Data-mining the B73 genome sequence for carotenoid biosynthesis gene candidates.

--Stinard, PS

Many of the genes associated with classical carotenoid-deficient endosperm mutants of maize have been cloned and characterized (e.g. *y1* (phytoene synthase; Buckner *et al.* 1990. Plant Cell 2:867-876); *vp5* (phytoene desaturase; Li *et al.* 1996. Plant Molecular Biology 30:269-279); *y9* (zeta-carotene isomerase; Li *et al.* 2007. Plant Physiology 144:1181-1189); *vp9* (zeta-carotene desaturase; Matthews *et al.* 2003. J Exp Bot 54:2215-2230); *ps1* (lycopene beta-cyclase; Singh *et al.* 2003. Plant Cell 15:874-884); and *vp2* (4-hydroxyphenylpyruvate dioxygenase; Matthews *et al.* 2003. J Exp Bot 54:2215-2230). However, to date, many carotenoid-deficient loci have eluded association with steps in the carotenoid biosynthetic pathway. The list of uncharacterized genes includes *lw1*, *lw2*, *lw3*, *lw4*, *w3*, *y8*, *y10*, and *cl1*. We report here the association of these loci (with reasonable confidence) to specific gene products. Our technique was to identify characterized *Arabidopsis* orthologs of carotenoid biosynthetic genes and perform BLAST searches against the maize B73 genome (version 2) using the MaizeGDB genome browser tools. The results are summarized in Figures 1 and 2, and Tables 1 and 2.

With the exception of *vp2*, the characterized genes involve steps in the direct pathway leading from geranylgeranyl diphosphate to beta-carotene. *vp2*, however, is implicated in the biosynthetic pathway for plastoquinone (Figure 1), an electron receptor involved in the desaturation steps between phytoene and lycopene. We first examined steps in the plastoquinone biosynthetic pathway in *Arabidopsis*. The PDS1 gene in *Arabidopsis* encodes 4-hydroxyphenylpyruvate dioxygenase, involved in the conversion of 4-hydroxyphenylpyruvate to homogentisic acid (Norris *et al.* 1995. Plant Cell 7:2139-2149). The Genbank sequence for PDS1 (NCBI Reference Sequence: NM_100536.3) was used to BLAST against the maize genome and picked up homology to gene model GRMZM2G088396 (Chr5:83859479..83861633), which is located on 5S near the estimated location of *vp2* (Chr5:78386141..80842741), and which encodes a putative 4-hydroxyphenylpyruvate dioxygenase. This is consistent with the data of Matthews *et al.* (2003).

The PDS2 gene in *Arabidopsis* encodes homogentisate solanesyltransferase, involved the conversion of homogentisic acid to 2-demethyl-plastoquinol-9 (Tian *et al.* 2007. Planta 226:1067-1073). The Genbank sequence for PDS2 (NCBI Reference Sequence: NM_001161137.1) was used to BLAST against the maize genome and picked up homology to gene model GRMZM2G113476 (Chr2:206847694..206863769), which is located on 2L near the estimated location of *w3* (Chr2:204481904..205710630), and which encodes a putative prenyltransferase/ zinc ion binding protein with high sequence homology to the *Arabidopsis* homogentisate solanesyltransferase gene. Thus the maize *w3* locus is an excellent candidate for the gene encoding maize homogentisate solanesyltransferase. A UniformMu line (UFMu-02780) carrying an insert (*mu1031674*) in this gene model segregates for a white endosperm viviparous mutant allele of *w3*.

Although this result is suggestive, confirmation that the *w3* locus encodes homogentisate solanesyltransferase will require molecular analysis.

The remaining uncharacterized genes were placed in the biosynthetic pathway leading from 1-deoxy-D-xylulose-5-P (DOXP) to isopentenyl-diphosphate (IPP), part of the plastidial DOXP/MEP pathway (Figure 2; reviewed in Lichtenthaler 2004. Proceedings of the 16th International Plant Lipid Symposium, Budapest, Hungary, pp. 11-24). Whereas most of the reduced carotenoid mutations in genes involved in the later, purely plastidial parts of the carotenoid biosynthetic pathway exhibit vivipary due to reduced synthesis of ABA, mutants in genes of the MEP pathway might be expected to exhibit a less severe phenotype due to shuttling of intermediates from the alternative cytosolic MVA pathway (Rodríguez-Concepción 2006. Phytochemistry Reviews 5:1-15). Thus, mutants in MEP pathway genes might be expected to produce low levels of endosperm carotenoids and exhibit dormancy, i.e. a "lemon white" phenotype. Such mutants include *lw1*, *lw2*, *lw3*, *lw4*, *cl1*, and *y10*.

The DXS gene in *Arabidopsis* encodes DOXP synthase, involved in the conversion of pyruvate and glyceraldehyde-3-P to 1-deoxy-D-xylulose-5-P (DOXP). Vallabhaneni and Wurtzel (2009. Plant Physiology 150:562-572) and Cordoba *et al.* (2011. J Exp Bot 62:2023-2038) report three DXS genes in maize, *dxs1*, *dxs2*, and *dxs3*. These correspond to maize gene models GRMZM2G137151 (Chr6:146378393..146382661), GRMZM2G493395 (Chr7:14077852..14081075), and GRMZM2G173641 (Chr9:20462059..20467072) respectively. Cordoba *et al.* indicate that of these three DXS genes, *dxs1* is expressed the most in leaves, and *dsx2* and *dsx3* are expressed the most in yellow endosperms, with *dsx2* expressed more highly than *dsx3*. The *y8* gene is estimated to be at Chr7:14027268..14618739, which overlaps the *dsx2* location and is therefore a candidate gene for *dxs2*. Although *y8* mutants are homozygous viable and therefore not traditional "lemon whites," the expression pattern of the three DXS genes may explain how a knockout in *dsx2* could result in the *y8* mutant phenotype. It is possible that a knockout of *dsx2* might not be fully compensated for by *dxs3* expression in the endosperm, leading to the pale yellow *y8* mutant phenotype. A fully functional *dsx1* gene would allow normal carotenoid production in the rest of the plant (i.e. a fully viable green plant). On the other hand, if only the *dxs1* gene were knocked out, one would expect a yellow endosperm albino seedling mutant. *w14* (estimated to be at Chr6:148253633..148506034) is a possible classical maize gene candidate for the *dxs1* locus.

The DXR gene in *Arabidopsis* encodes 1-deoxy-D-xylulose 5-phosphate reductoisomerase, involved in the conversion of 1-deoxy-D-xylulose-5-P (DOXP) to 2-C-methyl-D-erythritol-4-P (MEP). The Genbank sequence for DXR (NCBI Reference Sequence: NM_125674.2) was used to BLAST against the maize genome and picked up homology to gene models GRMZM2G056975 (Chr3:30226804..30233358) and GRMZM2G036290 (Chr8:8094442.. 8101055), both of which encode maize DXR protein and show high homology with each other and the *Arabidopsis* gene. These are excellent candidates for the duplicate genes *cl1* (Chr8:33707329.. 33742708) and *Clm1* (chromosome 8S, location unknown). Note that mutants at *cl1* lead to a reduction in both

endosperm and plant carotenoids. Variants at the *Clm1* locus are able to compensate for the reduction in plant carotenoids in *cl1* mutants, but not for the reduction in endosperm carotenoids. This could be due to tissue-specific differences in expression of the two DXR genes.

The CDPMEK gene in *Arabidopsis* encodes 4-diphosphocytidyl-2-C-methyl-D-erythritol kinase, involved in the conversion of 4-diphosphocytidyl-2-C-methylerythritol to 4-diphosphocytidyl-2-C-methyl-D-erythritol 2-phosphate. The Genbank sequence for CDPMEK (NCBI Reference Sequence: NM_128250.3) was used to BLAST against the maize genome and picked up homology to gene model GRMZM5G859195 (Chr3:187922271..187927591), which is located on 3L and which encodes 4-diphosphocytidyl-2-C-methyl-D-erythritol kinase with high sequence homology to the *Arabidopsis* gene. The maize *y10* locus is estimated to be at Chr3:205199570..205264647, which seems a little far from the location of GRMZM5G859195. However, the genetic map of chromosome 3 places *y10* close to *na1* (Chr3:184214701..185318488). Thus, the maize *y10* locus is an excellent candidate for the gene encoding maize 4-diphosphocytidyl-2-C-methyl-D-erythritol kinase.

The ISPF gene in *Arabidopsis* encodes 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase, involved in the conversion of 4-diphosphocytidyl-2-C-methyl-D-erythritol 2-phosphate to 2-C-methyl-D-erythritol-2,4-cyclodiphosphate. The Genbank sequence for ISPF (NCBI Reference Sequence: NM_180640.2) was used to BLAST against the maize genome and picked up homology to gene models AC209374.4_FG002 (Chr5:196279295..196281037) and GRMZM5G835542 (Chr4:155830779..155832786), both of which encode maize 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase and show high homology with each other and the *Arabidopsis* gene. These are excellent candidates for the duplicate factor loci *lw3* (Chr5:188462959..190607852) and *lw4* (Chr4:155828832..155834753).

The HDS gene in *Arabidopsis* encodes 4-hydroxy-3-methylbut-2-enyl diphosphate synthase, involved the conversion of 2-C-methyl-D-erythritol-2,4-cyclodiphosphate to 4-hydroxy-3-methylbut-2-enyl diphosphate. The Genbank sequence for HDS (NCBI Reference Sequence: NM_125453.6) was used to BLAST against the maize genome and picked up homology to gene model GRMZM2G137409 (Chr5:182124005..182130631), which is located on 5L near the estimated location of *lw2* (Chr5:174149224..175478743), and which encodes 4-hydroxy-3-methylbut-2-enyl diphosphate synthase with high sequence homology to the *Arabidopsis* gene. Thus, the maize *lw2* locus is an excellent candidate for the gene encoding maize 4-hydroxy-3-methylbut-2-enyl diphosphate synthase.

Finally, the HDR gene in *Arabidopsis* encodes 4-hydroxy-3-methylbut-2-enyl diphosphate reductase, involved the conversion of 4-hydroxy-3-methylbut-2-enyl diphosphate to isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP). The Genbank sequence for HDR (NCBI Reference Sequence: NM_119600.3) was used to BLAST against the maize genome and picked up homology to gene model GRMZM2G027059 (Chr1:272936836 to 272940502), which is located on 1L near the

estimated location of *lw1* (Chr1:271108631..273434076), and which encodes 4-hydroxy-3-methylbut-2-enyl diphosphate reductase with high sequence homology to the *Arabidopsis* gene. Thus the maize *lw1* locus is an excellent candidate for the gene encoding 4-hydroxy-3-methylbut-2-enyl diphosphate reductase.

Thus, gene candidates can be assigned to nearly all of the loci associated with reduced endosperm carotenoids. Mutants, many of which are derived from populations carrying active transposable elements, exist for all of these loci, so it should be a simple matter to determine whether these mutants are due to lesions at the candidate loci. However, there are still genes in the carotenoid biosynthetic pathway for which mutants have not yet been identified. One possible explanation is that some of these genes occur as duplicate loci in maize for which two or more genes would need to be knocked out in order to observe a mutant phenotype. One such example is the genes homologous to the *Arabidopsis* gene ISPD (Figure 2; NCBI Reference Sequence: NM_126305.2), encoding 4-diphosphocytidyl-2-C-methyl-D-erythritol synthase. The *Arabidopsis* gene picks up homology with maize gene models GRMZM5G856881 (Chr3:170115790..170118780) and GRMZM2G172032 (Chr8:164748939..164752371). These genes encode a putative 4-diphosphocytidyl-2-C-methyl-D-erythritol synthase with homology to each other and to the *Arabidopsis* gene. We predict that if both genes were knocked out, a reduced endosperm carotenoid mutant phenotype would result. This and other examples of predicted duplicate genes are summarized in Table 2. Reverse genetics tools such as the UniformMu project may someday identify knockouts in these individual genes that may then be combined to test this hypothesis.

Figure 1. Plastoquinone biosynthetic pathway. Classical maize gene candidates are listed at the left of each step. *?* = uncharacterized duplicate factor loci. *Arabidopsis* genes are in parentheses.

4-hydroxyphenylpyruvate

vp2 ⬇ 4-hydroxyphenylpyruvate dioxygenase (PDS1)

homogentisic acid

w3 ⬇ homogentisate solanesyltransferase (PDS2)

2-methyl-6-solanyl-1,4-benzoquinol (2-demethyl-plastoquinol-9)

? ? ⬇ 2-methyl-6-solanyl-1,4-benzoquinone methyltransferase (VTE3)

plastoquinol-9

Figure 2. DOXP/MEP pathway. Classical maize gene candidates are listed at the left of each step. *?* = uncharacterized duplicate factor loci. *Arabidopsis* genes are in parentheses.

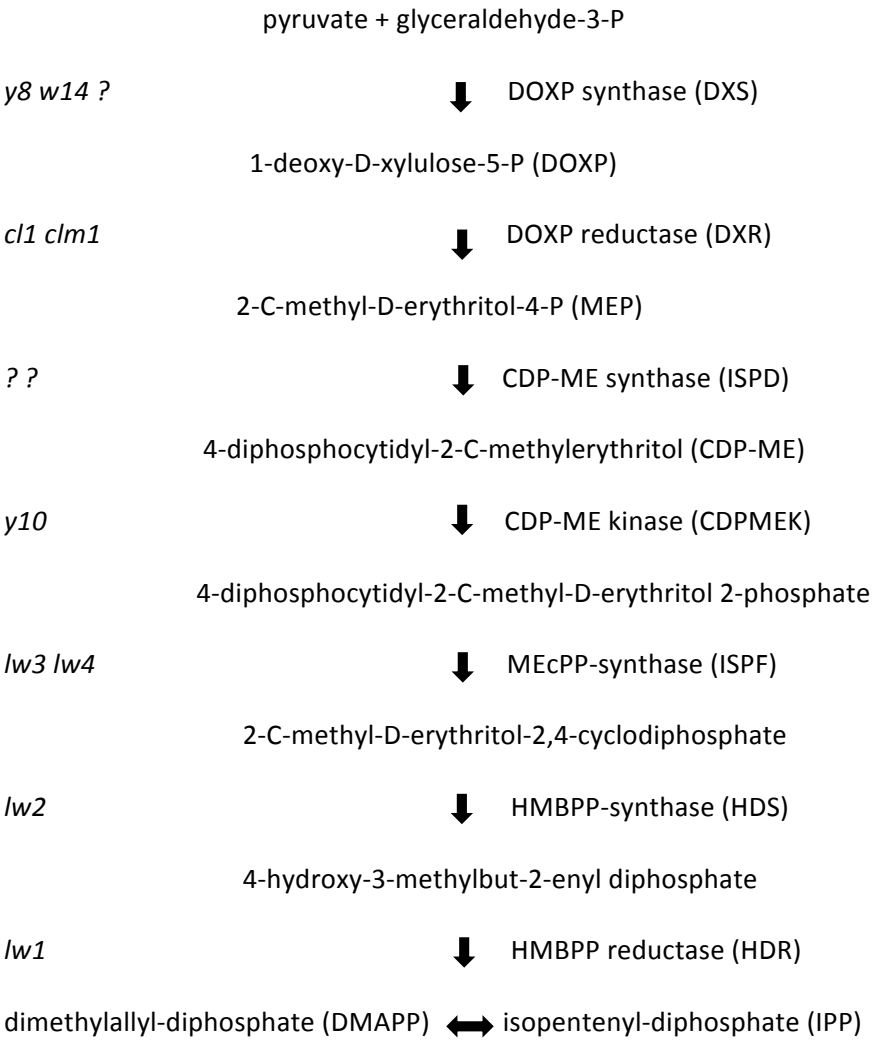

Table 1. Classical maize carotenoid genes and predicted gene models.

Classical Maize Gene	Location	Arabidopsis Gene Candidate	Orthologous Maize Gene Model
vp2	5S (5.04)	AT1G06570[1] (PDS1)	GRMZM2G088396
w3	2L (2.08)	AT3G11945 (PDS2)	GRMZM2G113476
y8	7S (7.01)	AT4G15560 (DXS)	GRMZM2G493395
w14	6L (6.05)	AT4G15560 (DXS)	GRMZM2G137151
cl1	3S (3.04)	AT5G62790 (DXR)	GRMZM2G056975
Clm1	8S	AT5G62790 (DXR)	GRMZM2G036290
y10	3L (3.07)	AT2G26930 (CDPMEK)	GRMZM5G859195
lw3	5L (5.06)	AT1G63970 (ISPF)	AC209374.4_FG002
lw4	4L (4.06)	AT1G63970 (ISPF)	GRMZM5G835542
lw2	5L (5.05)	AT5G60600 (HDS)	GRMZM2G137409
lw1	1L (1.10)	AT4G34350 (HDR)	GRMZM2G027059

[1] TAIR locus name (from www.arabidopsis.org).

Table 2. Predicted duplicate factor maize carotenoid genes and gene models.

Arabidopsis Gene	Orthologous Maize Gene Model	Location
AT3G63410[1] (VTE3)	GRMZM2G082998	1L
AT3G63410 (VTE3)	GRMZM2G099206 (pseudogene?)	3S
AT4G15560 (DXS)	GRMZM2G137151	6L
AT4G15560 (DXS)	GRMZM2G493395	7S
AT4G15560 (DXS)	GRMZM2G173641[2]	9S
AT2G02500 (ISPD)	GRMZM5G856881	3L
AT2G02500 (ISPD)	GRMZM2G172032	8L

[1] TAIR locus name (from www.arabidopsis.org).
[2] Data from Cordoba *et al.* 2011. J Exp Bot 62:2023-2038.