Novel genomic marker for the *Alm* locus in barley identified based on transcriptome analysis

N.A. Shmakov[1]*, A.Yu. Glagoleva[1], G.V. Vasiliev[1], D.A. Afonnikov[1], E.K. Khlestkina[2]

[1] *Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*
[2] *Vavilov Institute of Plant Genetic Resources, RAS, St. Petersburg, Russia*

DOI 10.18699/ICG-PlantGen2019-52

* e-mail: shmakov@bionet.nsc.ru

Abstract: Plastids are semi-autonomous organoids that give plant cells the ability to photosynthesize. They retain their own genome which works in tight coordination with the nuclear genome of the plant cells. Many aspects of such coordination are still unclear. A fitting model to study specifics of nucleus-plastid interactions are plants with partial albinism. The near-isogenic barley line i:BwAlm with partial albinism was studied using RNA-seq technology. De novo transcriptome reconstruction was performed, and transcriptomes of lines i:BwAlm and 'Bowman' were compared. A contig was identified that appears in i:BwAlm, but not in the isogenic line 'Bowman'.

Key words: plant albinism; transcriptome reconstruction; transcript annotation; transcript localization.

1. Introduction

Chlorophylls are green plant pigments that play a key part in photosynthesis. In plant cells, chlorophylls a and b are located in plastids, on the membrane of thylakoids, where they form complex structures with a large number of so-called Chlorophyll Binding Proteins and other photosynthesis-related proteins. The plastid genome ('plastome') is highly reduced and usually contains 100–120 genes (Börner et al., 2015), while the proteome of plastids may contain around 3000 proteins (Zoschke, Bock, 2018). Thus, the majority of the proteins presented in the plastids are encoded by nuclear genes (Khan et al., 2013). This requires a precise coordination of nuclear and plastid genomes for the proper functioning of the photosynthetic machinery (Liebers et al., 2017). It is known that the plastid-to-nucleus communication is mediated by signaling molecules, for example, Mg-protoporphyrin IX (Chan et al., 2016). However, the details of plastid-to-nucleus crosstalk is poorly understood.

A promising model for studying specific aspects of such crosstalk are plants with partial albinism. They provide materials for studying signaling components and pathways between plastid and nucleus (Arisha et al., 2015). The near-isogenic line (NIL) i:Bw*Alm* of barley (*Hordeum vulgare* L.) is a plant model of this kind. Plants of this line have chlorophyll-deficient lemma and pericarp and nodes. i:Bw*Alm* contains a recessive mutation in the *Alm* gene, which is located on chromosome 3HS (Costa et al., 2001). The *Alm* gene itself has not yet been identified, and its protein product, molecular function and mechanism of action are unknown.

Recently we performed an analysis of the transcriptome of the barley NILs i:Bw*Alm* (NGB20419) with partial albinism of spike and stem nodes and its parental NIL 'Bowman' (NGB22812) with normal phenotype. Using an approach based on read alignment to the reference genome, we identified several genes encoded in the nuclear genome and related to photosynthesis with differential expression between the lines i:Bw*Alm* and 'Bowman' (Shmakov et al., 2016). In this work, we extended our analysis to identify possible genes related to the plastid-to-nucleus communication using *de novo* assembled transcripts from a previous RNA-seq experiment.

2. Materials and methods

2.1 Plant material

The barley NILs i:Bw*Alm* (NGB20419) with partial albinism of spike and stem nodes and its parental cultivar 'Bowman' (NGB22812) were used in the RNA-seq analysis. The lines were provided by the Nordic Gene Bank (NGB, www.nordgen.org). These lines were previously genotyped by microsatellite markers. The only chromosome segment different between the NILs is a segment in chromosome 3HS that contains the *Alm* gene. To localize the contig of interest in the *H. vulgare* genome, a set of wheat-barley addition lines and the parental wheat 'Chinese Spring' and barley 'Betzes' cultivars were used.

2.2 Bioinformatic analysis: libraries preprocessing

Six short-read libraries were obtained by IonTorrent sequencing as described in Shmakov et al. (2016). The libraries were filtered using PrinSeq-lite v 0.20.4 (Schmieder, Edwards, 2011). Reads shorter than 50 nucleotides, longer than 270 nucleotides, and reads with mean quality below 20 were removed. Non-coding RNA contamination was identified using read alignment to non-coding RNA sequences of *H. vulgare* (Ensembl plants database, v. 42) by Bowtie2 v. 2.3.4 (Langmead, Salzberg, 2012): reads that successfully mapped to the ncRNA sequences were discarded. Clean libraries were mapped to the genome of *H. vulgare* (Ensembl plants database, v. 42) using Dart v. 1.3.2 (Lin, Hsu, 2018). These alignments were later used to perform genome-guided transcriptome assembly.

2.3 Bioinformatic analysis: transcriptome assembly

De novo assembly was performed for the libraries from two lines separately. Transcripts were assembled using three tools: rnaSpades v. 3.12.0 (Bushmanova et al., 2018) with default parameters, Trinity v. 2.2.0 (Grabherr et al., 2013) with default

Table 1
Metrics of the six barley libraries before and after preprocessing and mapping

Library	Raw reads, millions	Clean reads, %	Mapped, %
i:Bw*Alm*1	4.6	84.3	98.7
i:Bw*Alm*2	3.0	86.6	98.8
i:Bw*Alm*3	5.8	92.6	98.9
Bowman1	4.1	92.0	99.1
Bowman2	4.0	59.6	97.8
Bowman3	6.9	96.6	99.0

Table 2
Metrics for line-specific and unified transcriptomes

	Metrics of assembled transcriptomes		
Assembly	No. of raw contigs	No. of nr contigs	N50
i:Bw*Alm*	110,387	49,186	1026
Bowman	106,078	44,326	1050
Unified	93,512	58,049	940

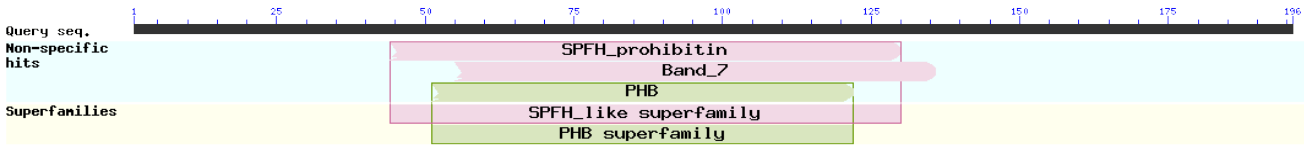

Figure 1. The domain structure of the putative protein encoded by contig DN5639c0g1t1 from line i:Bw*Alm*. The ruler shows amino acid numbering. Domains are shown schematically on the yellow bar.

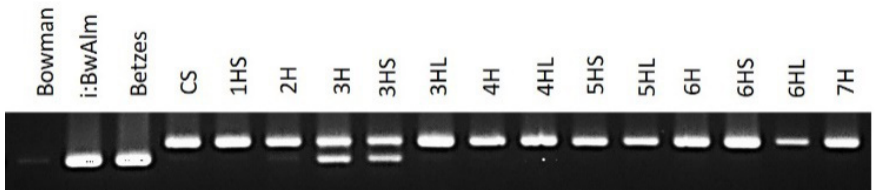

Figure 2. PCR profile of the DN contig in genomic DNA of barley NILs and wheat-barley addition lines.

parameters and trans-ABySS v. 2.0.1 (Robertson et al., 2010) with k-mer values set at 24, 48 and 64. Three assemblies obtained with trans-ABySS were then merged together using the transabyss-merge utility. Additionally, genome-guided assembly was performed using Trinity, and read alignments produced by Dart were used for assembly. The tr2aacds.pl tool of Evidential Gene pipeline v. '18may07' (Gilbert, 2013) was used to remove redundancy in the assemblies and to identify open reading frames and amino acid sequences encoded by the contigs. Contigs without ORFs or those encoding amino acid products less than 30 aa in length were excluded from further analysis.

Two barley line transcriptomes were obtained by merging transcripts from different assembling methods and removing redundancy. Finally, the unified transcriptome was built by merging line transcriptomes and removing redundant sequences. To evaluate the quality of the contigs, they were analyzed with BUSCO software (Simão et al., 2015) v. 3.0.2.

Kallisto software v. 0.45.0 (Bray et al., 2016) was used to quantify the expression values of contigs. The unified transcriptome was used as a reference. Contigs with expression values of less than 1 TPM (Transcripts Per Million reads, normalized measure of expression) were excluded from further analysis.

Two unified line-specific assemblies and a unified transcriptome assembly were analyzed using rnaQUAST v 1.5.1 software (Bushmanova et al., 2016) using *H. vulgare* genome v. 41 as a reference. This tool performs transcript to reference alignment by GMAP (Wu, Watanabe, 2005) and makes it possible to identify RNA sequences absent in the current version of the barley genome. Putative protein products of these contigs were then aligned to the NCBI nr protein database using the ublast tool from Usearch software (Edgar, 2010) v 8.1.1756_i86linux32. The e-value threshold for significant homology identification was set at 10^{-50}. Contigs with the best hit to the sequences of other than plant species origins were removed as contaminants. The remaining contigs were analyzed more closely.

2.4 Contig localization

To localize the contig of interest on barley chromosomes, the set of wheat-barley addition lines (Islam et al., 1981) and the

parental wheat cv. 'Chinese Spring' and barley cv. 'Betzes' were used. The primers (forward 5'GAGGACTTGGAT GAGAG3' and reverse 5'GCATTCCTGTTATCTTG3') were constructed using online service IDT PrimerQuest software (http://eu.idtdna.com/PrimerQuest/Home/).

3. Results and discussion

3.1 Library preprocessing

Filtering of the libraries removed ~15 % of all reads. Additionally, ~8 % of the remaining reads were removed as potential rRNA contamination. Of the remaining reads, ~98.5 % were successfully mapped to the *H. vulgare* genome. Table 1 contains metrics of library filtering and mapping.

After length and quality filtering and ncRNA contamination removal, a total of 24,913,867 reads remained in the six libraries, of which 22,636,345 were mapped to the reference genome and were later used for genome-guided transcriptome assembly.

3.2 Transcriptome reconstruction

Assembly was performed for two lines in separate, then a unified assembly of the *H. vulgare* lemma transcriptome was obtained. Table 2 shows several metrics of assembled transcriptomes for both lines separately and for the unified transcriptome. BUSCO analysis demonstrated that all transcriptomes contain 43–47 % of transcripts which have full-length alignment with BUSCO sequences, 25–27 % of them aligned partially and 26–30 % were not found among the BUSCO sequence set. Unified assembly has a greater percentage of full and fragmented BUSCO sequences than both line-specific assemblies.

54,875 contigs in the unified assembly have expression levels > 1 TPM. rnaQuast analysis identified 508 contigs from the i:Bw*Alm* assembly, 405 contigs from the 'Bowman' assembly and 788 contigs from the unified assembly which are absent in the barley reference genome. The search of homologs for the unaligned sequences from the unified assembly in the NCBI nr database yielded similar sequences for 15 contigs. One of them is a DN5639c0g1t1 contig, which originated from the i:Bw*Alm* transcriptome, and no significant homology to this transcript was found among the 'Bowman' transcripts. It has a length of 691 nucleotides, and an amino acid product 196 aa in length is predicted to be encoded by this contig; it has expression levels of ~6 TPM in the i:Bw*Alm* libraries. The best hit for the putative protein sequence encoded by DN5639c0g1t1 in the NCBI nr database is sequence BAK08282.1 from *H. vulgare* (e-value 4e-77). This is a predicted protein with an unknown function. The domain structure of the putative protein product of DN5639c0g1t1 is shown in Figure 1 as identified by the CDD/SPARCLE NCBI online service (Marchler-Bauer et al., 2017). It contains the SPFH_prohibitin (e-value = 4.2e-13), Band_7 (e-value = 9.8e-5) and PHB (e-value = 2e-3) domains in the middle part of the amino acid sequence. The translated amino acid sequence of contig DN5639c0g1t1 also has a homology to the *Solanum pennellii* prohibitin-1, mitochondria-like protein (NCBI protein accession number XP_015060913.1). The e-value of this homology is $3 \cdot 10^{-20}$.

Contig localization in the barley genome

The presence of length polymorphism between barley (amplicon length, 356 bp) and wheat (amplicon length, ~400 bp) in the amplified region allows us to localize DN5639c0g1t1 in the barley genome on chromosome 3HS (Figure 2) using wheat-barley addition lines. Since the only genome fragment that differs between the 'Bowman' and i:Bw*Alm* genomes is situated on the short arm of chromosome 3H, it can be speculated that this gene is situated inside this genome fragment and, thus, close to the *Alm* gene.

Resequencing of the contig from genomic DNA of line i:Bw*Alm* was performed. The sequenced fragment has a length of 311 nucleotides and contains an insert 102 nt in length, which presumably is an intron. Aside from this insert, its sequence is identical to the sequence of contig DN5639c0g1t1. This implies that the designed pair of primers is specific to the transcript of interest.

4. Conclusions

A contig was identified through RNA-seq analysis that is present in the NIL i:Bw*Alm* genome. At the same time, it either is absent in the genome of the isogenic line 'Bowman' or has a polymorphic region that forbids amplification of a fragment from the designed pair of primers. This contig is located on the short arm of barley chromosome 3H. The translated amino acid sequence of the contig has a weak homology to *S. pennellii* prohibitin-1 protein. Since this contig is present in the 'Betzes' genome, it is unlikely that it has any effect on the formation of the specific *Alm* phenotype. However, this gene can still be used to further narrow down the *Alm* locus in future experiments with segregating populations.

References

Arisha M.H., Shah S.N.M., Gong Z.-H., Jing H., Li C., Zhang H.-X. (2015). Ethyl methane sulfonate induced mutations in M2 generation and physiological variations in M1 generation of peppers (*Capsicum annuum* L.). *Frontiers Plant Sci*. https://doi.org/10.3389/fpls.2015.00399

Börner T., Aleynikova A.Y., Zubo Y.O., Kusnetsov V.V. (2015). Chloroplast RNA polymerases: Role in chloroplast biogenesis. *Biochimica Biophysica Acta*. https://doi.org/10.1016/j.bbabio.2015.02.004

Bray N.L., Pimentel H., Melsted P., Pachter L. (2016). Near-optimal probabilistic RNA-seq quantification. *Nature Biotechnol*. https://doi.org/10.1038/nbt.3519

Bushmanova E., Antipov D., Lapidus A., Przhibelskiy A.D. (2018). rnaSPAdes: a de novo transcriptome assembler and its application to RNA-Seq data. *BioRxiv*. https://doi.org/10.1101/420208

Bushmanova E., Antipov D., Lapidus A., Suvorov V., Prjibelski A.D. (2016). RnaQUAST: A quality assessment tool for de novo transcriptome assemblies. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btw218.

Chan K.X., Phua S.Y., Crisp P., McQuinn R., Pogson B.J. (2016). Learning the Languages of the Chloroplast: Retrograde Signaling and Beyond. *Ann Review Plant Biol*. https://doi.org/10.1146/annurev-arplant-043015-111854.

Costa J.M., Corey A., Hayes P.M., Jobet C., Kleinhofs A., Kopisch-Obusch A., … Wolfe R.I. (2001). Molecular mapping of the Oregon Wolfe Barleys: A phenotypically polymorphic doubled-haploid population. *Theoretical Applied Gen*. https://doi.org/10.1007/s001220100622.

Edgar R.C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btq461.

Gilbert D.G. (2013). Gene-omes built from mRNA-seq not genome DNA. *7th Annual Arthropod Genomics Symposium*. https://doi.org/10.7490/f1000research.1112594.1

Grabherr M.G. ., Haas B.J., Yassour M., Levin J.Z. ., Thompson D.A., Amit I., … Regev A. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature Biotechnol.* 2013;29(7):644–652. https://doi.org/10.1038/nbt.1883.Trinity

Islam A.K.M.R., Shepherd K.W., Sparrow D.H.B. (1981). Isolation and characterization of euplasmic wheat-barley chromosome addition lines. *Heredity*. https://doi.org/10.1038/hdy.1981.24.

Khan N.Z., Lindquist E., Aronsson H. New Putative Chloroplast Vesicle Transport Components and Cargo Proteins Revealed Using a Bioinformatics Approach: An Arabidopsis Model. *PLoS ONE*. 2013;8(4). https://doi.org/10.1371/journal.pone.0059898.

Langmead B., Salzberg S.L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*. https://doi.org/10.1038/nmeth.1923.

Liebers M., Grübler B., Chevalier F., Lerbs-Mache S., Merendino L., Blanvillain R., Pfannschmidt T. (2017). Regulatory Shifts in Plastid Transcription Play a Key Role in Morphological Conversions of Plastids during Plant Development. *Frontiers Plant Sci.* 1–8. https://doi.org/10.3389/fpls.2017.00023.

Lin H.N., Hsu W.L. DART: A fast and accurate RNA-seq mapper with a partitioning strategy. *Bioinformatics*. 2018;34(2):190–197. https://doi.org/10.1093/bioinformatics/btx558.

Marchler-Bauer A., Bo Y., Han L., He J., Lanczycki C.J., Lu S., … Bryant S.H. (2017). CDD/SPARCLE: Functional classification of proteins via subfamily domain architectures. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkw1129.

Robertson G., Schein J., Chiu R., Corbett R., Field M., Jackman S.D., … Birol I. (2010). De novo assembly and analysis of RNA-seq data. *Nature Methods*. https://doi.org/10.1038/nmeth.1517.

Schmieder R., Edwards R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*. 2011;27(6):863–864. https://doi.org/10.1093/bioinformatics/btr026.

Shmakov N.A., Vasiliev G.V., Shatskaya N.V., Doroshkov A.V., Gordeeva E.I., Afonnikov D.A., Khlestkina E.K. Identification of nuclear genes controlling chlorophyll synthesis in barley by RNA-seq. *BMC Plant Biol.* 2016;16(Suppl 3). https://doi.org/10.1186/s12870-016-0926-x.

Simão F.A., Waterhouse R.M., Ioannidis P., Kriventseva E.V., Zdobnov E.M. (2015). BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btv351.

Wu T.D., Watanabe C.K. (2005). GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/bti310.

Zoschke R., Bock R. (2018). Chloroplast translation: structural and functional organization, operational control, and regulation. *Plant Cell*. https://doi.org/10.1105/tpc.18.00016.

Acknowledgements. This work was supported by RSF project No. 18-14-00293.

Conflict of interest. The authors declare no conflict of interest.