-3.

The summary tables for identified k-mers (Table S2-3) suggest that the cistrome size provided by *metaRE* depends on the parameter settings. However, the most significant cis-elements are always the same. Noteworthy, to detect any motif associated with down-regulation, we had to get rid of the threshold for fold-change to identify DEGs only by FDR. Despite a more stringent multiple testing correction applied for heptamers and octamers, *metaRE* found more of them in this study, comparing to the number of significantly overrepresented hexamers (Table S2-3). This was not the case in other meta-analyses performed by *metaRE* [18, 25]. We can explain this fact by significant enrichment of many degenerated A/T-rich motifs in the transcriptional response to cold (Figure S1; discussed below). To sum up, we recommend performing a preliminary analysis under different settings to define the most appropriate one. Below we discuss only the results obtained under the stringent Bonferroni criterium for hexamers.

3.3. Analytics on cold stress-responsive cistrome for Arabidopsis thaliana

We detected 95/43 and 10/26 hexamers associated with up- and down-regulation in the early/late cold stress response (Table 1). A strong bias in a cistrome diversity was detected towards the early activatory response, but apparently, it correlates with many AT-rich elements found overrepresented in the upstream regions of early cold-responsive genes (even more AT-rich motifs were found in septamers and octamers; Figure S1; Table S2, Table S3). Another trend is that cold-responsive cistrome has much less cis-elements associated with down-regulation than with up-regulation. With only one exception, G-box CACGTG, hexamers were explicitly associated with either up- or down-regulation.

Table 1. Summary of predicted hexamers associated with cold stress response

	Early response (<6 h)	**Late response (>12 h)**
Up	95	43
Down	10	26
	Without A/T-rich hexamers	
Up	25	40
Down	10	26

Next, we applied the TOMTOM tool [24] to annotate detected by *metaRE* cis-elements associated with early and late cold-response. We were able to annotate more than 65% of detected cis-elements, however, many AT-rich elements and elements related to down-regulation remained unidentified (Tables S2-3). Many of the hexamers associated with the cold stress response significantly match the binding sites of known cold response regulators from CAMTA, AP2/ERF, bHLH, MYB, and bZIP families (Figure 2A) and this fitness confirms the adequacy of *metaRE* pipeline.

The binding sites for C-REPEAT BINDING FACTORs (CBFs) transcription factors from AP2/ERF family (CCGACA, ACCGAC; GCCGAC, CCGACC) were expected to be found as associated with the transcriptional cold response, as CBFs are the major regulator of cold acclimation [34-37]. However, CBF binding sites were not the most abundant and significant in early response (Table S2). The most significant motifs for early response to cold stress appeared to be: (1) ACGCGT (adjusted meta-*p*-value = 5.96E-84), the potential binding sites for CAMTA; (2) CACGTG (p = 1.52E-54), the G-box bound by bHLH and bZIP transcription factors; (3) ACACGT (p = 2.3E-53), the motif bound by NAC, BES, bZIP, and bHLH transcription factors; (4) ACGTGG (p = 2.65E-52), potential binding site for bZIP and bHLH; and (5) a group of AT-rich elements (3.47E-11<p<2.89E-50). The involvement of transcription factors bound to (1)-(4) with the cold response was known beforehand [29, 38-43]. However, the fact that they are more relevant to early cold response comparing CBF-binding sites is tempting, as CBF factors were recently shown to be involved in freezing not chilling resistance and may not be essential to survive in response to low positive temperature [31, 44]. Potential binding sites for CBFs were found the most significant for the late response to cold (Table S3).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 13 May 2020 doi:10.20944/preprints202005.0227.v1

Peer-reviewed version available at Genes 2020, 11, 634; doi:10.3390/genes11060634

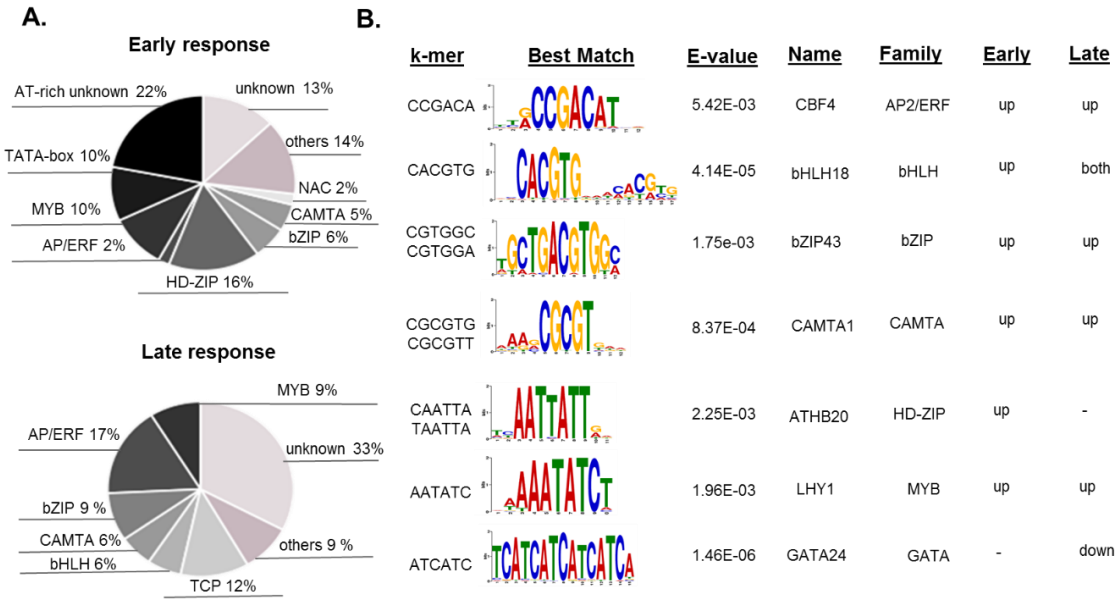

Figure 2. Cis-regulatory elements predicted with *metaRE* as systematically enriched in upstream regulatory regions of cold-induced genes. **(a)** Annotation of the hexamers to the known binding sites of *Arabidopsis thaliana* with the help of the TOMTOM tool [24]. Only significant best matches (E-value<0.05, one per hexamer) were calculated to build the round diagram. **(b)** Annotation details for particular hexamers associated with early, late, or both early and late responses. The best significant matches of the hexamers with the known binding sites associated with down-regulation in response to cold stress.

However, most of the detected AT-rich elements remained unknown; some of these sequences significantly match (TOMTOM, E-value<0.05) the known binding sites for HD-ZIP and MYB families (Table S3, Figure 2B). Although it is not clear if the detected association with HD-ZIP transcription factors is relevant, the involvement of LHY1 and CCA1 MYB transcription factors, into cold-stress has been discussed in several works [38, 45-47]. The motifs associated with down-regulation were also poorly annotated. Among the rare examples of annotated motifs associated with down-regulation are GATGAT/ATCATC, the potential binding site of GATA transcription factors (Figure 2B), and a family of potential TCP-binding motifs (Table S3). These results demonstrate the perspectives of *metaRE* usage in the study of the cis-regulatory code behind transcriptional reprogramming in complex reactions. It allows not only predicting the diversity of involved cis-elements and respective transcription factors but also rank them and clarify their role in certain phases of transcriptional response.

4. Discussion

4.1. metaRE tool for identification of cis-regulatory elements repertoire

The main idea behind the method implemented in *metaRE* R package is that if the cis-regulatory elements are involved in a transcriptional response, then they should be overrepresented in the promoters of differentially expressed genes. This idea is not new, and there are many approaches facilitating the analysis for cis-elements overrepresentation within upstream regions of pre-compiled gene sets, e.g., in [6, 48-50]. What was missing, the pipeline that allowed analyzing cis-elements overrepresentation systematically, summarizing the output taken from many independent datasets, the task that *metaRE* has solved.

The novelty of the *metaRE* method lies in: (1) taking into account a large number of comparable transcriptome experiments, and (2) the consideration of enrichment significance for an individual cis-element. Usually, authors evaluate the enrichment of cis-elements in one or more gene lists independently; the results of enrichment between the lists are not compared [4, 5]. In this case, information about differences in the degree of enrichment of the same cis-element in different

datasets is leveled, which can lead to over- and under-predictions. The method underlying *metaRE* solves this problem.

Separate studies showed that systematic analysis of transcriptome datasets is powerful in the identification of the cistrome behind a complex reaction [7,8,19]. The basic assumption in these studies, as well as in the *metaRE* algorithm, is that only robust and significant cis-element association with transcriptional response will be detected across multiple, diverse transcriptomic datasets that test similar experimental variables. This could be considered both as an advantage and as a disadvantage of the systematic analysis. On the one hand, analysis of several datasets excludes a bias that could be caused by separate experiments (tissue sampling, treatment duration, concentration, growth conditions, quality of data, etc.). Thus, meta-analysis would detect the major cis-elements that operate under a variety of conditions. On the other hand, this approach will miss rare and condition-specific cis-elements. The latter could be solved by separate analysis of the datasets from experiments performed on different tissues, so one can have a tissue-specific cistrome. E.g. in this study for cold-stress responsive cistrome, as well as in [18] for auxin-regulated citrome, we saw apparent differences in time-resolved results. If the number of transcriptomes allowed, these differences would be detected for tissue- and condition-specific reactions.

Cis-elements enrichment analysis is especially powerful when performed using the position weight matrices (PWM) for known transcription factors. E.g., using Homer [49], one can yield the list of exact regulators which binding sites are overrepresented in the upstream regions of candidate genes. However, in *metaRE* we intentionally used a simpler consensus model for identification of overrepresented elements, making it more versatile and applicable for more organisms. First, for almost all organisms, including the model ones, the binding sites of most transcription factors remain unknown. Moreover, only very few organisms have PWMs for at least a hundred transcription factors (e.g., *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Mus musculus*, *Homo sapiens* [51]. Second, *metaRE* could be applied not only to the upstream regions but to any sequences associated with the genes to find the signals unrelated to transcription factor -DNA binding and not described by PWMs. For example, analyzing the 3'UTR *metaRE* could help identify the sites for the miRNA seeds binding. Third, in the present study of cold-responsive cis-elements, consensus search in *metaRE* with the subsequent analysis of identified sequences using PWMs for known transcription factors in TOMTOM [24] was shown to be very fruitful, with more than 65% of the elements annotated. We believe that the hybrid approaches with preliminary screening for enriched consensuses and subsequent annotation and re-analysis of the data using more powerful models are in need. Like an approach used in the study to annotate transcription factor binding sites in Nannochloropsis spp. microalgae [52].

4.2. Identification of cold-stress responsive cistrome

Here we employ *metaRE* in the study of widely studied processes, in which molecular mechanisms are still full of gaps, cold stress response. Large-scale transcriptome studies showed that the CBF1-3, the major regulators of cold acclimation, in fact, regulate only a small portion of cold-responsive genes [28, 31, 44, 53] what means that other regulators exist. Here we see that CBFs binding sites are, indeed, not overrepresented in early cold stress response as the potential binding sites for other transcription factors (Table S2). CBFs binding sites appeared to be the most overrepresented in the late response (Table S3), what explains why only a small portion of cold-responsive genes are CBF-regulated.

The most significantly enriched cis-element in early cold stress response detected by *metaRE* was the potential binding site for CAMTA (Figure 1, Table S1). CAMTA1-3 are known upstream regulators of CBF1-3, they increase freezing tolerance via activation of ~15% cold-responsive genes [29, 39].

Park et al. (2015) found that, in parallel with CBF genes, 27 other "first-wave" transcription factor genes were highly up-regulated at an early stage of cold treatment. Analysis of gene expression in transgenic plants overexpressing eleven of these first-wave transcription factors identified four transcription factors from bZIP family (ZAT12, ZF, ZAT10, and CZF1) and heat-shock factor HSFC1

involved in the regulation of cold stress-responsive genes [28, 44, 54]. *metaRE* identified bZIP transcription factors binding sites as ones of the most significantly enriched in promoters of early responsive to cold genes (Figure 2; Tables S2), however, their impact was not that big in the late response.

In addition to the potential binding sites of other known regulators of cold-stress, like MYB, NAC and bHLH transcription factors, *metaRE* identified many unknown motifs, and the binding sites of transcription factors which role in the cold stress is unknown. For example, we found the overrepresentation of A/T-rich sequences in the promoters of cold-responsive genes. Earlier, we found a similar result for auxin-regulated cistrome in Arabidopsis [18], but not dioxin-responsive cistrome in human [25]. The role of A/T-rich sequences can be different: they might be the parts of A/T-rich transcription factors binding sites (e.g., CarG-boxes, or ATHB-binding sites), or they might be the TATA-box sequences, or they might be a part of chromatin landscape. The half of A/T-rich sequences were annotated by TOMTOM either as HD-ZIP binding sites or as TATA-boxes. Another half still can predict a specific epigenetic landscape, as cold-induced genes show enhanced chromatin accessibility, and a large number of active genes in cold-stored tubers are associated with a bivalent H3K4me3-H3K27me3 mark [55].

Another interesting result relates to the cis-elements overrepresented in the promoters of down-regulated by cold genes, which mechanisms of functioning are completely unknown. Here we found potential binding sites for GATA and TCP transcription factors, as well as many unknown motifs. A further experimental study is required to clarify the role of these candidate genes in the cold stress response.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, Figure S1: The percentage of A/T-rich motifs among predicted k-mers (n=6-8) detected in promoters of differentially expressed genes with different settings (nFC - no threshold for fold change; FC1.5 - the threshold is fold change 1.5; FC2 is fold change 2)., Table S1: The list of cold-stress responsive transcriptome datasets taken for meta-analysis with *metaRE*, Table S2:Cis-elements associated with early cold stress response, Table S3: Cis-elements associated with late cold stress response.

Authors' contributions: Methodology, V.V.M., D.D.N., P.A.C.; software, P.A.C.; validation and investigation, V.V.M., D.D.N., P.A.C. and Y.G.S.; writing—original draft preparation, D.D.N., V.V.M.; writing—review and editing, V.V.M., D.D.N., P.A.C. and Y.G.S.; visualization, D.D.N., P.A.C. and Y.G.S.; supervision, V.V.M.; funding acquisition, V.V.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Russian Foundation for Basic Research, grant number 18-04-01130, and by Russian State Budget, project number 0259-2019-0008-C-01. D.D.N. was funded by a PhD sandwich fellowship from Wageningen Graduate School. The study of cold stress response was funded by Russian Science Foundation, grant number 18-74-10008.

Acknowledgments: We thank Prof. Ivo Grosse, Nadya Omelyanchuk and Prof. Dolf Weijers for fruitful discussions and inspiration. We are grateful to Dr. Viktor Levitsky for consultations on the technical part of the algorithm.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results**.**

References

1. Rhodes, D.R.; Yu, J.; Shanker, K.; Deshpande, N.; Varambally, R.; Ghosh, D.; Barrette, T.; Pandey, A.; Chinnaiyan, A.M. Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proceedings of the National Academy of Sciences* **2004**, *101*, 9309–9314, doi:10.1073/pnas.0401994101.
2. Phan, J.H.; Young, A.N.; Wang, M.D. Robust Microarray Meta-Analysis Identifies Differentially Expressed Genes for Clinical Prediction. *The Scientific World Journal* **2012**, *2012*, 1–9, doi:10.1100/2012/989637
3. Shafi, A.; Nguyen, T.; Peyvandipour, A.; Draghici, S. GSMA: an approach to identify robust global and test Gene Signatures using Meta-Analysis. *Bioinformatics* **2019**, btz561, doi:10.1093/bioinformatics/btz561.
4. Bargmann, B.O.R.; Vanneste, S.; Krouk, G.; Nawy, T.; Efroni, I.; Shani, E.; Choe, G.; Friml, J.; Bergmann, D.C.; Estelle, M.; et al. A map of cell type-specific auxin responses. *Mol Syst Biol* **2013**, *9*, 688, doi:10.1038/msb.2013.40.
5. Vogel, J.T.; Zarka, D.G.; Van Buskirk, H.A.; Fowler, S.G.; Thomashow, M.F. Roles of the CBF2 and ZAT12 transcription factors in configuring the low temperature transcriptome of Arabidopsis: Arabidopsis low temperature transcriptome. *The Plant Journal* **2004**, *41*, 195–211, doi:10.1111/j.1365-313X.2004.02288.x
6. He, H.; Cao, S.; Niu, T.; Zhou, Y.; Zhang, L.; Zeng, Y.; Zhu, W.; Wang, Y.; Deng, H. Network-Based Meta-Analyses of Associations of Multiple Gene Expression Profiles with Bone Mineral Density Variations in Women. *PLoS ONE* **2016**, *11*, e0147475, doi:10.1371/journal.pone.0147475.
7. Ament, S.A.; Blatti, C.A.; Alaux, C.; Wheeler, M.M.; Toth, A.L.; Le Conte, Y.; Hunt, G.J.; Guzman-Novoa, E.; DeGrandi-Hoffman, G.; Uribe-Rubio, J.L.; et al. New meta-analysis tools reveal common transcriptional regulatory basis for multiple determinants of behavior. *Proceedings of the National Academy of Sciences* **2012**, *109*, E1801–E1810, doi:10.1073/pnas.1205283109.
8. Schwarz, B.; Azodi, C.B.; Shiu, S.-H.; Bauer, P. Putative cis-Regulatory Elements Predict Iron Deficiency Responses in Arabidopsis Roots. *Plant Physiol.* **2020**, *182*, 1420–1439, doi:10.1104/pp.19.00760.
9. Eddelbuettel, D.; François, R. **Rcpp**: Seamless *R* and *C++* Integration. *J. Stat. Soft.* **2011**, *40*, doi:10.18637/jss.v040.i08.
10. Eddelbuettel, D. *Seamless R and C++ Integration with Rcpp*; Springer New York: New York, NY, 2013; ISBN 978-1-4614-6867-7.
11. Davis, S.; Meltzer, P.S. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. *Bioinformatics* **2007**, *23*, 1846–1847, doi:10.1093/bioinformatics/btm254
12. Ritchie, M.E.; Phipson, B.; Wu, D.; Hu, Y.; Law, C.W.; Shi, W.; Smyth, G.K. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* **2015**, *43*, e47–e47, doi:10.1093/nar/gkv007.
13. Robinson, M.D.; McCarthy, D.J.; Smyth, G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **2010**, *26*, 139–140, doi:10.1093/bioinformatics/btp616.
14. McCarthy, D.J.; Chen, Y.; Smyth, G.K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research* **2012**, *40*, 4288–4297, doi:10.1093/nar/gks042.
15. Edgar, R. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research* **2002**, *30*, 207–210, doi:10.1093/nar/30.1.207.
16. Pagès, H. Aboyoun, P. Gentleman, R. DebRoy, S. R package version 2.56.0. *Biostrings: Efficient manipulation of biological strings*. **2020**.
17. Durinck, S.; Moreau, Y.; Kasprzyk, A.; Davis, S.; De Moor, B.; Brazma, A.; Huber, W. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics* **2005**, *21*, 3439–3440, doi:10.1093/bioinformatics/bti525.
18. Cherenkov, P.; Novikova, D.; Omelyanchuk, N.; Levitsky, V.; Grosse, I.; Weijers, D.; Mironova, V. Diversity of cis-regulatory elements associated with auxin response in Arabidopsis thaliana. *Journal of Experimental Botany* **2018**, *69*, 329–339, doi:10.1093/jxb/erx254.
19. Zemlyanskaya, E. V., Wiebe, D. S., Omelyanchuk, N. A., Levitsky, V. G., & Mironova, V. V. Meta-analysis of transcriptome data identified TGTCNN motif variants associated with the response to plant hormone auxin in Arabidopsis thaliana L. *Journal of bioinformatics and computational biology* **2016**, *14*(02), 1641009, doi: 10.1142/S0219720016410092.

20. Holm, S. A simple sequentially rejective multiple test procedure, *Scandinavian Journal of Statistics* **1979**, *6*.
21. Hochberg, Y. A sharper Bonferroni procedure for multiple tests of significance, *Biometrika* **1988**, *75*.
22. Benjamini, Y. Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* **1995**, *57*, 289-300, doi: 10.1111/j.2517-6161.1995.tb02031.
23. Benjamini, Y. Yekutieli, D. The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics* **2001**, *29*, 1165-1188.
24. Gupta, S.; Stamatoyannopoulos, J.A.; Bailey, T.L.; Noble, W. Quantifying similarity between motifs. *Genome Biol* **2007**, *8*, R24, doi:10.1186/gb-2007-8-2-r24.
25. Oschepkova, E.A.; Sizentsova, Ya.G.; Mironova, V.V.; Kolchanov N.A. Meta-analysis of transcriptome data detected new potential players in response to dioxin exposure in human. *Archives of Toxicology*. (under review)
26. Guan, Q.; Wu, J.; Zhang, Y.; Jiang, C.; Liu, R.; Chai, C.; Zhu, J. A DEAD Box RNA Helicase Is Critical for Pre-mRNA Splicing, Cold-Responsive Gene Regulation, and Cold Tolerance in *Arabidopsis*. *Plant Cell* **2013**, *25*, 342–356, doi:10.1105/tpc.112.108340.
27. Chiba, Y.; Mineta, K.; Hirai, M.Y.; Suzuki, Y.; Kanaya, S.; Takahashi, H.; Onouchi, H.; Yamaguchi, J.; Naito, S. Changes in mRNA Stability Associated with Cold Stress in Arabidopsis Cells. *Plant and Cell Physiology* **2013**, *54*, 180–194, doi:10.1093/pcp/pcs164.
28. Park, S.; Lee, C.-M.; Doherty, C.J.; Gilmour, S.J.; Kim, Y.; Thomashow, M.F. Regulation of the Arabidopsis CBF regulon by a complex low-temperature regulatory network. *Plant J* **2015**, *82*, 193–207, doi:10.1111/tpj.12796.
29. Kim, Y.; Park, S.; Gilmour, S.J.; Thomashow, M.F. Roles of CAMTA transcription factors and salicylic acid in configuring the low-temperature transcriptome and freezing tolerance of Arabidopsis. *Plant J* **2013**, *75*, 364–376, doi:10.1111/tpj.12205.
30. Lee, B.; Henderson, D.A.; Zhu, J.-K. The *Arabidopsis* Cold-Responsive Transcriptome and Its Regulation by ICE1. *Plant Cell* **2005**, *17*, 3155–3175, doi:10.1105/tpc.105.035568.
31. Jia, Y.; Ding, Y.; Shi, Y.; Zhang, X.; Gong, Z.; Yang, S. The *cbfs* triple mutants reveal the essential functions of *CBFs* in cold acclimation and allow the definition of CBF regulons in *Arabidopsis*. *New Phytol* **2016**, *212*, 345–353, doi:10.1111/nph.14088.
32. Kilian, J.; Whitehead, D.; Horak, J.; Wanke, D.; Weinl, S.; Batistic, O.; D'Angelo, C.; Bornberg-Bauer, E.; Kudla, J.; Harter, K. The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses: AtGenExpress global abiotic stress data set. *The Plant Journal* **2007**, *50*, 347–363, doi:10.1111/j.1365-313X.2007.03052.x.
33. Schlaen, R.G.; Mancini, E.; Sanchez, S.E.; Perez-Santángelo, S.; Rugnone, M.L.; Simpson, C.G.; Brown, J.W.S.; Zhang, X.; Chernomoretz, A.; Yanovsky, M.J. The spliceosome assembly factor GEMIN2 attenuates the effects of temperature on alternative splicing and circadian rhythms. *Proc Natl Acad Sci USA* **2015**, *112*, 9382–9387, doi:10.1073/pnas.1504541112.
34. Thomashow, M.F. PLANT COLD ACCLIMATION: Freezing Tolerance Genes and Regulatory Mechanisms. *Annu. Rev. Plant. Physiol. Plant. Mol. Biol.* **1999**, *50*, 571–599, doi:10.1146/annurev.arplant.50.1.571.
35. Gilmour, S.J.; Zarka, D.G.; Stockinger, E.J.; Salazar, M.P.; Houghton, J.M.; Thomashow, M.F. Low temperature regulation of theArabidopsisCBF family of AP2 transcriptional activators as an early step in cold-inducedCORgene expression. *Plant J* **1998**, *16*, 433–442, doi:10.1046/j.1365-313x.1998.00310.x.
36. Zarka, D.G.; Vogel, J.T.; Cook, D.; Thomashow, M.F. Cold Induction of Arabidopsis *CBF* Genes Involves Multiple ICE (Inducer of *CBF* Expression) Promoter Elements and a Cold-Regulatory Circuit That Is Desensitized by Low Temperature. *Plant Physiol.* **2003**, *133*, 910–918, doi:10.1104/pp.103.027169.
37. Novillo, F.; Alonso, J.M.; Ecker, J.R.; Salinas, J. CBF2/DREB1C is a negative regulator of CBF1/DREB1B and CBF3/DREB1A expression and plays a central role in stress tolerance in Arabidopsis. *Proceedings of the National Academy of Sciences* **2004**, *101*, 3985–3990, doi:10.1073/pnas.0303029101.
38. Yamaguchi-Shinozaki, K.; Shinozaki, K. Organization of cis-acting regulatory elements in osmotic- and cold-stress-responsive promoters. *Trends in Plant Science* **2005**, *10*, 88–94, doi:10.1016/j.tplants.2004.12.012.
39. Doherty, C.J.; Van Buskirk, H.A.; Myers, S.J.; Thomashow, M.F. Roles for *Arabidopsis* CAMTA Transcription Factors in Cold-Regulated Gene Expression and Freezing Tolerance. *Plant Cell* **2009**, *21*, 972–984, doi:10.1105/tpc.108.063958

40. Kidokoro, S.; Maruyama, K.; Nakashima, K.; Imura, Y.; Narusaka, Y.; Shinwari, Z.K.; Osakabe, Y.; Fujita, Y.; Mizoi, J.; Shinozaki, K.; et al. The Phytochrome-Interacting Factor PIF7 Negatively Regulates *DREB1* Expression under Circadian Control in Arabidopsis. *Plant Physiol.* **2009**, *151*, 2046–2057, doi:10.1104/pp.109.147033.
41. Lee, C.-M.; Thomashow, M.F. Photoperiodic regulation of the C-repeat binding factor (CBF) cold acclimation pathway and freezing tolerance in Arabidopsis thaliana. *Proceedings of the National Academy of Sciences* **2012**, *109*, 15054–15059, doi:10.1073/pnas.1211295109.
42. Maruyama, K.; Todaka, D.; Mizoi, J.; Yoshida, T.; Kidokoro, S.; Matsukura, S.; Takasaki, H.; Sakurai, T.; Yamamoto, Y.Y.; Yoshiwara, K.; et al. Identification of Cis-Acting Promoter Elements in Cold- and Dehydration-Induced Transcriptional Pathways in Arabidopsis, Rice, and Soybean. *DNA Research* **2012**, *19*, 37–49, doi:10.1093/dnares/dsr040.
43. Jiang, B.; Shi, Y.; Zhang, X.; Xin, X.; Qi, L.; Guo, H.; Li, J.; Yang, S. PIF3 is a negative regulator of the *CBF* pathway and freezing tolerance in *Arabidopsis*. *Proc Natl Acad Sci USA* **2017**, *114*, E6695–E6702, doi:10.1073/pnas.1706226114.
44. Zhao, C.; Zhang, Z.; Xie, S.; Si, T.; Li, Y.; Zhu, J.-K. Mutational Evidence for the Critical Role of CBF Genes in Cold Acclimation in Arabidopsis. *Plant Physiol.* **2016**, pp.00533.2016, doi:10.1104/pp.16.00533.
45. Agarwal, M.; Hao, Y.; Kapoor, A.; Dong, C.-H.; Fujii, H.; Zheng, X.; Zhu, J.-K. A R2R3 Type MYB Transcription Factor Is Involved in the Cold Regulation of CBF Genes and in Acquired Freezing Tolerance. *J. Biol. Chem.* **2006**, *281*, 37636–37645, doi:10.1074/jbc.M605895200.
46. Dong, M.A.; Farre, E.M.; Thomashow, M.F. CIRCADIAN CLOCK-ASSOCIATED 1 and LATE ELONGATED HYPOCOTYL regulate expression of the C-REPEAT BINDING FACTOR (CBF) pathway in Arabidopsis. *Proceedings of the National Academy of Sciences* **2011**, *108*, 7241–7246, doi:10.1073/pnas.1103741108.
47. Kidokoro, S.; Yoneda, K.; Takasaki, H.; Takahashi, F.; Shinozaki, K.; Yamaguchi-Shinozaki, K. Different Cold-Signaling Pathways Function in the Responses to Rapid and Gradual Decreases in Temperature. *Plant Cell* **2017**, *29*, 760–774, doi:10.1105/tpc.16.00669.
48. van Helden, J., André, B., Collado-Vides, J. A web site for the computational analysis of yeast regulatory sequences. *Yeast* **2000**, *16*, 177-187. doi: 10.1002/(SICI)1097-0061(20000130)16:2<177::AID-YEA516>3.0.CO;2-9.
49. Heinz, S.; Benner, C.; Spann, N.; Bertolino, E.; Lin, Y.C.; Laslo, P.; Cheng, J.X.; Murre, C.; Singh, H.; Glass, C.K. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell* **2010**, *38*, 576–589, doi:10.1016/j.molcel.2010.05.004.
50. Liu, T.; Ortiz, J.A.; Taing, L.; Meyer, C.A.; Lee, B.; Zhang, Y.; Shin, H.; Wong, S.S.; Ma, J.; Lei, Y.; et al. Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biol* **2011**, *12*, R83, doi:10.1186/gb-2011-12-8-r83.
51. Khan, A.; Fornes, O.; Stigliani, A.; Gheorghe, M.; Castro-Mondragon, J.A.; van der Lee, R.; Bessy, A.; Chèneby, J.; Kulkarni, S.R.; Tan, G.; et al. JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Research* **2018**, *46*, D260–D266, doi:10.1093/nar/gkx1126.
52. Hu, J.; Wang, D.; Li, J.; Jing, G.; Ning, K.; Xu, J. Genome-wide identification of transcription factors and transcription-factor binding sites in oleaginous microalgae Nannochloropsis. *Sci Rep* **2015**, *4*, 5454, doi:10.1038/srep05454.
53. Chinnusamy, V.; Zhu, J.; Zhu, J.-K. Cold stress regulation of gene expression in plants. *Trends in Plant Science* **2007**, *12*, 444–451, doi:10.1016/j.tplants.2007.07.002.
54. Zhao, C.; Lang, Z.; Zhu, J.-K. Cold responsive gene transcription becomes more complex. *Trends in Plant Science* **2015**, *20*, 466–468, doi:10.1016/j.tplants.2015.06.001.
55. Zeng, Z.; Zhang, W.; Marand, A.P.; Zhu, B.; Buell, C.R.; Jiang, J. Cold stress induces enhanced chromatin accessibility and bivalent histone modifications H3K4me3 and H3K27me3 of active genes in potato. *Genome Biol* **2019**, *20*, 123, doi:10.1186/s13059-019-1731-2.