Whealbi

Wheat and barley Legacy for Breeding Improvement

Grant agreement number: FP7-613556

Collaborative Project
SEVENTH FRAMEWORK PROGRAMME

Deliverable 5.2

List of validated candidate genes regulating seed storage proteins genes in response to nutrient supply

Due date: M60

Actual submission date: M60

Project start date: January 1st, 2014 **Duration:** 60 months

Workpackage concerned: 5

Concerned workpackage leader: UZH

Dissemination level: PU

Table of contents

Glossary and Definitions

SSP	Seed Storage Proteins
TFs	Transcription Factors
Y1H	Yeast One Hybrid Assay
SPA	Storage Protein Activator
SHP	SPA Heterodimerizing Protein
CCRR	Conserved Cis-Regulatory Region

Summary

Objectives Wheat (*Triticum aestivum*) end-use quality depends mainly on seed storage protein (SSP) concentration and composition. SSP are synthetized by genes, which are mainly regulated at transcriptional level by transcription factors (TFs). This first level of the transcriptional regulation (i.e. TFs directly interacting with the promoter of SSP genes) has been established. These TFs are called SPA, SHP, PBF, SAD, GAMYB, MYBS3, MCB1 and FUSCA3. Currently, knowledge on the transcriptional regulation of these TFs (second level of regulation) is poor. In this context, the objectives of task 5.1 are

1/ to identify genes of TFs (named candidate TFs) that contribute to this second level of regulation,

2/ to validate (or not) their effect on seed storage proteins synthesis.

Rationale:
To detect candidate TFs, we used two approaches: 1/ *In vitro* approach to identify DNA-proteins interactions by Yeast-One Hybrid (Y1H) screening and 2/ *In silico* approach based on regulatory network using RNAseq data.
To validate the effects of the novel candidate TFs, we also used two strategies. Candidate genes revealed by *in vitro* approach were tested by a functional analysis based on transient expression assay. All candidate genes coding for TFs involved in the second level of regulation (i.e. able to regulate SPA, SHP, PBF, SAD, GAMYB, MYBS3, MCB1 and FUSCA3, here are called target FTs), were used for allele mining. Their allelic diversity was deduced from the wheat exome of 500 wheats available in the project. Polymorphisms were used for association analysis in a part of the Whealbi collection for wheat, which has been previously phenotyped for SSP synthesis (Plessis et al, 2013).

Teams involved: INRA

Bread wheat (*Triticum aestivum*) is mainly used after transformations. All these transformations require a given end-use quality, which depends on seed storage protein (SSP) concentration and composition. Expression of SSP genes is mainly regulated by transcription factors (TFs), which specifically bind short conserved DNA motifs in the promoter region of coding genes (called cis-motifs). This first level of transcriptional regulation of SSP genes implies TFs able to bind cis-motifs included in their promoter. These TFs are SPA, SHP, PBF, SAD, GAMYB, MYBS3, MCB1 and FUSCA3. These TFs could be also regulated by transcriptional proteins (second level of regulation). Currently, knowledge on the TFs able to modulate the expression of genes coding for SPA, SHP, PBF, SAD, GAMYB, MCB1, MYBS3 and FUSCA3 is poor. Therefore, we aimed at identifying the TFs involved in the second level of regulation (called candidate TFs), studying their polymorphism using exome data provided by Whealbi project (allele mining). Polymorphisms were then used to study their effects on the SSP content and composition. In this report, we identified TFs putatively involved in the second level of regulation (candidate TFs). Allele mining performed to find polymorphisms within the sequences of these candidate TFs allowed genotype-phenotype associations to statistically validate their role on SSP synthesis.

I. Identification of candidate transcription factors

A wet approach was used. It was completed by an *in-silico* analysis.

1.1. *In vitro approach*: DNA-protein interactions by Yeast one Hybrid (Y1H) screening

1.1.1 Material and methods

SPA and *SHP* gene promoters (523 pb upstream of the start codon of the B homeolog *SPA* gene and 2297 pb upstream of the start codon of the three homoeologous *SHP* genes) were annotated using a home-made software PlantPAD (Ravel et al., 2014) to detect cis-motifs potentially able to bind TFs. For SHP, we focused on regulatory regions conserved between homoeologous promoters and their barley orthologous *BLZ1* promoter because of the assumption that conserved cis-regulatory regions (CCRRs) might be involved in the global regulation of *SHP*.

The Y1H assay involves two components (1) a reporter construct with DNA of interest cloned upstream of a gene encoding a reporter protein, called the ‘bait’; and (2) an expression construct that generates a fusion between a TF of interest and a yeast transcription activation domain (AD), called the ‘prey’.

The promoter of *SPA* was cloned in the reporter vector pHIS-I and introduced into yeast reporter strain EGY48. CCRRs of the *SHP* promoter were cloned into pINT1HISNB vector and introduced into yeast strain YM4271. For each CCRR (containing one or two cis-motifs surrounded by 10 to 15 bp), two tandem repeated copies were used as baits for Y1H screening against the wheat cDNA expression library. Each bait was tested by yeast autoactivation test with 3AT, used to eliminate auto-activation of the bait.

Two type of prey were used: 1/ a cDNA expression library constructed with mRNA from *T. aestivum* (cv. Récital) whole grains collected at 80, 160 and 220 °Cdays after anthesis. This cDNA entry library was cloned into the pGADT7-REC vector (Clontech) *via* an LR recombination reaction directly in yeast cells of Y187 strain (Bednarek, 2012). 2/ a cDNA expression library composed of 1,200 Arabidopsis TFs developed by the Regia consortium (Par-Ares et al, 2002). These TFs were cloned into the pDEST22 vector and introduced into yeast strain YM4271.

The Y1H screening was performed by mating of the EGY48 or YM4271 yeast reporter strain with the YM4271 or Y187 strain containing the prey sequences respectively on a selective medium lacking histidine and in presence of the adequate 3-AT concentration (determined for each bait). Each positive clone from the cDNA of Recital was sequenced. The resulting sequences were used in a blastn search on NCBI or Ensembl database, then the corresponding protein sequence was regarded via Gene Ontology to select TFs. One by one Y1H experiments were also performed to confirm DNA-protein interactions. Wheat orthologs to Arabidopsis TFs able to bind *SPA* promoter have to be identified.

1.1.2 Results

Two putative Arabidopsis TFs were identified by the Y1H screening of the *SPA* gene promoter with the cDNA *Arabidopsis* TFs library: a TF of the NAC (NAM (No Apical Meristem) – ATAF (Arabidopsis Transcription Activation Factor) – CUC (Cup-shaped Cotyledons)) family (NAC42, At3g12910) and an Homeobox gene1 (ATH1, At4g32980). In accordance with these results, *SPA* promoter includes the putative cis-motifs able to bind these TFs. One by one Y1H experiment revealed that only NAC42 interacts with *SPA* gene promoter.

The *in silico* cis-motifs annotation of the orthologous gene promoters (the three copies of *SHP* and the barley *BLZ1* promoters) revealed eight Conserved Cis-Regulatory Regions (CCRR1 to 8, Figure 1).

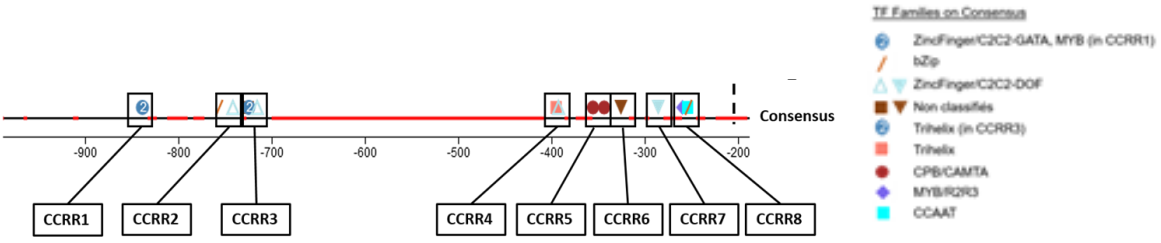

Figure 1: ***In silico* annotation of the consensus of *SHP* and *BLZ1* promoters.**
Positions are indicated relatively to the start site. Transcription Start Site (TSS) was positionned at -202 bp. Cis-motifs found into the four analysed sequences generate the consensus which is composed of eight Conserved Cis-Regulatory Regions (CCRR1 to 8) studied in the Yeast-one-Hybrid experiment.

Two tandem repeated copies of the each CCRR (containing one or two cis-motifs surrounded by 10 to 15 bp) were used as baits for Y1H screening against the cDNA expression library from *T. aestivum*. The yeast autoactivation test with 3AT eliminated the baits CCRR4 and 7 for which it was impossible to found a selective

medium that prevent their high level of autoactivity. For the others CCRR, positive clones on selective medium were obtained for the CCRR1 and CCRR5 DNA baits.

Two TFs belonging to the homeobox-leucine zipper family (HDZip) were identified to interact with the CCRR1: Homeobox-leucine zipper protein (ROC8) and Homeodomain-leucine zipper transcription factor (HDZipI-1). For the screen with CCRR5, most of clones were false positives. On 1,000 clones, only nine have a Gene Onthology-Molecular function related to DNA binding. Among these 9 clones, we found cDNAs of two TFs: *T. aestivum* ethylene responsive transcription factor 5a (ERF5a) and Zinc finger CCCH domain-containing protein 54 (C3H).

1.2 *In silico* approach: regulatory network construction

1.2.1 Material and methods

RNAseq data were obtained in relationship with the program BreedWheat, which also provided data of proteomics and metabolomics. mRNAs were extracted from grains of *T. monococcum*, a diploid wheat model species. Grains were harvested at different developmental stages (300, 400, 500 and 600°Cday after anthesis) from plants cultivated under different nitrogen (N) and sulfur (S) nutritions (Figure 2). These data were used to infer gene networks with the RulNet platform (http://rulnet.isima.fr). RulNet allows the connection between different -omic entities according to rules defined by the scientists with a biological meaning. Rulnet makes also possible to focus on rules involving attributes of special interested, which have to be declared as central attributes.

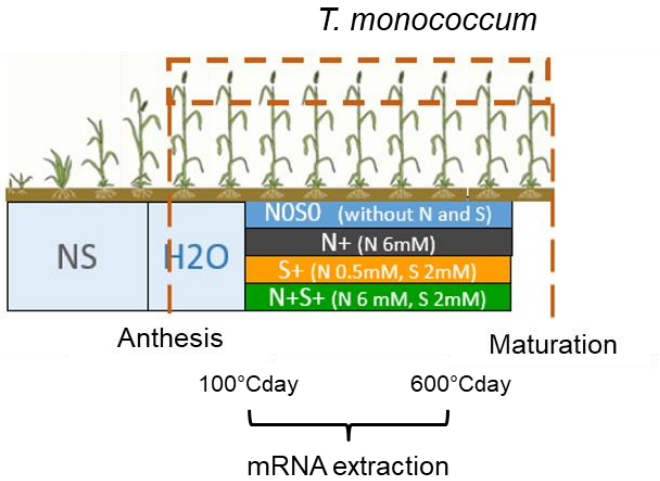

Figure 2: Experimental design of *T monococcum* plants cultivated under controlled conditions.
Four nutritions were applied between 200 and 700°Cd after anthesis (Control N0S0, N+, S+ and N+S+). For RNA extraction, grains were collected every 100°Cd after anthesis from 300°Cd to 600°Cd after anthesis.

In a first analysis, differentially expressed genes were identified. They were used to infer a co-expression network between co-expressed/co-accumulated genes, metabolites and proteins. In a second analysis, genes coding for TFs were extracted from the set of genes expressed in the grain by blast against PlantTFDB, which contains 3,606 wheat TFs (http://planttfdb.cbi.pku.edu.cn/). The rule used was: if the expression level of a gene coding for a candidate TF (TF belonging to the second level of regulation) was superior (or inferior) to a given threshold, then the difference of expression for target genes (the gene coding for a TF of the first level i.e. SPA,

SHP, SAD, MYBS3, PBF, MCB1, GAMYB, FUSCA3, which were considered as central attributes) between two following stages has to be positive or negative.

1.2.2 Results

RNAseq analysis showed that 22,910 genes out of 32,047 annotated genes of *T. monococcum* were expressed in grain. Three hundred eighty six out of the 22,910 genes were differentially expressed in the grain according to treatment. The set of differentially expressed genes contained 8 genes coding obviously for TFs, none of them were target TFs. It was thus not possible to study rules between differentially expressed and target TFs. It was decided to conserve differentially expressed genes coding TFs for allele mining as they could be implied in the SSP in conditions of nutrition stress. In addition, results showed a strong impact of a high-N supply without any added S. These results are described in the manuscript “Integrative analysis of the einkorn grain reveals new mechanisms of response to sulfur deficiency” by Bonnot *et al*. submitted to Plant Journal.

A second analysis specifically focusses on TFs. The blast search against the 3,606 *T. aestivum* TFs of PlantTFDB revealed that 1,159 TF genes are expressed in the grain of *T. monococcum*. In the following, we focussed on 580 of these TFs (50%, representing 53 TF families), which have a level above the median level calculated from the expression data of all TFs. Rules generated were applied in the set of these 580 TFs. The networks (see figure 3 for an example) obtained resulted in a list of 182 putative interactor genes (TFs candidates) belonging to 36 TFs families.

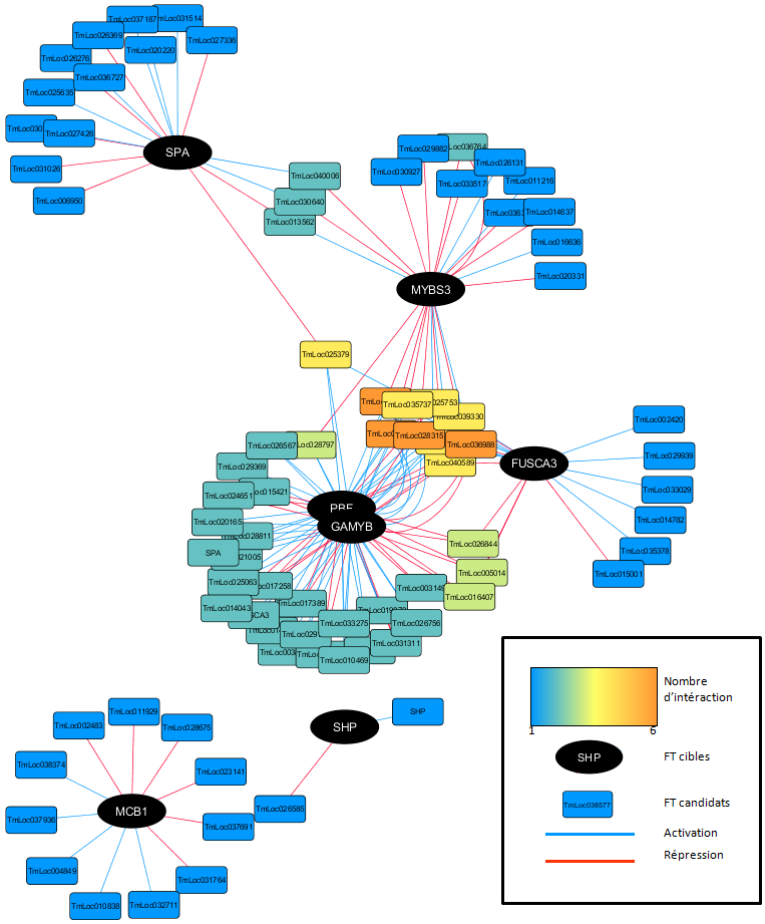

Figure 3: A network of TF transcripts during einkorn grain filling.
Rules were used between TF candidates and TF targets. The difference expression of TF targets between two following stages is positive or negative. The TF candidate expression is > or < of a threshold. This rule generates activation or repression links between TFs

These two approaches resulted in a list of eight plus 182 candidate TFs. Two of them were found by both approached. Therefore, we obtained 188 candidates TFs.

II. Validation of candidate genes

2.1 Functional validation of candidate TFs able to bind *SHP* promoter by transient expression assay

The functional relevance of the *in vitro* interactions between *SHP* promoter (CCRR1 and CCRR5) and its putative regulators (HdZip1, ROC8, ERF5 and C3H) was investigated *in vivo* by transient expression assay on wheat endosperms.

2.1.1 Material and methods

The promoter of the *SHP* gene from the A genome, named pSHP, was used as reporter for particle bombardment and contained 1,182 bp upstream region of the start codon. This sequence was amplified from DNA of wheat cultivar Chinese Spring and was cloned directly in the entry clone of Gateway system pDONRP4-P1. The complete CDS (coding Sequence) of the four putative regulators (synthesized by Sigma) of *SHP* gene were used as effector constructs, and cloned under the control of the maize *Ubiquitin* promoter plus the first intron of the *Ubiquitin* gene. All constructs used for transient expression assays were obtained using Gateway technology (Invitrogen). Three entry clones were used (pDONRP4-P1R, pDONR221 and pDONRP2R-P3) to obtain the expression vector pDESTR4-R3. pDONRP4-P1R contained the maize *Ubiquitin* promoter plus the first intron of the *Ubiquitin* gene. pDONR221 contained the reporter genes (either *GUS* or *GFP*) or the ORF (*Roc8, HdZip1, C3H and ERF5*). pDONRP2R-P3 contained the 3'-terminator *nopaline synthase* gene (3'-NOS). Six pDESTR4-R3 based expression vectors (pSHP with *GUS* gene reporter, pGFP, pRoc8, pHdZip1, pC3H and pERF5) were thus created (Figure 4a).

Immature Endosperms from cv. Récital were collected at 225 °Cdays after anthesis from plants grown in a controlled culture chamber averaging 19 °C per day. Gold particle coating and bombardments were performed according to Ravel *et al.* (2014). After bombardment, endosperms were incubated for 24 h in the dark at 30 °C in a Murashige and Skoog medium supplemented with 3% (w/v) sucrose. *GUS* and *GFP* expression were quantified according to Ravel *et al.* (2014). The pGFP construct was used to determine the efficiency of bombardment. The expression results were normalized by dividing the number of GUS foci by the number of GFP foci. For each combination of reporter/effector, at least three independent bombardments of three Petri dishes containing eight endosperms each were performed.

2.1.2 Results

Immature endosperms of wheat (Cv Récital) were transiently transformed by particle bombardment of the reporter alone (pSHP) or in combination with the effectors. Co-bombardments of pSHP with the pRoc8, pHDZipI-1 increased significantly GUS activity ($P < 0.001$ and $p < 0.05$) compared with that driven by pSHP alone, demonstrating an activation of *SHP* (Figure 4b) by these two TFs. Co-transfections of pSHP with pC3H, pERF5 did not modify significantly GUS expression. This result suggested that C3H and ERF5 do not regulate *SHP* expression in these conditions.

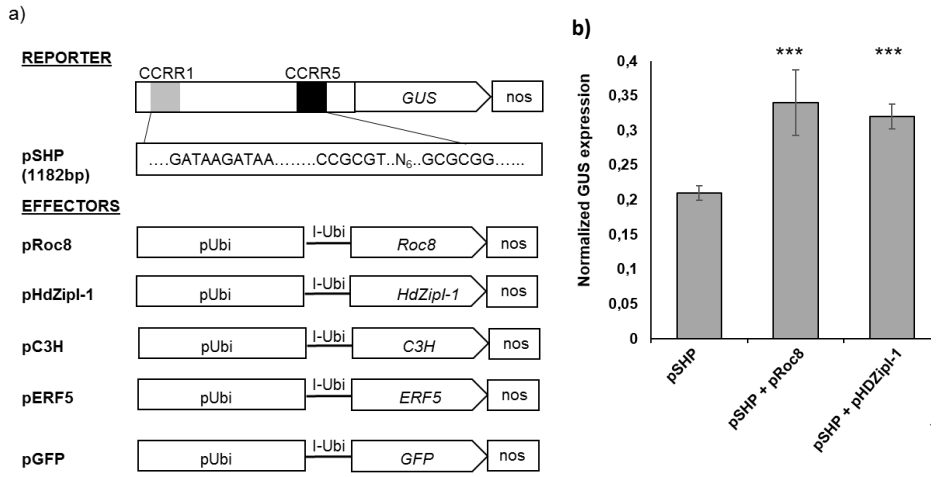

Figure 4: Transient expression assays of SHP promoter activity with ROC8, HdZIpI-1, C3H and ERF5 in wheat developing endosperms.

(a) Schematic representation of the reporter and effector constructs.

(b) Transient expression assays in developing wheat endosperms co-bombarded with the pSHP reporter with equimolar ratio of the TF effectors. The number of bombarded endosperms from at least three independent particle bombardments varied between 34 and 84. Asterisks above the data indicate significant differences between expression values for the reporter with the effector and the reporter without effector (***, $P < 0.001$; **, $P < 0.01$; *, $P < 0.05$).

2.2 Allele mining and association to validate candidate TFs

Y1H assays provided four wheat candidate TFs. Integrative biology provided 188 candidate TFs from *T. monococum*. Genetic association was performed in wheat using polymorphisms included in the sequence of all these candidate TFs to statistically validate their effect on SSP synthesis.

2.2.1. Material and methods

The three orthologs of wheat candidates were blasted on wheat pseudomolecule. The wheat orthologs/paralogs of *T. monococcum* candidates were searched by a blast analysis using sequences of the genes coding these TFs against the wheat pseudomolecule (The International Wheat Genome Sequencing Consortium, 2018). The blast results were analysed to find the coordinates of the orthologs and paralogs of candidate TFs on the wheat pseudomolecule. These

positions were used to extract polymorphisms within the wheat sequences of all candidate TFs from the variant file produced by Whealbi. Genetic association was performed in a part of the wheat collection studied in Whealbi, which was phenotyped for SSP content and composition by Plessis et al (2013) in three environments (at Clermont-Ferrand, Le Moulon with a high level of nitrogen fertilisation and Le Moulon with a low level of nitrogen fertilisation). Briefly, the traits analysed concerned total SSP, and all SSP fraction i.e. total gliadins and glutenins, high and low molecular weight glutenins, □/□-, □-, □1-2 and □5 gliadins. These traits were expressed in different units: % of grain N, µg of N allocated in a given fractions (more details in Plessis et al, 2013). The analysis was performed using the polymorphism with a MAF> 2.5 and the mixed model (Yu et al, 2006) comprising the kinship matrix (K) that accounts for relatedness among accessions to limit spurious associations. The significance of associations was tested with an F-test. False discovery rate was performed to correct for multiple testing. Calculations were done with the GWAS function included in the R package rrBLUP (Endelman, 2011). Each location was analysed separately.

2.2.2. Results

The candidate TFs from Y1H and from integrative biology corresponded to 21 and 639 locations (genes) on the pseudo-molecule, respectively. Allele mining was then performed for 660 genes. In this set, 262 genes (40%) contained 1,386 polymorphisms (SNPs) and 120 insertion-deletions (size >1 nucleotide) i.e. 1 polymorphism for about 313 nucleotides. Among candidate TFs from Y1H, only the A and D copies of *ERF5* contained polymorphisms and were thus included in the set of 262 genes.

Association could be performed in a set of 108 lines included in the Whealbi collection and phenotyped for SSP composition. Due to missing data, association analysis could be performed in 105 out of 108 accessions.

Eighteen candidate TFs were found to be associated at Pvalue=0.001 with at least one trait concerning SSP composition (Table 1).. The two copies of ERF5 were not significantly associated with traits related to SSP composition.

As shown in table 1, nine candidate TFs were associated at 0.001 with traits related to gliadins. The B copy of the C3H TF, homolog to TmLoc040006, contained a SNP at position 329,359,906 on the chromosome 7B, which is associated with several traits related to gliadins in the three locations, with a quite low log10(*p*-value) of -3.8 . Moreover,as these associations were found in all the environment for similar variable, they can be considered as robust. Similarly, the association found for the A copy of the homologous gene to Tm039820 (C2H2 family) located on the chromosome 1A appeared to be robust. This gene contained many polymorphisms. Several of them were associated to SSP, gliadin and glutenin quantities in the grain in several environments. The SNP located on 1A at 9,582,470 is probably the most interesting mutation as associated with several traits with a maximum score of 3.9. These two genes giving robust “associations” were also associated to the gliadin to glutenin ratio or to the high molecular weight to low molecular weight glutenin ration, which were important for the technological quality of wheat.

Table 1. Associations at *p*-value<0.001 between genes coding candidate TFs identified thanks to T. *monococcum* and traits related to SSP synthesis. Associations was analysed between polymorphisms within bread wheat sequence of genes homologous to genes coding for TFs of *T. monococum*.
Blue indicates TFs associated with gliadins. * indicates associations found in several locations.
The TF indicated in bold characters is significantly associated with traits related to N content, gliadins and glutenins synthesis.

T. monococcum candidate TF	Family	Chromosomal location of its wheat homolog
TmLoc039820	C2H2 family protein	chr1A
TmLoc030640	HB-PHD family protein	chr1A
TmLoc014043	bZIP family protein	chr1B
TmLoc026756	Trihelix family protein	chr1B
TmLoc023107	C2H2 family protein	chr1B
TmLoc041873	NAC family protein	chr3A
TmLoc013010	HB-PHD family protein	chr3A
TmLoc041017	WRKY family protein	chr3B
TmLoc010838	GRAS family protein	chr3B
TmLoc033275	C3H family protein	chr3B
TmLoc023394	C2H2 family protein	chr4A
TmLoc021005	TALE family protein	chr5A
TmLoc038798 *	ARR-B family protein	chr6B
TmLoc027274 *	bZIP family protein	chr7A
TmLoc040006 *	C3H family protein	chr7B
TmLoc020220 *	SBP family protein	chr7B
TmLoc013461	MYB_related family protein	chr7B
TmLoc012338	C3H family protein	chr7B

Conclusion

Our analysis was performed in a set of 105 accessions for which SSP composition was available. Therefore, associations reported has to be considered with precaution and need to be confirmed using larger samples. Nethertheless, some of them were found in several location and for several related traits and can be considered as robust. The association with the C3H TF identified by Y1H could not tested as the genes coding for this TF located on the homoeologous chromosomes 2 contained no polymorphism. However, association analysis revealed that three TFs belonging to the C3H family out of 18 were statistically associated with the gliadin synthesis. One of these C3H TF appeared important as being associated with several traits related to gliadins in the three environments studied. Therefore, our results suggest that this family could be involved in the regulation of SSP genes.

In addition, the Whealbi exome data do not reflect the entire wheat exome as only 50% of the genes were actually captured. Therefore, we could not find polymorphisms in more than 50% of our candidate genes. Other sources of polymorphisms should be used to complete genotyping data as markers from Breadwheat.

Publication submitted: Integrative analysis of the einkorn (*Triticum monococcum*) grain response to nitrogen and sulphur supply. Bonnot et al. Plant J.

References

Bednarek J (2012) Analyse fonctionnelle de TaGW2, une E3 ligase de type RING, dans le développement du grain de blé tendre (*Triticum aestivum*). Sciences agricoles. Université Blaise Pascal - Clermont-Ferrand II, 2012. Français. <NNT: 2012CLF22309>. <tel-00857341>

Endelman JB (2011) Software for genomic prediction with the RR-BLUP mixed model. <doi:10.3835/plantgenome2011.08.0024>)

The International Wheat Genome Sequencing Consortium (IWGSC) (2018) Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science 36: Issue 6403, eaar7191. DOI: 10.1126/science.aar7191

Par-Ares J, REGIA consortium (2002) REGIA, an EU project on functional genomics of transcription factors of Arabidopsis thaliana. Comp Funct Genomics 3: 102-108.

Plessis A., Ravel C., Bordes J., Balfourier F. and Martre, P. (2013) Association study of wheat grain protein composition reveals that gliadin and glutenin composition are trans-regulated by different chromosome regions. J. Exp. Bot. 64, 3627–3644

Ravel C, Fiquet S, Boudet J, Dardevet M, Vincent J, Merlino M, Michard R, Martre P (2014) Conserved cis-regulatory modules in promoters of genes encoding wheat high-molecular-weight glutenin subunits. Front Plant Sci 5: 621. doi: 10.3389/fpls.2014.00621

Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics 38, 203e208