Genome-wide analysis of a wheat transcription factor family: the power of bioinformatics resources

27 May 2020

— — —

Q&A session

Presenters

- **Susanne Schilling**
  Assistant Professor, School of Biology and Environmental Science, University College Dublin, Ireland
- **Rainer Melzer**
  Assistant Professor, School of Biology and Environmental Science, University College Dublin, Ireland

References:

- Schilling S, Kennedy A, Pan S, Jermiin LS, Melzer R. Genome-wide analysis of MIKC-type MADS-box genes in wheat: pervasive duplications, functional conservation and putative neofunctionalization. *New Phytol*. 2020;225(1):511-529. https://doi.org/0.1111/nph.16122

- Schilling S, Pan S, Kennedy A, Melzer R. MADS-box genes and crop domestication: the jack of all traits. *J Exp Bot*. 2018;69(7):1447-1469. https://doi.org/10.1093/jxb/erx479

The webinar recording is available on the IWGSC YouTube channel at https://youtu.be/efbph7f2jXU

Q: how can we check for the recombination rate?

Answer Given during webinar – Check recording at 51'12" mark

Q: How can we annotate genes using bioinformatics approaches except Blast2go? is there any other software?

There are a multitude of different programs to be doing this. It really depends on what your data is and to which level it is annotated. One frequently used program for identifying genes in HMMER, for gene annotation AUGUSTUS is used for genome wide level. I recommend searching for publications which describe similar analysis that you want to do and reading the methods section carefully.

Q: For expression analysis you used gene investigator. we have RNA seq data of wheat genes in different tissue available online. how can it be used? Do we need any software to get fpkm value and how do we use it?

If you have raw sequence data, you have to map it to the reference genome and go through the steps of analysis. I recommend checking out the galaxy platform for that, if you are new to RNA seq.

Q: have you investigated MADS- box gene in Triticum urartu?

No, this has not been done. But it would be really interesting to study the genes in the wheat progenitor species.

Q: Is response to biotic stress a novel function for MADS transcription factors, or are there any previously known MADS genes that respond to biotic stress?

Answer Given during webinar – Check recording at 54’20” mark

Q: Can the genes containing Simple sequence repeats can be the candidate genes for crop improvement? I have filtered out the sequences containing SSRs, can u suggest how to perform phylgeny of those?

If the SSR is within a protein coding sequence this can have "fine-tuning" effects to the protein function (e.g. modulating the activity of a transcription factor). If the SSR is part of a regulatory region this can influence the transcription level of the gene. I am sure you can build phylogies with SSRs, but I am not expert on this.

Q: Could the duplication of these genes be related to the fact that wheat is Hexaploid (so we would normally find homeologs), or that wheat duplicated this gene due to its evolutionary importance?

we do find homoelogs, and indeed we also find high homoelog retention. But the number of genes is still higher than expected, most probably caused by additional gene duplications .

Q: What biotic/abiotic stresses do you look for in heatmap data? Have you found that any stresses always cause larger expression changes in most genes?

I am looking in general without bias. There is no stress which cause alteration in most genes.

Q: Are there any reports of genetic engineering of MADS box genes for abiotic/biotic stress tolerance in any crop

Indeed, check out our 2018 Review, Schilling et al. 2018, Journal of Experimental Botany

Q: Did you check the expression of AGL17?

Yes indeed. These genes are expressed in wheat in the root, but also in some other tissues and during abiotic stress response (you can find more about that in the paper).

Q: is there any role in disease resistance?

Potentially yes, but further investigations are necessary.

Q: Can heat sock protein gene be marked by conserved gene analog markers ?

I am sorry, I don’t know.

Q: Could you please explain again how MADS box genes are involved in domestication?

Check out our 2018 Review, Schilling et al. 2018, Journal of Experimental Botany

Q: Does sample alignment of the genes of the family with the genes of other species enough?

MADS-box genes have been extensively studied in other organisms. We used two well studied species (Arabidopsis and rice) as comparisons. Genes from both species were sorted into the expected common subclades, hence we decided it was sufficient to sort wheat genes as well.

Q: Does InDels should be removed.

InDels in the alignment get removed. We used the program Alistat to select for sites in the alignments with little to no gaps (see methods)

Q: I have used the Triticum aestivum v2.2 from Phytozome https://phytozome-next.jgi.doe.gov/info/Taestivum_v2_2 as a reference genome for RNA-seq data that I have, is it possible to publish results with this version of genome?

This genome appears to be from 2014, it is the former IWGSC genome. I would consider using the up to date version 1.1 or even 2.0 for analyses.

Q: Have you done any functional validation in wheat (knockout/overexpression in wheat)?

Not yet. We have another project looking into specific subclades where we are also working on mutants.

Q: how to make synteny plot online plz suggest any you tube video

https://www.uni-giessen.de/fbz/fb08/Inst/bioinformatik/software/EDGAR/Features/synplots
https://genomevolution.org/wiki/index.php/Syntenic_dotplot
https://es.coursera.org/lecture/comparing-genomes/synteny-block-construction-Z2rfu

Q: By using genotyping by sequencing we can go for phylogenetic study of MADS box gene in wheat or any research going on by using genotyping by sequencing technology to find different type of MADS box gene in wheat crop

Answer given during webinar – Check recording at 52'15" mark

Q: Gene mining and identification in the wheat genome may be tricky sometimes because of polyploidy. Do you have any tips or specific suggestions for this?

The IWGSC genome is very well annotated, hence we did not have problems identifying genes and distinguishing different homoeologs.

Q: How can we find the relationship between transcription factor each other's? Or by using which tool?

Phylogenetic relationship? Build an alignment and then a tree using a maximum likelihood method or bayesian. I am using MAFFT and Iqtree, links in the PDF

Q: Could you please elaborate on how to group genes on phylogeny? did you check for gene structure for all the genes?

We compared the wheat genes with previously identified genes from other species. Because MADS-box genes have been studied in depth for 2 decades, we could rely on previously defined subclades.

Q: If genes of same family are highly diverse and resulting in low bootstraps, what option would you go for? is trimming sequences is a good option?

We used the program alistat to select for sites with fewer gaps (see methods of our paper and alistat documentation).

Q: How can I shortlist MADS box genes of rice, from my RNA-seq data?

Sure, look into Arora et al (2007, BMC Genomics) for a comprehensive analysis of MADS-box genes in rice.

Q: How the authors were able to find FLC-like genes in wheat? In phylogenetic tree of the reference article (Schilling et al. 2020), putative wheat FLC genes were grouped with rice MIKC*-type gene (OsMADS39, Arrora et al. 2007) and Arabidopsis FLC genes were grouped into separate distinct clade. The putative FLC genes could be MIKC*-type genes in wheat as indicated by phylogentic tree.

Good question. The topology we find in the tree is actually expected as it is documented that Ath FLC-like genes have been duplicated in the lineage leading to Ath. Hence we expect all Ath FLC-like genes in a sister clade to all the grass FLC-like genes. Grass FLC-like genes do have a non-canonical K-domain which is not always easily identifiable. That is one of the reasons that at the time of publication of Arora et al. (2007) it was not known that grasses also have FLC like genes, it can be challenging to identify them. However, grasses turned out to have FLC-like genes, including OsMADS51 and OsMADS37 (Ruelens et al 2013). Hence, we sorted the wheat genes clustering with them into the FLC subclade. The expression pattern is in accordance with a role in flowering time, while MIKC*-type genes would be expected to have a very narrow expression profile restricted to pollen ripening.

Q: How can one join the IWSC and the sequencing effort?

There are three ways to be involved in the IWGSC - as a general member, you may register at www.wheatgenome.org; if you want to lead a project within the IWGSC you may contact me (Kellye Eversole, eversole@eversoleassociates.com) to discuss the possibility of becoming a member of our scientific coordinating committee; or you may become a sponsor by contacting me as well.

Q: Does the use of WheatMine open for all users or one has to be a member

It is not necessary to be a member of the IWGSC to utilize WheatMine.

Q: Did you observe any preferential expression patter of MADS-Box genes (under native or specific treatment) from different sub-genomes?

yes, while some homoelogs were very balanced, we also observed some imbalanced homeologs, see figure 5C of the paper and suppl material for this analysis (triangular plot).

Q: Have you done the homeologs specific expression analysis?

yes, while some homeologs were very balanced, we also obrserved some imbalanced homeologs, see figure 5C of the paper and suppl material for this analysis (triangular plot).

Q: What is the best way to start with GWA of gene family in wheat?

I am unfortunately not an expert on this.

Q: please elaborate how to use RNA seq data directly from bio projects of NCBI or EBI, in case we have specific treatment to look into?I mean the easy to use tools for such an analysis using public raw RNASeq data

I would recommend using the Galaxy platform if you are not familiar with RNA seq analysis. They have free storage space, you can import data from NCBI and do analysis on the platform. They have a lot of tutorials on how to do RNA seq on the platform.
Check also recording at 55’45” mark

Q: is sra data is same for all gene or they are specific and how to get sra data for expression analysis?

Expression patterns can differ between homeologs (see our publication figure 5C. NCBI sequence archive for raw data.

Q: In which subgenome MADS box genes are prominently found and why????

Answered during webinar – Recording at 48’15’’ mark

Q: What can be precluded from MADS genes in telomeric location ?

This subtelomeric regions are hot spots of recombination and have higher gene density. We also touched on this during the Q&A in the live webinar.

Q: What is the fate of MAD box genes in vegetative parts like root?

Answered during webinar – Recording at 50' mark

Q: What is the role of MAD box genes in NUE?

I don't know what NUE is, sorry.

Q: Are we able to transform a specific transcription factor say MADS gene to increase the efficiency of transcription in other related organisms ..

Heterologous expression of MADS-box genes in other species has been done multiple times, however usually this is done because the species under investigation cannot easily be transformed. Hence scientists use model plants like arabidopsis and rice for expression. Some examples for crop improvement can be found in Schilling et al 2018, JXB.

Q: Does the wheat mine website include the updated annotations? Do the LC genes also included in the wheat mine categorizations?

I think they are using IWGSC Ref1.1, including HC and LC genes.

Q: How do you determine the paralog/homoelog relationship for members in the MADS subfamily, especially those subfamily underwent extensive duplication/deletion? Do you apply any criteria or use some software for define homoelogs in the traids?

We do specify this in our methods section in the paper. We used a combined approach of phylogenetic support, previous characterization and synteny

Q: are these data published in Schilling et al 2020? is it possible to get the list of MADS genes coordinates?

Yes, we have a detailed list of genes in the supplementary material (Table S2)

Q: how to perform tissue specific expression of mads genes and how to remove duplicates

start with wheat-expression.com

Q: Is there any web based program to identify target genes of our candidate transcription factors??

I don't know about that, unfortunately. You can analyze co-expression in genevestigator. But target genes are difficult to identify with a purely bioinformatics approach.

Q: what to do when the transcriptome data is available with older nomenclature of genes or if there are different studies but with different nomenclature of genes?

It depends on how old the identifiers. IWGSC Ref 1.0 and 1.1 are both on wheatmine, which you can use to export a table with both IDs for one gene. Old genome versions (TGAC) are also on Ensembl plants and the jbrowser on https://wheat-urgi.versailles.inra.fr/Tools also has the TGAC genes annotated, so you might be able to convert your IDs there. Maybe you can find more information here: https://www.researchgate.net/publication/338031552_A_roadmap_for_gene_functional_characterisation_in_wheat