Supplementary Information for

Genome-wide association identifies a missing hydrolase for tocopherol synthesis in plants
Elise Alberta, Sungsoo Kimber, Maria Magallanes-Lundbacka, Yan Baoa, Nicholas Deasona, Benoit Daniloa, Di Wuc, Xiaowei Lic, Joshua C. Woodb, Michael A. Gorec, C. Robin Bueell, Dean DellaPennaa*

*Corresponding author: DellaPenna, D.; Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI, (517) 432-9284
Email: dellapen@msu.edu

This PDF file includes:

Supplementary text
Figures S1 to S14
SI References
Supplementary Information Text

Supplemental Material and Methods

Plant material, experimental design and growth conditions
*Arabidopsis* was grown in growing chambers at 22°C under a 16h-photoperiod (120 μmol m⁻² s⁻¹).
A total of 814 fully sequenced *Arabidopsis thaliana* accessions the 1001 Genomes Project dataset produced sufficient seeds for analyses (1). One replicate per accession was grown in an augmented randomized incomplete-block design repeated during two consecutive years. Col-0 (CS76778) and Ler-0 (CS77020) accessions were included as checks in each 16-plant flat to control for spatial variation. Seed set ranged from 7 weeks to 6 months depending on accessions, and harvested, dry seeds were and stored in a desiccator at room temperature for at least six weeks prior to HPLC analysis. Mutant lines used in this study include for AT5G38210, Atcrr7-1 (EMS mutant, (2)) and Atcrr7-3 (SALK_073978), for AT5G39220, Atvte7-1 (SALK_064238), Atvte7-3 (SALK_020115), Atvte7-4 (SALK_055998), Atvte7-5 (SALK_064958) and AT1G3820-1 (GK-453H07). Homozygous mutants were identified by PCR genotyping using primers in SI Appendix, Dataset S11. Seeds for mutant lines were obtained from the *Arabidopsis* Biological Resource Center (ABRC), excepted *Atcrr7-1* and *Atcrr7-3*, which were kindly provided by Toshiharu Shikanai at Kyoto University. A maize Mu insertion mutant in the 5’ UTR of *ZmVTE7* (Mu1052364) was obtained from the UniformMu source (3), homozygous progeny identified by PCR (SI Appendix, D S11) and evaluated in the greenhouse with W22 controls. A total of 5 plants for each genotype were self-pollinated and dry kernels used for HPLC analysis.

Tocochromanol analysis of *Arabidopsis* and maize dry seed.
Tocochromanolks were extracted from 10-12 mg of mature, dry *Arabidopsis* seeds and 15 mg of ground maize kernel were extracted as described previously (4, 5). Tocopherols and tocotrienol analyses were performed as previously described (4) and quantified relative to calibration curves generated with commercial standards γ-, α-, and δ-tocopherol (Supelco) and γ-, α-, and δ-tocotrienol standards (Cayman Chemical Company).

Statistical analysis of the phenotypic data
Phenotypic data from the two panel outgrowths was screened for outliers by examining the Studentized deleted residuals obtained from mixed linear models fitted with year and genotype as random effects. Following outlier removal, a Box-Cox procedure (6) was implemented to find the most appropriate transformation that corrected for non-normality of the error terms and unequal variances (SI Appendix, Fig. S1). Then, for each trait, the best linear unbiased predictor (BLUP) for each accession was predicted from a mixed model fitted across years accounting for the genotype by year interaction in ASREML version 4.0. A model selection procedure was fitted to select the optimal model for each trait. The variance components obtained from the optimal models were used to estimate heritability on a line-mean basis and the standard errors of these estimates were approximated using the delta method (7). The aforementioned procedures were conducted in R 3.5.3.

Genome-wide and pathway level association analysis
We used the imputed SNP matrix released by the 1001 Genomes Project (release v3.1) and selected 1,802,374 polymorphic SNPs with a minor allele frequency MAF ≥ 0.05 in our 814 accessions. A genome-wide association analysis was performed within easyGWAS (8) using the Efficient Mixed-Model Association (EMMAX) algorithm, including three principal coordinate axes and a genomic relationship matrix to account for population structure and kinship, respectively. Q-Q plots, Manhattan plots and summary statistics are available at the easyGWAS website and as SI Appendix, Fig. S2. To correct for multiple-testing, Benjamini-Hochberg corrections were applied. SNPs were declared significantly associated with a corrected *P*-value < 0.05. SNP impacts on proteins were annotated with SnpEff (release 4.1L) and the SnpEff *A. thaliana* database (release 2015-01-08) (9). Linkage disequilibrium (LD) between pairs of SNPs was estimated by using squared allele-frequency correlations (r²) in Plink! (10). Additionally, we conducted a pathway-level analysis that used prior knowledge of the tocochromanol biosynthetic pathway to select 73 *a priori*
candidate genes (SI Appendix, Dataset S3). Benjamini-Hochberg corrections for a subset of 74,399 SNPs located +/- 40 kb of these *a priori* genes were performed for each trait.

**Phylogenetic analysis and structural variations in *Zea mays***

The AtVTE7 (AT5G39220.1) protein sequence was used to identify homologous proteins in other species by BLASTP searches of the National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/) or the Joint Genome Institute (https://www.jgi.doe.gov) databases. Phylogenetic analysis was performed using the PhyML+SMS pipeline in https://npgphylogeny.fr. The multiple sequence alignment was generated with MAFFT (11) and curated using BMGE (12) with default settings. The maximum likelihood tree was obtained with PhyML+SMS (13) using the AIC criterion to select the best evolutionary model and keeping default settings for other parameters. Branch support values are based on 1000 bootstrap replicates. The phylogenetic tree was drawn with iTOL (https://itol.embl.de/) (14). VTE7 genes and their 10kb flanking regions in the 26 NAM founder lines were retrieved from National Center for Biotechnology Information (https://www.ncbi.nlm.nih.gov/) and aligned against the B73 reference (v4) using minimap2 v2.22 to identify potential structural variations (15). Alignment was visualized with the R package pafr (https://dwinter.github.io/pafr/; v0.0.2).

**Generation of AtVTE7 and ZmVTE7 overexpressing lines**

A full-length AtVTE7 cDNA was obtained from the Arabidopsis Information Resource (TAIR, www.arabidopsis.org, stock U50129). A full-length ZmVTE7 cDNA was amplified from developing B73 maize embryos. cDNAs were mobilized by NoI digestion and ligated into the pMLBART and its modified version for C-terminal YFP fusions (16), both driven by the 35S CaMV promoter. Five constructs (35S:AtVTE7:YFP, 35S:AtVTE7, 35S:ZmVTE7:YFP, 35S:cTP:ZmVTE7:YFP and 35S:ZmVTE7) were used to transform *Atvte7-1 Arabidopsis* plants by the floral dip method (17). 35S:cTP:ZmVTE7:YFP was generated by deleting the predicted 16 amino acid chloroplast targeting sequence of ZmVTE7 and replacing it with the *Arabidopsis* (dicot) RecA chloroplast transit peptide (cTP) (18).

A total of 71 independent transformants were selected by glufosinolate (120 mg L⁻¹) resistance and their pooled T2 seed analyzed for tocopherol content. T2 seed from lines exhibiting elevated tocopherol levels were subject to segregation analysis and segregating 3:1 for glufosinolate resistance were carried through to homozygosity. Three 35S:AtVTE7:YFP lines and five 35S:cTP-ZmVTE7:YFP were selected for further analysis. Plants overexpressing 35S:AtVTE2 (line 11) (19) were crossed with two homozygous 35S:AtVTE7:YFP lines and F1 seed and leaves expressing both transgenes analyzed for tocopherol levels.

**Confocal microscopy**

To assess the subcellular localization of AtVTE7 and ZmVTE7, in homozygous, five-day-old 35S:AtVTE7-YFP, 35S:cTP-ZmVTE7-YFP and 35S:YFP transgenic seedlings was analyzed using a Nikon A1R1 confocal microscope, with 520 nm and 550 nm wavelengths for excitation and emission, and chlorophyll autofluorescence was detected at 643 nm to 730 nm. The empty 35S:YFP transgenic plants served as control. The same constructs and conditions were used for transient expression and imaging in *Arabidopsis* protoplasts.

**RT-Q-PCR and RNAseq of developing seed and leaves of *Zea mays* and *Arabidopsis***

RNA was extracted from 1 cm long green *Arabidopsis* siliques, 22DAP maize embryos and fully extended maize 5th leaves using the hot borate method (20) and treated with the TURBO DNA-free kit I (Invitrogen). First strand cDNAs were synthesized using oligod(T) and Superscript III enzyme, mixed into PowerUp SYBR Green Master Mix (Thermo Fisher) and loaded into a QuantStudio 7 Flex Real-Time PCR system (Thermo Fisher) with negative controls, dilution points and reference genes. Thermal cycling conditions were 30 min at 48°C, 10 min at 95°C followed by 40 cycles of 15 sec at 95°C and 1 min at 60°C. RT-Q-PCR results were captured and analyzed using the QuantStudio Real-Time PCR Software (Thermo Fisher) and relative expression levels obtained by normalization from the reference gene using the 2-ΔCT method. Primers pairs were designed using Primer3 (21) (SI Appendix, Dataset S11). RNA from 9, 14, and 18 DAP seeds was isolated in replicate from *Col-0* and *Atvte7-7* plants and Illumina TruSeq Stranded mRNA libraries (Illumina,
San Diego, CA) prepared. Libraries were then sequenced in single end 50 nt mode on an Illumina HiSeq 4000 (Illumina, San Diego, CA). RNAseq reads were cleaned using Cutadapt v1.18 (22) to trim adapters and low-quality sequences. Alignment of the reads to the Arabidopsis reference genome (TAIR 10) (23) was completed using HISAT2 v2.1.0 (24) in stranded mode using parameters (--min-intronlen 20 --max-intronlen 60000) and sorted using SAMTools v1.9 (25). HTSeq v0.6.1p1 (26) was used to generate counts using the Arabidopsis annotation (ARAPORT 11) (27) using the parameters; --stranded=reverse, --minaqual=10, and --mode=union. Counts were analyzed using DESeq2 (28) and were first normalized using the relative log expression method to correct for differences among library sizes. Weakly expressed genes were filtered out using HTSFilter package (29). Per-gene dispersions were estimated by incorporating data-driven prior distributions and negative binomial generalized linear models were fitted to estimate moderated log2 fold changes in gene expression levels between genotypes at each stage. Per-gene Wald test statistics were computed to identify significantly differentially expressed genes, with an FDR threshold of 5%. NCBI SRA accessions for RNAseq data will be made public upon acceptance. Correlation between expression abundance of VTE7 homologs and totalT in the NAM founders was assessed across kernel development by Spearman rank correlation tests using data from (30).

Analysis of Chlorophylls, Chlorophyll Derivatives and Tocomoenoenols

Arabidopsis seeds at development stages based on days after pollination were dissected from siliques and 150 seeds per biological replicate harvested and seed and leaf samples were immediately frozen in liquid nitrogen. Leaf tissue was extracted as described previously (4). Developing seeds were extracted as described previously (30) except that 360 µL of ethyl acetate and 540 µL of water were added to the homogenized acetone extract to allow phase separation. For samples destined for LC-MS analysis, telmisartan was also added as a LC-MS internal standard to the master extraction buffers at a final concentration of 104 nM. The lower organic phase (chloroform) for leaf extractions and the upper organic phase (ethyl acetate) for seed extraction was divided into three equal aliquots, evaporated, dissolved and analyzed as follows for the different compound classes: for tocopherol and tocomoenoenol analysis an aliquot was dissolved in 50 µL of 96:4 hexane: methyl tert-butyl ether and separated using a Diol column as described previously (31). For analysis of chlorophylls, chlorophyllides and pheophytins an aliquot was dissolved in 50 µL of 80:20 (v/v) Methanol:acetone and resolved on a Shimadzu Prominence HPLC equipped with a 5 µm Spherisorb ODS-2 column (250 x 4.6 mm) (Orochem Technologies) and separated as described in (32) except that the gradient was from 0 to 100% solvent B over 7.5 min, held at 100% solvent B for an additional 7.5 min and then returned to 100% solvent A for 3 min to re-equilibrate. Compounds were identified by retention time and spectral characteristics and quantified at 450 nm (carotenoids), 660 nm (Chl a and chlorophyllide a), 645 nm (Chl b) and fluorescence at 408 nm excitation and 670 nm emission for pheophytin a. Chlorophylls a and b commercial standards (Sigma-Aldrich) were used to generate standard curves for quantification of Chl a and b and chlorophyllides a and b when present. A pheophytin a standard was prepared by acidification of chlorophyll a (33).

For analysis of partially reduced chlorophyll intermediates (esterified to geranylgeranio, dihydroygeranylgeraniol or tetrahydrogeranylgeraniol instead of phytol) an aliquot was dissolved in 500 µL of isopropanol containing 1mg/ml butylatedhydroxytoluene and 10 µL analyzed using a Waters Xevo G2-XS QToF mass spectrometer fed by a Waters Acquity UPLC system using a Waters Acquity UPLC BEH C18 Column (130A, 1.7um, 2.1mm x 100mm) at 55°C with solvent A: 60:40 (v/v) acetonitrile:water with 10 mM ammonium formate and solvent B: 90:9:1 (v/v/v) isopropanol:acetoniitrile:water with 10 mM ammonium formate. The flow rate was 0.4 mL/min and the solvent gradient was (A:B) initially 80:20 with linear ramps to 57.43 at 2 minutes, to 46:54 at 12 minutes, to 30:70 at 12.10 minutes, to 1:99 at 18 minutes, to 60:40 at 18.10 min and held at 60:40 until 20 min. Mass spectra were acquired using positive mode electrospray ionization with dynamic range extension over m/z 50 to 2,000 with a mass resolution of 20,000. Two parallel functions were used with 0.2 s per function, TOF1 (parents) had a collision energy of 6 eV and TOF2 (products) had a continuously ramping Collision Energy (CE) from 20 to 80 eV. Other parameters include capillary voltage of 3.0 kV, cone voltage of 35 V, desolvation temp of 350°C, source temp of 100°C. Cone gas (N2) was at 40 L/hour and desolvation (N2) at 600 (L/hour). For statistical analyses of
non-targeted metabolites MS² mode was used to collect full-scan data of both precursor and product ions by alternating low CE and high CE scans in the same experiment. Statistical data analysis was performed using the Progenesis QI software package (Progenesis QI Version 2.4, Nonlinear Dynamics).

Expression and assay of recombinant fusion proteins
Full length cDNAs encoding AtVTE7, ZmVTE7 and AT1G13820 were subcloned as C-terminal His-tagged fusions into pET28a and transformed into BL21 star (DE3) star E. coli cells. Expression and purification of His-tagged fusion proteins was as per the manufacturer’s protocol (Thermo Fisher) except that induction was with a final isopropyl-β-D-thiogalactopyranoside concentration of 0.5mM at 23°C for 8h. Crude protein extracts or purified fusion proteins were assayed for hydrolase activity at 37°C in 400 µL of reaction buffer containing 10 mM HEPES pH 8.0 with 0.05% (v/v) Triton X-100 containing 2 micrograms of purified His-tagged proteins or 100mM HEPES pH 7.5 with 0.5% (v/v) Triton X-100 containing 100 µL of crude protein extracts. Synthetic substrates containing acyl groups esterified to 4-nitrophenol (4-Nitrophenyl-acetate (C₂), -butyrate (C₄), octanoate (C₈), -decanoate (C₁₀), -dodecanoate (C₁₂), -myristate (C₁₄), -palmitate (C₁₆) or -stearate(C₁₈)) were dissolved at 50 mM in ethanol (C₄-C₈), isopropanol (C₁₀) or at 25 mM in isopropanol (C₁₂-C₁₈) and added to reactions at a final concentration of 0.2mM. Release of 4-nitrophenol product was measured at 405 nm. One unit of activity was defined as one µmol of 4-Nitrophenol released per minute at 37 °C, pH 8.0. Chlorophylls a and b, pheophytin a, tocopherol acetate, a mixture of α- and γ-tocopherol fatty acid esters or phytol fatty acid esters were dissolved in acetone and tested as substrates under a variety of concentrations and final reaction conditions including ranging from pH 5.5-9.0, 5-20% acetone, 0.05-2% v/v Triton X-100 or NP40 or 17 mM n-dodecyl-b-D-maltoside.
Fig. S1. Distribution of untransformed BLUPs for tocopherol traits evaluated in the 814 *Arabidopsis thaliana* accessions and estimated broad sense heritabilities. BLUPs: Best Linear Unbiased Predictors. The white dots and whiskers inside the violin plots indicate median and first and third quartiles, respectively. $H^2$: Broad-sense heritability estimate on a line-mean basis across two outgrowths. SE: Standard error of the heritability estimate approximated using the delta method. SD: Standard deviation of the untransformed BLUPs. Range: Minimal and maximal values of untransformed BLUPs (in ng per mg dry seeds). $\lambda$: Lambda values used in Box-Cox transformations.
\gamma T

1,802,374 SNPs - whole genome FDR

74,399 SNPs in 73 pathway genes - pathway level FDR
Fig. S2. Manhattan plots from genome wide association analyses and pathway-level analyses performed for aT, dT, gT and totalT BLUP estimates. In each upper panel, the whole genome $-\log_{10}$ FDR $P$-values obtained for SNPs from the EMMAX GWA analyses performed in easyGWAS are plotted against their genomic positions. Corresponding interactive Manhattan plots and qqplots are available online at the easyGWAS website (https://easygwas.ethz.ch/). In each lower panel, the pathway-level $-\log_{10}$ FDR $P$-values obtained for a total of 74,399 SNPs located within +/- 40 kb of 73 a priori genes (SI Appendix, Dataset S3) are plotted against their genomic positions. In both panels, the black dashed lines indicate a FDR significance threshold at $\alpha = 0.05$. A priori pathway genes located within +/- 40 kb of statistically significant SNPs are indicated in both panels: white letters in a black box when significant at both the genome wide and pathway levels, white letters in a grey box when significant only at the pathway-level. VTE7 is highlighted by white letters in a red box. Genomic information for all significant SNPs is available in SI Appendix, Dataset S2 (whole genome EMMAX analysis) and S4 (pathway-level analysis).
Fig. S3. Linkage disequilibrium decay in the panel. Linkage disequilibrium (LD) in the panel was computed for 50,000 random single nucleotide polymorphisms (SNPs). Black lines show distribution at different percentile cutoffs (indicated at the left). The red line shows the average distribution of the variants. Percentiles and average were computed on 100 bp intervals.
Fig. S4. *AtVTE7* (AT5G39220), *AtCRR7* (AT5G39210) and *AT1G13820* expression across *Arabidopsis thaliana* development and *ZmVTE7* (GRMZM5G898684) across maize development. Tissue-specific expression was visualized in the eFP Browser (http://bar.utoronto.ca/efp/cgi-bin/efpWeb.cgi).
Fig. S5. (A) *AtVTE7* (AT5G39220) gene structure and location of the four *Atvte7* T-DNA alleles. Black and white boxes represent coding sequences and UTRs, respectively. The narrow gray boxes indicate the locations of T-DNA insertions. (B) Normalized expression of *AtVTE7* in 1 cm long green siliques measured by qPCR with Actin 2 (AT3G18780) used as the reference gene for normalization. Primers are available in SI Appendix, Dataset S1. ND indicates no expression detected. (C) Leaf tocopherol content in various *Atvte7* alleles and Col-0 wild type measured in leaves of 4-week-old plants. Means and standard deviations are from four biological replicates. Mutant means significantly different from Col-0 in Dunnett t-tests are indicated by asterisks: *** $P$-value < 0.001; ** 0.001 ≤ $P$-value < 0.01; * 0.01 ≤ $P$-value < 0.05.
Fig. S6. Maximum likelihood phylogenetic tree of AtVTE7 and related alpha/beta hydrolase proteins from plants, algae and cyanobacteria. The tree with the highest log likelihood (-6721.66) is shown. Branch support values obtained from 1,000 bootstraps are indicated when higher than 70%. Five clades are indicated with clades I (VTE7), II (Pheophytinase) and IV (CLD1 dephytylase) highlighted. The scale bar represents number of substitutions per site. Sequence data can be found at GenBank/EMBL with the following identification numbers: Arabidopsis thaliana (At): AtVTE7, NP_198738 (AT5G39220); AT1G13820, NP_001078682; AtPPH, NP_196884 (AT5G13800); AtCLD1, NP_001078682 (AT5G38520); AT5G19850, NP_568381; AT4G36530, NP_195371; Theobroma cacao (Tc): TcVTE7, EOY03666; Tc 01714, EOY01714; TcPPH, EOX96583; TcCLD1, EOX98976; Tc 32292, EOY32292; Tc 46891, XP_007046891; Zea mays (Zm): ZmVTE7, XP_008667052; ZmPPH, NP_001141976; ZmCLD1, NP_001359425; Zm 53276, XP_008653276; Zm 40734, NP_001140734; Oryza sativa (Os): OsVTE7, EEE67540; OsPPH, XP_015643750; OsCLD1, XP_015640194; Os 46902, XP_015646902; Os 46553, KAF2946553; Picea sitchensis (Ps): PsCLD1, ABK26500; Ps 24784, ABK24784; Physcomitrella patens (Pp): PpVTE7, XP_024404092; PpPPH, XP_024365762; Pp CLD1, XP_024384194; Pp 56405, XP_024356405; Pp 69572, XP_024369572; Chlamydomonas reinhardtii (Cr): CrVTE7, PNW84646; CrPPH, PNW75147; CrCLD1, PNW71679; Cr 75838, PNW75838; Cr 98367, XP_001698367; Synechococcus sp PCC 7002 (Sco): Sco 93039, ACA98039; Sco 06796, WP_012306796; Sco 07877, WP_012307877.
Fig. S7. (A) Diagram of AT1G13820 and location of the At1g13820-1 T-DNA allele. Black and white boxes represent coding sequences and UTRs, respectively. The narrow gray box indicates the location of the T-DNA insertion. (B) Unlike Col-0, AT1G13820 transcript was undetectable by RT-PCR in At1g13820-1 mutant leaves. Actin 8 (AT1G49240) is shown as a reference gene. Primers are available in SI Appendix, Dataset S1. (C) Seed and (D) leaf tocopherol content in Col-0 wild type, two Atvte7 alleles, At1g13820-1 and corresponding Atvte7 At1g13820-1 double mutants. Tocopherols were measured in dry seeds and in leaves harvested from 4-week-old plants. Means and standard deviations were computed from four biological replicates which for leaves is four, pooled 4-week-old rosettes and for seeds is from individual plants. Means significantly different from each other in Tukey’s Honest Significant Difference test are indicated with letters at each stage (P-value < 0.05).
Fig. S8. (A) Diagram of the ZmVTE7 gene (GRMZM5G898684) and location of the Zmvte7-1 Mu insertion. Black and white boxes represent coding sequences and UTRs, respectively. The narrow gray box indicates the location of the Mu insertion. Normalized ZmVTE7 expression measured by qPCR from (B) 22 days after pollination (DAP) embryos isolated from homozygous Zmvte7-1 and WT segregants, and (C) fully extended 5th leaves from homozygous Zmvte7-1 and the inbred parental line W22. GRMZM2G027378 (UBIE2) was used as the reference gene for normalization. Primers are available in SI Appendix, Dataset S1. Means and standard deviations were computed from five biological replicates taken from individual plants. Statistically significant difference between Zmvte7-1 and WT in t-test is indicated using asterisks: *** P-value < 0.001.
Fig. S9. Alignment of 10kb flanking the $VTE7$ locus in B73 (y-axis) vs the NAM founders. Alignment was performed using minimap2 (v2.22) and visualized with the R package pafr (https://dwinter.github.io/pafr/; v0.0.2).
Fig. S10. Correlation of VTE7 mRNA expression abundance across a developmental time course of kernels from 12 days after pollination (DAP) to 36 DAP, roots, and shoots with tocopherol levels in these tissues in the NAM founders. Blue and red hues represent positive and negative correlations, respectively, and correlation strength is depicted by both opacity and linearization of the inner circles. Data from (30).
Fig. S11. (Upper panels) Chloroplast localization of 35S:AtVTE7:YFP, 35S:cTP-ZmVTE7:YFP and 35S:YFP in homozygous 5-day old Arabidopsis seedlings and (Lower panels) 35S:AtVTE7:YFP and 35S:FBN2:YFP (fibrillin2, AT2G35490) by transient expression in Arabidopsis protoplasts. Fluorescence images were taken with a confocal laser scanning microscope. For each transformed construct, four images are displayed: YFP (green), chlorophyll autofluorescence (red), merged YFP and chlorophyll autofluorescence signals, and bright field. Scale bars indicate 20 µm. Fibrillin2 is a control that is targeted to plastoglobuli and thylakoids (34, 35).
Fig. S12. Total tocopherols in segregating T2 seed pools from individual AtVTE7 and cTP-ZmVTE7 overexpression lines in the Atvte7-1 mutant. N indicates the number of independent transgenic lines obtained for each construct: 35S:AtVTE7:YFP, 35S:AtVTE7, 35S:cTP-ZmVTE7:YFP and 35S:cTP-ZmVTE7. Black lines indicate means and red dots show lines segregating for a single transgene locus and for which homozygous transgenic seed and/or leaf metabolite contents are displayed in main text figures and SI Appendix, Datasets S6 and S8. Group means significantly different from each other in Mann-Whitney pairwise tests are indicated with letters (FDR P-value < 0.05). Data are available in SI Appendix, Dataset S5.
Fig. S13. Whole plant phenotypes of homozygous 4-week old wild-type, Atvte7-1 and lines over-expressing 35S:AtVTE7:YFP or 35S:cTP-ZmVTE7:YFP in the Atvte7-1 mutant. Lines labeled with red letters are those with seed and leaf metabolite contents displayed in main text figures and SI Appendix, Datasets S6 and S8 while data for lines with black lettering are shown only in SI Appendix, Dataset S6.
Fig. S14. (A) Total chlorophylls, (B) chlorophyllide a and (C) pheophytin a levels in developing seeds of Col-0, Atvte7-1 and lines overexpressing 35S:AtVTE7:YFP or 35S:cTP-ZmVTE7:YFP in the Atvte7-1 mutant. Seeds were analyzed at four stages: linear (9 days after pollination, DAP), mature green (14 DAP), yellowing (18 DAP) and brown (24 DAP). Means and standard deviations were computed from four biological replicates of 150 seeds taken from individual plants and expressed in ng in 150 seeds. At each stage, means significantly different from each other in Tukey’s Honest Significant Difference test are indicated with letters (P-value < 0.05).
List of Supplemental Datasets (Individual supplemental datasets are uploaded separately as excel files)

Dataset S1. Seed tocopherol content collected for 814 accessions selected from the 1001 Genomes Project dataset. Raw phenotypic means, Best Linear Unbiased Predictors (BLUPs, transformed) and untransformed BLUPs are displayed. \( \delta T \), delta-tocopherol; \( \gamma T \), gamma-tocopherol; \( \alpha T \), alpha-tocopherol; totalT, total tocopherols (\( \delta T + \gamma T + \alpha T \)).

Dataset S2. Association analysis for tocopherol BLUP estimates in the GWA panel using the Efficient Mixed-Model Association (EMMAX) algorithm. Bold indicates peak SNP. MAF, Minor Allele Frequency. \( \delta T \), delta-tocopherol; \( \gamma T \), gamma-tocopherol; \( \alpha T \), alpha-tocopherol; totalT, total tocopherols (\( \delta T + \gamma T + \alpha T \)).

Dataset S3. Genomic information for the 73 a priori pathway genes. Genes that were identified in this study in the whole genome and/or pathway-level approaches are displayed in bold (FDR \( P\)-value \( \leq 0.05 \)). MAF: Minor Allele Frequency.

Dataset S4. Significant marker-trait associations identified in the pathway-level analysis using 74,399 SNP +/- 40 kb from the 73 a priori candidate genes listed in Supplemental Dataset 3. A FDR threshold of 5% was applied to call significant SNPs. Bold denotes SNPs that also passed the genome-wide significance threshold of 5% FDR in the EMMAX GWA analysis. MAF: Minor Allele Frequency.

Dataset S5. Tocopherol content of segregating T2 seed pools for overexpression constructs in \( \text{Atvte7-1} \). Four different constructs were transformed: 35S:AtVTE7:YFP, 35S:AtVTE7, 35S:cTP-ZmVTE7:YFP and 35S:cTP-ZmVTE7. Values are expressed in ng tocopherols per mg dry seeds. YES: indicates lines containing a single transgene locus that were made homozygous and selected for detailed analyses. \( \delta T \), delta-tocopherol; \( \gamma T \), gamma-tocopherol; \( \alpha T \), alpha-tocopherol; totalT, total tocopherols (\( \delta T + \gamma T + \alpha T \)).

Dataset S6. Tocopherol and chlorophyll levels in leaves of \( \text{Col-0} \) wild-type, \( \text{Atvte7-1} \) and several lines over-expressing \( \text{ZmVTE7} \) in \( \text{Atvte7-1} \) mutant background. Means and standard deviations are computed from five biological replicates and expressed in ng per mg. One biological replicate corresponded to pooled rosettes from five 4-week-old plants. Means significantly different from each other in Tukey's Honest Significant Difference test are indicated with letters (\( P\)-value < 0.05).

Dataset S7. Activity of purified AtVTE7, ZmVTE7 and AT1G13820 proteins against synthetic substrates with esterified acyl chains of the indicated length.

Dataset S8. Chlorophyll content in leaves of \( \text{Arabidopsis thaliana} \) lines overexpressing \( \text{AtVTE7} \) and \( \text{ZmVTE7} \) in the \( \text{Atvte7-1} \) mutant background. Means and standard deviations are computed from five biological replicates and expressed in ng per mg. One biological replicate corresponded to pooled rosettes from five 4-week-old plants. Means significantly different from each other in Tukey's Honest Significant Difference test are indicated with letters (\( P\)-value < 0.05).

Dataset S9. Gene expression analysis of 9, 14 and 18 DAP developing seeds for the indicated pathways, including the 73 a priori genes in Supplemental Dataset 3, VTE7 and additional
chlorophyll biosynthetic and degradation genes. Three biological replicates each for 9, 14 and 18 DAP seeds were harvested from Col-0 and Atvte7-1 plants and used for RNA seq analysis. One biological replicate is composed of seeds harvested from 4-8 individual plants. Statistical analysis was done for each WT/mutant pair at each DAP. VTE7 mRNA levels were reduced ~100-fold at 9 and 14 DAP in Atvte7-1 in comparison to WT. Only one other gene, shikimate kinase2 at 18 DAP, was significantly different between Col-0 and Atvte7-1. The seven genes that were identified in this study using whole genome and/or pathway-level approaches are displayed in bold.

Dataset S10. Mass/time peaks of compounds identified by untargeted LC-MS analysis of Atvte7-1 and WT seeds at 9, 14 and 18 days after pollination (DAP). A total of eleven mass/time peaks passed a statistical cutoff of > 2-fold change and P-value < 0.05 between genotypes.

Dataset S11. List of primers used.

SI References
1. C. Alonso-Blanco, et al., 1,135 Genomes reveal the global pattern of polymorphism in Arabidopsis thaliana. Cell 166, 481–491 (2016).
2. M. Kamruzzaman Munshi, Y. Kobayashi, T. Shikanai, Identification of a novel protein, CRR7, required for the stabilization of the chloroplast NAD(P)H dehydrogenase complex in Arabidopsis. The Plant Journal 44, 1036–1044 (2005).
3. D. R. McCarty, et al., Steady-state transposon mutagenesis in inbred maize. Plant Journal 44, 52–61 (2005).
4. S. Gonzalez-Jorge, et al., Carotenoid cleavage dioxygenase4 is a negative regulator of β-carotene content in Arabidopsis seeds. Plant Cell 25, 4812–4826 (2013).
5. A. E. Lipka, et al., Genome-wide association study and pathway-level analysis of tocochromanol levels in maize grain. G3: Genes, Genomes, Genetics 3, 1287–1299 (2013).
6. G. E. P. Box, D. R. Cox, An analysis of transformations. Journal of the Royal Statistical Society: Series B (Methodological) 26, 211–243 (1964).
7. J. B. Holland, W. E. Nyquist, C. T. Cervantes-Martinez, Estimating and Interpreting Heritability for Plant Breeding: An Update. Plant Breeding Reviews 22, 9–112 (2003).
8. D. G. Grimm, et al., easyGWAS: A Cloud-Based platform for comparing the results of genome-wide association studies. The Plant Cell 29, 5–19 (2017).
9. P. Cingolani, et al., A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff; SNPs in the genome of Drosophila melanogaster strain w 1118; iso-2; iso-3. Fly 6, 80–92 (2012).
10. S. Purcell, et al., PLINK: A tool set for whole-genome association and population-based linkage analyses. The American Journal of Human Genetics 81, 559–575 (2007).
11. K. Katoh, D. M. Standley, MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Molecular Biology and Evolution 30, 772–780 (2013).
12. A. Criscuolo, S. Gribaldo, BMGE (Block Mapping and Gathering with Entropy): A new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evolutionary Biology 10 (2010).
13. F. Lemoine, et al., NGPhylogeny.fr: new generation phylogenetic services for non-specialists. Nucleic acids research 47, W260–W265 (2019).
14. I. Letunic, P. Bork, Interactive Tree Of Life (iTOL): An online tool for phylogenetic tree display and annotation. Bioinformatics 23, 127–128 (2007).
15. H. Li, Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34, 3094–3100 (2018).
16. A. P. Gleave, A versatile binary vector system with a T-DNA organizational structure conducive to efficient integration of cloned DNA into the plant genome. Plant Molecular Biology 20, 1203–1207 (1992).
17. S. J. Clough, A. F. Bent, Floral dip: A simplified method for Agrobacterium-mediated transformation of Arabidopsis thaliana. *Plant Journal* **16**, 735–743 (1998).
18. R. H. Kohler, Exchange of Protein Molecules Through Connections Between Higher Plant Plastids. *Science* **276**, 2039–2042 (1997).
19. E. Collakova, D. DellaPenna, Homogentisate phytyltransferase activity is limiting for tocopherol biosynthesis in Arabidopsis. *Plant Physiology* **131**, 632–642 (2003).
20. C. Y. Wan, T. A. Wilkins, A modified hot borate method significantly enhances the yield of high-quality RNA from cotton (Gossypium hirsutum L.). *Analytical Biochemistry* **223**, 7–12 (1994).
21. A. Untergasser, *et al.*, Primer3-new capabilities and interfaces. *Nucleic acids research* **40**, e115 (2012).
22. M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet journal* **17**, 10 (2011).
23. P. Lamesch, *et al.*, The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research* **40** (2012).
24. D. Kim, J. M. Paggi, C. Park, C. Bennett, S. L. Salzberg, Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nature Biotechnology* **37** (2019).
25. H. Li, *et al.*, The sequence alignment/map format and SAMtools. *Bioinformatics (Oxford, England)* **25**, 2078–9 (2009).
26. S. Anders, P. T. Pyl, W. Huber, HTSeq: a Python framework to work with high-throughput sequencing data. *Bioinformatics (Oxford, England)* **31**, 166–9 (2015).
27. C. Cheng, *et al.*, Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *The Plant Journal* **89** (2017).
28. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome biology* **15**, 550 (2014).
29. A. Rau, M. Gallopin, G. Celzou, F. Jaffrézic, Data-based filtering for replicated high-throughput transcriptome sequencing experiments. *Bioinformatics* **29**, 2146–2152 (2013).
30. C. H. Diepenbrock, *et al.*, Novel loci underlie natural variation in vitamin E levels in maize grain. *Plant Cell* **29**, 2374–2392 (2017).
31. L. Mène-Saffrané, A. D. Jones, D. DellaPenna, Plasto-chromanol-8 and tocopherols are essential lipid-soluble antioxidants during seed desiccation and quiescence in Arabidopsis. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 17815–17820 (2010).
32. M. Zapata, F. Rodríguez, J. L. Garrido, Separation of chlorophylls and carotenoids from marine phytoplankton: A new HPLC method using a reversed phase C8 column and pyridine-containing mobile phases. *Marine Ecology Progress Series* **195**, 29–45 (2000).
33. S. J. Schwartz, S. L. Woo, J. H. Von Elbe, High-performance liquid chromatography of chlorophylls and their derivatives in fresh and processed spinach. *Journal of Agricultural and Food Chemistry* **29**, 533–535 (1981).
34. I. Bouchnak, *et al.*, Unraveling hidden components of the chloroplast envelope proteome: opportunities and limits of better MS sensitivity. *Molecular & Cellular Proteomics* **18**, 1285–1306 (2019).
35. P. K. Lundquist, *et al.*, The Functional Network of the Arabidopsis Plastoglobule Proteome Based on Quantitative Proteomics and Genome-Wide Coexpression Analysis. *Plant Physiology* **158** (2012).