In Silico SNP Detection for Anthocyanin Metabolism Genes in *Vitis*

G. Dequigiovanni, P.S. Ritschel, J.D.G. Maia and V. Quecini
Embrapa Uva e Vinho
Bento Gonçalves RS
Brazil

Keywords: berry color, flavonoid, genetic marker, genetic variation, grapevine

Abstract

Grape flavonoids, especially anthocyanins, are important contributors to color, taste, antioxidant and nutraceutical properties – such as protection against cardiovascular diseases and cancer – for fresh fruit and processed products. Breeding programs and wild *Vitis* germplasm banks include a wide genetic variation in anthocyanin biosynthesis, and metabolism-associated traits, indicating a complex genetic control of the process throughout plant development. Expressed sequence tag (EST) and genomic databases were investigated using bioinformatic tools to identify SNP markers present in structural and regulatory genes associated with anthocyanin metabolism in *Vitis*. We have identified more than 1,000 putative SNPs in nine structural and twelve regulatory genes of the anthocyanin metabolism in *Vitis*. The data suggest that the identified genetic variation is sufficient to distinguish among *V. vinifera* cultivars, hybrids and wild species. Regulatory genes were shown to be more rapidly evolving than structural enzyme-coding sequences. The evolutionary rate of the sequences associated with regulatory factors also appear to be differential, with higher levels of conservation found in the MYB family of transcriptional regulators associated with anthocyanin regulation. After experimental validation, the predicted polymorphisms will provide tools for further mapping, genome structure and functional studies.

INTRODUCTION

The dissection of complex traits, controlled by several distinct genomic regions, into their individual components, which can be molecularly identified, requires extensive genome characterization (Tanksley and Fulton, 2007). Anthocyanin production in higher plants is coordinately controlled by developmental morphogenic and environmental factors (Koes et al., 2005). Flavonoid accumulation is associated to several important processes in grapevine and other species, such as berry color, flavor and ecological features such as pathogen resistance and photodamage protection (Winkel-Shirley, 2001). *Vitis* germplasm and breeding program banks exhibit a wide range of genetic variation concerning anthocyanin accumulation and metabolism. In order to investigate the genetic determinants of the available phenotypic diversity, the present study aimed to identify SNP in the structural and regulatory genes of the anthocyanin metabolism to develop tools for whole-genome scale studies of the flavonoid metabolism in *Vitis*.

MATERIAL AND METHODS

Sequences from publicly available *Vitis* databases were queried by tBLASTx and tBLASTn searches (Altschul et al., 1997) using as bait functionally characterized protein sequences from model species corresponding to structural and regulatory genes of anthocyanin metabolism. The hits retrieved from *Vitis* databases were filtered by sequence quality (Phred score ≥ 20), reverse BLAST and e-value ($\leq 1.e^{-25}$) and sequences failing to reach one of the criteria were excluded from further analyses. The sequences were aligned with ClustalX (Thompson et al., 1997), bootstrapped 1,000 times and majority rule consensus trees were determined from maximum likelihood and neighbor joining trees using PHYLIP v.3.6 software (Felsenstein, 2005). Putative single nucleotide polymorphisms were identified comparing the alignments against the inbred 'Pinot Noir' genome (Jaillon et al., 2007) using the cns2snp MAQ command and the SNPfilter Perl

script in the software MAQ version 0.7.1 (Li et al., 2008), accepting called SNPs with at least 100 reads; base quality higher than 20; copy number of the flanking region in the reference genome lower than 1.0; quality of the 3 bp flanking region around putative SNP higher than 10; quality of sequence read alignment across SNP higher than 60. The results are represented as SNPs per 1000 bp for *Vitis vinifera*, hybrids between *V. vinifera* × *V. labrusca* and other species (including wild and cultivated species). Divergence threshold (DT) and hidden Markov model islands (HMMI) methods were used to identify conserved regions in structural and regulatory genes employing a maximum percent variation of 20% and a minimal length of 80 amino acids and the probabilities of e*S* =0.75, e*F*=0.60, and *T*=0.1, respectively.

RESULTS AND DISCUSSION

The sequences corresponding to nine structural genes of the anthocyanin biosynthesis and twelve transcriptional regulators associated to the metabolite accumulation were compiled from *Vitis* genome and expressed sequence tag (EST) databases. The sequences were retrieved by querying public-access databases using functionally characterized sequences from model species as bait. The sequences were aligned and investigated for the presence of single nucleotide polymorphisms against grapevine reference genome employing the SNP calling algorithm of the MAQ version 0.7.1 software. A total of 1,419 SNPs (622 in structural genes and 797 in the coding sequences of transcriptional regulators) were identified (Tables 1 and 2), including sequences carrying one or more mismatches (Fig. 1). The identification of SNPs based on EST sequence data generally displays a rate of false-positives ranging from 15 to 50%, although the rate is much lower when a reference genome sequence is available (Ganal et al., 2009). The stringent criteria employed for SNP calling and the availability of a *Vitis* reference genome suggest that the false-positive rates in the current work are likely to be closer to the lower-end of the range (15%), which would correspond roughly to 1,200 useful SNP markers. A recent study has demonstrate that a 9K SNP array provides sufficient resolution to distinguish among *V. vinifera* cultivars, between *V. vinifera* and wild *Vitis* species, and even among diverse wild *Vitis* species (Myles et al., 2010). Thus, our in silico approach may provide informative tools for the investigation of anthocyanin metabolism in *Vitis*. In *Eucalyptus*, a recent work employing next-generation sequencing has identified similar SNP frequencies in the structural genes of flavonoid biosynthesis (Külheim et al., 2009). Structural genes generally exhibit higher levels of sequence conservation in comparison to regulatory factors (Tables 1 and 2). The coding sequences for the structural genes *CHS* and *FLS* exhibited higher SNP density (Fig. 2). In contrast, lower levels of sequence divergence were observed for ANS and UFGT (Fig. 2). For the regulatory factors, higher frequencies of SNPs were found in regulatory genes from the bHLH and MADS family whereas higher degrees of sequence conservation were found in genes of the MYB family. These results were confirmed by further evolutionary phylogenetic analysis, which detected slow-evolving domains outside the functional DNA-binding MYB domain in *Vitis* orthologs of *Arabidopsis* production of anthocyanin Pigment1 (Borevitz et al., 2000) by HMMI and DT methods (Fig. 4). Although the sequence conservation in the functional domains MADS and K box were detected by evolutionary phylogenetic analyses in *Vitis* species, the evolution rate was indistinct for all investigated sequence regions, demonstrated by the absence of HMMI and DT even when less stringent parameters were employed.

CONCLUSIONS

The current work has performed an extensive in silico survey of the single nucleotide polymorphisms in genes involved in anthocyanin metabolism in *Vitis*, exploiting publicly available data and using bioinformatic tools. Our analyses have demonstrated the existence of considerable genetic variation in sequences associated to anthocyanin metabolism, within *V. vinifera* cultivars and among hybrids and wild species. The frequencies of single nucleotide polymorphisms were lower in structural genes, in

comparison to that identified in sequences corresponding to regulatory factors. Among the structural genes, *CHS* and *FLS* exhibited higher levels of polymorphism, whereas *ANS* and *UFGT* displayed smaller frequencies of SNPs. For the regulatory factors, higher frequencies of SNPs were found in regulatory genes from the bHLH and MADS family whereas higher degrees of sequence conservation were found in genes of the MYB family. The differential distribution of polymorphic sites over evolutionary distance, using divergence threshold and hidden Markov model islands methods, suggests that the DNA-binding domain of bHLH and MADS proteins has a faster evolutionary rate in comparison to MYB regulators. The characterization of a large number of common SNPs for the anthocyanin metabolism in *Vitis* will allow further genome-wide association studies to dissect the complex genetic control of anthocyanin metabolism.

ACKNOWLEDGEMENTS

The authors thank the Daniela Dal Bosco and Iraci Sinski for excellent technical assistance. The work was financed by a Macroprograma 2 grant (02.05.02.09.00.05) and AgroVerde - Grapevine Germplasm Bank from Embrapa to P.S.R.

Literature Cited

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acid Res. 25:3389-3402.

Borevitz, J.O., Xia, Y., Blount, J., Dixon, R.A. and Lamb, C. 2000. Activation tagging identifies a conserved MYB regulator of phenylpropanoid biosynthesis. Plant Cell 12: 2383-2394.

Felsenstein, J. 2005. PHYLIP (Phylogeny Inference Package), Version 3.6. Available at http://evolution.genetics.washington.edu/phylip.html.

Ganal, M.W., Altmann, T. and Röder, M.S. 2009. SNP identification in crop plants. Curr. Opin. Plant Biol. 12:211-217.

Jaillon, O., Aury, J.M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., Choisne, N., Aubourg, S., Vitulo, N., Jubin, C., Vezzi, A., Legeai, F., Hugueney, P., Dasilva, C., Horner, D., Mica, E., Jublot, D., Poulain, J., Bruyère, C., Billault, A., Segurens, B., Gouyvenoux, M., Ugarte, E., Cattonaro, F., Anthouard, V., Vico, V., Del Fabbro, C., Alaux, M., Di Gaspero, G., Dumas, V., Felice, N., Paillard, S., Juman, I., Moroldo, M., Scalabrin, S., Canaguier, A., Le Clainche, I., Malacrida, G., Durand, E., Pesole, G., Laucou, V., Chatelet, P., Merdinoglu, D., Delledonne, M., Pezzotti, M., Lecharny, A., Scarpelli, C., Artiguenave, F., Pè, M.E., Valle, G., Morgante, M., Caboche, M., Adam-Blondon, A.F., Weissenbach, J., Quétier, F. and Wincker, P. 2007. French-Italian Public Consortium for Grapevine Genome Characterization. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449:463-467.

Koes, R., Verweij, W. and Quattrocchio, F. 2005. Flavonoids: a colorful model for the regulation and evolution of biochemical pathways. Trends Plant Sci. 10:236-242.

Külheim, C., Yeoh, S.H., Maintz, J., Foley, W.J. and Moran, G.F. 2009. Comparative SNP diversity among four *Eucalyptus* species for genes from secondary metabolite biosynthetic pathways. BMC Genomics 10:452.

Li, H., Ruan, J. and Durbin, R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 18:1851-1858.

Myles, S., Chia, J.-M., Hurwitz, B., Simon, C., Zhong, G.Y., Buckler, E. and Ware, D. 2010. Rapid genomic characterization of the genus *Vitis*. PLoS ONE 5(1): e8219. doi: 10.1371/journal.pone.0008219.

Ovcharenko, I., Boffelli, D. and Loots, G.G. 2004. eShadow: a tool for comparing closely related sequences. Genome Res. 14:1191-1198.

Tanksley, S.D. and Fulton, T.M. 2007. Dissecting quantitative trait variation – examples from the tomato. Euphytica 154:365-370.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F. and Higgins, D.G. 1997. The CLUSTALX windows interface: Flexible strategies for multiple sequence alignment

aided by quality analysis tools. Nucleic Acids Res. 25:4876-4882.
Winkel-Shirley, B. 2001. Flavonoid biosynthesis. A colorful model for genetics, biochemistry, cell biology, and biotechnology. Plant Physiol. 126:485-493.

Tables

Table 1. Summary of identified SNPs for nine structural genes of the anthocyanin biosynthesis in *Vitis*.

Gene	aa identity (%)[a]	Length (bp)			Putative SNPs	Coverage[b]
		Intron	Exon	Total		
CHS	85.2	1	2	1179	147	*Vvi*, hyb, w
CHI	83.9	3	4	741	31	*Vvi*, hyb, w
DFR	76.8	5	6	1128	67	*Vvi*, hyb, w
F3H	81.8	2	3	1077	96	*Vvi*, hyb, w
F3´H	-	2	3	1530	53	*Vvi*, hyb, w
F3´5´H	-	2	2	1527	46	*Vvi*, hyb, w
FLS	65.5	2	3	1011	133	*Vvi*, hyb, w
ANS/LDOX	84.7	2	3	462	17	*Vvi*, hyb, w
UFGT	-	1	2	1371	32	*Vvi*, hyb, w
Total					622	

[a] Deduced amino acid identity in comparison to *Arabidopsis thaliana* model gene.
[b] Coverage within the genus: *Vvi*; *Vitis vinifera*, hyb: *V. vinifera* × *V. labrusca* hybrids, w; other species.

Table 2. Summary of the identified SNPs for six families of regulatory genes of the anthocyanin biosynthesis in *Vitis*.

Gene family	aa identity (%)[a]	Length (bp)			Putative SNPs	Coverage[b]
		Intron	Exon	Total		
WD40						
TTG1	77.5	1	1	1008	193	*Vvi*, hyb, w
HLH						
GL3/EGL3	61.8	8	7	1962	171	*Vvi*, hyb, w
TT8	62.3	8	7	834	87	*Vvi*, hyb, w
MADS						
TT16	59.2	5	6	768	38	*Vvi*, hyb, w
MYB						
TT2	60.8	2	3	969	17	*Vvi*, hyb, w
PAP1	58.7	2	3	720	32	*Vvi*, hyb, w
GL1/WER	73.5	2	3	648	18	*Vvi*, hyb, w
MYB61	55.6	2	3	1317	23	*Vvi*, hyb, w
CPC/TRY	89.2	2	3	264	15	*Vvi*, hyb, w
WRYK						
TTG2	71.0	4	4	1314	79	*Vvi*, hyb, w
Zn C2H2						
TT1	83.5	2	3	969	124	*Vvi*, hyb, w
Total					797	

[a] Deduced amino acid identity in comparison to *Arabidopsis thaliana* model gene.
[b] Coverage within the genus: *Vvi*: *Vitis vinifera*, hyb: *V. vinifera* × *V. labrusca* hybrids, w: other species.

Figures

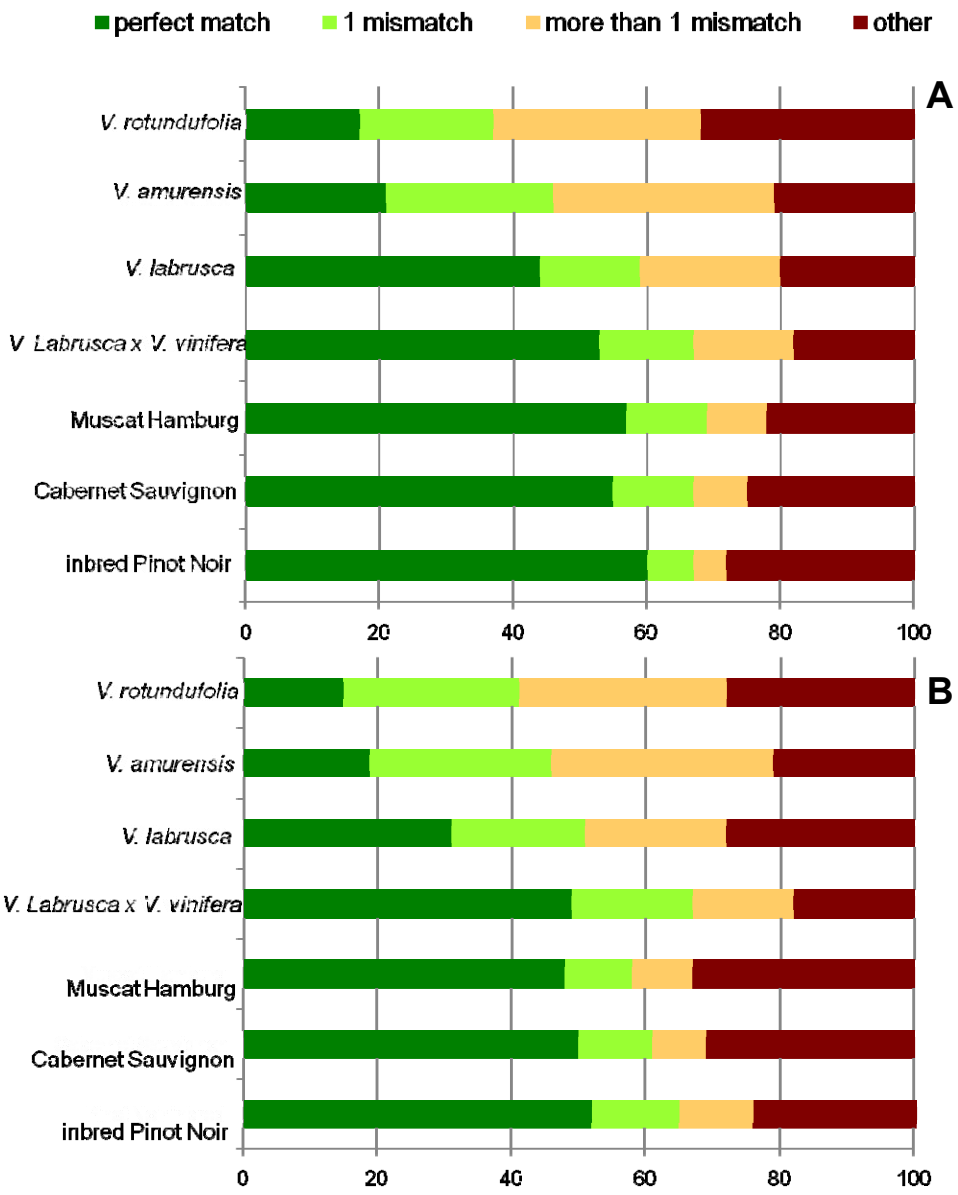

Fig. 1. Alignment results of the database reads to the grapevine reference genome for anthocyanin metabolism structural (A) and (B) regulatory genes The upper bars in the plot indicate the proportion of reads belonging to each of the categories in the legend. Other represents reads mapping to repetitive regions or with no match (discarded).

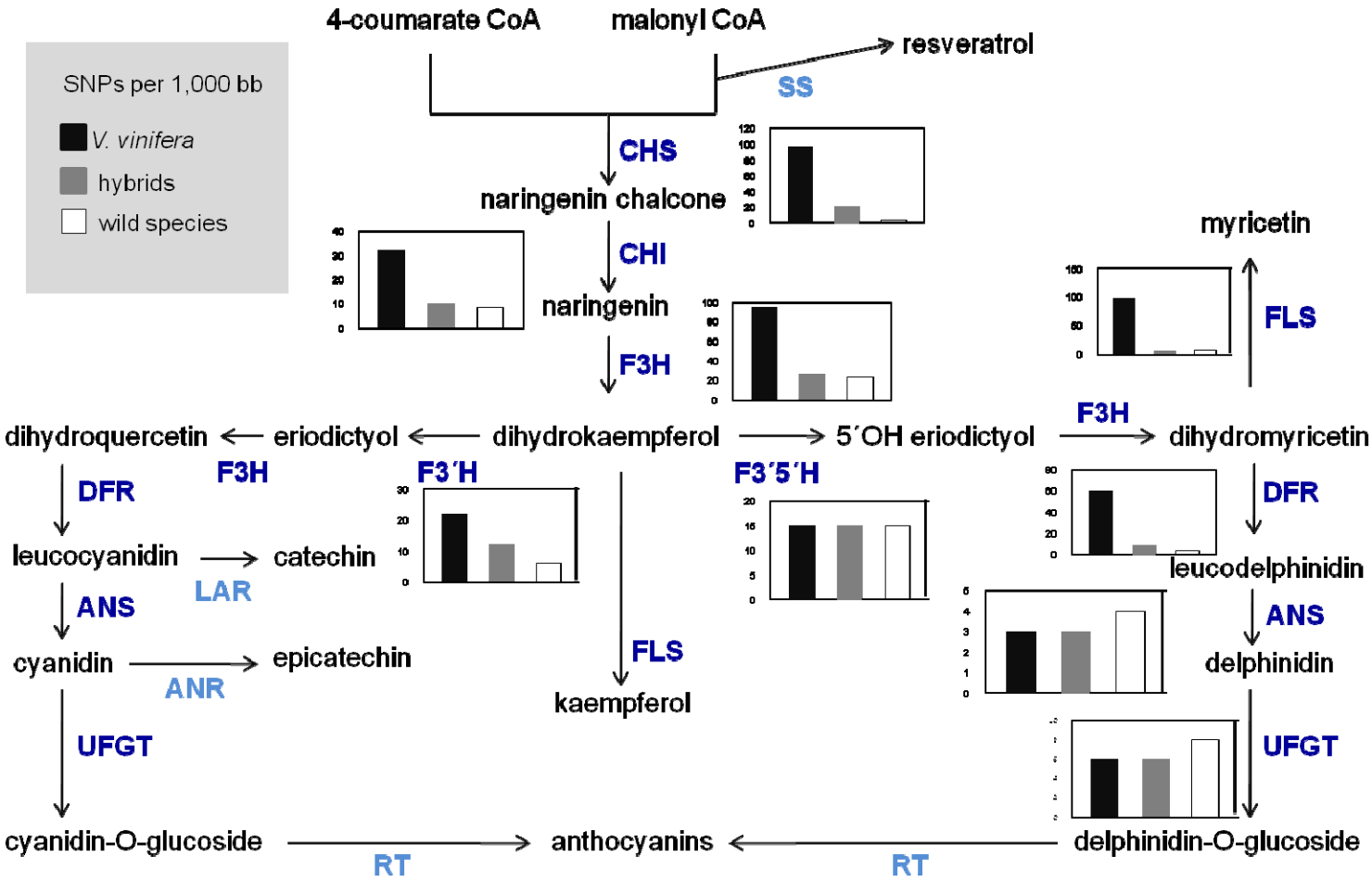

Fig. 2. SNP frequency in the coding sequences of genes involved in anthocyanin biosynthesis in *Vitis*. Scheme represents the consensus pathway for *Vitis*. Enzymes are represented in blue lettering: dark blue; the investigated enzymes with SNP frequency graphs; light blue: non-investigated enzyme sequences. The data represents normalized SNP frequency per 1000 bp for *Vitis vinifera*, *Vitis* hybrids (*V. vinifera* × *V. labrusca*) and wild species (*V. amurensis* and *V. rotundifolia*).

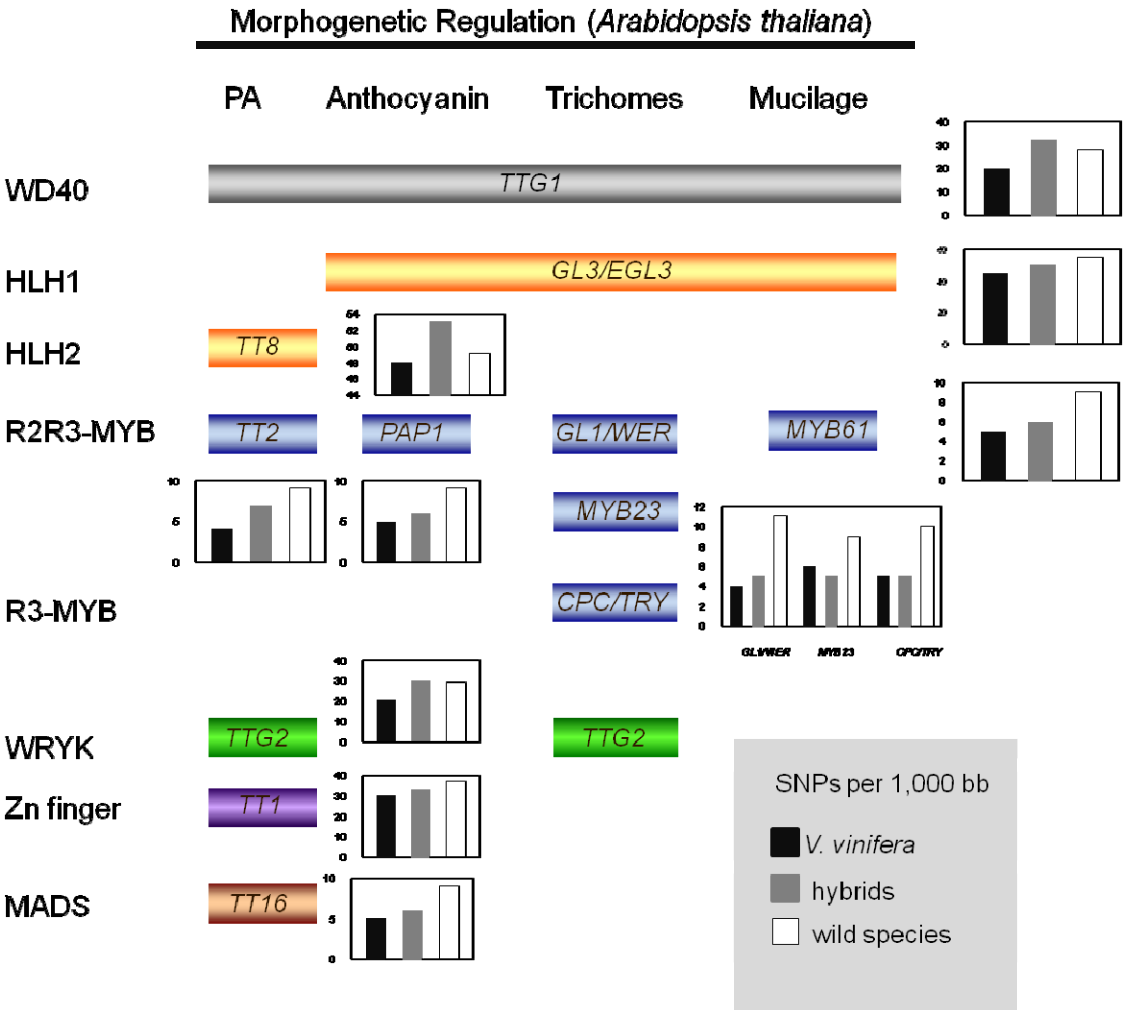

Fig. 3. SNP frequency in the coding sequences of the putative regulatory genes of anthocyanin metabolism in *Vitis*, according to functional evidence from model species *Arabidopsis thaliana*. The scheme represents gene families with SNP frequency graphs. The data represents normalized SNP frequency per 1,000 bp for *Vitis vinifera*, *Vitis* hybrids (*V. vinifera* × *V. labrusca*) and wild species (*V. amurensis* and *V. rotundifolia*).

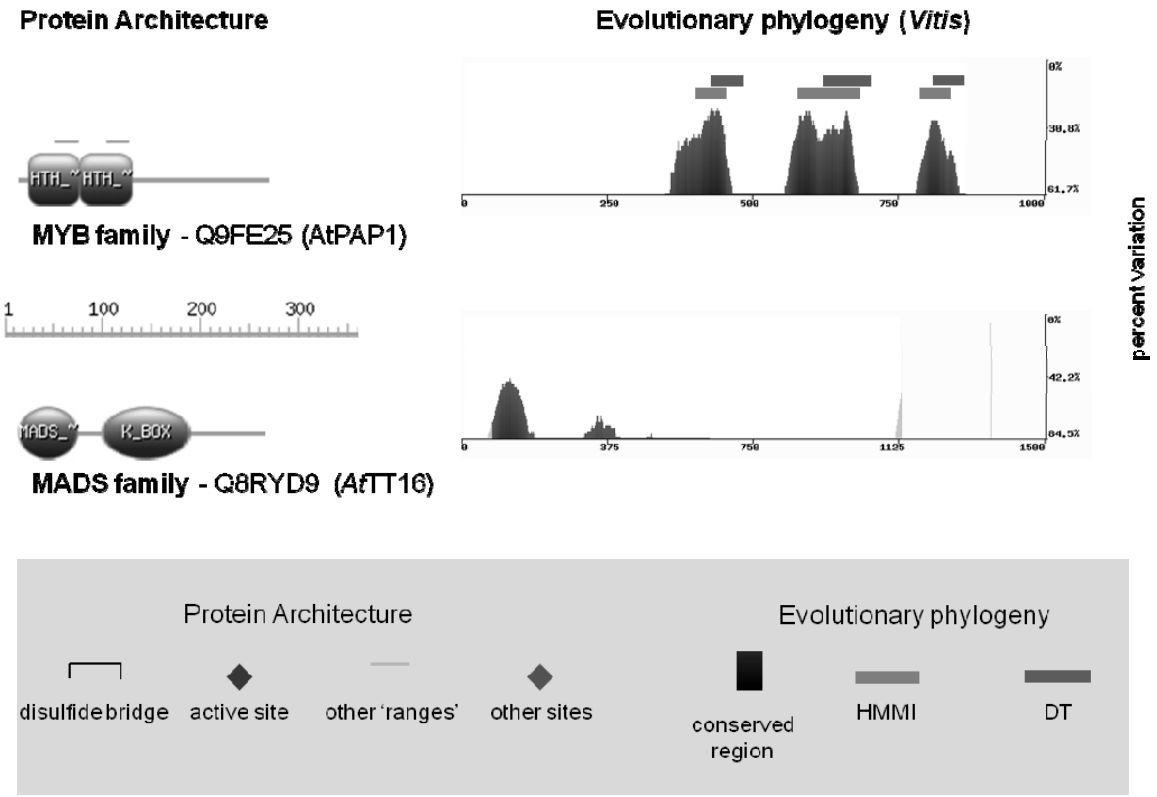

Fig. 4. Evolutionary phylogenetic analysis of anthocyanin regulatory sequences in *Vitis*. Functional domains are schematically represented as protein architecture (according to ExPaSy Prosite analysis tool). The percent variation is shown by nucleotide position for seven *Vitis* PAP1 and TT16 sequences. Regions determined to be slow-evolving are indicated by light gray (HMMI) or dark grey bars (DT).