FRONT MATTER

Title

- The *Eruca sativa* genome and transcriptome: A targeted analysis of sulfur metabolism and glucosinolate biosynthesis pre and postharvest

Authors

Luke Bell [1*], Martin Chadwick [2], Manik Puranik [2], Richard Tudor [3], Lisa Methven [2], Sue Kennedy [3], Carol Wagstaff [2]

Affiliations

[1] School of Agriculture, Policy & Development, PO Box 237, University of Reading, Whiteknights, Reading, Berkshire. RG6 6AR. UK.

[2] School of Chemistry Food & Pharmacy, PO Box 226, University of Reading, Whiteknights, Reading, Berkshire. RG6 6AP. UK.

[3] Elsoms Seeds Ltd., Pinchbeck Road, Spalding, Lincolnshire. PE11 1QG. UK.

[*] luke.bell@reading.ac.uk

Abstract

Rocket (*Eruca sativa*) is a source of health-related metabolites called glucosinolates (GSLs) and isothiocyanates (ITCs) but little is known of the genetic and transcriptomic mechanisms responsible for regulating pre and postharvest accumulations. We present the first *de novo* reference genome assembly and annotation, with ontogenic and postharvest transcriptome data relating to sulfur assimilation, transport, and utilization. Diverse gene expression patterns related to sulfur metabolism and GSL biosynthesis are present between inbred lines of rocket. A clear pattern of differential expression determines GSL abundance and the formation of hydrolysis products. One breeding line sustained GSL accumulation and hydrolysis product formation throughout storage. Copies of *MYB28*, *SLIM1*, *SDI1* and *ESM1* orthologs have increased and differential expression postharvest, and are associated with GSLs and hydrolysis product formation. Two glucosinolate transporter gene orthologs (*GTR2*) were found to be associated with increased GSL accumulations.

MAIN TEXT

Introduction

Sulfur (S) is a critical macronutrient that plants require for growth and development (*1*). Sulfate (SO_4^{2-}) is utilized as a primary means of synthesizing numerous S-containing metabolites, such as amino acids (cysteine and methionine), glutathione (GSH), and glucosinolates (GSLs) (*2*). GSL compounds are present in species of the order Brassicales, and are abundant in many commonly consumed vegetables and condiments worldwide, such as rapeseed (*Brassica napus*), Chinese cabbage (*Brassica rapa*), cabbage (*Brassica oleracea* var. *capitata*), and broccoli (*B. oleracea* var. *italica*) (*3*). GSLs are also found in the leafy vegetable *Eruca sativa* (commonly known as "salad" rocket, rucola, or arugula), which has gained significant popularity amongst consumers over the last ten years (*4*). Rocket salad is known for its distinctive flavour, aroma, and pungency, and can be eaten raw without the need for cooking (*5*).

GSLs are synthesised as part of plant defense mechanisms against pests and diseases (*6*), and can also act as important S storage molecules (*1*). They can be grouped into three classes:

aliphatic (derived from methionine), indolic (derived from tryptophan), and aromatic (derived from phenylalanine) (*7*). Salad rocket contains all three of these classes, with aliphatic GSLs making up the largest part of the GSL profile in leaf tissues. Up to 18 GSLs are regularly reported for *E. sativa* (*8*). Compounds such as glucosativin (4-mercaptobutyl GSL; GSV) and glucorucolamine (4-cystein-S-yl)butyl GSL; GRL) are unique to the genera *Eruca* and *Diplotaxis* ('wild' rocket) (*9*). Rocket is also unique in that they are the only species reported to contain dimeric GSLs. Glucosativin can exist in a dimer form (dimeric 4-mercaptobutyl GSL; DMB), and diglucothiobeinin (4-(β-D-glucopyranosyldisulfanyl)butyl GSL; DGTB) is a unique GSL dimer that is also present (*10*).

Despite the advances being made in elucidating the *Arabidopsis thaliana* and *B. oleracea* GSL pathways, very little novel gene discovery has taken place outside of these species. The reason for this is the lack of genome sequence available for *E. sativa,* and reliance upon knowledge about related species, which is not able to account for the large differences in the GSLs of the species.

Much is now known about the 'core' GSL biosynthesis pathway in *Arabidopsis* and the regulatory mechanisms that respond to different biotic and abiotic stimuli (*11*). Six main R2R3 MYB transcription factors (TFs) have been identified as regulators of GSL synthesis. Aliphatic GSLs are regulated by *MYB28*, *MYB29*, and *MYB76*, and indolic GSLs by *MYB34*, *MYB51*, and *MYB122* (*2*). These MYBs are in turn regulated by basic helix-loop-helix (bHLH) transcription factors such as *MYC2*, which are involved in plant defense response (*12*). Other transcriptional regulators, such as *SLIM1* (*SULFUR LIMITATION 1*) and *SDI1* (*SULFUR DEFICIENCY INDUCED 1*) also interact with MYB transcription factors to regulate the use and efficiency of sulfur within the plant. As GSLs contain significant amounts of sulfur (up to 30% of total plant S-content) the synthesis and breakdown management of these compounds is crucial in times of stress. This is why GSL biosynthesis is so closely related to both primary and secondary sulfur metabolism pathways (Figure 1) (*13*).

Individual downstream GSL biosynthesis genes are regulated in response to a wide range of stress stimuli in response to changes in both MYB and MYC activity. Some of the most well studied are genes encoding methylthioalkylmalate synthase (MAM) enzymes (regulators of GSL side chain lengths) (*19*), genes encoding CYP79 enzymes (catalysts of the conversion of chain elongated amino acids into their respective aldoximes) (*20*), and genes encoding CYP83 enzymes (that convert indolic aldoximes into corresponding thiohydroxymates) (*16*). Few studies have attempted to understand such complex regulation in the context of common horticultural practice, or indeed how transcriptomic regulation and response to the stresses imposed by harvesting, washing, processing, and storage differs between cultivars in such a scenario. *E. sativa* is a crop with great potential for enhancement of nutritional value, and it is therefore essential to understand how GSL biosynthesis and sulfur metabolism are regulated in order to direct breeding programs.

GSLs themselves are not bioactive, and are hydrolysed by myrosinase enzymes (TGGs) when tissue damage takes place. They form numerous breakdown products including isothiocyanates (ITCs), which are of foremost interest for their anticarcinogenic effects in humans (*17*). Salad rocket produces the ITC sulforaphane (SF; a breakdown product of glucoraphanin; 4-methylsulfinylbutyl GSL, GRA), which has been well documented for its potent anticarcinogenic properties (*18*). SF is abundant in broccoli, however its hydrolysis from GRA is often inhibited or prevented due to high cooking temperatures employed by consumers, which denatures myrosinase at temperatures >65°C (*19*). As rocket does not have a cooking requirement, it is an ideal crop for improvement and the enhancement of GRA content through selective breeding.

A previous study by Bell et al. (*20*) observed that both GSL and ITC concentrations increased significantly in rocket salad post-processing, but that this varied according to cultivar. The authors proposed that in response to the harvesting and washing process, stress responses within leaf tissues were initiated, leading to the increase in synthesis of GSLs and subsequent hydrolysis into ITCs. Sugar content, by comparison, showed little dynamic change and little

reduction in the same samples, which could have implications for sensory perceptions and consumer acceptance (*5*). The rejection of rocket leaves based on taste and flavour could be a significant barrier to the ingestion of health-related compounds such as SF.

We present a *de novo E. sativa* reference genome sequence, and report on the specific effects harvest, wash treatment, and postharvest storage have on GSL biosynthesis and sulfur metabolism gene expression through RNA sequencing (RNAseq) in three elite inbred lines. We also report on the degree of global transcriptomic change between each line, and the implications these data have for plant breeding selections and optimisation of GSL/ITC composition. We also present evidence of transcriptomic changes between first and second cuts of rocket plants, and how this in turn leads to elevated concentrations of both GSLs and ITCs. We hypothesised that each rocket line would vary in their ability to retain and synthesize GSLs post washing and during shelf life cold storage; as well as vary in their relative abundances between first and second cuts. This study represents a significant advancement of the genetic knowledge for this species.

Results

E. sativa genome assembly and annotation

De novo reference genome sequence was produced by interleaving Illumina MiSeq and HiSeq2500 sequence data (Illumina Inc., San Diego, CA, USA). An elite breeding line (designated **C**) was selected for PCR-free paired-end sequencing and long mate-pair end sequencing and assembled into 49,933 contigs (>= 500 bp). The resulting assembly was approximately 851 Mb in size (Supplementary Table S1).

Transposable elements (TEs) within the *E. sativa* genome comprise approximately 66.3% of its content. The majority of TEs are long terminal repeat (LTR) retrotransposons (37.3%), with long interspersed nuclear elements (LINEs; 3.3%) and short interspersed nuclear elements (SINEs; 0.3%) having lower relative abundance. Approximately 18.2% of all TEs identified were of unknown classification (Supplementary Table S2).

A combined method of *de novo* prediction, RNAseq (of leaf, stem, and root tissue), and homology with related species' genomes was used to predict gene numbers in *E. sativa* (Figure 2a). A total of 45,438 protein-coding genes were identified within the assembly, with an average length of 1,889.6 bp, and an average of 4.76 exons per gene. This genome size is smaller than that predicted for radish (*Raphanus sativus*), and larger than *Arabidopsis lyrata* (Supplementary Table S3), and is therefore consistent with what is known of Brassicales phylogeny (*21*). Using publically available databases, 98.3% of predicted genes were found to have homology with other plant species (Figure 2b, Supplementary Table S4).

RNAseq analysis of *E. sativa* plants

RNA sequencing and bioinformatics was conducted on 18 plants from three elite inbred lines designated **A**, **B**, and **C**; giving a total of 54 plant samples. Leaf material was collected at six separate time points, in triplicate. These corresponded to three commercially relevant harvest times ('early harvest' at 22 days after sowing, **EH**; 'harvest' at 30 days after sowing, **H**; 'second cut', **SC;** leaves harvested from the same **H** plants 43 days after sowing), and three consecutive postharvest time points (harvested at 30 days after sowing and designated: 'pre-wash', **PW**; 'day 0' of shelf life, 1 day post wash, **D0**; and 'day 7' of shelf life, **D7**). See Supplementary Figure S1 for a schematic of the experimental design.

After sample QC and clean-up over 2.6 billion clean paired-end reads were produced, averaging ~49 million reads per sample. Q20 (<1% error rate) averaged 96.3%, Q30 (<0.1% error rate) averaged 90.8%, and GC content ranged from 44.5% to 47.4%.

Global differential gene expression

Thousands of significant Differentially Expressed Genes (DEGs) were found between each ontogenic and shelf life time point. The total numbers of DEGs for line **A, B** and **C** are presented in Figure 2c, 2d and 2e, respectively. Relatively few significant DEGs were observed between **EH** and **H** samples for lines **A** and **B** (<333; Figure 2c and 2d), whereas they were observed at a higher rate in **C** (2,234; Figure 2e). This indicates a high level of plasticity of **C** across growth stages.

This trend was reversed at **PW**, where only 180 DEGs were observed compared to **H** in **C**, and 1,343 were observed in **A**. During shelf life (**D0** and **D7**) **C** expressed a greater number of DEGs compared to **H**, than **A** or **B** (2,340 at **D0** and 3,075 at **D7**, respectively). By contrast, DEGs at **SC** were much less variable between the three lines (330 – 676) indicating a greater degree of uniformity of expression. In terms of DEGs different from **H** across all sample points, **A, B** and **C** had 480, 308, and 270, respectively. These particular genes may provide candidate expression markers in future studies, as a means of quantifying the degree of transcriptional change observed between cultivars postharvest.

Figure 3 presents the numbers of DEGs at each sample point between cultivars, along with the degree of spatial variation and separation using Principal Component Analysis (PCA). The close clustering observed is indicative of the high degree of reproducibility of expression seen between each biological replicate for each line tested (Supplementary Figure S2). It also emphasizes the differences in global expression patterns at any given ontological growth stage or postharvest time point between lines **A, B** and **C**. The numbers of global DEGs therein make up a sizeable portion of the total number of genes within the genome. This ranged from 17,167 at **PW** (37.78% of total genes; Figure 3b) to 22,482 at **EH** (49.48% of total genes; Figure 3a).

PCA of each of the DEGs (Figure 3) was used to scrutinize genes responsible for the highest degree of spatial separation according to factor scores. This analysis yielded 1,568 genes. These in turn were filtered for common differential expressions at more than one of the sample points. Of particular note are several genes related to sulfur metabolism and GSL biosynthesis: *SLIM1a, MYB28b, GSTF11, IGMT4a, TGG1c, NIT2c, ESM1b, ESM1d, GTR2a, GTR2d, DHAR3, SiRb, SULTR2;2,* and *SULTR3;5a*.

Sulfur and phytochemical composition of *E. sativa*

Sulfur content of *E. sativa*

Total sulfur content for each of the breeding lines is presented in Figure 4a. No significant differences were observed between lines and sample time points ($P = 0.434$). As will be discussed in the subsequent sections and the context of GSL biosynthesis, this observation is highly important.

Glucosinolate profiles and contents of *E. sativa*

Nine GSL compounds were observed and quantified within rocket leaf tissues, which was consistent with previous observations (*8*) (Figure 4b). These were: GRA, glucoalyssin (GAL), DGTB, GSV, 4-hydroxyglucobrassicin (4OHB), DMB, glucoerucin (GER), 4-methoxyglucobrassicin (4MOB), and neoglucobrassicin (NGB). As previously reported (*20*), GSL concentrations were highly dynamic in ontogenic and postharvest shelf life samples. Numerous significant differences were observed between the three rocket lines and each respective time point (see Supplementary Data File S1 for ANOVA and Tukey's HSD significances).

There are two broad, but key, observations of note within these data. The first is the difference in GSL concentrations between the first (**H**) and second cuts (**SC**). For each of the cultivars an increase in total GSL concentrations was observed, primarily due to elevations of GSV (**A**, a 1.4-fold increase, $P<0.0001$; **B**, a 1.6-fold increase, $P<0.0001$; **C**, a 1.8-fold increase, $P<0.0001$) and GRA (**B**, a 2.6-fold increase, $P<0.0001$; **C**, a 1.8-fold increase, $P<0.0001$; Supplementary Data File S1). Anecdotal evidence has previously suggested that second cuts are more pungent in flavour and aroma, and these data show that this may theoretically be due to significant elevations in GSL concentrations between sequential harvests from the same plants, and thus greater hydrolysis to form pungent glucosinolate hydrolysis product (GHP) compounds. Line **C** produced the highest total concentrations of GSLs in **SC** (a 1.7-fold increase; $P<0.0001$), and line **B** also saw significant elevations compared to **H** (a 1.6-fold increase; $P<0.0001$).

The second key observation relates to the retention of GSL concentrations over the shelf life period. Line **A** consistently contained the greatest GSL concentrations compared to **B** and **C**, up until **D7** where content declined significantly (a 0.6-fold decrease compared to **D0**, $P<0.0001$; Supplementary Data File S1). **C** by comparison contained high concentrations of GSLs during shelf life, peaking at **D0**, with a non-significant decrease at **D7** (0.3-fold reduction). This line did not demonstrate the same decline in GSLs towards the end of shelf life as in the other two, and displays a propensity for maintaining GSLs for longer into the shelf life period. We hypothesised that this was due to fundamental differences in expression of GSL biosynthesis genes pre and postharvest.

Glucosinolate hydrolysis product profiles and contents of *E. sativa*

Glucosinolate hydrolysis product (GHP) concentrations are presented in Figure 4c (see Supplementary Data File S1 for ANOVA and Tukey's HSD significances). As with previous studies of rocket (*22*), three main GHPs were detected: sativin (a 1,3-thiazepane-2-thione; hydrolysis product of GSV; SAT), erucin (ITC of GER), and SF. The fluctuations in total GHP concentration largely mirror those observed for GSLs, however the increases between **H** and **SC** are much less pronounced, with no significant differences between cuts. Concentrations are more variable at each sample point compared to GSLs, which is partly due to their relative molecular instability and volatility.

As with GSLs, line **B** displayed the lowest concentrations of GHPs, whereas the differences between lines **A** and **C** are less apparent. The trend of reduction of GHPs over shelf life is also visible for lines **A** and **B**, though only significant in **B** (a 0.9-fold reduction, $P<0.0001$). Concentrations remained relatively higher in line **C** (1.2 $mg.g^{-1}$ dw, a 0.6-fold reduction from **D0**).

Monosaccharide profiles and contents of *E. sativa*

Monosaccharides are important in terms of sensory attributes and the masking of bitter and pungent sensory attributes in rocket (*5*) altering consumer perception and preference. Glucose is also known to influence stress responses and interact with MYB transcription factors (*23*) (Figure 1b). The concentrations of sugars observed in *E. sativa* lines are presented in Figure 4d (see Supplementary Data File S1 for ANOVA and Tukey's HSD significances).

Unlike previous reports (*20*) the changes in sugar concentrations in this study were highly dynamic across each of the respective time points. Both lines **A** and **B** contained relatively low total concentrations compared to line **C**. Line **B** displayed relatively consistent concentrations, with no significant differences observed. **A** showed a similar trend to GSL and GHP concentrations by declining at the end of shelf life (**D7**; a 0.5-fold decrease from **D0**, $P<0.0001$). Sugar concentrations displayed a trend of decline at **SC**, though not significantly.

Line **C** is distinct from the others in terms of its sugar profile and the relative differences between sample points. Concentrations increased postharvest (**D0** and **D7**; a 1.4 and 1.2-fold increase relative to **H**, respectively), perhaps owing to a breakdown of stored carbohydrate to facilitate respiration. Line **C** sugar content consists primarily of glucose, whereas **B** tended to have greater concentrations of galactose, and **A** was composed of similar amounts of each monosaccharide. This may have implications for sensory attributes, but also the ability for leaves to sustain metabolic processes postharvest.

Sulfur assimilation and glucosinolate biosynthesis pathway gene expression analysis

General

A total of 177 genes related to both sulfur and GSL metabolism pathways were selected from RNAseq data (NovoFinder; Novogene Co. Ltd., Hong Kong). Prior to Principal Component Analysis (PCA) analyses of sulfur metabolism and GSL-related genes, gene expression data were correlated with metabolite data using Pearson's correlation coefficient analysis (Supplementary Data File S1). All significances quoted hereafter were at the $P<0.001$ significance level. Genes significantly correlated with GSL and GHP concentrations are presented in Figure 5. For full expression data for all genes analysed see Supplementary Data File S2.

Sulfate assimilation gene expression

Figure 5a presents differential gene expression within the sulfate assimilation pathway of *E. sativa*. In the primary stages of sulfur metabolism, assimilated sulfate is activated via adenylation to adenosine-5'-phosphosulfate (APS), catalyzed by ATP sulfurylase (ATPS) (*24*). In *E. sativa* four ATPS-encoding genes were identified: *APS1a, APS1b, APS2,* and *APS3* (Figure 1a). Very few significant DEGs were observed between sample points for each respective rocket line (see Supplementary Data File S2 for full values and statistics of each sample comparison). However,

between **H**, **SC**, and **PW**, each respective line did show significant differential expression of ATPS genes.

In the second stage of the pathway, APS is reduced to sulfite by adenosine-5'-phosphosulfate reductase (APR) genes (*25*). Four APRs were identified (*APR1a, APR1b, APR2a,* and *APR2b*) as well as six ARP-like genes (*APRL4, APRL5a, APRL5b, APRL5c, APRL7a,* and *APRL7b*). *APR1a* and *APR2a* showed significant differential expression across multiple samples and time points (Figure 4b). Line **B** displayed low relative expression of these genes compared to **A** and **C**. Line **C** exhibits significantly higher expression postharvest compared to **H**; 2.2 log2-fold (**D0**) and 2.7 log2-fold (**D7**) increases of *ARP1a*, and 0.9 log2-fold (**D0**) and 1.1 log2-fold (**D7**) increases of *APR2a* were observed. We hypothesise that this may be indicative of a greater ability to remobilize sulfate via APS genes to facilitate and maintain secondary metabolite biosynthesis for longer into shelf life.

Two copies of genes encoding sulfite reductase (SIR; *SiRa* and *SiRb*) were identified. *SiRa* showed significantly higher levels of expression in line **C** (Figure 5b). Line **C** had no significant change in activity of this gene relative to time point **H**, however both lines **A** and **B** had significantly lower expression postharvest (Figure 5b).

Sulfur metabolism transcription, regulation, and transport gene expression

Six sulfur deficiency and limitation transcription factors were identified within the transcriptome of *E. sativa*: three copies of *SDI1* (*SDI1a, SDI1b,* and *SDI1c*) and three copies of *SLIM1* (*SULFUR LIMITATION 1*, aka *ETHYLENE INDUCED 3*; *SLIM1a, SLIM1b*, and *SLIM1c*). These genes are thought to play critical roles in the management and use-efficiency of sulfur in plants, and have been linked with optimization of GSL biosynthesis under S-limited conditions in *A. thaliana* (*13*).

SDI1a and *SDI1c* were differentially expressed between each line (Figure 5b, Supplementary Data File S2), with **C** having the highest levels of expression postharvest. It might be expected that each line would see a similar trend of expression over the course of shelf life, as additional sulfur is not obtainable; however only line **C** displayed this (Figure 5b). As shown in Figure 4a, sulfur content was not significantly different between any experimental stages or between breeding lines. As will be discussed in a subsequent section, expressions of MYB28 orthologs were not negatively associated with expression of *SDI1* gene copies. Previous research has shown that the SDI1 protein binds directly to MYB28, inactivating expression and reducing GSL biosynthesis (*26*). In *E. sativa* the opposite appears to be true, with significant positive correlations between respective expression of two MYB28 copies and *MYB29* with SDI1 copies (*SDI1a* and *MYB29*, $r = 0.72$; *SDI1b* and *MYB28a*, $r = 0.507$; *SDI1c* and *MYB28c*, $r = 0.459$; Figure 6). At **D7**, both **A** and **B** had significantly lower expression levels compared with **H** (a 2.2 and 3.7 log2-fold reduction of *SDI1a*, respectively; and a 3.9 and 3.2 log2-fold significant reduction of *SDI1c*, respectively).

A similar pattern was observed for *SLIM1b*, where expression in **C** was significantly higher at time points **D0** and **D7** relative to **H** (a 1.4 and 1.1 log2-fold significant increase, respectively). Expression of *SLIM1c* by comparison was not significantly different for each respective plant line between time points, but there were clear and significant differences in expression between lines (Supplementary Data File S2). Line **C** had highest expression of this gene, followed by **A**; with **B** having significantly lower expression (Figure 5b). Previous studies have shown that SLIM1 down regulates APK gene expression and GSL biosynthesis as a way of conserving sulfur for primary metabolism (*26*). Our data suggest that this is only the case between *SLIM1a* and *APK3* ($r = -0.597$, $P<0.001$; Figure 6). *SLIM1a* expression was positively (and significantly) correlated with *APK* expression ($r = 0.521$), and *SLIM1b* and *SLIM1c* with *APK4* ($r = 0.575$ and 0.698, respectively; Figure 6). This suggests *E. sativa* has a complex and interacting network of sulfur metabolism genes, where functions are not analogous to those found in *A. thaliana*.

Sulfur transport genes also showed significant differences of expression between each rocket line. 16 sulfur transport (*SULTR*) genes were identified; of note were *SULTR1;2a*, *SULTR2;1a*, *SULTR2;1b*, *SULTR4;1a*, and *SULTR4;2*. *SULTR1;2a* has previously been associated with the uptake of environmental sulfate in root tissues (Supplementary Data File S2), but low levels of expression were detected in leaf tissues used in this analysis. Postharvest, line **C** had differential expression of this gene compared to **A** and **B** in **D7** samples (Figure 4b). This was more pronounced for *SULTR2;1a* and *SULTR2;1b*, and both **A** and **B** had significant reductions in expression at **D0** and **D7** relative to **H**. **SC** samples showed significant increases relative to **H**, with the exception of *SULTR2;1a* in **B**.

The relatively less well understood *SULTR4;1* and *SULTR4;2* genes also had distinct patterns of expression between lines. *SULTR4;1a* saw significant increases in expression in postharvest samples relative to **H** (Figure 5b). Line **A** had higher expression of *SULTR4;2* during growth before declining rapidly and significantly post-wash (**D0**). The opposite trend was seen in **C**, where gene expression peaked at **D7**. These data are suggestive of more active intra-leaf sulfur transport in line **C** postharvest, and may be associated with the higher expression of *APR*, *SIR*, *SDI1* and *SLIM1* genes to facilitate more efficient S utilisation during this period.

Glutathione synthesis

Three genes encoding glutamate-cysteine ligase (GSH1) and three encoding glutathione synthetase (GSH2) enzymes were identified in *E. sativa*. These genes are essential for the synthesis of glutathione and antioxidant response. Informative patterns were observed in the expression of four genes: *GSH1b*, *GSH1c*, *GSH2a*, and *GSH2b*.

With the exception of *GSH2b*, each was most highly expressed in rocket line **B**, with significant increases observed postharvest (Figure 5b). Line **A** and **C** were relatively unchanged between sample points for these genes, but had a marked difference in expression for *GSH2b* relative to each other. **B** had negligible levels of *GSH2b* expression.

As both glutathione and secondary S-containing metabolites, such as GSLs, have been associated with antioxidant responses (*26*) the differences observed between each of the lines in terms of both GSL concentrations and glutathione-related gene expression, may be indicative of different adaptive metabolic strategies for dealing with oxidative stress postharvest. Lines **A** and **C** favor secondary sulfur metabolism and the synthesis of GSLs, and **B** favors primary sulfur metabolism and glutathione synthesis.

Glucosinolate-related transcription factors

MYC and MYB TFs are master regulators of GSL biosynthesis, and there is much interest surrounding their activity and expression under varying environmental circumstances. Three copies of *MYC2* were found within the *E. sativa* genome, and 12 GSL regulating MYB TFs. Salad rocket has three copies of *MYB28*, and one copy of *MYB29*, regulating aliphatic GSL biosynthesis. Intriguingly, no homologous copies of *MYB76* were found within the annotation, perhaps indicating a loss of function. Although rocket typically contains relatively low concentrations of indolic GSLs compared to aliphatic, there are three *MYB34*, three *MYB51*, and two *MYB122* encoding genes present in our annotation.

MYC2a and *MYC2c* were most highly expressed in line **A**, and had highly uniform patterns of relative expression. **SC** had the highest expression values for this line, suggesting a general response to mechanical wounding and stress, however this was not significantly different from **H**. The only significant difference for *MYC2c* between **H** and **SC** was in line **B** (a 0.7 log2-fold increase; Figure 5c).

MYB28a and *MYB28b* display high degrees of differential expression between each rocket line. While **A** has high expression of *MYB28a* in samples **EH, H, SC**, and **PW**, it is has by comparison lower expression of *MYB28b* compared to **C** (Figure 5c). **C** on the other hand has

relatively high expression for both of these TFs, and displays significantly higher expression postharvest, up to and including **D7**. Combined with what is known about theses TFs in other Brassicaceae species, it is highly likely that the differences in GSL concentrations observed postharvest are closely linked to the differential expression of *MYB28a* and *MYB28b* between the respective lines.

Glucosinolate biosynthesis

As shown in Figure 1b, the aliphatic GSL pathway begins with BCAT4, and deaminates methionine to 4-methylthio-2-oxobutyrate. Rocket contains two genes encoding BCAT4, and two genes encoding BCAT3; converting 2-oxo acids to homomethionine and dihomomethionine. *BCAT3-1a* displayed no significant variation between lines during growth, but saw significant increases for all (compared to **H**) at **D0** and **D7** (Supplementary Data File S2). The most marked and significant increase was in **B**. It is unclear exactly how this 'preference' for BCAT3 activity over BCAT4 is regulated or affects the synthesis pathway, but the relative and respective activity of these genes is highly correlated with GSL content.

Chain elongation of the GSL molecule is determined by MAM genes. In rocket, only orthologs of *MAM1* were identified, and each of the three copies had differing expression patterns (Figure 5d). Line **A** displayed higher relative expression of *MAM1a*, whereas **C** had greater levels for *MAM1b* and *MAM1c*. **B** however maintained relatively low expression for all three of these genes. **A** had reduced expression activity during shelf life, whereas in **C**, levels were significantly higher compared to **H** (Figure 5d).

Only one *CYP79F1* homolog was found in rocket, with no expression found for a corresponding *CYP79F2* gene. The encoded enzymes are responsible for the catalytic conversion of chain-elongated methionine to their respective aldoximes. The lack of a *CYP79F2* homolog in rocket may be suggestive of a loss of function, and/or redundancy with other enzymes. Of note for *CYP79F1* expression was the significant differences observed between **EH** and **H**; indicating that earlier harvests of rocket leaves may have a reduced ability for GSL biosynthesis compared with later ones and second cuts (**SC**). Expression was significantly greater in **C** during shelf life. Similarly, in the conversion of aldoximes to nitrile oxides, *CYP83A1* expression was uniformly higher in **A** and **C** than **B**, with line **C** having relatively and significantly higher expression in shelf life samples (Figure 5d).

Further down the aliphatic GSL pathway (Figure 1b), nine SOT genes (that encode sulfotransferase enzymes) were identified: one for *SOT16* and *SOT17*, respectively, and seven *SOT18* homologs. It is unclear why rocket has so many copies of these sulfotransferase enzymes, however each showed independent expression patterns between rocket lines (e.g. *SOT17a* and *SOT18f*; Figure 5d, Supplementary Data File S2). There is a high degree of dynamic change at each sample point for all three lines in *SOT17a*, whereas for *SOT18f*, expression is low but consistent (with the exception of line **C** where there is a significant reduction postharvest).

Glucosinolate hydrolysis

Myrosinases (*TGG1* and *TGG2*) and nitrilases (*NIT2*) are involved in GSL breakdown and recycling. It is becoming increasingly apparent that GSL synthesis and storage is part of a dynamic system that is in a state of constant flux *in planta*, and it stands to reason that genes controlling the synthesis of enzymes within this system are in turn dynamic in their expression. This is well illustrated within the data (Figure 5e) and expression differs according to cultivar and leaf age, as genes encoding TGG1 and TGG2 enzymes vary significantly.

Numerous TGG genes were identified in rocket: 11 *TGG1* orthologs, and three *TGG2* (Supplementary Data File S2). Some of these genes appear to have differential expression according to ontogeny and shelf life, with some copies expressed at **EH** with virtually none during postharvest (e.g. *TGG1h*, *TGG2a*, and *TGG2c*; Figure 5e). Others however display the inverse of this, with

increased relative expression postharvest (*TGG1a*; Supplementary Data File S2). It is unclear how these differences of expression are regulated, or what the purpose/advantage of having different active myrosinase gene copies may be at any given time in the life cycle and/or tissue type. It is known that myrosinases are functionally redundant, however it has also been noted that their activity and specificity is linked with developmental processes, and may explain some of the high levels of expression observed at **EH**.

Myrosinase action upon GSLs primarily yields ITCs as hydrolysis products under pH-neutral conditions. The action of co-enzymes (such as nitrile specifier proteins; NSPs) can further modify these products to form nitriles. In the case of rocket, only one NSP was identified (*NSP5*). Expression values for this gene were orders of magnitude greater than myrosinases, which would suggest a high amount of nitrile hydrolysis products would be formed. This is however not the case, as the main hydrolysis products observed were ITCs (the ITC of sativin is unstable and undergoes spontaneous tautomeric rearrangement post-hydrolysis), as is evidenced in Figure 4c. No nitriles were detected in these samples above trace levels, indicating that *NSP5* expression is not a proxy for determining the kinds or amounts of hydrolysis products that may be produced in *E. sativa*. It may also be suggestive of other functions of this gene, such as in indole-3-acetic acid (IAA) catabolism. Expression is significantly greater during the shelf life period, suggesting involvement in protein degradation and senescence.

Another explanation may be that the high expression of *NSP5* is inhibited by the five *EPITHIOSPECIFIER MODIFIER* 1 (*ESM1*) orthologs found in the rocket genome. These proteins inhibit the action of NSPs and promote ITC formation. Expression was significantly greater in line **C** for *ESM1b* (Figure 5e) at all sample points, and fits with the observed pattern of sustained GHP formation postharvest. The lower activities in **A** and **B** did not however correspond to a reciprocal decrease in the relative concentrations of GHPs, and neither were nitrile concentrations at anything above trace levels. This is suggestive of other mechanisms of regulation. Much further work is needed to satisfactorily explain the genetic regulation of GHP formation in rocket and the unexpectedly high expression of *NSP5*.

Glucosinolate transporters

Eight GSL transporter genes were identified in the rocket annotation; four *GTR1* and four *GTR2* homologs. These genes are involved in leaf distribution and long-range phloem GSL distribution, respectively. Expression of *GTR2a* and *GTR2d* were significantly correlated with GSL abundance and GHP formation in the analysed leaf tissues. Of particular note is that **B** had no expression of *GTR2d* at any of the sample points, indicating that the gene may be non-functional, and contribute to the overall low concentrations of GSLs observed. The exact site of GSL synthesis has not been established in *E. sativa*, however it is known that transport from the root tissues to leaves occurs in other Brassicaceae. If this transport system is impaired in **B**, this would partially explain the significantly lower abundance of GSLs observed in leaves (Figure 4b). Coupled with the high expression of glutathione-related genes and similar sulfur content of **B** compared to lines **A** and **C**, the inactivity of this gene copy may have significant effects on leaf sulfur transport, metabolism, and antioxidant response. The lower GSL content in leaves may therefore be compensated by increased glutathione synthesis.

Principal component analysis of sulfur and glucosinolate metabolism genes

General

PCA was conducted with regression of supplementary data for GSL, hydrolysis products, and monosaccharides (Figure 4). As no significant differences were observed for respective sulfur content in each sample, these data were not included in the analysis. The Kaiser-Meyer-Olkin test value for the data was 0.721, indicating an acceptable level of sampling. PCA with Varimax rotation produced 51 principal components (PCs), with PCs 1 (19.72%), 2 (13.23%), and 3 (12.61%)

containing >10% explanatory variation (45.56% cumulative). As such, these were selected for biplot presentation (Figure 6). Only correlations significant at the $P<0.001$ level are presented and discussed.

Aliphatic glucosinolates and hydrolysis products

SULTR4;1a was significantly correlated with GRA concentrations (r = 0.577), which is closely associated with shelf life samples for lines **A** and **C** (Figure 6b, cluster **II**). Figure 6a and 6b show a distinct separation between ontogenic and shelf life samples along PC1. The increased expression of sulfur transport genes such as this postharvest may provide some explanation as to why GSL concentrations increase in the initial stages shelf life (**PW**), as S may be re-mobilized to facilitate biosynthesis. Efficient transport and storage of sulfur pre-harvest may also facilitate better retention and decreased degradation of GSLs postharvest. This can be seen in Figure 6b (**V**) where *SULTR3;1a* and *SULTR3;2* are highly associated with pre-harvest expression.

SF and SAT concentrations were significantly correlated with *APR2a* gene expression (Figure 6c **I** and **II**; r = 0.58, SF; r = 0.464, SAT) and associated in particular with **A** ontogenic samples and **PW**. *APR2* is known to contribute to sulfur accumulation and homeostasis, as well as facilitating cysteine synthesis, and is associated with increased myrosinase activity and GSL recycling. Line **A** (on average) contained the highest ontogenic concentrations of GRA, SF, GSV, and SAT (Figure 4b and 4c); this is supported by a significant correlation and association with GSL-related transcription factors *MYB28a* (r = 0.486, SF; r = 0.53, SAT), *MYC2a* (r = 0.596, SF; r = 0.626, SAT) and *MYC2c* (r = 0.584, SF; r = 0.583, GSV; r = 0.634, SAT; Figure 6c **I** and **II**), as well as a drought tolerance-related gene *SAL1b* (r = 0.595, GRA; r = 0.547, SF; r = 0.499, SAT; Figure 6a **II, III** and **IV**, 6c **II**). **A** was also associated with increased activity of *MAM1a* (Figure 6c **II**), potentially facilitating greater GRA biosynthesis through chain elongation. It may be that lines **A** and **C** have increased relative GRA concentrations at **EH** and **H**, but preferentially express *MYB28c* and *MYB28b*, respectively. It is unknown if the function of each MYB28 TF are redundant in rocket, but these data would suggest that there is some clear overlap of function, though the expression of *MYB28b* is associated with increased GSL biosynthesis postharvest (Figure 5c).

The lower relative expression in line **B** for many of these genes is consistent with its lower GSL and hydrolysis product concentrations, irrespective of sample point. GRA/SF, GSV/SAT, and GER concentrations were significantly and negatively correlated with *SPERMIDINE SYNTHASE 1c* (*SPDS1c*; r = -0.622, GRA; r = -0.614, SF; r = -0.6, GSV; r = -0.454, SAT; r = -0.604, GER), which had a high degree of co-separation in all **B** samples (Figure 6c **III**). This association may be related to increased primary S metabolism and reduced partitioning of methionine for secondary S metabolites (Figure 1a).

Indolic glucosinolates

Line **B** contained the highest concentrations of indolic GSLs; predominantly 4MOB (Figure 4b). While apparently only a negligible difference between the plants tested, the underlying gene activities correlated and associated with this trait are highly significant and have implications for the entire GSL profile. Indolic GSLs detected in rocket are negatively correlated with expression of numerous genes related to S metabolism, such as: *SULTR4;1b* (r = -0.577, 4MOB; Figure 6a **I**, 6c **IV**), *APR2a* (r = -0.535, 4OHB; Figure 6c **I** and **II**), *SAL1b* (r = -0.512, 4OHB; Figure 6a **II**, **III** and **IV**), *APRL5c* (r = -0.487, 4OHB; r = -0.655, 4MOB; Figure 6a **I**), *APRL7b* (r = -0.581, 4MOB), *SiRa* (r = -0.607, 4MOB; Figure 6a **I**, 6c **IV**), *GPX7a* (r = -0.57, 4MOB; Figure 6a **I**, 6c **IV**), *GPX8* (r = -0.659, 4MOB; Figure 6a **I**, 6c **IV**), and *GRa* (r = -0.587, 4OHB; r = -0.597, 4MOB; Figure 6a **I** and **II**, 6c **IV**). In addition, the TFs *SLIM1c* (r = -0.595, 4OHB; r = -0.509, 4MOB), *MYB28b* (r = -0.49, 4MOB; Figure 6a **I**, 6c **IV**), and *MYC2a* (r = -0.47, 4OHB; Figure 6b **I** and **IV**, 6c **I** and **II**) were also inversely correlated to indolic GSL content. Several of these genes have a diametric opposition to that seen in aliphatic GSL biosynthesis, indicating that differences in

relative expression strongly influence the ratio of aliphatic:indolic GSLs. Under reduced expression, these genes therefore infer an increase in indolic GSL biosynthesis, and a large reduction in aliphatic GSLs. Line **B** has relatively low expression of each of these genes (and **A** and **C** higher); as well as high expression of *BCAT3-1b*, *GTR1d*, *GSH1b*, *SPDS1c*, and myrosinases *TGG1g* and *TGG1i*. In combination, this suggests that with a reduced influence of *MYC2a* and *MYB28b*, indolic GSL synthesis and aliphatic GSL breakdown, which reduces overall concentrations within leaf tissue.

Free glucose concentrations

Whilst free monosaccharides are important components for determining taste and consumer acceptability of rocket, they are critical metabolites in their own right. The process of GSL biosynthesis is reliant upon glucose and its availability, and is known to have a promoting effect upon *MYB28* transcription (Figure 1b). PCA and correlation analysis (Figure 6) revealed a large number of highly significant associations between free glucose concentrations, sulfur metabolism, and GSL biosynthesis genes.

Line **C** contained significantly greater concentrations of glucose at each respective sample point. This is reflected in the expression of numerous genes associated with sulfur and GSL metabolism, particularly in the ontogenic sample points (**EH, H**, and **SC**). This potentially feeds into the increased abundances postharvest and during shelf life (**PW, D0,** and **D7**). These include significant correlations with: *SULTR4;1b* (r = 0.702; Figure 6a **I**, 6c **IV**), *APK4* (r = 0.631; Figure 6a **I** and **II,** 6c **IV**), *SLIM1c* (r = 0.545; Figure 6a **I** and **II,** 6c **IV** and **V**), *MYB28b* (r = 0.591; Figure 6a **I**, 6c **IV**), *MYB51* (r = 0.691; Figure 6a **I**, 6c **IV**), *UGT74B1a* (r = 0.624; Figure 6a **I** and **II**, 6c **IV**), *APRL5c* (r = 0.551; Figure 6a **I** and **II**, 6c **IV**), *APRL7b* (r = 0.625), *SiRa* (r = 0.607; Figure 6a **I**, 6c **IV**), *GPX8* (r = 0.455; Figure 6a **I**, 6c **IV**), *GRa* (r = 0.629; Figure 6a **I** and **II**, 6c **IV**), *GSTF10a* (r = 0.649; Figure 6a **I** and **II**, 6c **IV**), and *CYP83B1* (r = 0.59; Figure 6a **II**, 6b **II**). Clearly line **C** has a distinct transcriptomic profile under the tested conditions, which may be linked with the abundance of free glucose. This raises numerous questions about the genetic diversity, metabolism, and performance of rocket salad cultivars during shelf life, and if monosaccharide concentrations are linked with improved quality.

Discussion

Study limitations

As plants used in this study were grown under controlled environment conditions, the phytochemical and gene expression patterns may be different to field-grown plants and leaf material that is processed through commercial supply chains. The lines used were also inbred and genetically uniform, and so heterozygosity present within commercial cultivars of *E. sativa* is likely to produce much more variable responses in 'real-world' produce. That being said, these data provide a benchmark for future expression and phytochemical studies of rocket. The sequence and annotation data provided will also facilitate future genomic/transcriptomic studies of the species.

The complexity of the *E. sativa* genome and future novel gene discovery

The presented reference genome and annotation for *E. sativa* shows a huge amount of complexity. The transcriptomic evidence presented here also illustrates just how variable traits and expression can be between breeding lines under controlled environmental conditions. In the three inbred lines tested, global differential expression of genes was markedly different. This is exemplified in line **C** samples, which displayed unique patterns of expression, both during growth stages and postharvest (Supplementary Data File S2). Some studies within the literature assume cultivars of *E. sativa* are similar in terms of phytochemistry and postharvest quality (*27*); this study has definitively shown that this is not the case.

Clearly one drawback of this study is that annotations and pathway elucidations are almost entirely reliant upon comparison with *Arabidopsis,* and assumptions that gene functions are orthologous between the two genera. Development of these data in *Eruca* is however, a major step forward for a crop once considered to be niche, and now joins a growing list of minor crops in the genomic era. While this study has highlighted the orthologous genes that are likely to be involved with sulfur assimilation, GSL accumulation, and postharvest stress response, much further research will be needed to unpick novel gene functions.

E. sativa has a distinct and complex glucosinolate pathway

Evolutionary divergence between *E. sativa* and other Brassicales species has led to a unique GSL synthesis pathway, with many genes being duplicated. Aside from the duplications of *MYB28* found in salad rocket, one TF prominent in GSL biosynthesis has no homologous sequence or expression in the tested rocket lines: *MYB76*. Similarly, other genes, such as *MAM2*, *MAM3*, *CYP79F2*, *CYP71A13*, *GSTF9*, *UGT74C1*, and *NIT3*, are all absent from the reported rocket genome annotation. This of course may be amended with future annotation iterations and improvements, however it is conceivable that some of these genes may have been lost or made redundant over the course of evolutionary time and divergence with *A. thaliana*.

Conversely, some gene homologs have undergone numerous duplications in *Eruca*, and it is not presently clear what the function(s) of these numerous copies may be. It may be the result of segmental duplications within the genome, such as those observed in the *Brassica* A genome (*28*), and future, more detailed studies of the *Eruca* genome structure may reveal the nature and number of any such events. Genes such as SOTs, FMO $_{\text{GS-OX}}$s, myrosinases (TGGs), *ESM1*s, and GSL transporters all have several copies, and it has yet to be determined if these perform the same function as in *Arabidopsis,* or have evolved new ones that are responsible for the unique GSL profile of rocket.

Examples of this include increases in copy numbers compared to *B. rapa* and *B. oleracea*. In these two species only two copies of *SOT18* have been identified (*29*), whereas *E. sativa* has seven. *B. rapa* has two copies of FMO $_{\text{GS-OX}}$ genes, and salad rocket has at least ten. The closely related *Diplotaxis tenuifolia* ("wild" rocket) transcriptome has been reported to contain three copies of *MYB28* (*30*), and is consistent with the hypothesis that duplication occurred after their divergence with a common ancestor within the *B. oleracea* lineage.

Finally, both *Arabidopsis* and *B. rapa* have four myrosinase gene copies, and *B. oleracea* has six (*29*). Our data indicate that *Eruca* has at least 14 copies, as well as two copies encoding PEN2 myrosinases. Clearly there has been a massive diversification and duplication of these enzymes in rocket, but it has yet to be established if this is reflected in functionality and spatial expression.

One example of recent novel gene discovery outside of *Arabidopsis* and *Brassica* species GSL synthesis pathways is *GLUCORAPHASATIN SYNTHASE 1* (*GRS1*) in radish; which is thought to have evolved from a mutation in a 2-oxoglutarate-dependent dioxygenase (2OGD) (*31*). It is conceivable that similar mutations have occurred in *E. sativa* that have led to the evolution of GSV, GRL, DGTB, and DMB. Future work will elucidate the causative genes responsible for the synthesis of these GSLs, and the development of the genomic and transcriptomic resource of this study is an important first step in achieving this.

Specific genes are highly correlated with glucosinolate biosynthesis

PCA analyses highlighted several genes that are significantly correlated with the synthesis of GSLs in *E. sativa*. In terms of sulfur assimilation, the expression of genes *SULTR4;1b* and *APR2a* appear to be strongly associated with both aliphatic and indolic GSL biosynthesis. *SULTR4;1* facilitates transport of sulfate from cell vacuoles into the cytoplasm, and has been previously linked with the activity of *MYB28* and *MYB29* (*32*), which is supported by this study. It may be that

expression expedites the synthesis of GSLs by facilitating greater availability of sulfate. This is in turn linked with the activity of *APR2*, which is known to be responsible for regulating sulfur homeostasis (*33*). This gene has been associated with increased GSL recycling and myrosinase activity (*34*), and significant correlations with *TGG1f* (r = 0.638, *APR2a;* r = 0.798, *APR2b*) and *TGG2c* (r = 0.703, *APR2a*; r = 0.793, *APR2b*) were observed within our data. It seems likely that the transport of sulfate and its management in terms of recycling is pivotal for GSL abundance and flux at any given time. This also likely has a significant role in determining the duration of postharvest biosynthesis and retention of GSLs.

The natural strategy of the leaf is to remobilize sulfur around parts of the plant as required, such as in times of deficiency. The transcriptome of severed leaves in the postharvest context is an evolutionary dead end, and not subject to natural selection. As such, the differences we have observed could be attributed to different strategies for dealing with unexpected physiological stress, or as part of senescence response.

Glutathione metabolism competes with glucosinolate biosynthesis for sulfur

As shown in Figure 4a, the content of sulfur between the three tested breeding lines was not significantly different. In light of the observed differences in gene expression and GSL accumulations, we theorize that primary and secondary sulfur metabolism pathways 'compete' for assimilated environmental sulfur. As content was not significantly different postharvest (**PW, D0,** and **D7**) compared to pre-harvest first cut (**H**) in any of the breeding lines, this suggests that the degree of remobilization and ability to synthesize/recycle GSLs is under strict genetic control. The exact reasons why each line differs in this respect are unclear, but as shown in Figure 4b, the amount of total sulfur assimilated during growth is not reflected in the postharvest concentrations of GSLs. Line **B** displays this most clearly: it contains significantly no more or less sulfur than lines **A** or **C**, yet synthesizes far less GSLs. We hypothesise that the lack of expression of *GTR2d* impairs GSL transport from major sites of synthesis (possibly in the roots) thereby forcing leaf tissues to cope with oxidative stress via glutathione synthesis and preferential shunting of sulfur into the primary metabolism pathway. This is evidenced by the significantly higher expression of glutathione synthetase (*GSH2a*) and glutamate-cysteine ligase genes (*GSH1a* and *GSH1c*). Lines **A** and **C** by contrast cope with oxidative stress postharvest by sustaining GSL biosynthesis. As **C** contained greater concentrations of monosaccharides, we theorize that this facilitates GSL metabolism for longer, reducing oxidative stress, and delaying the onset of senescence.

Glutathione-related genes were significantly and negatively correlated with the accumulation of GSLs. This is significant for future breeding efforts, as it may dictate not only total GSL content, but also the ability of a cultivar to maintain synthesis and storage of GSLs postharvest. Increased internal oxidative stress responses are also likely to induce senescence as a result, and lead to a loss of product quality. Therefore, this could have a double-negative impact on both visual and health-related traits. The differential expression observed in these genes may be indicative of a key discriminating factor in primary sulfur metabolism between rocket breeding lines, and the ability to cope with oxidative stress postharvest.

Postharvest maintenance of glucosinolate content depends on more than MYC2 and MYB28 expression

The relative expression of *MYC2* and *MYB28* genes strongly influences the overall concentrations of aliphatic GSLs within the plant during growth. The promotion of GSL biosynthesis in Brassicaceae by these is widely known in the literature (*35*) and it seems that rocket is no exception in this regard. Under the dehydrating conditions of shelf life storage, the activity of genes such as *SAL1b* also likely improved the retention of GSLs by increasing transcription of SOTs (e.g. *SOT16* and *SOT17a*) and promoting biosynthesis through the degradation of PAPS (3'-phosphoadenosine-5'-phosphosulfate) as a pseudo-form of drought stress (*36*).

These data indicate that homologous copies of *MYB28* genes may be under independent regulatory control in rocket leaves, and similarly have independent regulatory functions themselves. While there is likely to be overlap and redundancy of function between the three copies, their respective levels of expression may subtly control GSL concentrations and shape profile compositions. With reduced *MYB28* expression, our data suggest that indolic GSLs are synthesized in greater amounts, though still only make up a minor proportion of the whole profile.

Second cut rocket has greater uniformity of gene expression and glucosinolate content than first cut

It is common growing practice for multiple cuts to be taken from rocket sowings. Anecdotal evidence has suggested that the second cut (**SC**) is more consistent in terms of taste and flavour than the first (**H**). For the first time we present transcriptomic and phytochemical evidence to support this assertion. All cultivars saw increases in GSL abundance in the second cut, and the shift is best exemplified by line **B**. First cut (**H**) samples were low in expression of numerous GSL biosynthesis related genes that later saw significant increases at **SC**. This had the result of making its expression profile more similar to that of **A** and **C**, along with comparable concentrations of GSLs.

Whilst hydrolysis products did not see as great an increase in abundance as GSLs, it should not be forgotten that GHP abundance alone is not responsible for flavour and pungency of rocket. When the respective monosaccharide profiles are taken into account, it can be seen that there is a general decrease in sugars at **SC**. This then leads to a shift in the ratio between monosaccharides and GHPs, potentially making the **SC** samples more pungent. Although flavour was not tested in this study, our previous work has consistently highlighted the importance of GHP:sugar ratios in determining pungency (*37*). The identification of five *ESM1* genes in rocket also likely influences the relative abundances of GHPs produced from GSLs, and could have a significant role in determining taste and flavour of a crop.

Materials and Methods

Plant material for genome sequencing

Three elite inbred lines of salad rocket were produced through self-pollination for five generations at Elsoms Seeds Ltd. (Spalding, UK) from 2010-2016, giving an estimated inbreeding coefficient of 0.969. Each line was derived from germplasm accessions obtained from the Leibniz-Institut für Pflanzengenetik und Kulturpflanzenforschung (IPK Gatersleben, Germany). For reasons of commercial sensitivity these lines (**A**, **B**, and **C**) and their lineage will not be identified.

For genome sequencing, plants of each line were grown under controlled growth room conditions, and leaf samples had DNA extracted and sent to the Earlham Institute for QC analysis. Samples were quantified using a Qubit fluorometer and dsDNA assay kit (ThermoFisher Scientific, Loughborough, UK) and assessed for quality using NanoDrop (ThermoFisher Scientific; according to 260/280 and 260/230 ratios).

Genome sequence library preparation and assembly

DNA sequencing and assembly was performed as a service by the Earlham Institute (Norwich, UK). *De novo* genome sequencing and assembly was performed using PCR free paired-end (PE) and LMP sequencing. After DNA sample QC, line **C** was selected for sequencing and reference genome assembly. One PCR free PE library was constructed from gDNA, and sequenced on one lane of an Illumina HiSeq2500 in rapid run-mode (v2) using 250 bp PE reads. LMP sequencing was also conducted using one set of Nextera libraries (Illumina) from gDNA, and sequenced on one lane of an Illumina MiSeq with 250 bp PE reads. After data QC and assembly of the high coverage PE library, LMP libraries were mapped to determine their suitability for assembly improvement. Three additional libraries were selected and re-sequenced to a higher depth of

coverage on a single lane of an Illumina HiSeq2500 in rapid run-mode, to again yield 250 bp PE reads.

Genome sequencing bioinformatics

FASTQ files were converted to BAM format using PicardTools (v1.84, http://broadinstitute.github.io/picard/; FastqToSam option) and then assembled using DISCOVAR *de novo* sequence assembler (build revision 52488). All LMP libraries were processed using NextClip to analyse and create a high quality read subset for scaffolding the DISCOVAR-assembled sequences. SOAP and SSPACE were used to scaffold the DISCOVAR assembly using data from three of the NextClip-processed LMP read libraries.

Genome annotation

Annotation was performed by Novogene Co. Ltd. (Hong Kong). A homology and *de novo*-based approach was taken in order to identify TEs. The homology-based approach used known repetitive sequence databases: RepBase, RepeatProteinMask, and RepeatMasker (http://www.repeatmasker.org/). *De novo* repeat libraries were created using LTR_FINDER, RepeatScout (http://www.repeatmasker.org/), and RepeatModeler (http://www.repeatmasker.org/RepeatModeler.html).

An integrated approach was taken to compute consensus gene structures, such as cDNA, proteins in related species, and *de novo* predictions. The homology-based approached used the related genomes of *A. lyrata, A. thaliana, B. napus, Boechera stricta, Capsella rubella,* and *R. sativus* (Supplementary Table S3) to compare against *E. sativa* to find homologous sequences, and predict gene structures (using BLAST and genewise). *Ab initio* statistical models were also used to predict genes and their intron-exon structures; e.g. Augustus, GlimmerHMM, and SNAP (http://homepage.mac.com/iankorf/). EVidenceModeler (EVM) software was then used to combine *ab initio* predictions, protein and transcript alignments, and RNAseq data into weighted consensus gene structures. Lastly, PASA was used to update the consensus predictions by adding UTR annotations and models for alternative splicing isoforms. All predicted proteins were functionally annotated using alignments to SwissProt, TrEMBL, KEGG, and InterPro (Figure 2b).

The full reference genome sequence and annotation can be found in the European Nucleotide Archive (Assembly accession no: GCA_902460325; Study ID: PRJEB34051; Sample ID: ERS3673677; Annotation accession number ERZ1066251).

Plant material growth and collection for RNA, elemental, and phytochemical analyses

For RNAseq analyses seeds were sown in a random order in seedling compost, and raised under controlled environment conditions in plastic trays inside a Weiss-Technik Fitotron cabinet (Weiss-Technik UK Ltd., Loughborough, UK). Daytime temperature was set to 20 °C, and nighttime temperatures to 14 °C (long day cycle; 16 h light, 8 h dark). Light intensity was set at 200 μmol m^{-2} s^{-1}. During a one-hour period of 'dawn' and 'dusk', light and temperature changes were ramped on a gradient. Humidity was ambient. After ten days of growth, seedlings were transplanted to one-litre pots in standard peat-based compost.

Postharvest (post sample **H**), leaves were stored for two days in a cold store (4 °C) (*20*). Samples for **D0** and **D7** were washed individually in mildly chlorinated water (sodium hypochlorite; 30 ppm) for two minutes, then rinsed for one minute with distilled water (all at >14 °C to avoid cold-shock). Leaves were dried of excess moisture for one minute using a kitchen salad spinner, then placed in fresh bags, sealed, and stored overnight at 4 °C. Shelf life leaves were stored in the cold and dark (4 °C) for seven days (**D7**) – typical of the use-by date of commercially bagged leaves.

All samples were taken between the hours of 1 – 3 pm to mitigate diurnal fluctuations in phytochemical content and gene expression (*38*). Immediately after each of the aforementioned

samples was taken, leaves were frozen using liquid nitrogen and ground into a fine powder using a pestle and mortar. Samples were stored at -80 °C in tubes and lyophilized prior to chemical analysis. A subset of non-lyophilized sample was kept aside for RNA extractions.

RNA extraction and quality control

RNA for RNAseq and qRT-PCR analyses was extracted using RNeasy Plant Mini Kit (Qiagen, Manchester, UK) according to the manufacturer 'Plants and Fungi' procedure. As part of the protocol, an on-column DNase digestion was incorporated according to the RNase-Free DNase Set (Qiagen) protocol. Samples were checked for degradation and contamination prior to sequencing using agarose gel electrophoresis (1%, TAE buffer), Qubit, and NanoPhotometer (Implen, CA, USA) methods. Briefly, ≥2 µg of total RNA was obtained for each sample at a minimum concentration of ≥50 $ng.\mu L^{-1}$. RNA integrity was determined and evaluated using an Agilent 2100 Bioanalyzer.

RNAseq library preparation and sequencing

After QC procedures, sequencing libraries were prepared using NEBNext Ultra RNA Library Prep Kit for Illumina (NEB, MA, USA) following the manufacturer's instructions, and index codes were added to attribute sequences to each sample. mRNA was purified from total RNA by using poly-T oligo-attached magnetic beads. Fragmentation was carried out using divalent cations under elevated temperature in NEBNext First Strand Synthesis Reaction Buffer (5x). First strand cDNA was synthesized using random hexamer primer and M-MuLV Reverse Transcriptase (RNase H-). Second strand cDNA synthesis was subsequently performed using DNA Polymerase I and RNase H. Remaining overhangs were converted into blunt ends via exonuclease/polymerase activities. After adenylation of 3' ends of DNA fragments, NEBNext Adaptor with hairpin loop structure were ligated to prepare for hybridization. In order to select cDNA fragments of 150 – 200 bp in length preferentially, the library fragments were purified with an AMPure XP system (Beckman Coulter, MA, USA). 3 µl USER Enzyme (NEB) was used with size-selected, adaptor-ligated cDNA at 37 °C for 15 min, followed by 5 min at 95 °C before PCR. PCR was performed with Phusion High-Fidelity DNA polymerase, Universal PCR primers and Index (X) Primer. Finally, PCR products were purified (AMPure XP system) and library quality was assessed on using an Agilent 2100 Bioanalyzer system.

The clustering of the index-coded samples was performed on a cBot Cluster Generation System using HiSeq PE Cluster Kit cBot-HS (Illumina) according to the manufacturer's instructions. After cluster generation, the library preparations were sequenced on an Illumina Hiseq platform and 125 bp/150 bp paired-end reads were generated.

RNAseq bioinformatics

Raw data (raw reads) of FASTQ format were firstly processed through Novogene Co. Ltd. perl scripts. Clean reads were obtained by removing reads containing adapter, reads containing ploy-N, and low quality reads from the raw data. At the same time, Q20, Q30 and GC content of the clean data were calculated.

An index of the reference genome was built using Bowtie (v2.2.3), and PE clean reads were aligned to the reference genome using TopHat (v2.0.12). TopHat was selected as the mapping tool as it can generate a database of splice junctions based on the gene model annotation file and thus a better mapping result is achieved than other non-splice mapping tools.

HTSeq (v0.6.1) was used to count the read numbers mapped to each gene. FPKM (Fragments Per Kilobase of transcript sequence per Millions base pairs sequenced) of each gene was calculated based on the length of the gene and reads count mapped to each gene. Differential expression analysis of each sample point/inbred line (three biological replicates) was performed using the DESeq R package (1.18.0). The resulting *P*-values were adjusted using Benjamini and

Hochberg's approach for controlling the false discovery rate. Genes with an adjusted *P*-value (<0.05) were assigned as being significantly differentially expressed.

RNAseq validation by qRT-PCR

Independent RNA extractions were conducted for qRT-PCR validation, and quality checked according to the same protocols and instrumentation as for RNAseq. cDNA synthesis was conducted using qPCRBIO cDNA Synthesis Kit (PCR Biosystems Ltd., London, UK) according to the manufacturer instructions. cDNA was then diluted 10x prior to analysis. All 54 biological samples were tested in triplicate.

PCR primers were designed using PRIMER3 (http://bioinfo.ut.ee/primer3/) using default settings. Ten genes related to GSL biosynthesis and transcription were selected at random for the validation analysis (*BCAT4, CYP83B1, MYB122-1a, MYB51a, SOT16, SUR1, TGG1b, TGG1d, TGG1j,* and *UGT74B1*), with *ACT11* used as a reference gene (*39*). Gene sequences of *E. sativa* were obtained using NovoFinder (Novogene Co. Ltd.), and primer annealing sites were designed to span intron-intron boundaries where possible (see Supplementary Table S5).

Analysis was performed using the $2^{-\Delta\Delta Ct}$ method on a Roche LightCycler 480 Instrument and the Advanced Relative Quantification protocol (v1.5.1). Primer efficiencies were determined by analyzing each primer set with log-fold dilutions of cDNA (Supplementary Table S5). 2x qPCRBIO SyGreen Blue Mix Lo-ROX (PCR Biosystems Ltd.) was used to prepare a master mix for all reactions. Reaction volumes totaled 10 µL, and the PCR method used was as per the manufacturer recommendations.

Data were normalized and expressed as the log2-fold change relative to *ACT11*. RNAseq data for each of the tested genes were similarly converted for direct comparison of the two methodologies (Supplementary Figure S3). An ANOVA test found no significant differences between the two data sets for each of the respective genes.

Intact glucosinolate extraction and analysis by LC-MS

Intact GSLs were extracted according to the protocol used by Bell et al. (*10*). Immediately before LC-MS analysis, samples were diluted with 4 mL of HPLC-grade water. Samples were analyzed in a random sequence with standards and QC samples. External standards of sinigrin (SIN; >99% TLC), GRA (99.86%, HPLC), GAL (98.8%, HPLC), 4OHB (96.19%, HPLC) and GER (99.68%, HPLC) were prepared for quantification of GSL compounds. SIN was used to quantify DGTB, GSV, and DMB, as no standards are available for these compounds. 4OHB was used to quantify the indole GSLs 4MOB and NGB. All standards with the exception of SIN (Sigma Merck, Gillingham, UK) were purchased from PhytoPlan (Heidelberg, Germany). Recovery of extracted GSLs was calculated by spiking six random samples (in duplicate) with SIN (0.06 µM). The average recovery was 104.8%, indicating excellent preservation of GSLs throughout the extraction process. Limits of detection (LOD) and quantification (LOQ) were established for the method by running serial dilutions of SIN (LOD = 2.14 $mg.kg^{-1}$; LOQ = 6.48 $mg.kg^{-1}$).

LC-MS analysis was performed in the negative ion mode on an Agilent 1260 Infinity Series LC system (Agilent, Stockport, UK) equipped with a binary pump, degasser, auto-sampler, column heater and diode array detector, coupled to an Agilent 6120 Series single quadrupole mass spectrometer. Separation of samples was achieved on a Gemini 3 µm C_{18} 110Å (150 x 4.6 mm) column (with Security Guard column, C_{18}; 4mm x 3mm; Phenomenex, Macclesfield, UK). GSLs were separated during a 40 min chromatographic run, with a 5 min post-run sequence. Mobile phases consisted of ammonium formate (0.1%; A) and acetonitrile (B) with the following gradient timetable: (i) 0 min (A-B, 95:5, v/v); (ii) 0-13 mins (A-B, 95:5, v/v); (iii) 13-22 mins (A-B, 40:60, v/v); (iv) 22-30 mins (A-B, 40:60, v/v); 30-35 mins (A-B, 95:5, v/v); (v) 35-40 mins (A-B, 95:5, v/v). The flow rate was optimized for the system at 0.4 mL min^{-1}, with a column temperature of 30

°C, and 20 µl of sample injected into the system. Quantification was conducted using a diode array detector at a wavelength of 229 nm.

MS analysis settings were as follows: Atmospheric pressure electrospray ionization was carried out in negative ion mode (scan range m/z 100–1500 Da). Nebulizer pressure was set at 50 psi, gas-drying temperature at 350 °C, and capillary voltage at 2,000 V. Compounds were identified using their primary ion mass [M-H]-, and comparison to authentic standards. Data were analyzed using Agilent OpenLAB CDS ChemStation Edition for LC-MS (vA.02.10). GSL concentrations from each time point were averaged over three biological replicates with two technical replicates of each (n = 6). This approach was also conducted for GHP and monosaccharide content.

Glucosinolate hydrolysis product extraction and analysis by GC-MS

GHPs were extracted according to the protocol presented by Ku et al. (*40*) with the following modification: samples were hydrolysed in d.H_2O for three hours at 30 °C before extraction with dichloromethane (DCM) for 21 hours. This duration was optimized for maximum yields of GHPs by comparison of extractions for three hours incubation at 30 °C with immediate DCM extraction, and three, nine, and 21 hours post incubation with DCM. GC-MS analysis and GHP identification was conducted according to the method presented by Bell et al. (*20*). Concentrations of all GHPs were calculated as equivalents of SF standard (Sigma).

Monosaccharide extraction and analysis by HPLC

Free monosaccharides were extracted according to the method presented by Bell et al. (*5*), with the exception that 0.2 g of lyophilized leaf powder was extracted. Extracts were analyzed on an Agilent 1100 series HPLC system equipped with a binary pump, degasser, and auto-sampler, with an external column heater (50 °C). A Bio-Rad Aminex HPX-87H (300 x 7.8 mm, 9 µm particle size) column with a Micro-Guard Cation H guard column (Bio-Rad, Watford, UK) was used to achieve separation with an isocratic gradient of 5 mM sulfuric acid, and a flow rate of 0.6 mL per min. A Polymer Laboratories ERC-7515 refractive index detector (Church Stretton, UK) was used to detect monosaccharides. Compounds were quantified using authentic standards and analyzed with Agilent ChemStation software (Santa Clara, CA, USA).

Sulfur content analysis by ICP-OES

Lyophilized samples were weighed into acid washed glass boiling tubes, pre-digested in 70% nitric acid for 24 hours, before being heated to 90 °C for two hours using a heat block. Once cooled, these were filtered through a 0.45 µM syringe filter, and diluted give an acid concentration of 3%. These samples were analysed using inductively coupled plasma-optical emission spectroscopy (ICP-OES) (Perkin Elmer Optima 7300 DV). Sulphur content was determined using the radial signal at 181.975 nm.

Statistical analyses

All statistical analyses (not included in bioinformatics sections) were performed using XL Stat (Addinsoft, Paris, France). Shapiro-Wilk normality tests were conducted for all variables and fit a normal distribution. ANOVA with post-hoc Tukey's Honest Significant Difference (HSD) tests were performed to generate multiple pairwise comparisons between sampling points for each cultivar (i.e. **H** vs. **D7** for cultivar **B**) and between cultivars at each respective time point (i.e. **A** vs. **B** for time point **H**) for phytochemical data (Supplementary Data File S1). All PCAs were performed using Pearson correlation coefficient analysis, n-1 standardization, Varimax rotation, and Kaiser Normalization. Phytochemical data were regressed onto the gene expression data as supplementary variables for the targeted analysis.

Supplementary Materials

Fig. S1. RNAseq experimental design and sampling diagram.
Fig. S2. RNAseq sample replicate Pearson correlation analysis.
Fig. S3. RNAseq gene expression validation by qRT-PCR.
Table S1. Summary of genome assembly and annotation statistics.
Table S2. Reference genome sequence composition.
Table S3. Predicted protein-coding genes within the *E. sativa* reference genome.
Table S4. Numbers of genes with homology or functional assignment within the *E. sativa* genome annotation.
Table S5. qRT-PCR primer sequences and amplification efficiencies.
Data file S1. Phytochemical Analyses of Variance (ANOVA) pairwise comparisons and RNAseq gene expression correlation analysis.
Data file S2. Sulfur metabolism and glucosinolate biosynthesis gene expression values and annotation descriptions.

References and Notes

1. S. Kopriva *et al.*, Editorial: Frontiers of Sulfur Metabolism in Plant Growth, Development, and Stress Response. *Front. Plant Sci.* **6**, (2016).
2. H. Frerigmann, T. Gigolashvili, Update on the role of R2R3-MYBs in the regulation of glucosinolates upon sulfur deficiency. *Front. Plant Sci.* **5**, 626 (2014).
3. X. F. Yan, S. X. Chen, Regulation of plant glucosinolate metabolism. *Planta.* **226**, 1343–1352 (2007).
4. L. Bell, C. Wagstaff, Glucosinolates, Myrosinase Hydrolysis Products, and Flavonols Found in Rocket (*Eruca sativa* and *Diplotaxis tenuifolia*). *J. Agric. Food Chem.* **62**, 4481–92 (2014).
5. L. Bell, L. Methven, A. Signore, M. Jose Oruna-Concha, C. Wagstaff, Analysis of Seven Salad Rocket (*Eruca sativa*) Accessions: The Relationships Between Sensory Attributes and Volatile and Non-volatile Compounds. *Food Chem.* **218**, 181–191 (2017).
6. I. Winde, U. Wittstock, Insect herbivore counteradaptations to the plant glucosinolate-myrosinase system. *Phytochemistry*. **72** (2011), pp. 1566–1575.
7. Z. Zhang, J. A. Ober, D. J. Kliebenstein, The gene controlling the quantitative trait locus EPITHIOSPECIFIER MODIFIER1 alters glucosinolate hydrolysis and insect resistance in *Arabidopsis*. *Plant Cell.* **18**, 1524–1536 (2006).
8. L. Bell, C. Wagstaff, Rocket Science: A Review Of Phytochemical and Health-related Research In *Eruca* and *Diplotaxis* Species. *Food Chem. X.* **1**, 100002 (2019).
9. S. J. Kim *et al.*, Structural elucidation of 4-(cystein-S-yl)butyl glucosinolate from the leaves of *Eruca sativa*. *Biosci. Biotechnol. Biochem.* **71**, 114–121 (2007).
10. L. Bell, M. J. Oruna-Concha, C. Wagstaff, Identification and quantification of glucosinolate and flavonol compounds in rocket salad (*Eruca sativa, Eruca vesicaria* and *Diplotaxis tenuifolia*) by LC-MS: highlighting the potential for improving nutritional value of rocket crops. *Food Chem.* **172**, 852–861 (2015).
11. M. Francisco *et al.*, The Defense Metabolite, Allyl Glucosinolate, Modulates *Arabidopsis thaliana* Biomass Dependent upon the Endogenous Glucosinolate Pathway. *Front. Plant Sci.* **7**, 774 (2016).
12. K. Kazan, J. M. Manners, MYC2: The Master in Action. *Mol. Plant*. **6**, 686–703 (2013).
13. F. Aarabi *et al.*, Sulfur deficiency–induced repressor proteins optimize glucosinolate biosynthesis in plants. *Sci. Adv.* **2** (2016).
14. I. E. Sønderby *et al.*, A systems biology approach identifies a R2R3 MYB gene subfamily with distinct and overlapping functions in regulation of aliphatic glucosinolates. *PLoS One*. **2** (2007).
15. M. D. Mikkelsen *et al.*, Modulation of CYP79 genes and glucosinolate profiles in *Arabidopsis* by defense signaling pathways. *Plant Physiol.* **131**, 298–308 (2003).
16. S. Bak, R. Feyereisen, The involvement of two p450 enzymes, CYP83B1 and CYP83A1, in auxin homeostasis and glucosinolate biosynthesis. *Plant Physiol.* **127**, 108–18 (2001).
17. K. S. Satyan *et al.*, Phenethyl isothiocyanate (PEITC) inhibits growth of ovarian cancer cells by inducing apoptosis: Role of caspase and MAPK activation. *Gynecol. Oncol.* **103**, 261–270 (2006).
18. I. Herr, M. W. Buechler, M. W. Büchler, Dietary constituents of broccoli and other cruciferous vegetables: Implications for prevention and therapy of cancer. *Cancer Treat. Rev.* **36**, 377–383 (2010).
19. V. Rungapamestry, A. J. Duncan, Z. Fuller, B. Ratcliffe, Effect of cooking brassica vegetables on the subsequent hydrolysis and metabolic fate of glucosinolates. *Proc. Nutr. Soc.* **66**, 69–81 (2007).
20. L. Bell, H. N. Yahya, O. O. Oloyede, L. Methven, C. Wagstaff, Changes In Rocket Salad Phytochemicals Within The Commercial Supply Chain: Glucosinolates, Isothiocyanates, Amino Acids And Bacterial Load

Increase Significantly After Processing. *Food Chem.* **221**, 521–534 (2017).
21. T. Arias, J. C. Pires, A fully resolved chloroplast phylogeny of the brassica crops and wild relatives (Brassicaceae: Brassiceae): Novel clades and potential taxonomic implications. *Taxon*. **61**, 980–988 (2012).
22. J. Fechner *et al.*, The major glucosinolate hydrolysis product in rocket (*Eruca sativa* L.), sativin, is 1,3-thiazepane-2-thione: Elucidation of structure, bioactivity, and stability compared to other rocket isothiocyanates. *Food Chem.* **261**, 57–65 (2018).
23. T. Gigolashvili, B. Berger, U.-I. Flügge, Specific and coordinated control of indolic and aliphatic glucosinolate biosynthesis by R2R3-MYB transcription factors in *Arabidopsis thaliana*. *Phytochem. Rev.* **8**, 3-13 (2009).
24. S. Samuilov *et al.*, Knock-down of phosphoserine phosphatase gene effects rather N- than S-metabolism in *Arabidopsis thaliana*. *Front. Plant Sci.* **9**, 1830 (2018).
25. M. Wang, Y. Jia, Z. Xu, Z. Xia, Impairment of Sulfite Reductase Decreases Oxidative Stress Tolerance in *Arabidopsis thaliana*. *Front. Plant Sci.* **7**, 1843 (2016).
26. K. X. Chan, S. Y. Phua, F. Van Breusegem, Secondary sulfur metabolism in cellular signalling and oxidative stress responses. *J. Exp. Bot.* **70**, 4237–4250 (2019).
27. M. V Selma *et al.*, Impact of Organic Soil Amendments on Phytochemicals and Microbial Quality of Rocket Leaves (*Eruca sativa*). *J. Agric. Food Chem.* **58**, 8331–8337 (2010).
28. C. Jiang *et al.*, Structural and functional comparative mapping between the *Brassica* A genomes in allotetraploid *Brassica napus* and diploid *Brassica rapa*. *Theor. Appl. Genet.* **123**, 927–941 (2011).
29. S. Liu *et al.*, The *Brassica oleracea* genome reveals the asymmetrical evolution of polyploid genomes. *Nat. Commun.* **5**, 499–507 (2014).
30. M. Cavaiuolo *et al.*, Gene expression analysis of rocket salad under pre-harvest and postharvest stresses: A transcriptomic resource for *Diplotaxis tenuifolia*. *PLoS One*. **12**, e0178119 (2017).
31. T. Kakizaki *et al.*, A 2-Oxoglutarate-Dependent Dioxygenase Mediates the Biosynthesis of Glucoraphasatin in Radish. *Plant Physiol.* **173**, 1583–1593 (2017).
32. I. E. Sønderby *et al.*, A Complex Interplay of Three R2R3 MYB Transcription Factors Determines the Profile of Aliphatic Glucosinolates in *Arabidopsis*. *Plant Physiol.* **153**, 348–363 (2010).
33. S. Kopriva, A. Calderwood, S. C. Weckopp, A. Koprivova, Plant sulfur and Big Data. *Plant Sci.* **241**, 1–10 (2015).
34. A. Maruyama-Nakashita, E. Inoue, A. Watanabe-Takahashi, T. Yamaya, H. Takahashi, Transcriptome profiling of sulfur-responsive genes in *Arabidopsis* reveals global effects of sulfur nutrition on multiple metabolic pathways. *Plant Physiol.* **132**, 597–605 (2003).
35. I. Mostafa *et al.*, New nodes and edges in the glucosinolate molecular network revealed by proteomics and metabolomics of *Arabidopsis myb28/29* and *cyp79B2/B3* glucosinolate mutants. *J. Proteomics*. **138**, 1–19 (2016).
36. K. X. Chan, M. Wirtz, S. Y. Phua, G. M. Estavillo, B. J. Pogson, Balancing metabolites in drought: the sulfur assimilation conundrum. *Trends Plant Sci.* **18**, 18–29 (2013).
37. L. Bell, L. Methven, C. Wagstaff, The influence of phytochemical composition and resulting sensory attributes on preference for salad rocket (*Eruca sativa*) accessions by consumers of varying TAS2R38 diplotype. *Food Chem.* **222**, 6–17 (2017).
38. S. Huseby *et al.*, Diurnal and light regulation of sulphur assimilation and glucosinolate biosynthesis in *Arabidopsis*. *J. Exp. Bot.* **64**, 1039–48 (2013).
39. R. Hu, C. Fan, H. Li, Q. Zhang, Y.-F. Fu, Evaluation of putative reference genes for gene expression normalization in soybean by quantitative real-time RT-PCR. *BMC Mol. Biol.* **10**, 93 (2009).
40. K.-M. Ku, M. J. Kim, E. H. Jeffery, Y.-H. Kang, J. A. Juvik, Profiles of Glucosinolates, their Hydrolysis Products, and Quinone Reductase Inducing Activity from 39 Arugula (*Eruca sativa* Mill.) Accessions. *J. Agric. Food Chem.* **64**, 6524–6532 (2016).

Acknowledgments:

General:

The authors would like to thank: the Vegetable Plant Breeding Team at Elsoms Seeds Ltd.; members of the Genomics Pipelines Group in the BBSRC National Capability in Genomics and Single Cell (BB/CCG1720/1) at Earlham Institute; Dr. Yunan Lin and Irene Wei for project and technical support for genome annotation, RNAseq, and bioinformatics at Novogene Co. Ltd.; Matthew Richardson for maintenance of controlled environment facilities at the University of Reading Controlled Environment Laboratory; and Dr. Marcia Boura for advice on qRT-PCR.

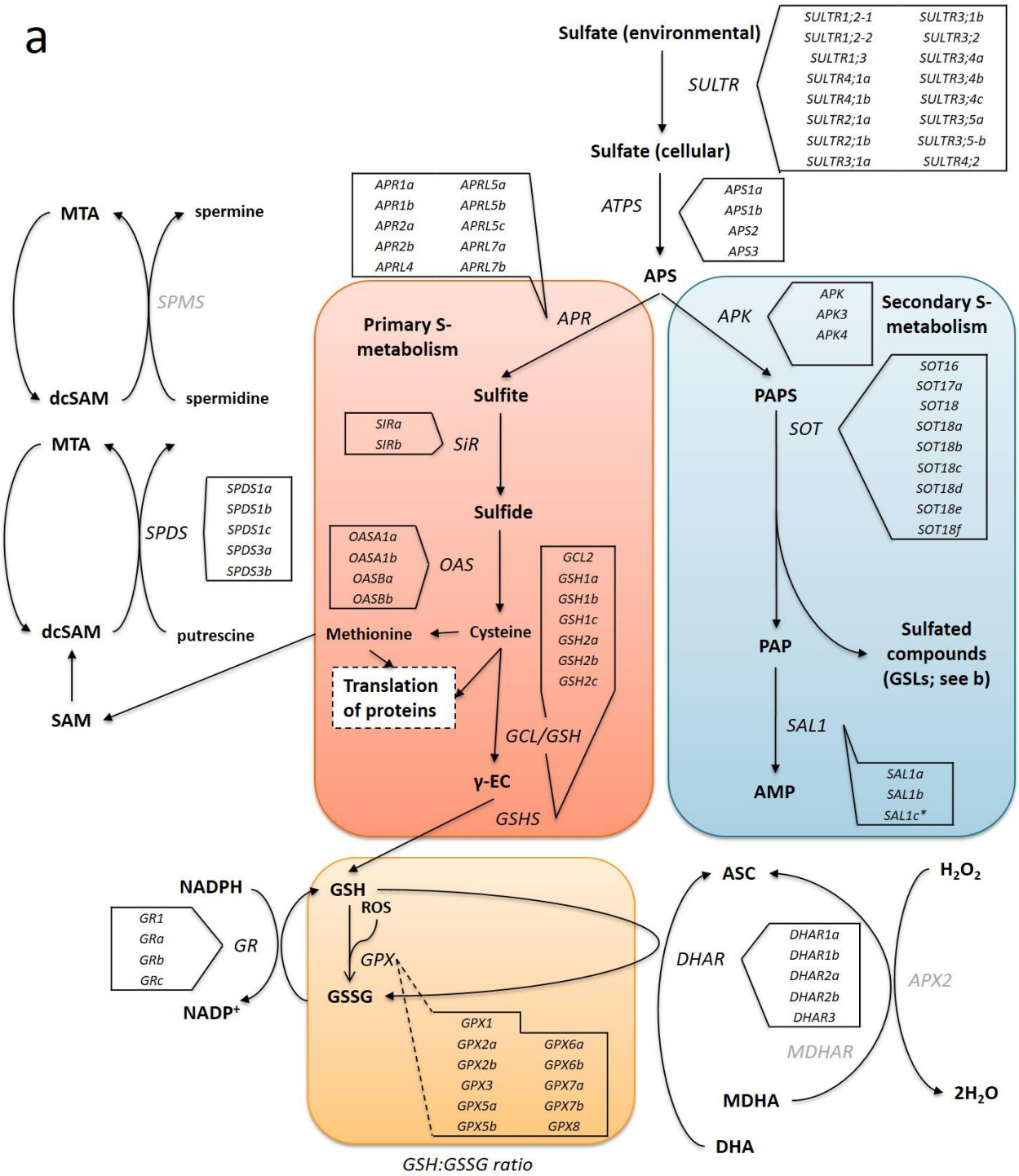
Funding:
Dr. Luke Bell was supported by a BBSRC Case Award (BB/J012629/1) in partnership with Elsoms Seeds Ltd. (Spalding, UK) and Bakkavor Group Ltd. (Spalding, UK) for *de novo* genome sequencing and assembly. Dr. Luke Bell, Dr. Martin Chadwick, and Manik Puranik were supported by a BBSRC LINK award (BB/N01894X/1) for all other work.

Author contributions:
LB and CW conceived and designed the experiment and analyses. RT and SK produced the breeding line seed material for genome sequencing and the RNAseq experiment. LB grew plants in controlled environment, performed RNA extractions and quality control, qRT-PCR validation, glucosinolate analysis by LC-MS, and hydrolysis product analysis by GC-MS. MP performed sugar analysis by HPLC. MC performed sulfur content analysis by ICP-OES. LB performed ANOVAs, Pearson's correlation analyses, and Principal Component Analyses of all phytochemical and gene expression data. LB wrote the paper, with contributions from MC, RT, SK, and CW. Funding was obtained by LB, LM, and CW.

Competing interests:
The authors declare no conflicts of interest.

Data and materials availability:
Full reference genome sequence and annotation can be found in the European Nucleotide Archive. Assembly accession no: GCA_902460325; Study ID: PRJEB34051; Sample ID: ERS3673677; Annotation accession number ERZ1066251.

Figures and Tables



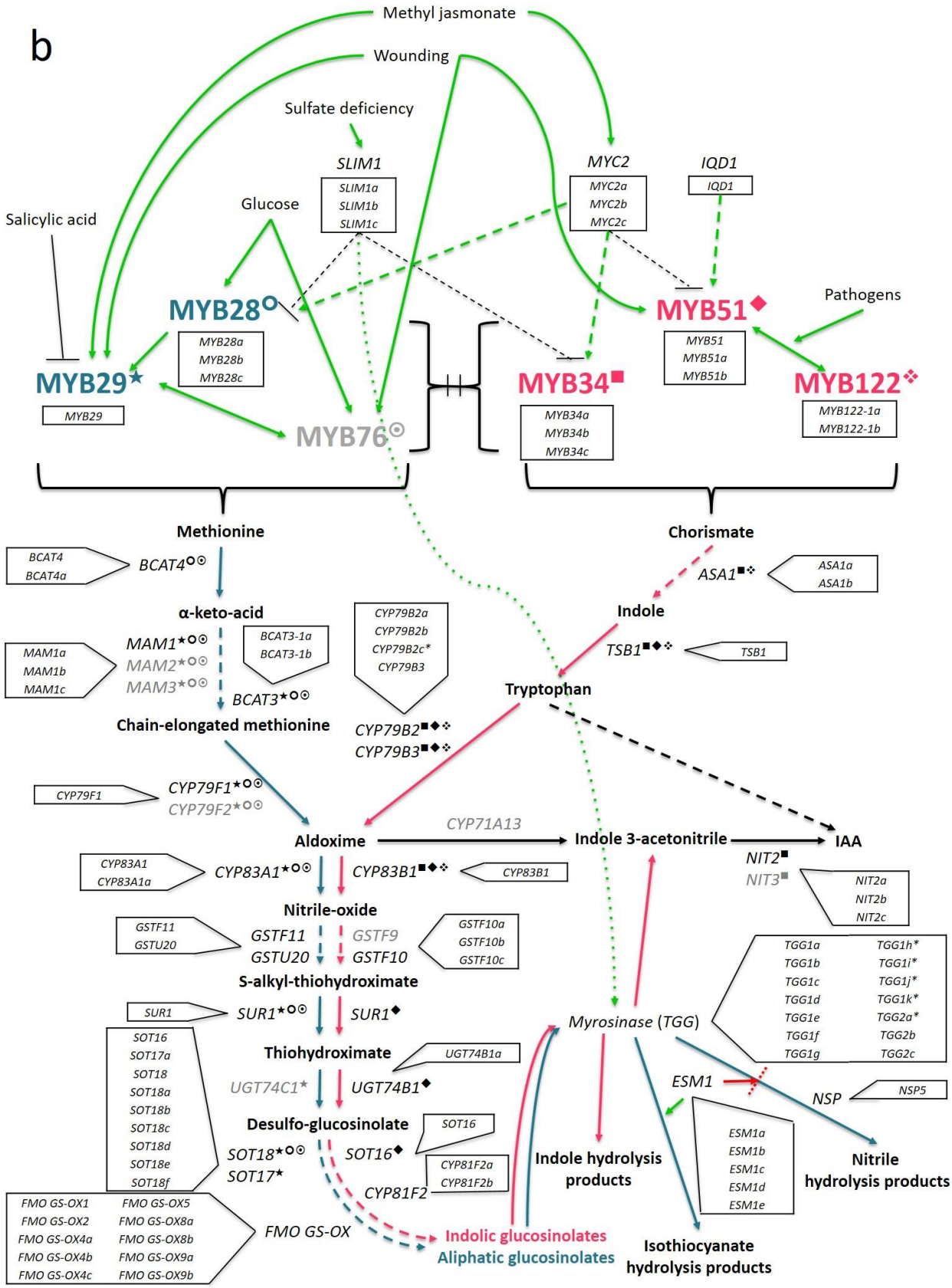
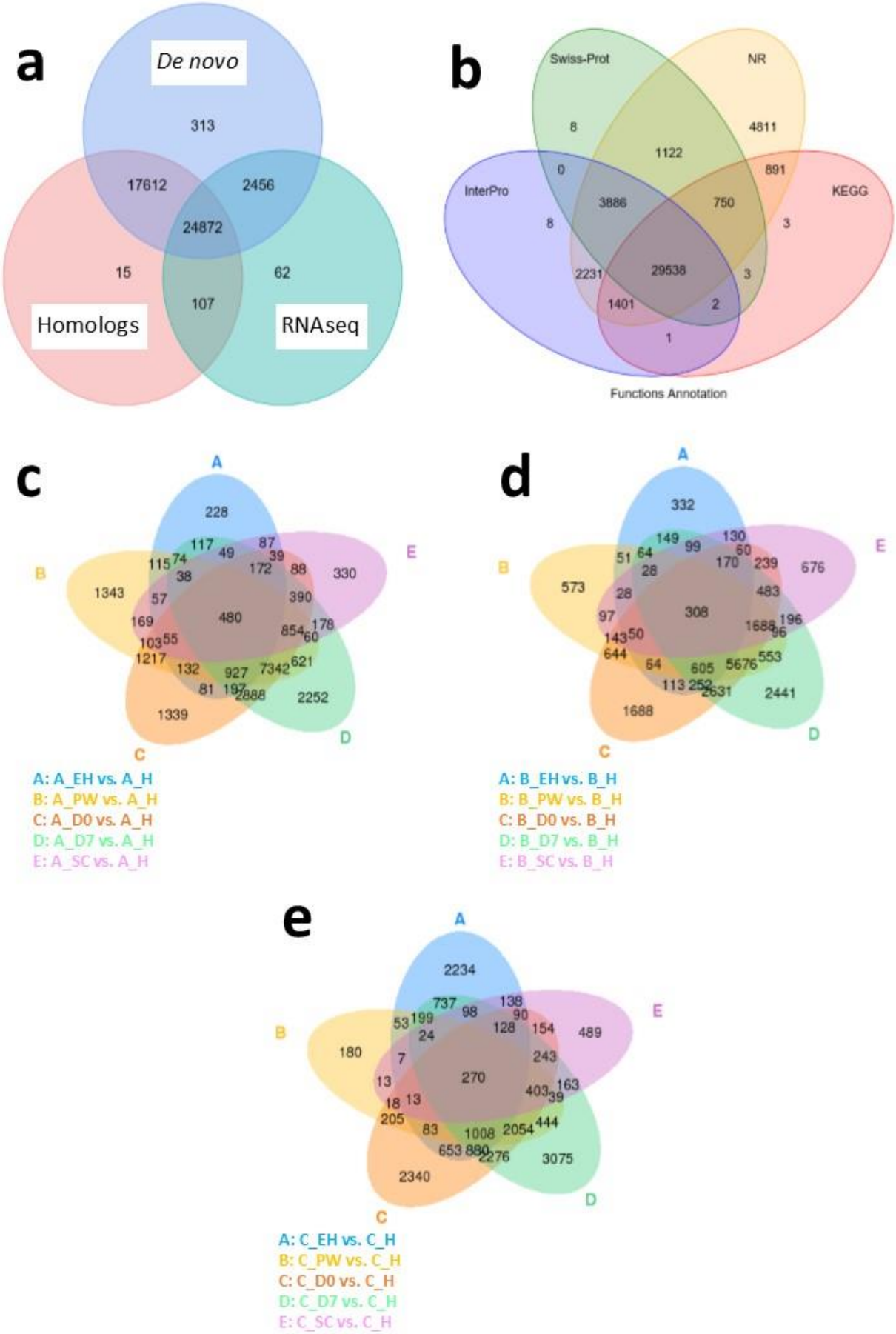


Fig. 1. Sulfur metabolism and glucosinolate biosynthesis pathways.
The primary (red box) and secondary (blue box) sulfur metabolism pathways (**a**) of *Arabidopsis thaliana* with identified homologous genes within the *Eruca sativa* genome

annotation (see boxed insets) adapted from Chan et al. (*36*). Environmental sulfur is assimilated and integrated into key amino acids (cysteine and methionine) and enzymes. Sulfur metabolism is also intrinsically linked with oxidative stress via glutathione synthesis. Under stress conditions 5'-phosphoadenosine-3'-phosphate (PAP), glutathione disulfide (GSSG), and reduced glutathione (GSH) direct sulfate towards GSH production. The GSH:GSSG redox state ratio is also known to influence sulfur assimilation rates (orange box). *SOT* (*sulfotransferase*) genes link secondary sulfur metabolism with the final sulfation step of GSL biosynthesis, and it is thought that *SAL1* plays an important role in regulating the activity of these genes through interaction with PAP. GSL biosynthesis (**b**; adapted from Gigolashvili et al.) (*23*) is initiated by a complex and interacting network of abiotic and biotic factors. Aliphatic synthesis pathway shown in teal, and the indolic pathway shown in pink, is regulated by R2R3-MYB transcription factors. Known interactions between MYBs and specific genes within each respective pathway are highlighted as follows: ○ = MYB28, ★ = MYB29, ⊙ = MYB76, ■ = MYB34, ◆ = MYB51, ❖ = MYB122. Genes with identified orthologs in the *E. sativa* genome annotation are written in black; those with no identified homologous sequence are written in grey. Abbreviations: *SULTR, sulfate transporter*; ATP, adenosine triphosphate; *ATPS, ATP sulfurylase*; *APR, APS reductase*; *APK, APS kinase*; *SiR, sulfite reductase*; *OASTL, O-acetylserine lyase*; *GCL, glutamate cysteine ligase*; γ-EC, γ-glutamyl-cysteine; *GSHS, GSH synthetase*; *GR, glutathione reductase*; *GPX, glutathione peroxidase*; ASC, ascorbate; DHA, dehydroascorbate; *DHAR, DHA reductase*; MDHA, monodehydroascorbate; *MDHAR, MDHA reductase*; *APX2, ascorbate peroxidase 2*; H_2O_2, hydrogen peroxide; PAPS, 5'-phosphoadenosine-3'-phosphosulfate; SAM, S-adenosyl methionine; dcSAM, decarboxylated SAM; MTA, methylthioadenosine; *SPDS, spermidine synthase*; *SPMS, spermine synthase*; *SLIM1, sulfur limitation 1*; *IQD1, protein IQ domain 1*; *BCAT, methionine aminotransferase*; *MAM, methylthioalkylmalate synthase*; *CYP, cytochrome P450*; *GST, glutathione-S-transferase*; *SUR1, C-S lyase 1*; *UGT, UDP-glycosyltransferase*; *FMO* $_{GS\text{-}OX}$*, flavin-containing monooxygenase*; *ASA1, anthranilate synthase alpha subunit 1*; *TSB1, tryptophan synthase beta chain 1*; IAA, indole-3-acetic acid; *NIT, nitrilase*; *ESM1, epithiospecifier modifier protein 1*; *NSP, nitrile specifier protein*.


a
De novo
313
17612
2456
24872
15
62
107
Homologs
RNAseq
b
Swiss-Prot
NR
8
4811
1122
0
891
InterPro
KEGG
3886
750
8
3
2231
29538
3
1401
2
1
Functions Annotation
c
A
B
C
D
E
228
117
87
115
74
49
39
38
172
88
330
1343
57
390
169
480
854
178
103
55
60
1217
132
927
7342
621
81
197
2888
2252
1339
A: A_EH vs. A_H
B: A_PW vs. A_H
C: A_D0 vs. A_H
D: A_D7 vs. A_H
E: A_SC vs. A_H
d
A
B
C
D
E
332
149
99
130
51
64
60
28
170
239
676
573
28
483
97
308
196
143
50
1688
96
644
64
605
5676
553
113
252
2631
2441
1688
A: B_EH vs. B_H
B: B_PW vs. B_H
C: B_D0 vs. B_H
D: B_D7 vs. B_H
E: B_SC vs. B_H
e
A
B
C
D
E
2234
737
138
53
199
98
90
24
128
154
489
180
7
243
13
270
163
403
39
18
13
205
83
1008
2054
444
653
880
2276
3075
2340
A: C_EH vs. C_H
B: C_PW vs. C_H
C: C_D0 vs. C_H
D: C_D7 vs. C_H
E: C_SC vs. C_H

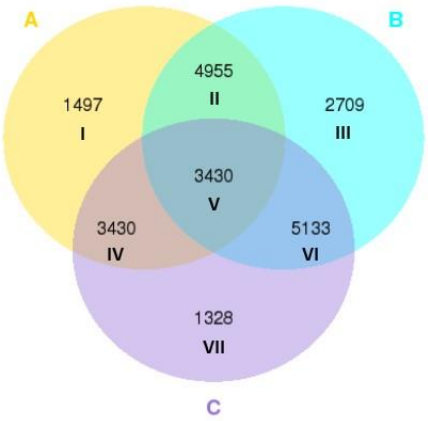
1115

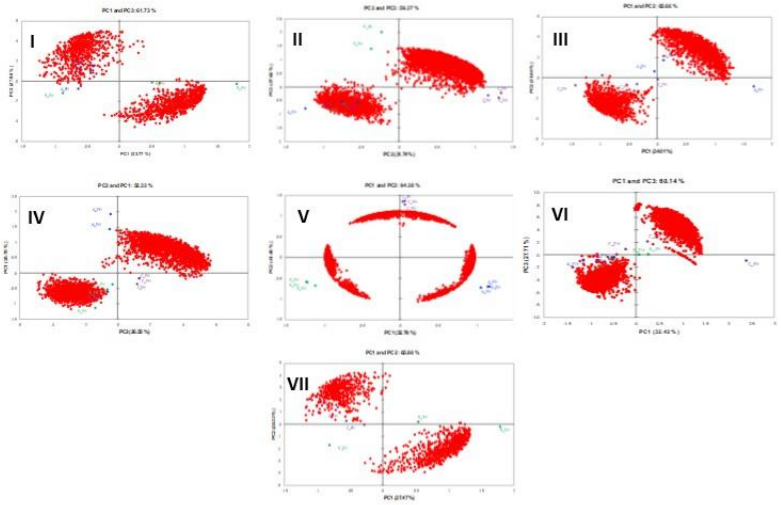
Figure 2. *Eruca sativa* genome annotation and differentially expressed gene number Venn diagrams for each inbred line

Venn diagrams of the *Eruca sativa* reference genome annotation gene identification sources (**a**) and functional annotation databases used to assign putative gene identities (**b**). Also shown are Venn diagrams of global differentially expressed genes (DEGs) at an early harvest (**EH**), second harvest (**SC**), pre-wash (**PW**), post-wash (**D0**), and seven-day shelf life (**D7**) time points relative to a first harvest (**H**) time point of three elite breeding lines: **A** (**c**), **B** (**d**), and **C** (**e**). The numbers of DEGs identified under each condition are contained within the ellipses and their overlaps.

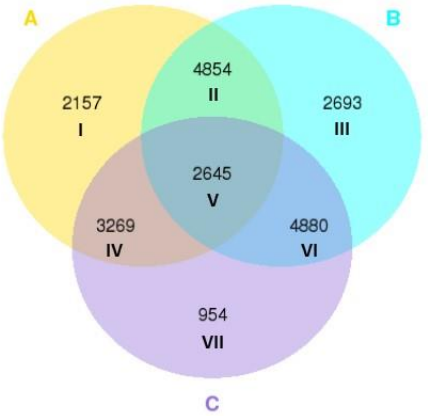
a

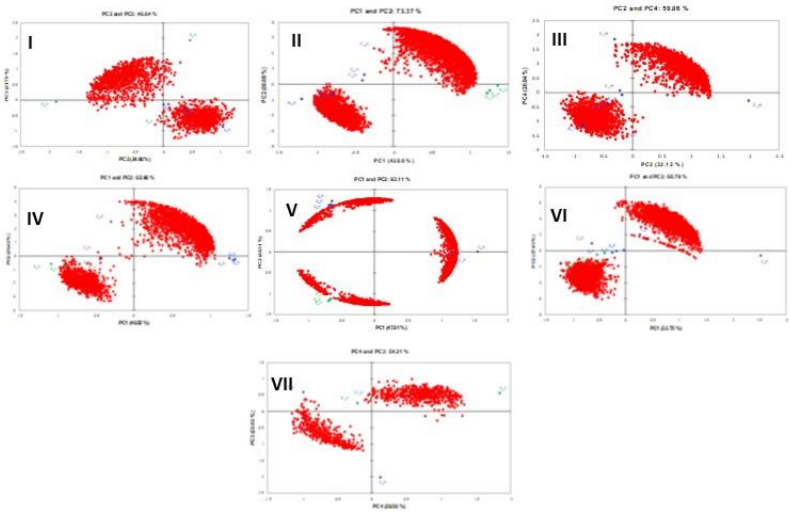
EH


A
B
C
1497
I
4955
II
2709
III
3430
V
3430
IV
5133
VI
1328
VII

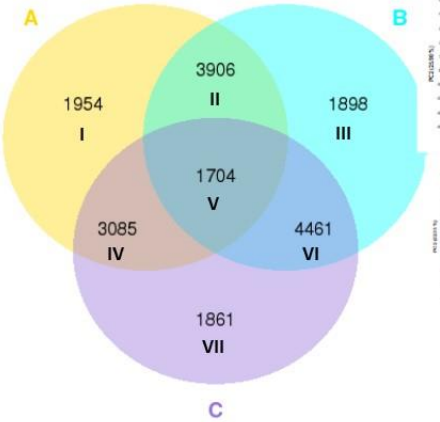

I
II
III
IV
V
VI
VII

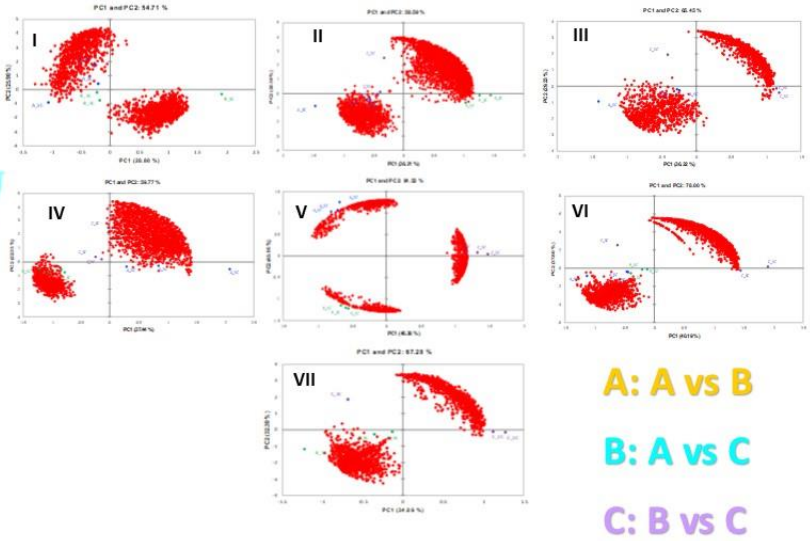
H


A
B
C
2157
I
4854
II
2693
III
2645
V
3269
IV
4880
VI
954
VII


I
II
III
IV
V
VI
VII

SC

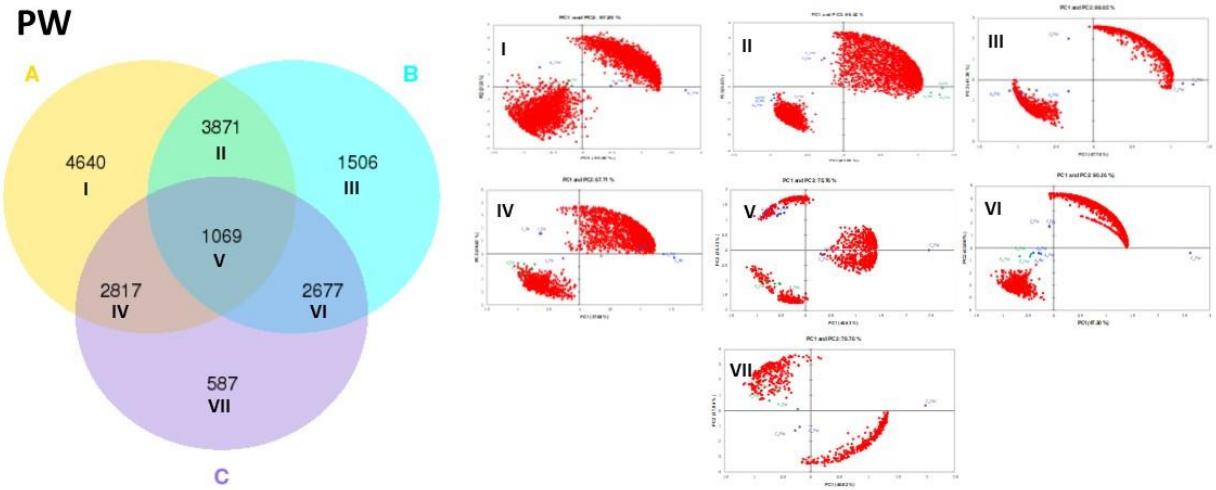

A
B
C
1954
I
3906
II
1898
III
1704
V
3085
IV
4461
VI
1861
VII


I
II
III
IV
V
VI
VII
A: A vs B
B: A vs C
C: B vs C

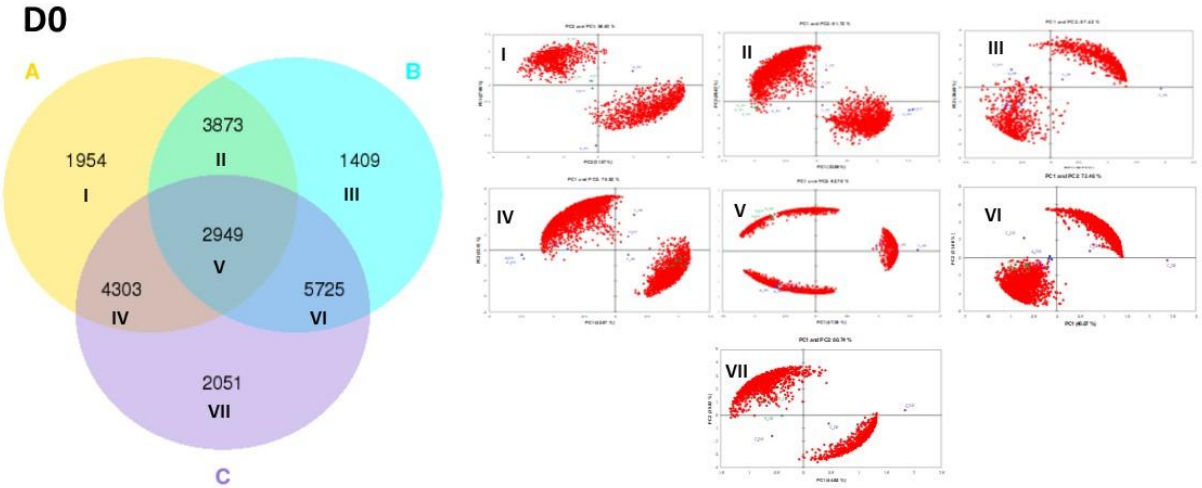
1126

b

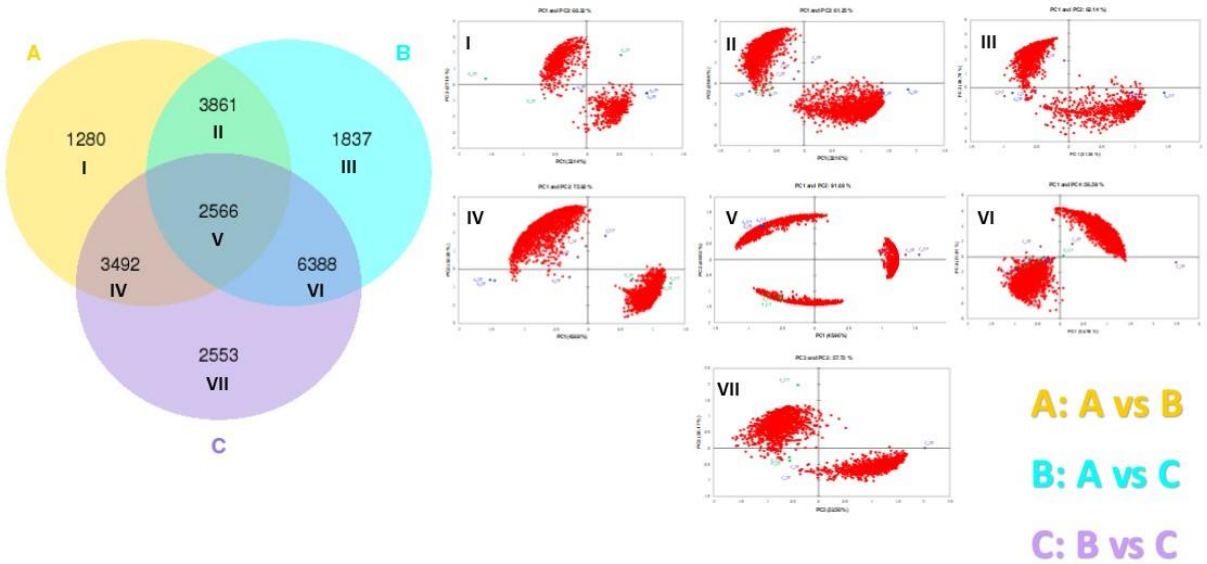
PW



D0



D7



1127

Figure 3. Venn diagrams of global differentially expressed genes between inbred lines with Principal Component Analyses

Venn diagrams of the numbers of differentially expressed genes (DEGs) found between three elite inbred lines of *Eruca sativa*. Different harvest ontogeny (**a**) and postharvest (**b**) time points are presented with accompanying Principal Component Analysis (PCA) plots that highlight the multidimensional separations of each DEG cluster between lines **A, B,** and **C**. Roman numerals are indicative of corresponding PCA analysis of each Venn diagram segment. Red data points within each loadings-scores biplot represent gene expression data (FPKM). Blue circles = **A**; green circles = **B**; purple circles = **C**. See insets for Venn diagram color coding. Abbreviations: early harvest (**EH**), harvest (**H**), second harvest (**SC**), pre-wash (**PW**), post-wash (**D0**), and seven-day shelf life (**D7**).

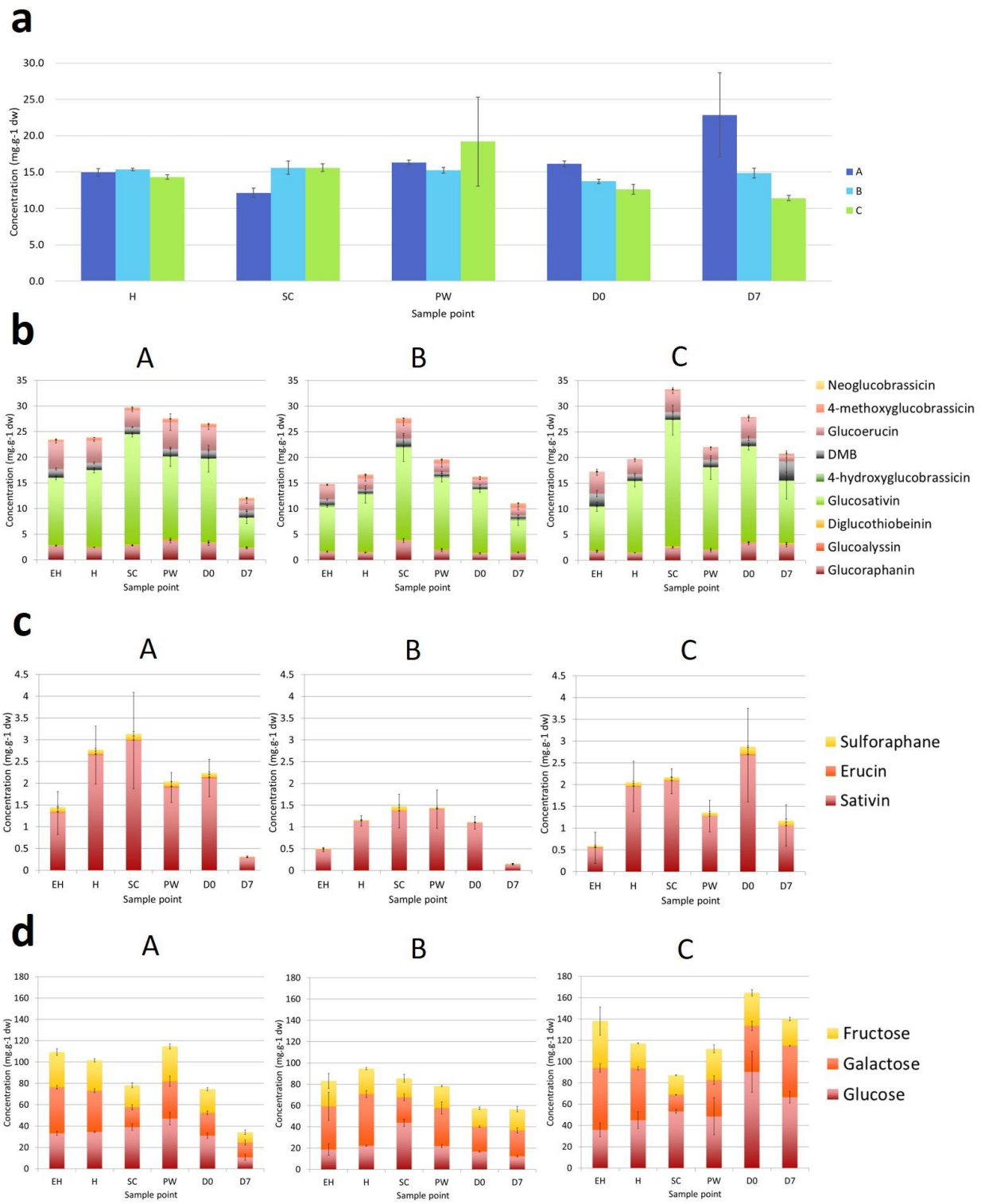
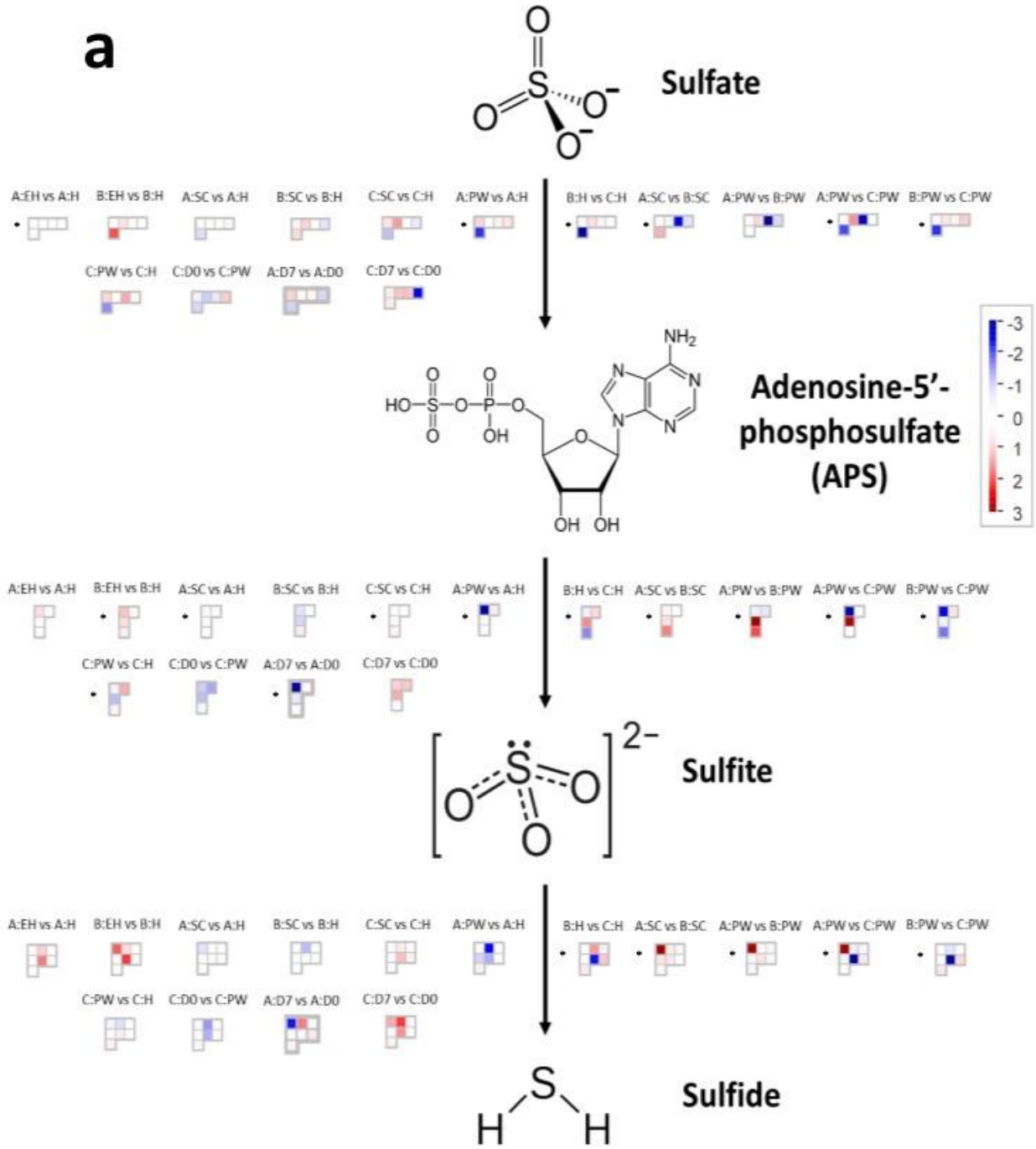
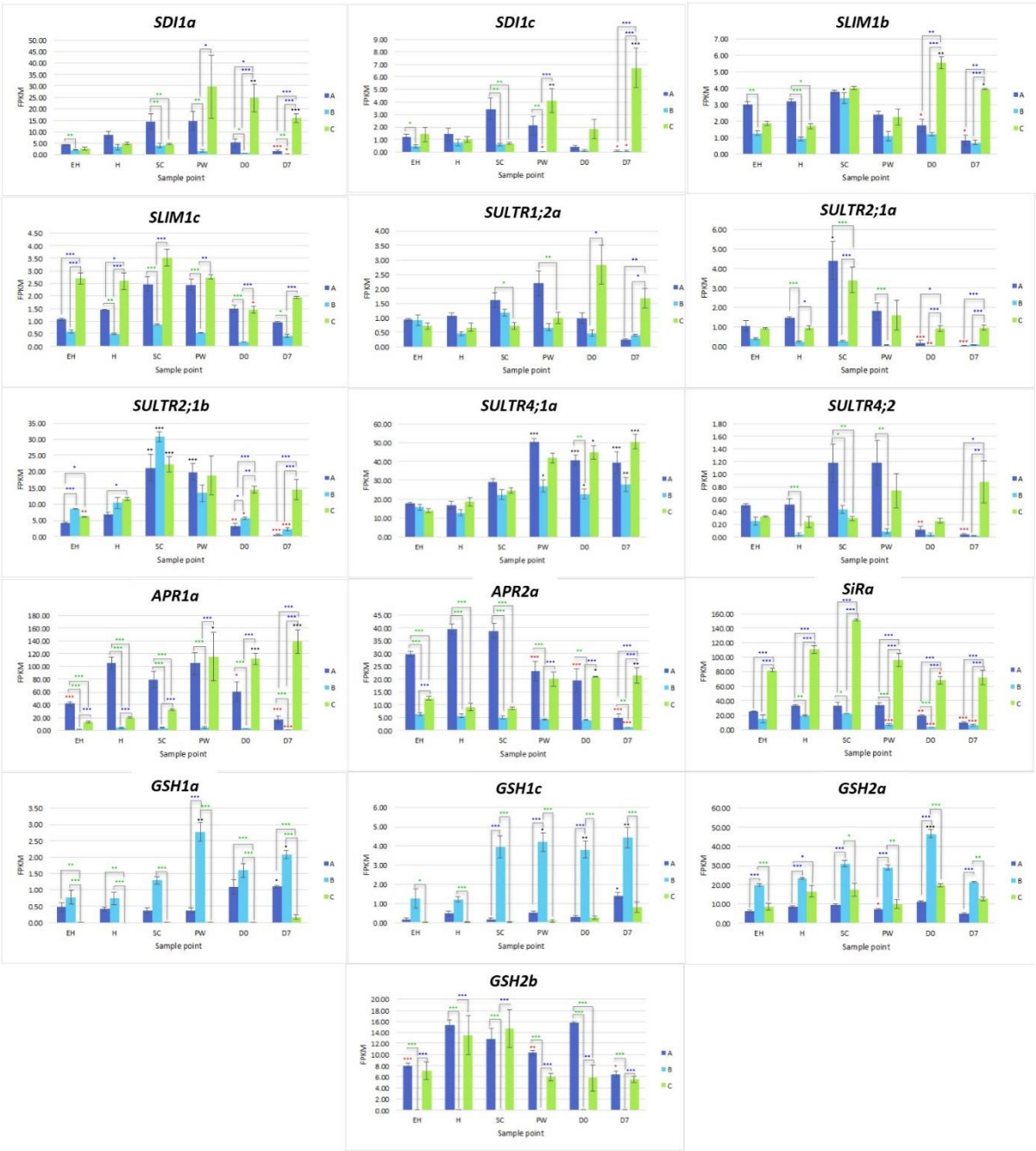


Figure 4. Elemental sulfur and phytochemical compositions of rocket inbred lines
Elemental sulfur (**a**), glucosinolate (**b**), glucosinolate hydrolysis product (**c**), and monosaccharide (**d**) concentrations observed in elite inbred lines of *Eruca sativa* (**A, B,** and **C**). Concentrations are expressed as $mg.g^{-1}$ of dry weight. Error bars represent standard error of the mean of each analyte detected. See insets for compound color coding. For ANOVA and Tukey's HSD pairwise significance values see Supplementary Data File S1.

Abbreviations: early harvest (**EH**), harvest (**H**), second harvest (**SC**), pre-wash (**PW**), post-wash (**D0**), and seven-day shelf life (**D7**).

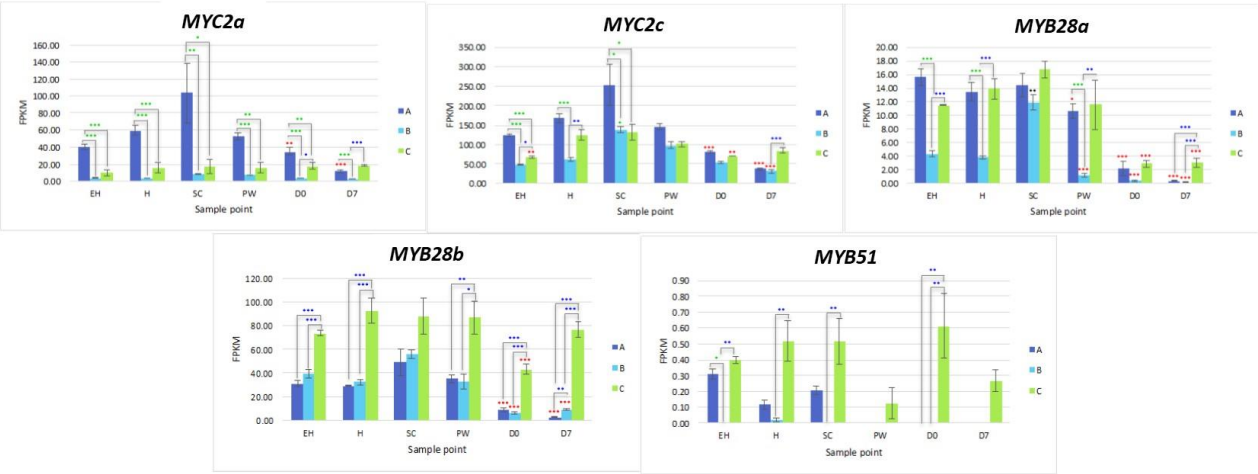


b


SDI1a
SDI1c
SLIM1b
SLIM1c
SULTR1;2a
SULTR2;1a
SULTR2;1b
SULTR4;1a
SULTR4;2
APR1a
APR2a
SiRa
GSH1a
GSH1c
GSH2a
GSH2b
FPKM
Sample point
EH
H
SC
PW
D0
D7
A
B
C

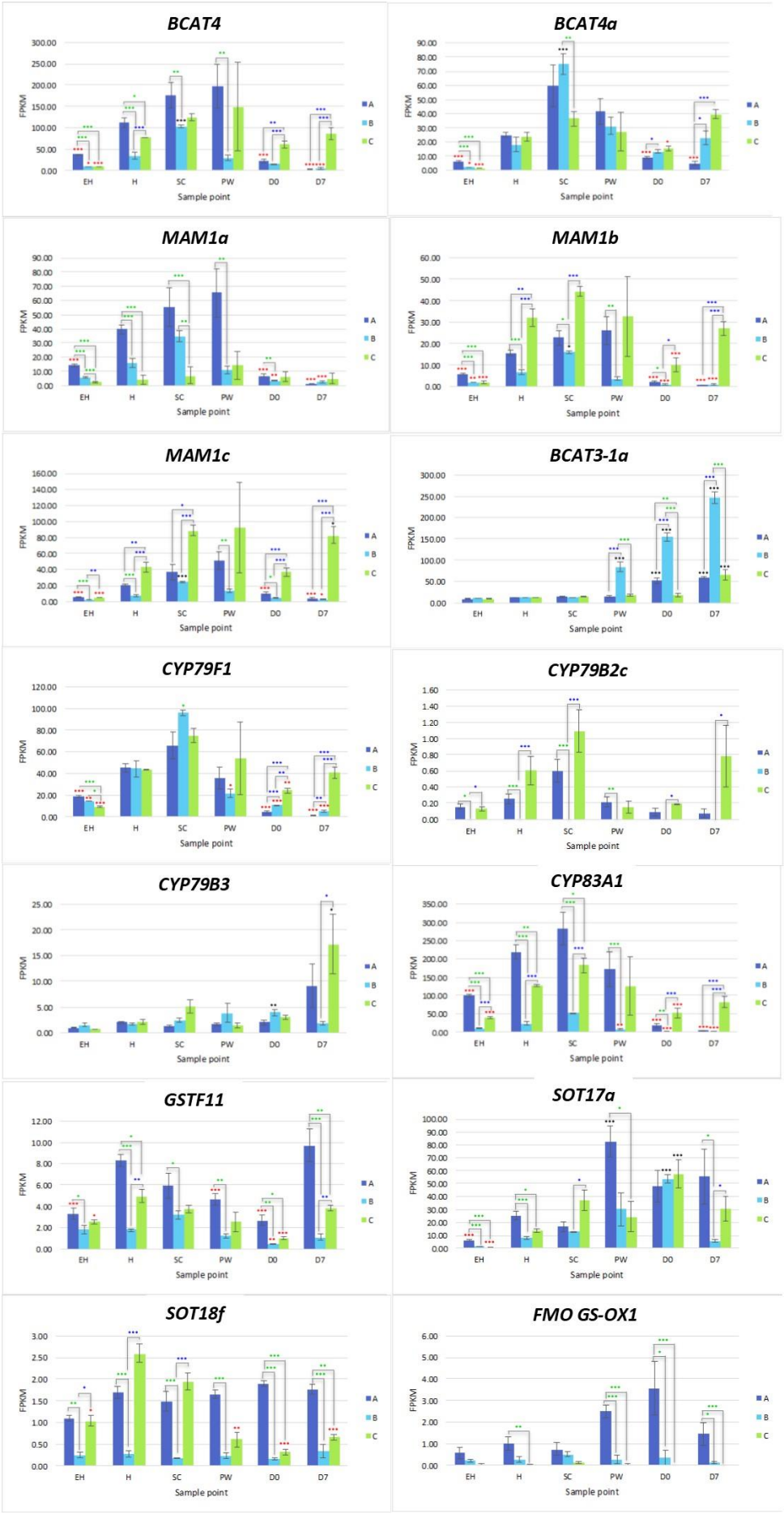
1152

C


MYC2a
MYC2c
MYB28a
MYB28b
MYB51
FPKM
Sample point
EH
H
SC
PW
D0
D7
A
B
C

1153

d


BCAT4
BCAT4a
MAM1a
MAM1b
MAM1c
BCAT3-1a
CYP79F1
CYP79B2c
CYP79B3
CYP83A1
GSTF11
SOT17a
SOT18f
FMO GS-OX1
FPKM
Sample point
EH
H
SC
PW
D0
D7
A
B
C

1154

e

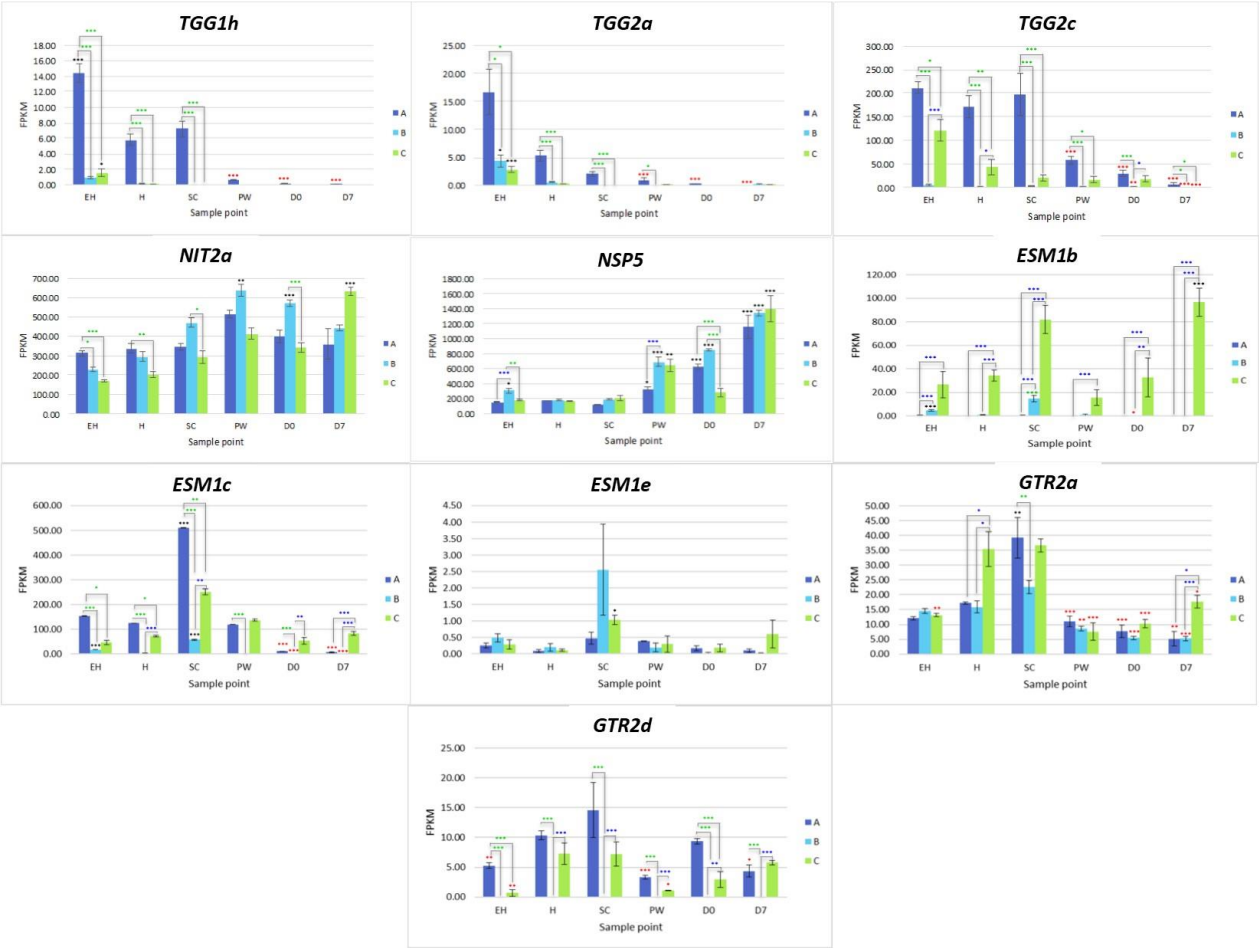


Figure 5. RNAseq expression data for sulfur metabolism and glucosinolate biosynthesis-related genes

RNAseq expression data (FPKM) for genes involved with sulfate assimilation (**a**), sulfate transport and redox response (**b**), glucosinolate transcription factors (**c**), glucosinolate biosynthesis (**d**), and glucosinolate hydrolysis and transport (**e**) in three elite inbred lines of *Eruca sativa* (**A** = dark blue; **B** = light blue; **C** = green). For (**a**) a custom MapMan (version 3.6.0RC1) annotation file of the *E. sativa* reference genome was created using Mercartor4 (version 1.0, plaBi dataBase, Institute of Biology, Aachen, Germany), and used to visualize differential expression of genes within the sulfate assimilation pathway. In (**b**) to (**e**), standard errors of the mean expression values are represented by error bars. Asterisks denote levels of significance of up and down regulation within sample points (between each inbred line) and relative to the point of harvest for each respective sample point: * = $P \leq 0.05$; ** = $P \leq 0.01$; *** = $P \leq 0.001$; green = significant up regulation between lines **A, B,** and **C**; blue = significant down regulation between lines **A, B,** and **C**; black = significant up regulation relative to **H**; red = significant down regulation relative to **H**. Abbreviations: early harvest (**EH**), harvest (**H**), second harvest (**SC**), pre-wash (**PW**), post-wash (**D0**), and seven-day shelf life (**D7**).

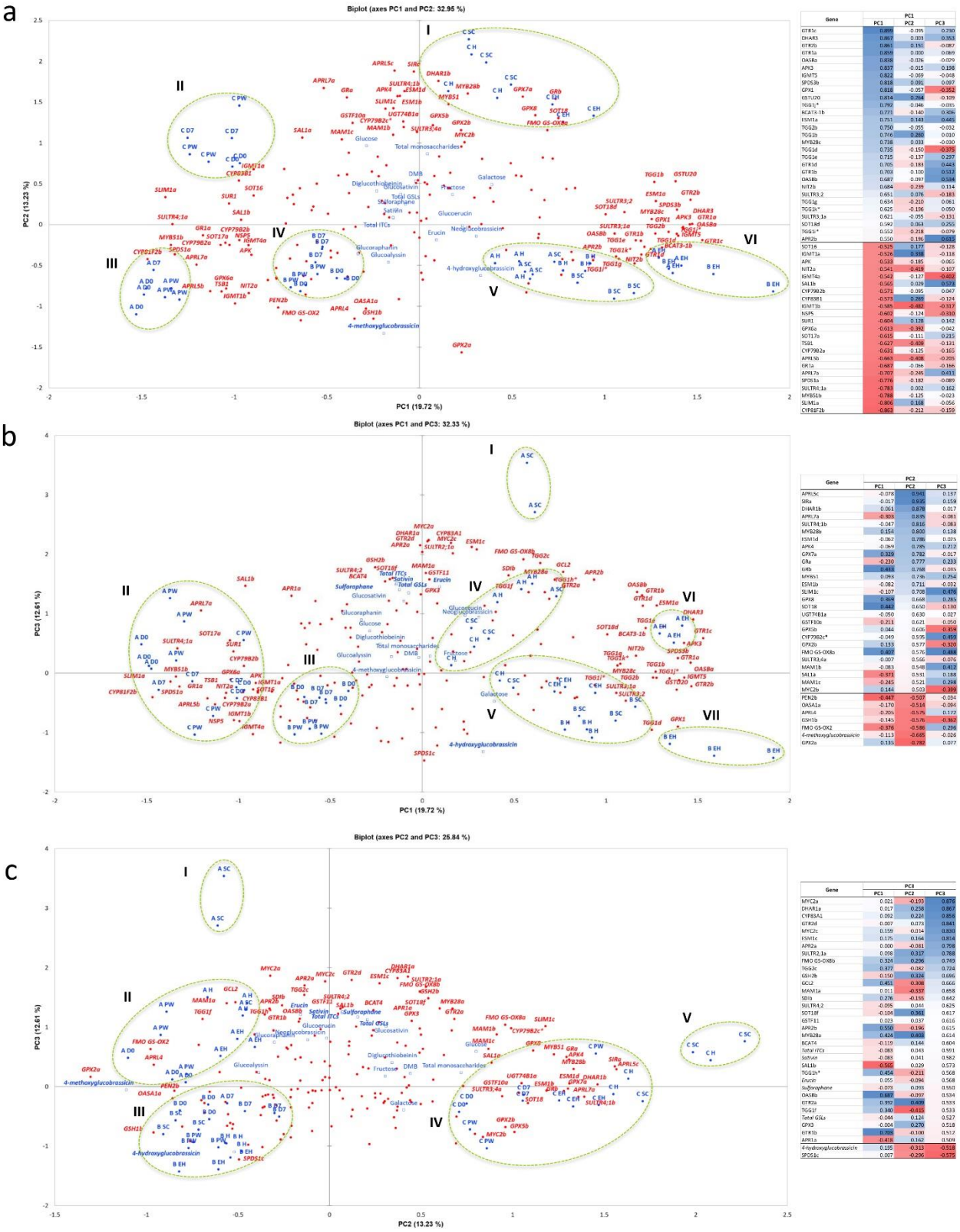


Figure 6. Principal Component Analysis of sulfur assimilation pathway, glucosinolate biosynthesis, and glucosinolate hydrolysis gene expression data

Principal Component Analysis of sulfur assimilation pathway, glucosinolate biosynthesis, and glucosinolate hydrolysis gene expression data (FPKM) for three *Eruca sativa* elite inbred lines (**A, B,** and **C**) across ontogenic and postharvest sample points. Biplot (**a**)

displays Principal Components (PCs) 1 and 2, which represent 33% of variation within the data. Biplot (**b**) displays PC1 and PC3, explaining 32.3% of variation within the data; and (**c**) displays PC2 and PC3, explaining 25.8% of the variability. The PCA plots presented are the results of Varimax rotation. Each biplot is accompanied by a factor loadings table sorted according to PC1 (**a**), PC2 (**b**), and PC3 (**c**); italics denotes a supplementary variable, and * = a putative novel gene within the reference annotation. Blue coloration denotes high factor loading scores, red denotes low. Only genes with loading values >0.5 were included, and each is represented within the biplots in red (bold italics). Red circles represent individual genes included in the analysis (*n* = 177). Blue circles represent sample point variables and have accompanying labels (blue bold). Blue squares denote phytochemical data regressed onto the PCA as supplementary variables. Bold data labels indicate phytochemical components with >0.5 factor loadings scores. Green dotted ellipses denote clusters of variables and are numbered using Roman numerals, which are quoted within the text. Abbreviations: early harvest (**EH**), harvest (**H**), second harvest (**SC**), pre-wash (**PW**), post-wash (**D0**), and seven-day shelf life (**D7**).

Supplementary Materials

Supplementary Table S1

Table S1. Summary of genome assembly and annotation of *Eruca sativa*

Genome assembly	>= 0 bp	>= 1,000 bp	Largest contig	Total (>= 500 bp)
Contig number	1,041,818	12,352	1,477,633	49,933
Total length	850,956,505	562,271,846		586,731,295

Assembly-related statistics

GC%	N50	NG50	N75	NG75	L50	LG50	L75	LG75
36.25	196,831	136,378	87,576	2,634	789	1,256	1,889	7,243

Supplementary Table S2

Table S2. Transposable elements content in the reference genome

Type	Denovo + Repbase*		TE proteins$		Combined TEs£	
	Length (bp)	**% in genome**	**Length (bp)**	**% in genome**	**Length (bp)**	**% in genome**
DNA	69,251,054	8.14	21,607,510	2.54	76,517,426	8.99
LINE	20,290,781	2.38	17,153,783	2.02	28,200,567	3.31
SINE	2,134,305	0.25	0	0	2,134,305	0.25
LTR	311,377,915	36.59	91,001,347	10.69	317,124,290	37.27
Other^	106,176	0.01	0	0	106,176	0.01
Unknown^^	155,033,031	18.22	0	0	155,033,031	18.22
Total	547,675,259	64.36	129,571,414	15.23	563,873,839	66.26

* = RepeatMasker based on the uclust algorithm combined with the known Repbase and *de novo* repeat library created by RepeatModeler / Repeat Scout / LTR_finder
$ = RepeatProteinMask based on Repbase
£ = The non-redundant set of results combining Denovo+Repbase TEs and TE proteins
^ = Repeats that can be classified by RepeatMasker, but not included by classes above
^^ = Repeats that could not be classified by RepeatMasker

Supplementary Table S3

Table S3. Predicted protein-coding genes within the *E. sativa* reference genome

Gene set		Number	Average gene length (bp)	Average CDS length (bp)	Average exons per gene	Average exon length (bp)	Average intron length (bp)
*De novo**	Augustus	50,179	1,701.56	1,024.89	4.48	228.88	194.57
	Glimmer HMM	73,989	1,335.36	725.53	3.03	239.16	299.86
	SNAP	80,264	1,231.13	728.3	4.02	180.98	166.27
	Geneid	100,165	2,127.25	585.82	3.07	191.03	745.87
	Genscan	71,813	3,942.04	785.32	3.92	200.11	1,079.4
Homolog^	*Arabidopsis lyrata*	32,667	1,867.18	1,084.09	4.86	223.12	202.93
	Arabidopsis thaliana	27,416	1,870.34	1,218.4	5.13	237.58	157.91
	Brassica napus	101,040	1,764.75	1,001.16	4.91	204.06	195.48
	Boechera stricta	27,416	2,006.68	1,181.2	5.09	231.86	201.61
	Capsella rubella	26,521	1,958.82	1,248.6	5.19	240.53	169.46
	Raphanus sativus	49,733	2,064.75	1,194.41	4.94	241.57	220.66
RNAseq^^	Cufflinks	43,200	2,848.16	1,723.58	6.03	285.65	223.41
	PASA	37,870	1,744.08	1,034.72	4.77	216.9	188.13
EVM		59,643	1,665.12	926.2	4.17	222.2	233.23
PASA-update		59,491	1,656.25	929.33	4.17	222.9	229.36
Final set		**45,438**	**1,889.6**	**1,069.44**	**4.76**	**224.81**	**218.3**

* = The combined results by EVM of *5-ab initio* gene predictions
^ = The combined results by EVM of homology-based gene prediction
^^ = = The combined results by EVM of transcriptome data sets

Supplementary Table S4

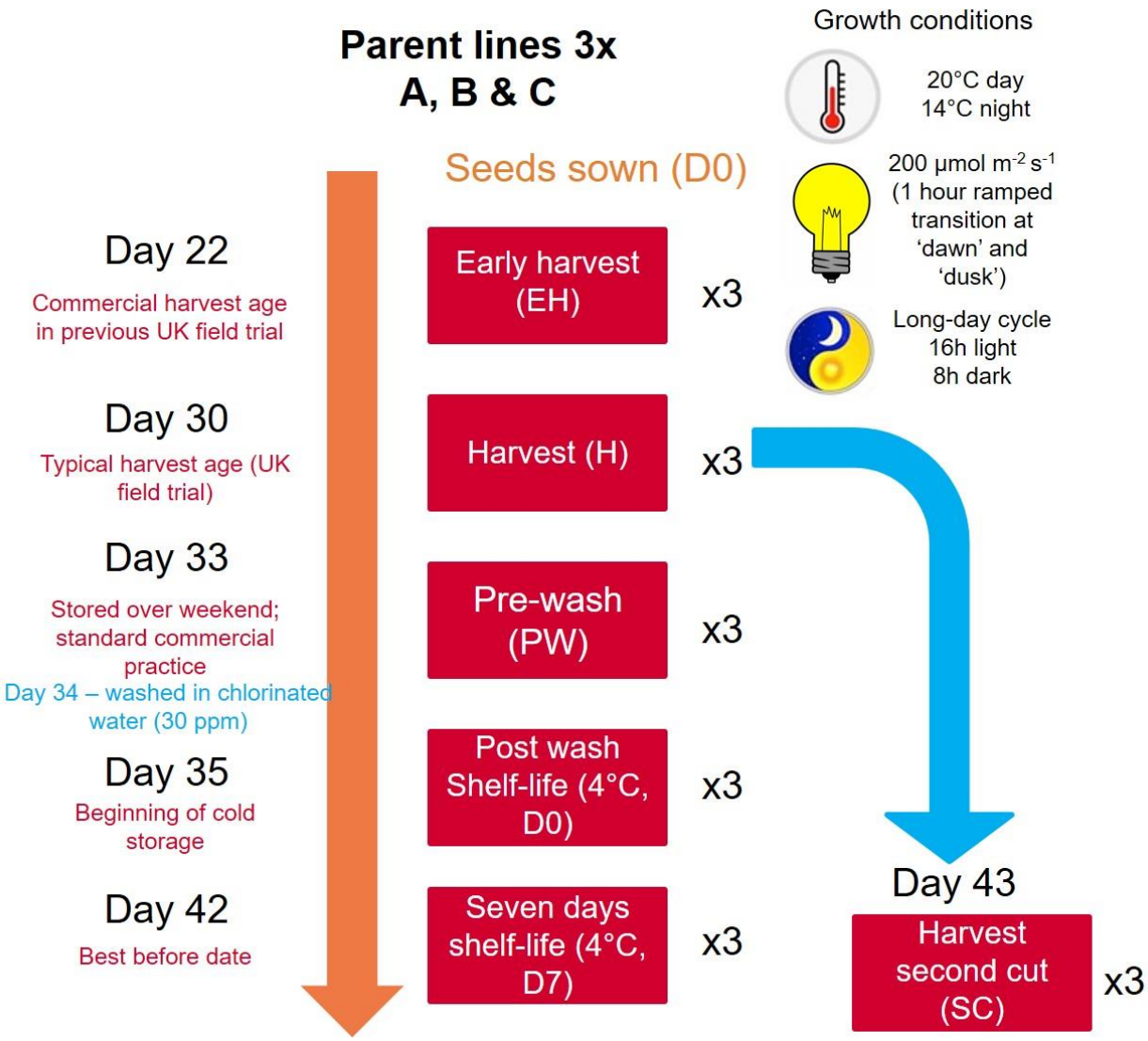
Table S4. Numbers of genes with homology or functional assignment

Database		Number of annotated genes	Gene annotation %
NR		44,630	98.2
Swiss-Prot		35,309	77.7
KEGG		32,589	71.7
InterPro	All	37,067	81.6
	Pfam	34,541	76
	GO	25,820	56.8
Annotated		44,655	98.3
Total		**45,438**	**-**

Supplementary Table S5

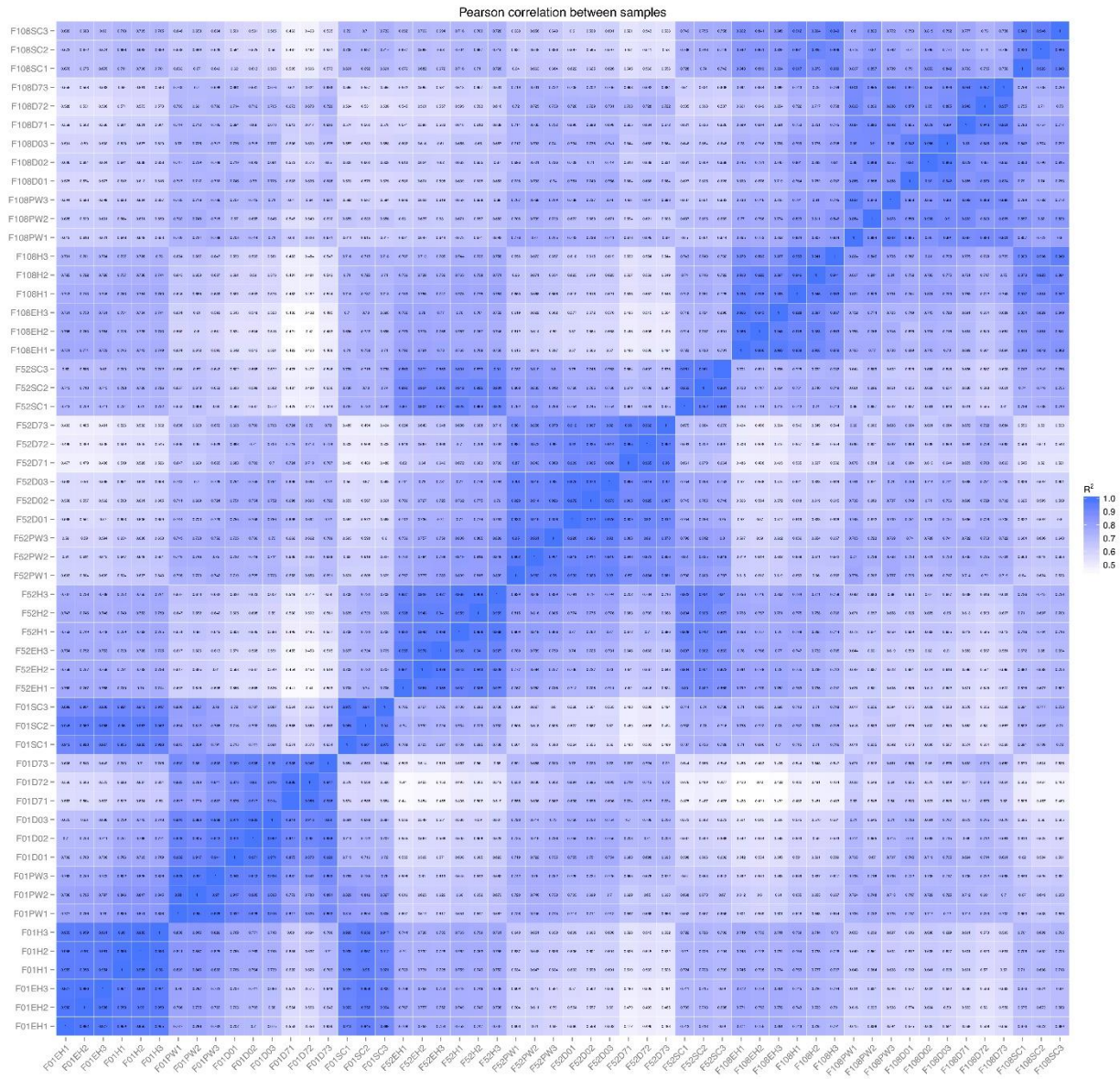
Table S5. qRT-PCR primers and efficiencies

Gene	Oligo	Length	tm	GC%	Sequence	Product size	Efficiency
MYB122-1a	LEFT	20	62.45	60.00	GGCGAACCCTACCCGACAAA	163	2.13
	RIGHT	23	60.31	43.48	GTGGACCATTTGTTGCCATGAAT		
MYB51a	LEFT	19	60.00	57.89	GGCGAACTCTCCCCGAAAA	164	1.93
	RIGHT	21	62.29	52.38	TGCAGCCCATTTGTTTCCGTG		
BCAT4	LEFT	20	62.35	55.00	TCGTCTCCGCCGTCAAACAA	198	1.91
	RIGHT	20	61.55	55.00	ACCCCGCGTTATCCTTGTGA		
SOT16	LEFT	21	63.06	57.14	CCCAACACAACGGACACTGGT	187	1.77
	RIGHT	20	62.74	60.00	AGAGGGTTCGTTGCGTCGTC		
CYP83B1	LEFT	23	59.73	43.48	TTTGGATATTGTTGTACCCGGGA	150	2.03
	RIGHT	20	59.12	55.00	TCTTCTTCCGAGACGTGTCC		
SUR1	LEFT	22	61.94	50.00	TCATTCAGGCTGCACTTCCTCA	105	1.9
	RIGHT	23	61.17	47.83	GCCTATCACACACCAAATCGACA		
TGG1b	LEFT	20	61.91	55.00	AGCTTCTTCATGGCCTCGCT	190	1.93
	RIGHT	22	63.20	59.09	CCCCTCCCTCCTTCAATCTGGT		
TGG1j	LEFT	21	61.51	52.38	TGGTACATGGAGCCGCTTACA	159	2.08
	RIGHT	21	61.76	57.14	TGACTGGGCGTACTGAGTGAC		
UGT74B1a	LEFT	20	62.99	65.00	CAGCATCGACGCCTACTCCG	247	1.71
	RIGHT	20	61.26	60.00	GGGAACAGCGGGGAGAGTAA		
TGG1d	LEFT	21	62.26	52.38	TCAGAAGACCGTTGCCAAGCT	104	1.84
	RIGHT	22	61.55	50.00	ACGTTGACACCCTTCTCCTTGA		
ACT11	LEFT	20	62.99	60.00	TTCACCACCACAGCAGAGCG	165	2.19
	RIGHT	20	62.59	65.00	CCTCTCCCCTCCGATGGTGA		



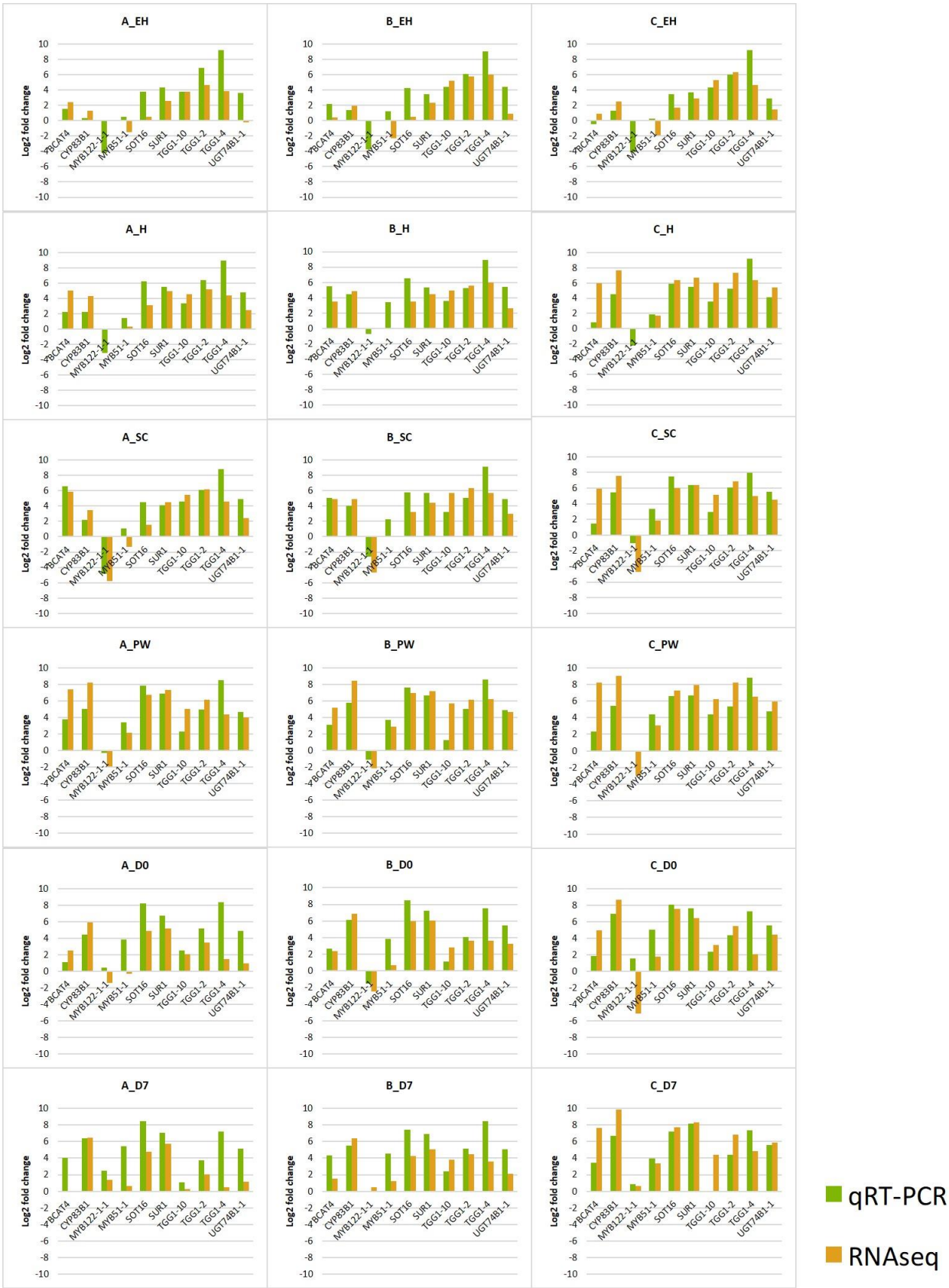
Supplementary Figure S1

RNAseq experimental design and sampling diagram. Three elite inbred lines of *Eruca sativa* were grown under controlled environment conditions and sampled at each of the six time points indicated (in triplicate). Abbreviations: early harvest (**EH**), harvest (**H**), second harvest (**SC**), pre-wash (**PW**), post-wash (**D0**), and seven-day shelf life (**D7**).



Supplementary Figure S2

Pearson correlation matrix of RNAseq biological sample replicate gene expression values. Replicates of each sample showed a high degree of correlation (r^2 = >0.884) indicating robust reproducibility of gene expression between the individual plants tested at each respective sample point.



Supplementary Figure S3

qRT-PCR (green) vs. RNAseq (orange) gene expression of ten randomly selected glucosinolate biosynthesis and hydrolysis-related genes. Data are expressed as the normalized log2 fold-change in expression relative to the reference gene *ACT11*. ANOVA

revealed no significant difference between the two data sets. Abbreviations: early harvest (**EH**), harvest (**H**), second harvest (**SC**), pre-wash (**PW**), post-wash (**D0**), and seven-day shelf life (**D7**).

Supplementary Data File S1

Analysis of Variance outputs with Tukey's HSD pairwise comparisons between sample points and each respective rocket breeding line: Tab 1 – glucosinolate analysis; Tab 2 – glucosinolate hydrolysis product analysis; Tab 3 – sugar analysis. Highlighted values are significant at the following levels: P = <0.05 (yellow), P = 0.01 (orange), and P = 0.001 (green). Tab 4 contains a Pearson's correlation analysis matrix for sulfur and glucosinolate-related gene expression values and phytochemical observations. Values in bold are significant correlations at the P = 0.001 threshold.

Supplementary Data File S2

RNAseq read counts, log2-fold changes, *P*-values, and adjusted *P*-values (padj) for sulfur metabolism, glucosinolate biosynthesis, hydrolysis, and transport genes for each of the three rocket lines and the respective sample points. Significant up/down regulation is denoted by green/red highlighting, respectively. KEGG annotation numbers and UniProt gene descriptions for orthologous genes in *Arabidopsis thaliana* are provided.