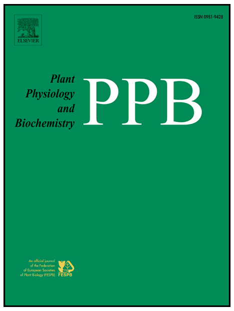


Exploring glucosinolates diversity in Brassicaceae: A genomic and chemical assessment for deciphering abiotic stress tolerance

Anyse Pereira Essoh, Filipa Monteiro, Ana Rita Pena, M. Salomé Pais, Mónica Moura, Maria Manuel Romeiras

PII: S0981-9428(20)30086-3

DOI: https://doi.org/10.1016/j.plaphy.2020.02.032

Reference: PLAPHY 6066

To appear in: *Plant Physiology and Biochemistry*

Received Date: 2 December 2019

Revised Date: 29 January 2020

Accepted Date: 20 February 2020

Please cite this article as: A.P. Essoh, F. Monteiro, A.R. Pena, M.Salomé. Pais, Mó. Moura, M.M. Romeiras, Exploring glucosinolates diversity in Brassicaceae: A genomic and chemical assessment for deciphering abiotic stress tolerance, *Plant Physiology et Biochemistry* (2020), doi: https://doi.org/10.1016/j.plaphy.2020.02.032.

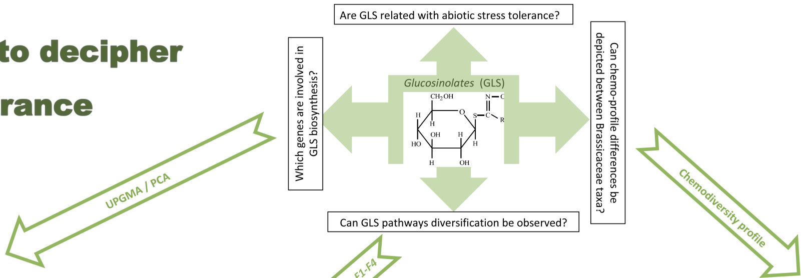
Brassicaceae diversity to decipher abiotic stress tolerance

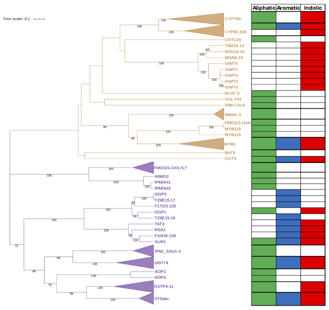
Are GLS related with abiotic stress tolerance?

Which genes are involved in GLS biosynthesis?

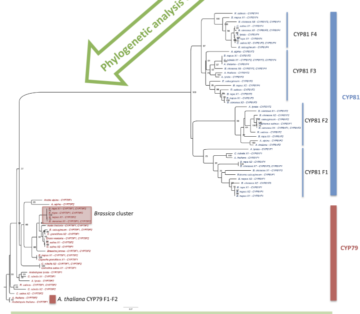
Can chemo-profile differences be depicted between Brassicaceae taxa?

Can GLS pathways diversification be observed?

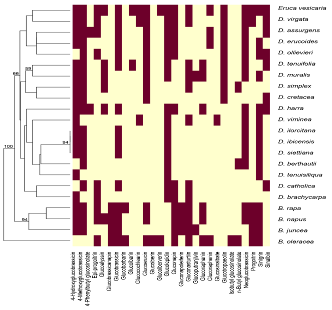




Separation between genes in GLS core structure and *CYP450/MYB* gene families.



Recent diversification of aliphatic genes and an earliest for indolic genes



Distinct GLS chemo-profile between *Brassica* crops and *Diplotaxis* species (wild relatives)

Exploring glucosinolates diversity in Brassicaceae: a genomic and chemical assessment for deciphering abiotic stress tolerance

Anyse Pereira Essoh[1,2,3*], Filipa Monteiro[1,4* †], Ana Rita Pena[4], , M. Salomé Pais[5], Mónica Moura[2], Maria Manuel Romeiras[1,4,5 †]

[1]Linking Landscape, Environment, Agriculture and Food (LEAF), Instituto Superior de Agronomia, Universidade de Lisboa, Lisboa, Portugal.
[2]Research Centre in Biodiversity and Genetic Resources (CIBIO), InBIO Associate Laboratory, Faculdade de Ciências e Tecnologia, Universidade dos Açores, Ponta Delgada, Portugal.
[3]Nova School of Business and Economics, 2775-405 Campus de Carcavelos, Portugal.
[4]Centre for Ecology, Evolution and Environmental Changes (cE3c), Faculdade de Ciências,Universidade de Lisboa, Lisboa, Portugal.
[5]Academia das Ciências de Lisboa, Rua Academia das Ciências 19, 1200-168 Lisboa, Portugal.
*** authors contributed equally to this work.**

[†] **Correspondence:**
fimonteiro@fc.ul.pt; mmromeiras@isa.ulisboa.pt

Running title: Brassicaceae diversity to decipher abiotic stress tolerance

ABSTRACT

Brassica is one of the most economically important genus of the Brassicaceae family, encompassing several key crops like *Brassica napus* (cabbage) and broccoli (*Brassica oleraceae* var. *italic*a). This family is well known for their high content of characteristic secondary metabolites such as glucosinolates (GLS) compounds, recognize for their beneficial health properties and role in plants defense. In this work, we have looked through gene clusters involved in the biosynthesis of GLS, by combining genomic analysis with biochemical pathways and chemical diversity assessment. A total of 101 Brassicaceae genes involved in GLS biosynthesis were identified, using a multi-database approach. Through a UPGMA and PCA analysis on the 101 GLS genes recorded, revealed a separation between the genes mainly involved in GLS core structure synthesis and genes belonging to the *CYP450*s and *MYB*s gene families. After, a detailed phylogenetic analysis was conducted to better understand the disjunction of the aliphatic and indolic genes, by focusing on *CYP79F1-F2* and *CYP81F1-F4*, respectively. Our results point to a recent diversification of the aliphatic *CYP79F1* and *F2* genes in *Brassica* crops, while for indolic genes an earliest diversification is observed for *CYP81F1-F4* genes. Chemical diversity revealed that *Brassica* crops have distinct GLS chemo-profiles from other Brassicaceae genera; being highlighted the high contents of GLS found among the *Diplotaxis* species. Also, we have explored GLS-rich species as a new source of taxa with great agronomic potential, particularly in abiotic stress tolerance, namely *Diplotaxis*, the closest wild relatives of *Brassica* crops.

Keywords: chemical diversity; genomic diversity; GLS; abiotic stress; *Brassica* crops; *Diplotaxis*

1. Introduction

The Brassicaceae is one of the world's most economically important plant family (Ishida et al., 2014). It includes important crop species such as *Brassica oleracea* (e.g., cauliflower, Brussels sprouts, cabbage, broccoli, and Kai Lan), *Brassica rapa* (e.g., pakchoi, choy sum, and Chinese cabbage), *Nasturtium officinale* (e.g., watercress), and *Raphanus sativus* (e.g., daikon radish and red cherry radish). Other species such as *Diplotaxis tenuifolia* and *Eruca vesicaria*, commonly referred as 'rocket salads', have also attracted a considerable interest as culinary vegetables because of their strong flavor and content of putative health-promoting compounds (Verkerk et al., 2010). These species and their crop wild relatives (CWR – taxa closely related to crops) grown primarily in the Euro-Mediterranean region, which contains the highest proportion of agronomically important plants representing an important reservoir of genetic resources for crop improvement (Kell et al., 2008). CWR are likely to contain a great genetic diversity necessary to combat climate change because of the diversity of habitats in which they grow and the wide range of conditions they are adapted to (Ford-Lloyd et al., 2011).

Among the most important chemical compounds produced by Brassicaceae species are the Glucosinolates (GLS), which proved to have health promoting effects and importance in abiotic stress tolerance (Cartea and Velasco, 2007). They are constituted by a common structure comprising a β-D-thioglucose group, a sulfonated oxime moiety and a variable side-chain derived either from methionine, tryptophan, phenylalanine, or from other branched chain amino acids. GLS are found in 16 dicotyledonous plant families where, at least, 130 different structures have been identified so far (Fahey et al., 2001; Collett et al., 2014).

GLS are present at different concentrations throughout the plant organs. They can reach 1% of the dry weight in some tissues of *Brassica* (Fahey et al., 2001). Within a single species, up to 4 different GLSs dominate the GLS occurrence in the plant (Verkerk et al., 2008). The type, concentration and distribution of the GLS in the plants of Brassicaceae family vary according to a high number of factors, namely species (Bellostas et al., 2004), variety (Choi et al., 2014), plant organ (Brown et al., 2003; Bellostas et al., 2004) or plant age (Fahey et al., 1997; Brown et al., 2003) and developmental cycle. Moreover, environmental conditions such as season (Cartea et al., 2007), biotic (Verkerk et al., 2008) or abiotic stress factors such as salinity or drought, are also known to play a role on the production and content of these compounds (Khan et al., 2011; Martínez-Ballesta et al., 2015).

Recent studies have revealed that GLS and their derivatives have beneficial effects on humans. They can help in suppressing tumor growth of various types of cancers namely: breast, brain, blood, bone, colon, gastric, liver, lung, oral, pancreatic and prostate (Zhang et al., 2003; Soundararajan and Kim, 2018). Significant reduction in plasma LDL-C levels has also been reported as being directly linked to consumption of GLS-rich broccoli (Armah et al., 2015). Some GLS derived products are reported to have antimicrobial effects and well documented health benefits (Cavaiuolo and Ferrante, 2014; Bischoff, 2016). Exclusive or excessive feeding of vegetables and/or seeds from the *Brassica* plants have been associated with toxic effects in livestock (VanEtten and Tookey, 1983; Tripathi and Mishra, 2007) and strategies have been explored to reduce GLS content in *Brassica* vegetables to increase their palatability for animal consumption (Verker et al., 2008).

The GLS biosynthetic pathway has been partially elucidated by studies on *Arabidopsis* (e.g. reviewed in Grubb and Abel 2006; Halkier and Gershenzon 2006). The GLS, synthesized from amino acids, are grouped in three subtypes according to their corresponding precursors: i) aliphatic GLS, derived from alanine, leucine, isoleucine, valine, and methionine; ii) indole GLS, derived from tryptophan; and iii) aromatic GLS, derived from phenylalanine and tyrosine (Fahey et al., 2001; Halkier and Gershenzon, 2006). Different authors have reported on aliphatic GLS accounting for 70–97% of the total GLS content in leaves of *Brassica oleracea* (Cartea et al., 2007), leaves and stems of *Brassica napus* (Cleemput and Becker, 2012), leaves and seeds of *Brassica juncea* (Gupta et al., 2012; Othmane, 2015), and sprouts and mature leaves of *Brassica rapa* (Wiesner et al., 2013). The formation of the GLS core structure involves the action of enzymes from different families, namely the CYP79 (Hansen et al., 2001; Chen et al., 2003), CYP83 (Bak and Feyereisen, 2001), UGT74 (Grubb et al., 2014), C-S-lyases (Mikkelsen et al., 2004) and of sulfotransferases (SOTs or STs) (Piotrowski et al., 2004). These enzymes are involved in the biosynthesis of basic GLS structures from elongated and non-elongated amino acids. The basic GLS structures are subjected to a range of secondary side chain modification and transformation pathways catalyzed by enzymes such as flavin monooxygenase (FMOOXs) (Hansen et al., 2007), GLS-AOPs (Mithen et al., 1995), GLS-OH (Hansen et al., 2008) and CYP81Fs (Pfalz et al., 2009; 2011) to generate different types of GLS structures, that are the last finalizing gene family involved in the indolic biosynthetic pathway (Clarke, 2010; Fahey et al., 2001).

The most important mechanism for the wide production of secondary metabolites as glucosinolates relies on whole-genome events, which occurred in Brassicaceae evolution history (Kliebenstein et al., 2001a,b; Kroymann, 2011). The availability of the whole-genome sequences gives an opportunity for using comparative genomics, which, in turn, can lead to a better understanding of the genome evolution in this family. Whole-genome sequences are available for more than 100 plant species (Tohge et al., 2014). The massive contribution, resulting from next-generation technologies, cannot be currently matched by metabolomics, especially if high-quality and species-optimized approaches are adopted (Fukushima et al., 2014). With the increasing number of whole-genome sequences and the freely available genomic resources, the opportunities for conducting an analysis based on comparative genomics is foreseen.

In this paper, we investigated gene clusters involved on the biosynthesis of GLS, by combining genome analysis with biochemical pathways and compound structure assessment. Considering the high diversity in GLS content in Brassicaceae species, we aim to: i) contribute to the global GLS gene inventory in Brassicaceae; ii) compare gene diversity within the three GLS sub-pathways; iii) assess a potential genetic basis for GLS divergence using 6 CYP genes (*CYP79F1-F2* and *CYP81F1-F4*), known to be key genes of indolic and aliphatic GLS biosynthetic pathways, respectively; and iv) increase the knowledge on the chemical diversity of GLS compounds in major *Brassica* crops compared to the CWR of the genus *Diplotaxis*. By combining chemical data with genomic sequences, we expect to provide information of interest for promoting the use of the neglected *Diplotaxis* genus as a potential viable CWR of economically important *Brassica* crops.

2. Materials and Methods

2.1. GLS biosynthetic genes: compilation and gene ontology annotation

QuickGO (https://www.ebi.ac.uk/QuickGO/, Binns et al., 2009), AmiGO (http://amigo.geneontology.org/amigo, Carbon et al., 2008) and MetaCyc (https://metacyc.org/, Caspi et al., 2017) databases were used to filter genes involved in GLS biosynthetic process (GBP) by searching the specific GO term (GO:0019761). Sequences representing the complete set of GLS biosynthetic genes in *Arabidopsis thaliana* were acquired from The Arabidopsis Information Resource (TAIR, www.arabidopsis.org, accessed on July 2019, Berardini et al., 2015), and further complemented with a set of genes listed as GLS genes in the Brassica database (BRAD, http://brassicadb.org, Wang et al., 2015), which is a web-based database of genetic data at the whole genome scale for important *Brassica* crops. After, a complete assessment of GLS biosynthetic genes in Brassicaceae species was retrieved, through searching of several public databases namely: Arabidopsis Information Resource (TAIR), BrassicaDB, and nucleotide blast (Blastn) at NCBI, restricting the search to orthologs within the Brassicaceae family. The genes sequences listed as GLS genes in BRAD, were subjected to nucleotide Blast (Blastn on TAIR), to identify *Arabidopsis thaliana* homologous genes with a threshold of E-value $\leq 10^{-10}$. The following step was to perform a complete assessment of GLS biosynthetic genes in Brassicaceae species by using the BLASTN algorithm in National Center for Biotechnology Information (NCBI) public database, restricting the search for orthologs within the Brassicaceae family, with a threshold of E-value $\leq 10^{-10}$ and 50% of query cover. Blast2GO v.5.2 (Götz et al., 2008) was used to assign GO terms to the sequences dataset, to allow unigene annotation according to three main Gene Ontology categories, i.e. Cellular Compartment, Molecular Function and Biological Process. A BlastX algorithm was used with the following parameters: a constant expectation value threshold of $1.0E^{-10}$, 20 Blast Hits, HSP length cutoff set at 33 and HSP Hit Coverage at 60. The different genomic information gathered from a multi-databasing approach was represented by an Euler diagram using the online generator tool available at https://www.meta-chart.com/. The resulting figure (Figure 2) was scaled, so that the area of the shape was proportional to the number of genes it contained, and the overlapping shapes represented the genes that were present in more than one database.

2.2. Gene clustering analysis

The collected GLS biosynthetic genes were used to perform a gene cluster analysis under two different approaches: unsupervised Principal Component Analysis (PCA) and a UPGMA. The PCA analysis was carried out using *factoextra* package in R version 3.6.1 through RStudio version 1.2.5001. To carry out the UPGMA analysis, a dataset containing the 78 GLS gene sequences assigned to each of the sub-pathways was analyzed using MEGA X version 10.0.5 (Kumar et al., 2018). A model assessment was performed to calculate the most adequate model to the dataset, and subsequently, a UPGMA analysis was constructed using 10000 bootstraps. Phenograms were edited using FigTree version 1.4.4 (Rambaut, 2009).

2.3. Phylogenetic analysis of CYP79F and CYP81F genes

Sequences from *Arabidopsis thaliana CYP79F1-F2* and *CYP81F1-F4* were retrieved from the TAIR database. Brassicaceae orthologs were assessed by blasting genes from *Arabidopsis thaliana* against the NCBI database using Blastn, with an E-value of $\leq 10^{-10}$ and 50% of query cover, restricted to the Brassicaceae family. A total of 101 sequences were retrieved and analyzed, where only 69 were marked as unique (i.e. not shared across genes). The final dataset comprised 25 sequences from *CYP79F1-F2* [*CYP79F1*: n=8, *CYP79F2*: n=7 and shared: n=10] and 44 from *CYP81F1-F4* [*CYP81F1*: n=9, *CYP81F2*: n= 9, *CYP81F3*: n= 10, *CYP81F4*: n= 9 and shared: n=7]. Sequences were aligned using MAFFT version 7 auto strategy (Katoh et al., 2017) and then trimmed using trimAl version 1.3 available at the Phylemon 2 suite (http://phylemon.bioinfo.cipf.es/) under the automated1 algorithm. Model calculations were carried out using PartitionFinder2 (Lanfear, 2017) and then phylogenetic estimations were made using RAxML version 8.2.10 through raxmlGUI version 1.5b2 using a ML+ rapid bootstrap, autoMRE, using *Arabidopsis thaliana* CYP79 genes as outgroups. Lastly, visualization and manipulation of the trees was done using FigTree version 1.4.4 (Rambaut, 2009).

2.4. GLS compounds assessment

Major agricultural brassica crops (i.e. *Brassica* sp., *Eruca vesicaria*) were selected and compiled for GLS compounds analysis through an exhaustive literature review (*Brassica rapa* – Cartea et al., 2012; *Brassica napus* – Velasco et al., 2008; *Brassica olearaceae* – Bhandari et al., 2015; *Eruca vesicaria*- D'Antuono et al., 2008). *Diplotaxis* species were also included in the GLS chemodiversity analysis as being probable precursors and wild relatives of *Brassica* crops (D'Antuono et al., 2008). A matrix of presence/absence was built and then projected as a heat map using the Heatmap tool freely available (http://www.hiv.lanl.gov/) using the Euclidean distance method with an average linkage clustering, and 10000 bootstraps.

3. Results

3.1. Genomic information on GLS genes

The species diversity assessment carried out on GLS genes available at public databases enabled the identification of 101 *Arabidopsis* genes that were blasted using Blastn (NCBI) restricted to *Brassicaceae*. From the results obtained, 36 species contain information on orthologous genes belonging to the GLS metabolic pathway. As expected, the most represented species was *Arabidopsis thaliana*, which accounted for 32% of the total GLS available genes. Other species, in particular the major crop species *Brassica napus*, *Brassica oleracea and Brassica rapa*, display 37% of the genomic information available at public databases. *Raphanus sativus* (radish) comprised 8% of the data, with other Brassicaceae model species, namely *Camelina sativa*, *Capsella rubella*, *Arabis alpina* and *Eutrema salsugineum* complementing the remaining genomic information available on GLS genes (Figure 1).

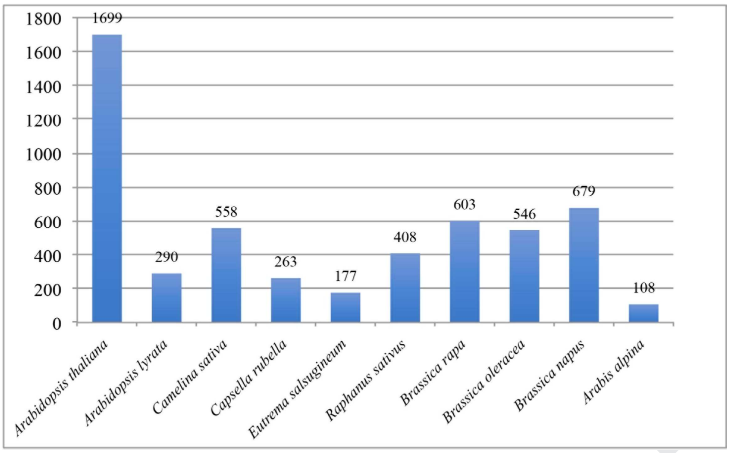


Figure 1- Genomic information of the GLS biosynthetic genes of the top ten Brassicaceae species registered at NCBI database. Number of sequences available per species is represented above each bar.

3.2. Global assessment of GLS biosynthetic genes identification

A global overview of the GLS biosynthetic pathway in the Brassicaceae family was developed using a multi-database approach. Although several studies have already been performed to achieve a similar pathway reconstruction analysis, we provide in our paper a global assessment of the GLS biosynthetic pathway using not only genes described for *Arabidopsis* but also for *Brassica* species. To do so, we retrieved all the genes belonging to the GLS biosynthetic pathway using its specific GO term (GO:0019761) (Supplementary Table 1). From this thorough inventory, a total of 101 genes were identified in *Arabidopsis thaliana* as being GLS biosynthetic genes: 52 from AmiGO, described as being involved in the GLS biosynthesis (GO:0019761); 52 from Brassica database (Brassicadb.org) classified as GLS genes and 67 from MetaCyc that were present in the GLS synthesis reaction cascade (Figure 2).

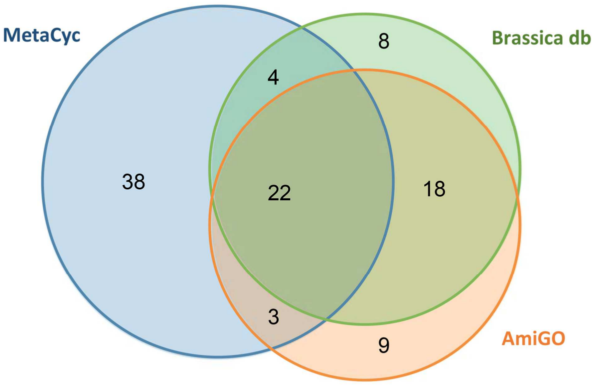


Figure 2- Euler diagram displaying GLS gene annotation gathered from a multi-database approach.

Using MetaCyc, it was possible to assign the genes according to each of the GLS sub pathways: aromatic, indolic and aliphatic. From the total 101 genes, 78 were assigned to each of the sub-pathways, while the remaining 23 were not reported as biosynthetic specific genes, and thus their assignment remains unclear. This is probably due to putative functions related to substrate diversity and regulation of GLS synthesis. Using this database, it was possible to identify 31 genes specific from aliphatic GLS, 26 genes specific from indolic GLS synthesis, and 6 genes specific from aromatic GLS biosynthesis (Table 1, for specific genes see Supplementary Table 2).

Table 1- GLS genes information according to sub-pathways of indolic, aliphatic and aromatic. Number of genes - total of genes annotated in each sub-pathway; Number of specific genes - genes exclusive to a given sub-pathway; Number of shared genes - genes shared in at least two sub-pathways.

	Aliphatic	Aromatic	Indolic	Combined unigenes of the 3 pathways
Nº. Genes	40	20	41	78
Nº. Specific Genes	31	6	26	-
Nº. Shared Genes	9	14	14	8

Gene Ontology (GO) assignment revealed a high diversity regarding the three multi-level categorizations: Biological Process (BP), Molecular Function (MF) and Cellular Component (CC) (Figure 3). On the first categorization level, the top-hits of biological process (with more than 45 sequences) were related to metabolic and cellular processes, followed by response to stimulus. In the molecular function level, binding and transcription regulation activities were the most represented after the catalytic activity; while in the cellular component level, genes were mainly grouped by membrane and/or organelle. These GO terms tie in with GLS biosynthetic functions, like the transcription regulation activities attributed to the MYB gene family, known to act as transcription factors/regulators of GLS unique to the GLS-synthesizing Brassicales (*MYB34*, *MYB51* and MYB122 in indolic pathway; *MYB28*, *MYB29* and *MYB76* in aliphatic GLS). These hints at possible unknown GLS functions need to be further explored to fully assign and determine the complete gene functions in the Brassicaceae GLS biosynthetic pathway.

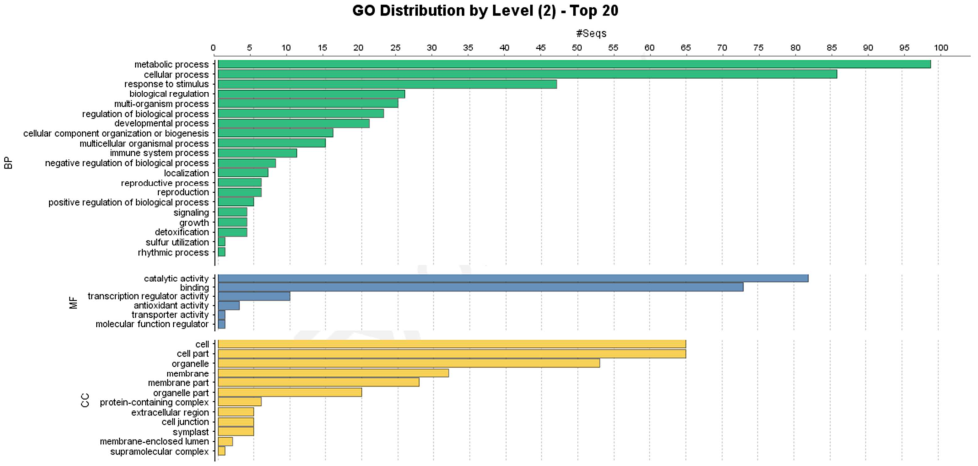


Figure 3- Gene Ontology (GO) terms assignment for the GLS biosynthetic genes. The graph displays the term enrichment levels of the annotated sequences along with the GO term hierarchy: Biological Process (BP, in green), Molecular Function (MF, in blue) and Cellular Component (CC, in yellow).

3.3. Uncovering GLS sub-pathways gene specificity

In order to detect sub-pathway specificity of GLS genes, two different methodologies have been applied: 1) UPGMA as a bottom-up hierarchical clustering method to evaluate gene clustering based on sequence alignment, regardless evolutionary features; 2) Principal Component Analysis (PCA) to assess gene grouping discrimination associated to each sub-pathway.

An UPGMA phenogram (Figure 4) allowed the discrimination of two main clusters: one highlighted in purple, which include many genes shared by the 3 subtypes pathways (aliphatic / indolic / aromatic) and mostly related to the synthesis of GLS core structure; and a second cluster in orange with many genes belonging to the *CYP450*s and *MYB*s gene families, which are essentially genes related to side chain elongation of GLS, regardless of being indolic or aliphatic, and which are known to be responsible for the great diversity of existing compounds.

PCA analysis of the GLS genes (Supplementary Figure 2) showed a shared membership with no discrimination between the three sub-pathways (indolic, aliphatic and aromatic). These results are corroborated by the UPGMA phenogram where no pathway-specific clustering was identified. The analysis of PCAs loading plots (Supplementary Figure 2), PCA1 reveals 26% of the total variation, while PCA2 accounts for 13.5%. The PCA1 variation appears to be connected with a group composed of CYP79 and CYP81 genes. Interestingly, these genes belong to different GLS pathways: *CYP81F1-F4* is indolic-specific while *CYP79F1-F2* is exclusive to the aliphatic pathway. Only *CYP83A1* and B1, and *CYP79A2* are shared within the three pathways.

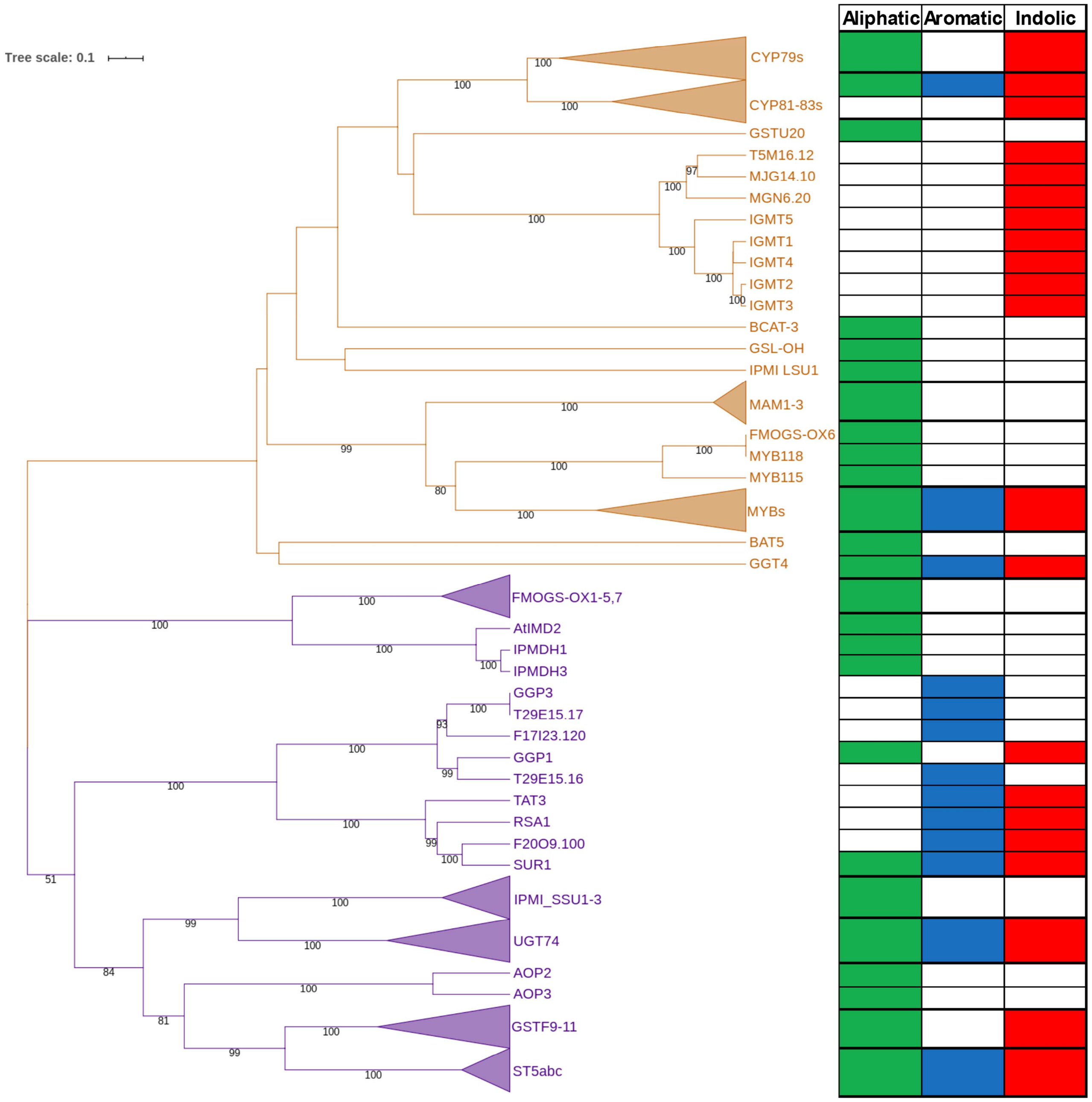


Figure 4- UPGMA phenogram of the 101 GLS biosynthetic genes. Bootstrap values above 50 are represented on the branches. Detailed UPGMA tree is available at Supplementary Figure 1.

3.4. Testing gene divergence as a baseline to GLS diversification

Considering the two approaches by UPGMA and PCA, which disclosed a potential clustering of *CYP79* and *CYP81* genes (*CYP79F1-F2* and *CYP81F1-F4*), a more detailed phylogenetic analysis was conducted to understand the disjunction of the aliphatic and indolic genes. Considering that side chain modifications of indolic GLS are controlled by four CYP81F enzymes (*CYP81F1- F4*) (Barco and Clay, 2019), while *CYP79F1* and *CYP79F2* are involved in the biosynthesis of aliphatic GLS in *Arabidopsis thaliana*, by exploring sequence diversification on Brassicaceae orthologs, a possible differentiation of those genes could be uncovered on *Brassica* crops, which display a higher diversity of aliphatic compounds than *Arabidopsis*. Overall, the obtained ML phylogeny is well supported, resulting in two main clades, which separate the genes associated with the aliphatic pathway (*CYP79F1* and *CYP79F2*) and those associated to the indolic pathway (*CYP81F1*, *CYP81F2*, *CYP81F3* and *CYP81F4*) (Figure 5). The phylogenetic analysis clearly splits the two types of CYPs analyzed, *CYP79* (in red) and *CYP81* (in blue), into two well-supported clusters. As such, it appears that the gene divergence between these CYPs underlies the basis of indolic and aliphatic GLS biosynthesis.

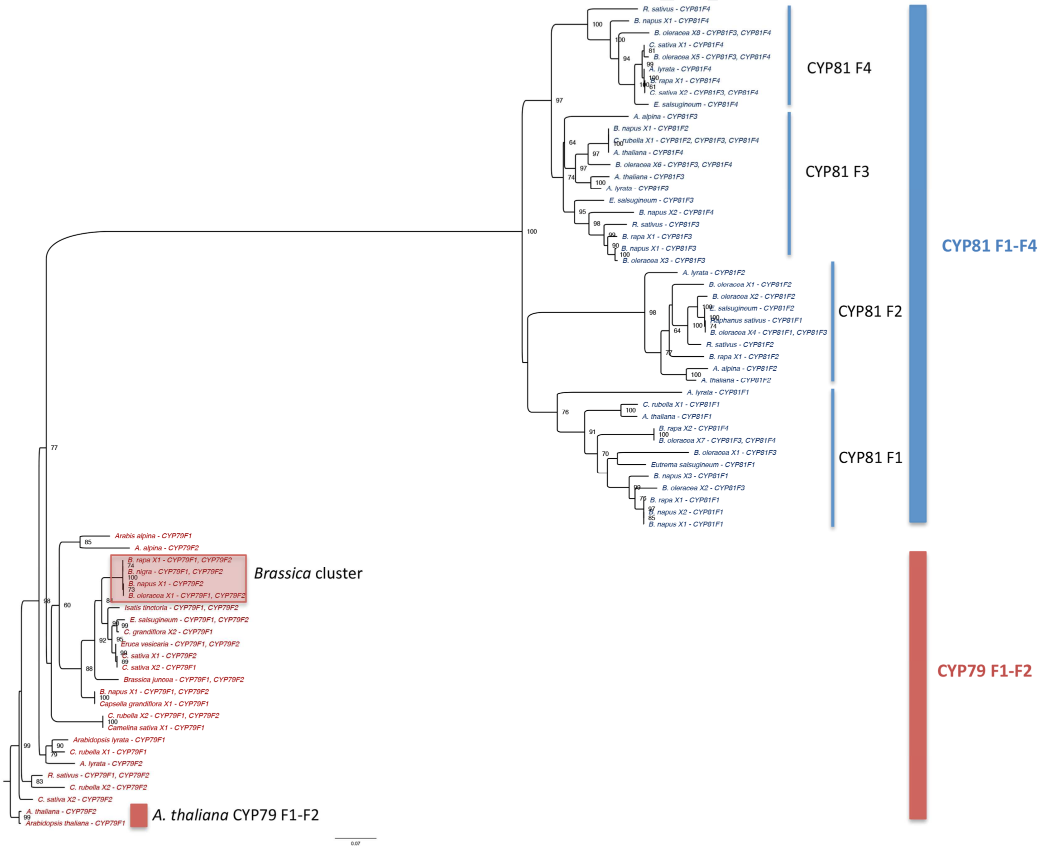


Figure 5- Phylogenetic tree from the Maximum Likelihood analysis of *CYP79F1-F2* and *CYP81F1-F4* genes in Brassicaceae with *A. thaliana CYP79F* genes as outgroups. Acronyms are present as the first letter of the genus and the second to species, e.g. At for *Arabidopsis thaliana*, and gene identification when possible. Upon lack of complete CYP annotation, accession numbers were used. Different copies of the same gene are identified by an "X" following sequential numbering, e.g. *A. thaliana* X1, *A. thaliana* X2. Only bootstrap values above 50 are presented. Accession numbers of the sequences analyzed are provided in Supplementary Table 3.

From the analysis of the *CYP79* genes, a cluster including all major *Brassica* crops (*Brassica rapa, Brassica juncea, Brassica olearacea*) is evident (Figure 5, highlighted as *Brassica* cluster), where no disjunction is observed from being *CYP79F1* or *CYP79F2*. It can be easily recognize that *Brassica* crops are usually grouped in the same cluster, which reveals a common diversification of indolic GLS that portrays *Brassica* chemotypes. Two apparent copies of *Brassica napa CYP79F2* are grouped while other *Brassica* sp. sequences were assembled in different tree branches disclosing a wide divergence on the *CYP79F1* and *F2* gene sequences which could be associated with the diversity of aliphatic GLS in *Brassica* s.l. Regarding *CYP81F1-F4*, four clusters were obtained matching essentially each of the *CYP81F* genes covered (Figure 5).

3.5. Snapshot on GLS chemodiversity: *Brassica* crops and *Diplotaxis*

By performing a snapshot of the GLS chemodiversity using an average linkage clustering method (Figure 6), a cluster including the *Brassica* crops (e.g. *Brassica olearacea*, *Brassica juncea*, *Brassica rapa*, and *Brassica napus*) can be depicted, while *Diplotaxis* species appear to have a more complex and diversified GLS chemical profile. Phylogenetic relationships indicate that *Diplotaxis* maintains most of the primitive morphological characters while *Brassica* presents the most evolved ones with *Erucastrum* occupying an intermediate position (Gómez-Campo and Tortosa, 1974; Gómez-Campo, 1980; Sánchez-Yélamo, 2009). By comparing the GLS chemotype diversification between *Brassica* and *Diplotaxis* species, the latter shows a distinct GLS profile. In what concerns rocket crops, collectively attributed to *Diplotaxis* and *Eruca,* the wild (*Diplotaxis tenuifolia*) and cultivated (*Eruca sativa* and *Eruca vesicaria*) rockets are clustered together sharing a common GLS profile.

The results obtained revealed that *Brassica* and *Diplotaxis* have distinct GLS chemo-profiles. Within *Brassica* species, a shared GLS profile is displayed, namely in what concerns aliphatic GLS such as progoitrin, gluconapin, glucobrassicanapin that are specific to *Brassica* chemo-lineage. In *Diplotaxis* and *E. vesicaria*, glucolepidin appears as the main distinctive GLS, followed by glucoerucin. Moreover, such GLS are more diverse among *Diplotaxis* species than in *Brassica* species, possibly as the result of crop selection events that have narrowed *Brassica* chemodiversity when compared to *Diplotaxis* species, in which few domestication events occurred and several species are in the wild exposed to habitat conditions and constraints.

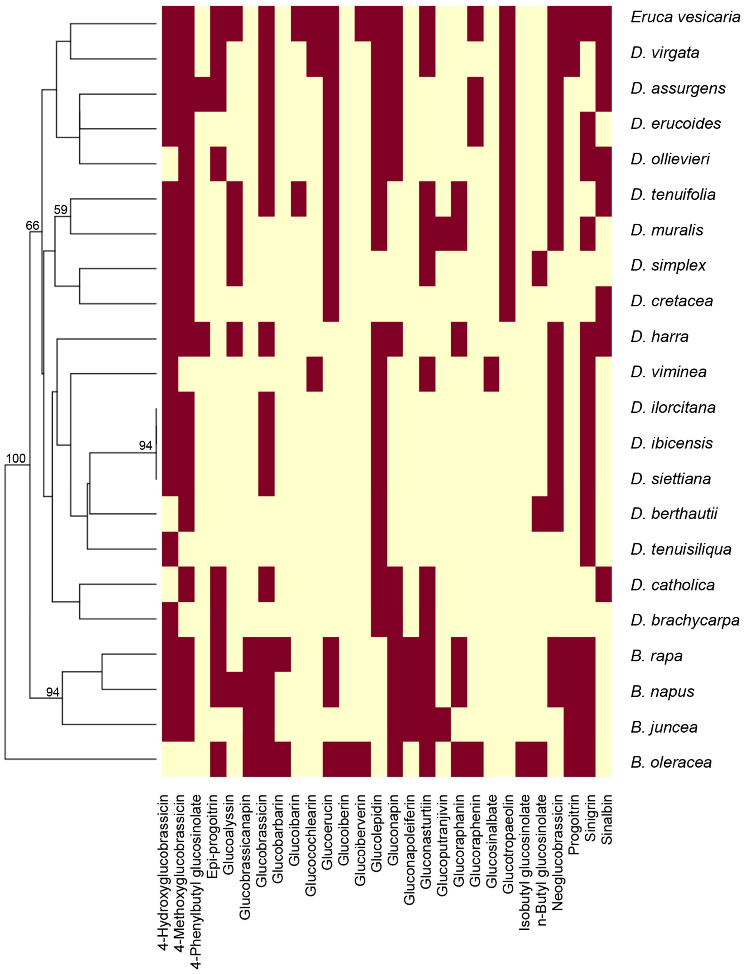


Figure 6– Chemodiversity profiling of GLS in *Brassica* and rocket species (*Diplotaxis* and *Eruca*). Data matrix of GLS chemodiversity is provided in Supplementary Table 4. Colors indicate presence (red) and absence (light yellow) of a glucosinolate compound. Bootstraps values above 50 are presented in the clustering phenogram resulted from the Eucledian distances method.

4. Discussion

4.1. GLS biosynthetic pathway: gene signature of aliphatic and indolic vias

In this study, we have performed a comprehensive assessment of the GLS biosynthetic pathways in Brassicaceae family. A reconstruction analysis of GLS pathway and a global assessment using genes described for *Arabidopsis* and for *Brassica* species were established, using a multi-database approach (i.e. TAIR, NCBI, Brassicadb, MetaCyc). From a total of 101 genes identified, about 78 previously identified genes in *Arabidopsis* were classified into the three sub-pathways of the GLS biosynthetic pathway (Supplementary Table 1), while the remaining 23 were not possible to assign to any of the three-specific pathway of GLS, and need further study to uncover their functional role in specific pathways. With the upcoming availability of more and more genomic resources from Brassicaceae species, a complete set of functional disclosure within each GLS is foreseen in a near future. Our clustering approach of the 101 GLS genes recorded, showed a clear separation between the genes involved mainly in GLS core structure synthesis shared between the 3 sub-pathways (aliphatic / indolic / aromatic), and other genes belonging to the CYP450s and MYBs gene families (Figure 4). The last cluster was therefore essentially composed of genes related to the side chain elongation process of GLS, which is responsible for the great diversity of GLS compounds, namely the biosynthesis genes (*CYP450*s) and the regulators of transcription (*MYB*s). In the PCA analysis, *CYP450*s genes are the highest contributors for explaining the variation within GLS biosynthetic pathway, namely: *CYP81 F1-F4*, which is indolic-specific, and *CYP79F1-F2*, unique to the aliphatic pathway. These overall results allowed us to conclude that the GLS biosynthetic pathway depends on upstream genes essentially involved in the core structure synthesis, while genes involved in the synthesis of aliphatic and indolic GLS, apparently specific of Brassicaceae, may depend on two set of gene clusters, known to be important for aliphatic - and indolic-specific pathways (*CYP79F1-F2* and *CYP81F1-F4*, respectively). Likewise, aliphatic and indolic GLS are the two most important types of GLS present in Brassicaceae.

4.2. Gene diversification in GLS: the case of *CYP79F1-F2* and *CYP81F1-F4*

Aliphatic and indolic GLS are derived from aliphatic (methionine, alanine, valine, leucine, and isoleucine) and indolic amino acids (tryptophan), respectively (Wittstock and Halkier, 2002). In *Arabidopsis thaliana* and *Arabidopsis lyrata*, aliphatic GLS are formed exclusively from methionine (Windsor et al., 2005). Species of the Brassicaceae have been useful models to understand the dynamics and impacts of ancient polyploidy (genome doubling), with the entire family having undergone a whole genome duplication (named At-α) and the *Brassica* crops suffered an additional genome triplication (Br-α) (Schranz and Mitchell-Olds, 2006; Thomas et al., 2006). Several authors have suggested that the genetic diversification of GLS in Brassicaceae is correlated with the polyploid occurrence in this family, with the At-β WGD event at 77.5 Mya where indolic glucosinolates appeared, and the At-α event at approximately 56 Mya (Kagale et al., 2014), where chain elongation of Met-derived aliphatic GLS is present (Schranz et al., 2011). Moreover, it has been pointed out that the diversity based on GLS composition in *Brassica* species could be related to A, B and C genomes (Ishida et al., 2014). The three ancestral *Brassica* species with diploid genome chromosomes: *Brassica nigra* (BB, 2n = 16) contain GLS with three carbon (C) side chains, derived from a single elongation reaction: *Brassica oleracea* (CC, 2n = 18) contains GLS with either 3C or 4C side chains; and *Brassica rapa* (AA, 2n = 20) contains GLS with either 4C or 5C side chains

(Ishida et al., 2014). Recently, studies have focused on genome evolution underlying the basis of GLS diversification. Bergh et al. (2016) reported that the genes that have undergone a high diversification process encode the MAM (Methylthioalkylmalate) enzymes and also the CYP81 side-chain modification enzymes responsible for a large part of the GLS chemotypes observed. MAM synthase enzymes are central for the diversification of aliphatic GLS structures in *Arabidopsis thaliana* and related species (Heidel et al, 2006); while *CYP81F* acts in the final step of the indolic GLS pathway and have been reported as responsible for a wide array of natural variation among *Arabidopsis thaliana* ecotypes (Pfalz et al., 2009). In this species, the biosynthesis of indolic GLS, hydroxylations are catalyzed by cytochromes P450 of the *CYP81F* subfamily (Pfalz et al., 2009; Barco and Clay, 2019), followed by methylation of the methyltransferases, IGMT1 and IGMT2 in *Arabidopsis thaliana* (Pfalz et al., 2011). In indolic GLS, the CYP81Fs family (CYP81F1-F3) has been identified as the encoder of the oxidizing enzyme that converts indolyl- 3-methyl GLS (I3M) to 4OH-I3M, while CYP81F4 acts in the hydroxylation at C1-position (Pfalz et al., 2011). Such secondary modifications can present high variability among species in nature and they are the main responsible for the diversity observed across more than 120 types of GLS that have been described to date (Kliebenstein et al., 2001a). CYP81F2 has been suggested to have neofunctionalized in plant innate immunity that subsequently was maintained in *Arabidopsis thaliana*, but lost in the ancestral Brassicaceae species. The phylogenetic analysis performed revealed four clusters, each of them associated to *CYP81F1* to *F4*. Since *Brassica* crops were grouped in the same cluster, it suggests a common diversification of indolic GLS that portrays *Brassica* chemotypes, which are present in less extent in aliphatic GLS.

In the aliphatic GLS pathway in Brassicaceae, *CYP79* is a key variable gene that has been considered as a driving force in GLS diversification. Several steps catalyzed by *CYP79F1* and *CYP79F2* result from gene duplication (Olson-Manning et al., 2013). CYP79F1 and CYP79F2 present slightly different substrate specificities: CYP79F1 uses both short- and long-chain substrates, whereas CYP79F2 tends to use only long-chain substrates. It has been considered that in *Brassica rapa*, like in *Arabidopsis thaliana* all the gene counterparts participate in the formation of the GLS core structure, except for *CYP79F2* (Wang et al., 2011). The absence of *CYP79F2* agrees well with the fact that all profiles of aliphatic GLS in *Brassica rapa* are composed of short-chain GLS. From the phylogenetic analysis we performed, it can be concluded that *Arabidopsis thaliana CYP79F1* and *F2* genes are in the basis of the diversification of the remaining Brassicaceae species (Figure 6). *Brassica* crops are grouped in a single cluster (highlighted in shaded red in Figure 6), which represents a common genetic basis of the *CYP79F1-F2* responsible for the GLS diversification and possibly links to the additional genome triplication (Br-α) event that these crops suffered throughout their evolution. The annotation of *CYP79F1* and *F2* genes in Brassicaceae is limited as only recently genome sequences are being released, pushed by the continuous lower costs of whole-genome sequencing technologies. With our study, we were able to determine the disjunction of *Arabidopsis thaliana CYP79F1-F2* with the remaining Brassicaceae species and in particular with *Brassica* crops, which were grouped together in a lineage associated with aliphatic GLS.

Our results revealed the most recent diversification of *CYP79F1* and *F2* genes in *Brassica* crops, where a single cluster including *Brassica* species is difficult to depict (Figure 6). This lack of clear clustering from *CYP79F1* and *CYP79F2*, in opposite with what is observed in *CYP81F1-F4*, may suggest the absence of a *CYP79F2* gene as reported for *Brassica rapa* (Wang et al., 2011), which may not be the case for other *Brassica* species. This may suggest a shared genetic basis underlying short-chain aliphatic GLS, since in *Arabidopsis thaliana* a *CYP79F2* knockout mutant presents a considerable reduction of long-chain aliphatic GLS (Chen et al., 2003). Moreover, future annotation efforts of Brassicaceae genes has to be performed as a way to clarify *CYP79F1* diversification within *Brassica* crops that should be linked to a higher production of short-chain aliphatic GLS.

4.3. Chemical diversity of GLS in Brassicaceae

GLS production by Brassicaceae plants is considered as being influenced by environmental factors such as soil, climate and cultivation conditions including fertilization, harvest time, and plant organ (Martínez-Ballesta et al., 2013). In general, the diversity of GLS profiles is higher in *Brassica oleracea* as opposed to *Brassica rapa* (Figure 7). The Brassicaceae plant tissues include one or more major GLS mostly composed of aliphatic GLS. In general, Brassicaceae vegetables GLS contain an alkyl side chain with 3–5 carbons (Ishida et al., 2014). From these ones, glucoiberin is present mostly in *Brassica oleracea* vegetables (cabbage, broccoli, and cauliflower) while, gluconapin and progoitrin are ubiquitous in many *Brassica* vegetables such as *Brassica rapa* (Chinese cabbage, mustard spinach, and turnip), *Brassica oleracea*, *Brassica juncea* (mustard green), and *Brassica napus* (rapeseed vegetable) (Ishida et al, 2014). Glucoerucin is mainly found in cultivated *Eruca sativa* and wild rockets (*Diplotaxis tenuifolia*, *Diplotaxis* sp.) rockets.

In general, *Diplotaxis* spp. emerges as an extremely GLS-rich species, revealing likely taxonomic affinities with taxa previously examined by other criteria suggesting a high potential for further exploitation. The disclosure of a distinct GLS chemo-profile between *Brassica* crops and *Diplotaxis* species (i.e. in *Brassica*, progoitrin, gluconapin, glucobrassicanapin are the most abundant GLSs, while in *Diplotaxis* glucolepidin and glucoerucin are the most distinctive), opens a new perspective for addressing more studies towards not only the characterization of new taxa from the later genus but also the quantification of such GLS, since many of them, in high amounts, are considered to be anti-nutritional even in vegetables (e.g. Augustine et al., 2013). GLS production and contents in Brassicaceae plants are influenced by environmental factors such as soil, climate and cultivation conditions including fertilization, harvest time, and plant position, besides its straight relation to both biotic and abiotic stresses (Martínez-Ballesta et al., 2013; Ishida et al., 2014). Despite several reports on a positive relationship between GLS production and abiotic stress, it is still unknown which are the mechanisms of resistance to drought and salinity conditions. Determining a chemodiversity profile associated with phenotypes adapted to extreme environmental conditions, such as drought and salinity, could be a good strategy for prospecting GLS compounds and contents and quantity for coping with abiotic stresses.

4.4. Abiotic stress and GLS crosstalk in Brassicaceae: wild rockets as emergent taxa

Variation in the amount and profile of GLS compounds has been correlated with abiotic stresses (Variyar et al., 2014). Among the most important, salinity and drought stresses are known to significantly affect crops productivity. Overall, GLS content increases markedly under salinity, drought, high temperature and nitrogen (N) deficiency (Martínez-Ballesta et al., 2015).
Extensive studies in Brassicaceae family showed a positive correlation between salt stress and GLS content, [e.g. in broccoli (López-Berenguer et al., 2009), canola (Khalifa, 2012), radish sprouts (Yuan et al., 2010), pakchoi (Keling and Zhujun, 2010)]. An increase in the signature of GLS content has also been reported for Brassicaceae taxa under drought stress, namey in *Brassica napus* (Champolivier and Merrien, 1997), *Brassica oleracea* (Radovich et al., 2005), *Brassica rapa* (Zhang et al., 2008), *Brassica juncea* (Tong et al., 2014), and *Brassica carinata* (Ngwene et al., 2017).
However, recent studies in wild rocket (*D. tenuifolia*), demonstrated that salinity conditions did not affect the total amount of GLS profile (Bonasia et al., 2017; Cocetta et al., 2018). Bonasia et al., (2017) showed that the aliphatic-GLSs proidrin, epiproidrin, and glucoerucin contents were unaffected by salt stress (Bonasia et al., 2017), with glucoerucin emerging as a GLS compound specific of *Diplotaxis*, of *Eruca vesicaria* and of *E. sativa* (Barillari et al., 2005). Furthermore, glucoerucin could be linked to a distinctive chemical signature of the *Diplotaxis*-*Eruca* lineage involved in salt tolerance, setting it apart from the *Brassica* crops chemo-lineage (Figure 6).
Under drought stress, indole glucosinolate biosynthetic genes revealed to be up-regulated in wild rocket (Cavaiuolo et al., 2017), which accounts for a possible tolerance mechanism as described for other brassicas under stress (Martínez-Ballesta et al., 2015). In this tolerance mechanism, *MYB* genes

(particularly MYB28 an MYB29) may play a role in variations of GLS contents. Salehin et al. (2019) confirmed that MYB28 and MYB29 are important transcription factors regulating the synthesis of indole GLS, where a *cyp79f1f2* double mutant revealed to be less tolerant to drought, probably due to the loss of aliphatic GLS compounds, corroborating former studies (Martínez-Ballesta et al., 2015). Moreover, Martínez-Ballesta et al. (2015) highlighted that pathways involved in the physiological responses to salt stress are connected to GLS metabolism. Under salt stress, an increase in short-chain aliphatic GLS was observed which has been further associated to a higher expression of aquaporins, involved on osmoregulation pathways (Martínez-Ballesta et al., 2014), and thus could contribute to water saving process (Martínez-Ballesta et al., 2015). Overall, short-chain aliphatic GLS may contribute to water saving under salt stress (Martínez-Ballesta et al., 2015), while under drought indolic GLS seems to be the most affected (Salehin et al., 2019).

When compared to Brassica crops, wild rockets seem to display a different GLS profile that could be associated to an abiotic stresses tolerance. Indeed, the neglected and underutilized rocket species, i.e. *Eruca sativa* (rocket), *Diplotaxis tenuifolia* and *Diplotaxis muralis* (wild rocket), as well as other wild taxa distributed and adapted to extreme ecological conditions (i.e. severe salinity and drought conditions), may be considered as potential targets to understand abiotic stress tolerance mechanisms. *Diplotaxis* is considered an unexplored Brassicaceae crop wild relative (CWR), with *Brassica* crops having evolved from the *Diplotaxis–Erucastrum* complex (Arias and Pires, 2012), which makes *Diplotaxis* species an important reservoir of genetic resources for crop improvement.

5. Conclusions

Overall, we have analysed gene clusters involved in the biosynthesis of GLS, by combining genome analysis with biochemical pathways and chemical diversity assessment. An integrated approach was performed by assessing a global GLS gene inventory in Brassicaceae and its diversity, analysing a potential genetic basis for GLS divergence using 6 CYP genes (CYP79F1-F2 and CYP81F1-F4), known to be key genes of indolic and aliphatic GLS biosynthetic pathways, linked to a chemical diversity evaluation of GLS compounds in major *Brassica* crops compared to the wild relative genus *Diplotaxis*. Our results point to a recent diversification of the aliphatic CYP79F1 and CYP79F2 genes in *Brassica* crops, while for indolic genes a clear separation is observed for CYP81F1-F4 genes, revealing an earliest divergence on this GLS sub-pathway. Chemical diversity assessment allowed recognizing that *Brassica* and *Diplotaxis* have distinct GLS chemo-profiles, highlighting that the latter genus includes extremely GLS-rich species. Considering the enormous potential of biodiversity for finding new traits useful in breeding programs, screening of GLS-enriched

Brassicaceae species is of particular interest. Despite that GLS profiles may vary among species and according to plant development and/or environmental factors, a highly diverse and unexplored chemodiversity has been recognized within *Diplotaxis*. The discovery of the genomic information behind such GLS diversity could constitute a potential for discovering new phytochemical and nutraceutical sources potentially transferable to *Brassica* crops. Also, understanding the relationship between Brassicaceae GLS genes and abiotic stress tolerance will be useful to contribute as source of genes for improving new Brassicaceae vegetable varieties to cope with effects of global climate changes.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Authors Contributions

Conceptualization, F.M. and M.M.R.; methodology, F.M., A.P.E., A.R.P., M.M.R.; Bioinformatic analysis, A.P.E., F.M and A.R.P.; Results analysis, A.P.E., F.M., A.R.P., M.M., M.M.R.; writing—original draft preparation, A.P.E., F.M., A.R.P., M.M.R.; writing—review and editing, A.P.E., F.M., A.R.P., M.S.P., M.M.R. and M.M. All authors have approved the submitted version of this manuscript.

Funding

This research was funded by Fundação para a Ciência e Tecnologia (FCT) and Aga Khan Development Network (AKDN) under the project CVAgrobiodiversity/333111699. Fellowships SFRH/BPD/114664/2016, SFRH/BD/135362/2017 to FM and APE, respectively, and research units: UID/AGR/04129/2019 (LEAF) and UID/BIA/00329/2019 (cE3c) were funded by Portuguese National Funds through FCT, Portugal.

Acknowledgments

The authors would like to acknowledge the supported provided by Fundação para a Ciência e Tecnologia (FCT) and Aga Khan Development Network (AKDN).

References

Armah, C. N., Derdemezis, C. Traka, M. H., Dainty, J. R., Doleman, J. F., Saha, S., et al. 2015. Diet rich in high glucoraphanin broccoli reduces plasma LDL cholesterol: evidence from randomised controlled trials. *Molecular Nutrition & Food Research* 59, 918–926. https://doi.org/10.1002/mnfr.201400863.

Augustine, R., Mukhopadhyay, A., Bisht, N. C. 2013. Targeted silencing myb28transcription factor gene directs development of low glucosinolate lines in oilseed (*Brassica juncea*). *Plant Biotechnology Journal* 11, 855–866. https://doi.org/10.1111/pbi.12078.

Bak, S., and Feyereisen, R. 2001. The involvement of two p450 enzymes, CYP83B1 and CYP83A1, in auxin homeostasis and glucosinolate biosynthesis. *Plant Physiology* 127, 108–118. https://doi.org/10.1104/pp.127.1.108.

Barco, B. and Clay, N.C. 2019. Evolution of glucosinolate diversity via whole-genome duplications, gene rearrangements, and substrate promiscuity. *Annual Review of Plant Biology* 70 (1): 585–604. https://doi.org/10.1146/annurev-arplant-050718-100152.

Barillari, J., Canistro, D., Paolini, M., Ferroni, F., Pedulli, G. F., and Iori, R. (2005).Direct antioxidant activity of purified glucoerucin, the dietary secondary metabolite contained in rocket (*Eruca sativa* Mill.) seeds and sprouts. *Food Chemistry* 53, 2475–2482. doi: 10.1021/jf047945a

Bellostas, N., Sorensen, J., Sorensen, H. 2004. Qualitative and quantitative evaluation of glucosinolates in cruciferous plants during their life cycles. *Agroindustria* 3, 267–272.

Berardini, T. Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., Huala, E. 2015. The *Arabidopsis* information resource: making and mining the ‘gold standard’ annotated reference plant genome. *Genesis* 53, 474–485. https://doi.org/10.1002/dvg.22877.

Bergh, E. van den, Hofberger, J. A., Schranz, M. E. 2016. Flower power and the mustard bomb: comparative analysis of gene and genome duplications in glucosinolate biosynthetic pathway evolution in Cleomaceae and Brassicaceae. *American Journal of Botany* 103, 1212–1222. https://doi.org/10.3732/ajb.1500445.

Bhandari, S., Jo, J., Lee, J. 2015. Comparison of glucosinolate profiles in different tissues of nine Brassica crops. *Molecules* 20, 15827–15841. https://doi.org/10.3390/molecules200915827.

Binns, D., Dimmer, E., Huntley, R., Barrell, D., O’Donovan, C., Apweiler, R. 2009. QuickGO: a web-based tool for gene ontology searching. *Bioinformatics* 25, 3045–3’46. https://doi.org/10.1093/bioinformatics/btp536.

Bischoff, K. L. 2016. Glucosinolates. *Nutraceuticals*, 551–554. https://doi.org/10.1016/b978-0-12-802147-7.00040-1.

Bonasia, A., Lazzizera, C., Elia, A., Conversa, G. 2017. Nutritional, biophysical and physiological characteristics of wild rocket genotypes as affected by soilless cultivation system, salinity level of nutrient solution and growing period. *Frontiers in Plant Science* 8:300. doi:10.3389/fpls.2017.00300.

Brown, P. D, Tokuhisa, J. G., Reichelt, M., Gershenzon, J. 2003. Variation of glucosinolate accumulation among different organs and developmental stages of *Arabidopsis thaliana*. *Phytochemistry* 62, 471–481. https://doi.org/10.1016/s0031-9422(02)00549-6.

Carbon, S., Ireland, A., Mungall, C. J., Shu, S. Q., Marshall, B., Lewis, S. 2008. AmiGO: online access to ontology and annotation data. *Bioinformatics* 25, 288–289. https://doi.org/10.1093/bioinformatics/btn615.

Cartea, M. E., and Velasco, P. 2007. Glucosinolates in Brassica foods: bioavailability in food and significance for human health. *Phytochemistry Reviews* 7, 213–229. https://doi.org/10.1007/s11101-007-9072-2.

Cartea, M. E., de Haro, A., Obregón, S., Soengas, P., Velasco, P. 2012. Glucosinolate variation in leaves of *Brassica rapa* crops. *Plant Foods for Human Nutrition* 67, 283–288. https://doi.org/10.1007/s11130-012-0300-6.

Caspi, R., Billington, R., Fulcher, C., Keseler, I. M., Kothari, A., Krummenacker, M., Latendresse, M. et al. 2017. The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Research* 46 (D1): D633–39. https://doi.org/10.1093/nar/gkx935.

Cavaiuolo, M., and Ferrante A. 2014. Nitrates and glucosinolates as strong determinants of the nutritional quality in rocket leafy salads. *Nutrients* 6, 1519–1538. https://doi.org/10.3390/nu6041519.

Cavaiuolo, M., Cocetta, G., Spadafora, N. D., Müller, C. T., Rogers, H. J., Ferrante, A. 2017.Gene expression analysis of rocket salad under preharvest and postharvest stresses: a transcriptomic resource for *Diplotaxis tenuifolia*. *PLoS ONE* 12(5): e0178119. https://doi.org/10.1371/journal.pone.0178119

Champolivier, L., and Merrien, A. 1997. Corrigendum to 'Effects of water stress applied at different growth stages to *Brassica napus L.* var. *oleifera* on yield, yield components and seed quality'. *European Journal of Agronomy* 6 (3–4): 309–10. https://doi.org/10.1016/s1161-0301(97)00003-8.

Charron, C. S., and Carl E. S. 2004. Glucosinolate content and myrosinase activity in rapid-cycling *Brassica oleracea* grown in a controlled environment. *Journal of the American Society for Horticultural Science* 129 (3): 321–30. https://doi.org/10.21273/jashs.129.3.0321.

Chen, S., Erich, G., Kirsten, J., Peter, N., Bodil, J., Olsen, C. et al. 2003. CYP79F1 and CYP79F2 have distinct functions in the biosynthesis of aliphatic glucosinolates in *Arabidopsis*. *The Plant Journal* 33(5): 923–937. https://doi.org/10.1046/j.1365-313x.2003.01679.x.

Choi, S., Park, S., Lim Y., Kim, S.,Park, J., and An, G. 2014. Metabolite profiles of glucosinolates in cabbage varieties (*Brassica oleracea* var. *capitata*) by season, color, and tissue position. *Horticulture, Environment, and Biotechnology* 55; 237–247. https://doi.org/10.1007/s13580-014-0009-6.

Clarke, D. 2010. Glucosinolates, structures and analysis in food. *Analytical Methods* 2; 310. https://doi.org/10.1039/b9ay00280d.

Cleemput, S., and Becker, H. 2011. Genetic variation in leaf and stem glucosinolates in resynthesized lines of winter rapeseed (*Brassica napus L.*). *Genetic Resources and Crop Evolution* 59; 539–46. https://doi.org/10.1007/s10722-011-9701-x.

Cocetta, G., Mishra, S., Raffaelli, A., Ferrante, A. 2018. Effect of heat root stress and high salinity on glucosinolates metabolism in wild rocket. *Journal of Plant Physiology* 231: 261-270, https://doi.org/10.1016/j.jplph.2018.10.003.

Collett, M.,Stegelmeier, B., and Tapper, B. 2014. Could nitrile derivatives of turnip (*Brassica rapa*) glucosinolates be hepato- or cholangiotoxic in cattle? *Journal of Agricultural and Food Chemistry* 62; 7370–7375. https://doi.org/10.1021/jf500526u.

D’Antuono, L. F., Elementi, S., and Neri, R. 2008. Glucosinolates in *Diplotaxis* and *Eruca* leaves: diversity, taxonomic relations and applied aspects. *Phytochemistry* 69 (1): 187–199. https://doi.org/10.1016/j.phytochem.2007.06.019.

Fahey, J. W., Zhang, Y., and Talalay, P. 1997. Broccoli sprouts: an exceptionally rich source of inducers of enzymes that protect against chemical carcinogens. *Proceedings of the National Academy of Sciences* 94 (19): 10367–72. https://doi.org/10.1073/pnas.94.19.10367.

Fahey, J.W., Zalcmann, A.T., and Talalay, P. 2001. The chemical diversity and distribution of glucosinolates and isothiocyanates among plants. *Phytochemistry* 56(1):5-51. https://doi.org/10.1016/S0031-9422(00)00316-2.

Ford-Lloyd, B. V., Schmidt, M., Armstrong, S. J., Barazani, O. Z., Engels, J., Hadas, R. et al. 2011. Crop Wild Relatives—Undervalued, Underutilized and under Threat? *BioScience* 61(7); 559–565. https://doi.org/10.1525/bio.2011.61.7.10.

Fukushima, A., Kusano, M., Mejia, R., Iwasa, M., Kobayashi, M.,Hayashi, N.,Watanabe-Takahashi, A., et al. 2014. Metabolomic characterization of knockout mutants in Arabidopsis: development of a metabolite profiling database for knockout mutants in *Arabidopsis*. *Plant Physiology* 165; 948–961. https://doi.org/10.1104/pp.114.240986.

Gómez-Campo, C., and Tortosa M. 1974. The taxonomic and evolutionary significance of some juvenile characters in the Brassiceae. *Botanical Journal of the Linnean Society* 69 (2): 105–124. https://doi.org/10.1111/j.1095-8339.1974.tb01619.x.

Gómez-Campo C.1980 Morphology and morphotaxonomy of the tribe Brassiceae. In: Tsunoda S, Hinata K, Gómez-Campo C. (ed) Brassica crops and wild allies. *Japan Scientific Societies Press*, Tokyo, Pp.3-31.

Götz, S., Garcia-Gomez, J. M.,Terol, J., Williams, T. D., Nagaraj, S. H.,Nueda, M. J.,Robles, M., Talon, M.,Dopazo, J., and Conesa, A. 2008. High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research* 36 (10): 3420–35. https://doi.org/10.1093/nar/gkn176.

Grubb, C. D., and Abel, S. 2006. Glucosinolate metabolism and its control. *Trends in Plant Science* 11 (2): 89–100. https://doi.org/10.1016/j.tplants.2005.12.006.

Grubb, C. D., Zipp, B. J., Kopycki, J., Schubert, M., Quint, M., Lim, E., Bowles, D.J., Pedras, M. S. C., and Abel, S. 2014. Comparative analysis of Arabidopsis UGT74 glucosyltransferases reveals a special role of UGT74C1 in glucosinolate biosynthesis. *The Plant Journal* 79 (1): 92–105. https://doi.org/10.1111/tpj.12541.

Gupta, S., Sangha, M. K ., Kaur, G., Atwa, A. K. A., Banga, S., and Banga, S. S.. 2012. Variability for leaf and seed glucosinolate contents and profiles in a germplasm collection of the *Brassica juncea*. *Biochemistry & Analytical Biochemistry* 01 (07). https://doi.org/10.4172/2161-1009.1000120.

Halkier, B. A., and Gershenzon, J. 2006. Biology and biochemistry of glucosinolates. *Annual Review of Plant Biology* 57 (1): 303–333. https://doi.org/10.1146/annurev.arplant.57.032905.105228.

Hansen, C. H., Wittstock, U., Olsen, C.E., Hick, A. J., Pickett, J. A., and Halkier, B.A.. 2001. Cytochrome P450 CYP79F1 from *Arabidopsis* catalyzes the conversion of dihomomethionine and trihomomethionine to the corresponding aldoximes in the biosynthesis of aliphatic glucosinolates. *Journal of Biological Chemistry* 276 (14): 11078–85. https://doi.org/10.1074/jbc.m010123200.

Hansen, B. G., Kliebenstein, D. J., and Halkier, B. A. 2007. Identification of a flavin-monooxygenase as the S-oxygenating enzyme in aliphatic glucosinolate biosynthesis in *Arabidopsis*. *The Plant Journal* 50 (5): 902–10. https://doi.org/10.1111/j.1365-313x.2007.03101.x.

Hansen, B. G., Kerwin, R. E., Ober, J. A., Lambrix, V. M., Mitchell-Olds, T., Gershenzon, J., Halkier, B. A., and Kliebenstein, D. J. 2008. A novel 2-oxoacid-dependent dioxygenase involved in the formation of the goiterogenic 2-hydroxybut-3-enyl glucosinolate and generalist insect resistance in *Arabidopsis*. *Plant Physiology* 148 (4): 2096–2108. https://doi.org/10.1104/pp.108.129981.

Heidel, A.J., Clauss, M. J., Kroymann, J., Savolainen O., and Mitchell-Olds, T. 2006. Natural Variation in MAM Within and Between Populations of *Arabidopsis lyrata* Determines Glucosinolate Phenotype. *Genetics* 173 (3): 1629–36. https://doi.org/10.1534/genetics.106.056986.

Ishida, M., Hara, M., Fukino, N., Kakizaki, T., and Morimitsu, Y. 2014. Glucosinolate metabolism, functionality and breeding for the improvement of Brassicaceae vegetables. *Breeding Science* 64 (1): 48–59. https://doi.org/10.1270/jsbbs.64.48.

Kell, S.P., Knüpffer, H., Jury, S.L., Ford-Lloyd, B.V. and Maxted, N. 2008. Crops and wild relatives of the Euro-Mediterranean region: making and using a conservation catalogue. In: Maxted, N., Ford-Lloyd, B.V., Kell, S.P., Iriondo, J., Dulloo, E. and Turok, J. (eds) Crop Wild Relative Conservation and Use. Pp. 69–109. CAB International, Wallingford, UK.

Khalifa, N.S. 2012. Protein expression after NaCl treatment in two tomato cultivars differing in salt tolerance. *Acta Biologica Cracoviensia Series Botanica* 54 (2). https://doi.org/10.2478/v10182-012-0020-0.

Khan, M.A.M., Ulrichs, C., and Mewis, I. 2011. Drought stress - impact on glucosinolate profile and performance of phloem feeding cruciferous insects. *Acta Horticulturae*, 917; 111–17. https://doi.org/10.17660/actahortic.2011.917.13.

Kliebenstein, D. J., Kroymann, J., Brown, P., Figuth, A., Pedersen, D.,Gershenzon, J., and Mitchell-Olds, T. 2001a. Genetic control of natural variation in *Arabidopsis* glucosinolate accumulation. *Plant Physiology* 126 (2): 811–25. https://doi.org/10.1104/pp.126.2.811.

Kliebenstein, D. J., Kroymann, J., Brown, P., Figuth, A., Pedersen, D.,Gershenzon, J., and Mitchell-Olds, T. 2001b. Gene duplication in the diversification of secondary metabolism: tandem 2-oxoglutarate-dependent dioxygenases control glucosinolate biosynthesis in *Arabidopsis*. *The Plant Cell* 13 (3): 681. https://doi.org/10.2307/3871415.

Kumar, S., Stecher, G., Li,M., Knyaz, C., and Tamura K. 2018. MEGA X: molecular evolutionary genetics analysis across computing platforms. Eds Battistuzzi, F. U. *Molecular Biology and Evolution* 35 (6): 1547–49. https://doi.org/10.1093/molbev/msy096.

Lanfear, R., Frandsen, P. B., Wright, A. M., Senfeld, T., and Calcott, B. 2016. PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. *Molecular Biology and Evolution* 34 (3): 772–73. https://doi.org/10.1093/molbev/msw260.

López-Berenguer, C., Martínez-Ballesta, M.C., Moreno, D. A., Carvajal, M., and García-Viguera, C. 2009. Growing hardier crops for better health: salinity tolerance and the nutritional value of broccoli. *Journal of Agricultural and Food Chemistry* 57 (2): 572–78. https://doi.org/10.1021/jf802994p.

Martínez-Ballesta, M. C., Moreno, D. A., Carvajal, M. 2013. The physiological importance of glucosinolates on plant response to abiotic stress in *Brassica*. *International Journal of Molecular Sciences* 14(6): 11607-11625. https://doi.org/10.3390/ijms140611607.

Martínez-Ballesta, M. C., Muries, B., Moreno, D. A., Domínguez-Perles, R., García-Viguera, C., and Carvajal, M. 2014. Involvement of a glucosinolate (sinigrin) in the regulation of water transport in *Brassica oleracea* grown under salt stress. *Physiologia Plantarum* 150, 145–160. doi: 10.1111/ppl.12082

Martínez-Ballesta, M. C., Moreno-Fernández, D. A., Castejón, D., Ochando, C., Morandini, P. A., and Carvajal M. 2015. The impact of the absence of aliphatic glucosinolates on water transport under salt stress in *Arabidopsis thaliana*. *Frontiers in Plant Science* 6:524. doi: 10.3389/fpls.2015.00524.

Merah, O. 2015. Genetic variability in glucosinolates in seed of *Brassica juncea*: interest in mustard condiment. *Journal of Chemistry*. 2015: 606142. http://dx.doi.org/10.1155/2015/606142.

Mikkelsen, M. D., Naur, P., and Halkier, B. A. 2004. *Arabidopsis* mutants in the c-s lyase of glucosinolate biosynthesis establish a critical role for indole-3-acetaldoxime in auxin homeostasis. *The Plant Journal* 37 (5): 770–77. https://doi.org/10.1111/j.1365-313x.2004.02002.x.

Mithen, R., Clarke, J., Lister, C., and Dean, C. 1995. Genetics of aliphatic glucosinolates. III. Side chain structure of aliphatic glucosinolates in *Arabidopsis thaliana*. *Heredity* 74 (2): 210–15. https://doi.org/10.1038/hdy.1995.29.

Moreira-Rodríguez, M., Nair ,V., Benavides, J., Cisneros-Zevallos. L., and Jacobo-Velázquez, D. 2017a. UVA, UVB light doses and harvesting time differentially tailor glucosinolate and phenolic profiles in broccoli sprouts. *Molecules* 22 (7): 1065. https://doi.org/10.3390/molecules22071065.

Moreira-Rodríguez, M., Nair ,V., Benavides, J., Cisneros-Zevallos. L., and Jacobo-Velázquez, D. 2017b. UVA, UVB light, and methyl jasmonate, alone or combined, redirect the biosynthesis of glucosinolates, phenolics, carotenoids, and chlorophylls in broccoli sprouts. *International Journal of Molecular Sciences* 18 (11): 2330. https://doi.org/10.3390/ijms18112330.

Ngwene, B., Neugart. S., Baldermann, S., Ravi, B. and Schreiner, M. 2017. Intercropping induces changes in specific secondary metabolite concentration in ethiopian kale (*Brassica carinata*) and african nightshade (*Solanum scabrum*) under controlled conditions. *Frontiers in Plant Science* 8 https://doi.org/10.3389/fpls.2017.01700.

Pfalz, M., Vogel, H., and Kroymann. J. 2009. The gene controlling the indole glucosinolate modifier1 quantitative trait locus alters indole glucosinolate structures and aphid resistance in *Arabidopsis*. *The Plant Cell* 21 (3): 985–99. https://doi.org/10.1105/tpc.108.063115.

Pfalz, M., Mikkelsen, M. D., Bednarek, P., Olsen, C. E., Halkier, B. A. and Kroymann., J. 2011. Metabolic engineering in *Nicotiana benthamiana* reveals key enzyme functions in *Arabidopsis* indole glucosinolate modification. *The Plant Cell* 23 (2): 716–29. https://doi.org/10.1105/tpc.110.081711.

Piotrowski, M., Schemenewitz, A., Lopukhina, A., Müller, A., Janowitz, T., Weiler, E. W. and Oecking, C. 2004. Desulfoglucosinolate sulfotransferases from *Arabidopsis thaliana* catalyze the final step in the biosynthesis of the glucosinolate core structure. *Journal of Biological Chemistry* 279: 50717–25. https://doi.org/10.1074/jbc.m407681200.

Radovich, T. J. K., Kleinhenz, M. D. and Streeter, J. G. 2005. Irrigation timing relative to head development influences yield components, sugar levels, and glucosinolate concentrations in cabbage. *Journal of the American Society for Horticultural Science* 130 (6): 943–49. https://doi.org/10.21273/jashs.130.6.943.

Rambaut A. 2009. FigTree version 1.3.1 [computer program] http://tree.bio.ed.ac.uk .

Salehin, M., Li, B., Tang, M., Katz, E., Song, L., Ecker, J. R., Kliebenstein, D. J., and Estelle, M. 2019. Auxin-sensitive Aux/IAA proteins mediate drought tolerance in *Arabidopsis* by regulating glucosinolate levels. *Nature Communications* 10 (1). https://doi.org/10.1038/s41467-019-12002-1.

Sánchez-Yélamo, M. D. 2009. Relationships in the *Diplotaxis–Erucastrum–Brassica* complex (Brassicaceae) evaluated from isoenzymatic profiles of the accessions as a whole. Applications for characterisation of phytogenetic resources preserved *ex situ*. *Genetic Resources and Crop Evolution* 56 (7): 1023–1036. https://doi.org/10.1007/s10722-009-9423-5.

Soundararajan, P., and Kim., J. 2018. Anti-carcinogenic glucosinolates in cruciferous vegetables and their antagonistic effects on prevention of cancers. *Molecules* 23: 2983. https://doi.org/10.3390/molecules23112983.

Tohge, T., Perez de Souza, L., and Fernie, A. R. 2014. Genome-enabled plant metabolomics. *Journal of Chromatography B* 966: 7–20. https://doi.org/10.1016/j.jchromb.2014.04.003.

Tong, Y., Gabriel-Neumann, E., Ngwene, B., Krumbein, A., George, E., Platz, S., Rohn, S., and Schreiner, M. 2014. Topsoil drying combined with increased sulfur supply leads to enhanced aliphatic glucosinolates in *Brassica juncea* leaves and roots. *Food Chemistry* 152 (June): 190–96. https://doi.org/10.1016/j.foodchem.2013.11.099.

Tripathi, M.K., and Mishra, A.S. 2007. Glucosinolates in animal nutrition: a review. *Animal Feed Science and Technology* 132 (1–2): 1–27. https://doi.org/10.1016/j.anifeedsci.2006.03.003.

VanEtten, C. H., and H. L. Tookey. 1983. Glucosinolates. In *Handbook of Naturally Occurring Food Toxicants*, 15–30. Boca Raton, Florida: CRC Press.

Variyar, P.S., Banerjee, A., Akkarakaran, J. J., and Suprasanna., P. 2014. Role of glucosinolates in plant stress tolerance. *Emerging Technologies and Management of Crop Stress Tolerance*, 271–91. https://doi.org/10.1016/b978-0-12-800876-8.00012-6.

Velasco, P., Soengas, P., Vilar, M., Cartea, M. E. and del Rio., M. 2008. Comparison of glucosinolate profiles in leaf and seed tissues of different *Brassica napus* crops. *Journal of the*

American Society for Horticultural Science 133 (4): 551–58. https://doi.org/10.21273/jashs.133.4.551.

Verkerk, R., S. Tebbenhoff, and M. Dekker. 2010. Variation and distribution of glucosinolates in 42 cultivars of *Brassica oleracea* vegetable crops. *Acta Horticulturae*, no. 856: 63–70. https://doi.org/10.17660/actahortic.2010.856.7.

Verkerk, R., Schreiner, M., Krumbein, A., Ciska, E., Holst, B., Rowland, I., De Schrijver, R., et al. 2008. Glucosinolates in Brassica vegetables: the influence of the food supply chain on intake, bioavailability and human health. *Molecular Nutrition & Food Research* 53: S219–S219. https://doi.org/10.1002/mnfr.200800065.

Wang, H., Wu, J., Silong Sun, Liu, B., Cheng, F., Sun, R. and Wang., X. 2011. Glucosinolate biosynthetic genes in *Brassica rapa*. *Gene* 487: 135–142. https://doi.org/10.1016/j.gene.2011.07.021.

Wang, X., Wu, J., Liang, J., Cheng, F. and Wang., X. 2015. Brassica database (BRAD) version 2.0: integrating and mining Brassicaceae species genomic resources. *Database* 2015: bav093. https://doi.org/10.1093/database/bav093.

Wiesner, M., Zrenner, R., Krumbein, A., Glatt, H. and Schreiner, M. 2013. Genotypic variation of the glucosinolate profile in pak choi (*Brassica rapa* ssp. *chinensis*). *Journal of Agricultural and Food Chemistry* 61: 1943–1953. https://doi.org/10.1021/jf303970k.

Windsor, A. J., Reichelt, M., Figuth, A., Svatoš, A., Kroymann, J., Kliebenstein, D. J., Gershenzon, J. and Mitchell-Olds., T. 2005. Geographic and evolutionary diversification of glucosinolates among near relatives of *Arabidopsis thaliana* (Brassicaceae). *Phytochemistry* 66: 1321–33. https://doi.org/10.1016/j.phytochem.2005.04.016.

Wittstock, U., and B. A. Halkier. 2002. Glucosinolate research in the *Arabidopsis* Era. *Trends in Plant Science* 7: 263–70. https://doi.org/10.1016/s1360-1385(02)02273-2.

Yuan, G., Wang, X., Guo, R. and Wang., Q. 2010. Effect of salt stress on phenolic compounds, glucosinolates, myrosinase and antioxidant activity in radish sprouts. *Food Chemistry* 121: 1014–1019. https://doi.org/10.1016/j.foodchem.2010.01.040.

Zhang, Y., Tang, L. and Gonzalez., V. 2003. Selected isothiocyanates rapidly induce growth inhibition of cancer cells. *Molecular Cancer Therapy* 2 (10): 1045–1052.

Zhang, H., Schonhof, I., Krumbein, A., Gutezeit, B., Li, L., Stützel, H. and Schreiner. M. 2008. Water supply and growing season influence glucosinolate concentration and composition in turnip root (*Brassica rapa* ssp. *rapifera L.*). *Journal of Plant Nutrition and Soil Science* 171: 255–265. https://doi.org/10.1002/jpln.200700079.

Figures

Figure 1- Genomic information of the GLS biosynthetic genes of the top ten Brassicaceae species registered at NCBI database. Number of sequences available per species is represented above each bar.

Figure 2- Euler diagram displaying GLS gene annotation gathered from a multi-database approach.

Figure 3- Gene Ontology (GO) terms assignment for the GLS biosynthetic genes. The graph displays the term enrichment levels of the annotated sequences along with the GO term hierarchy: Biological Process (BP, in green), Molecular Function (MF, in blue) and Cellular Component (CC, in yellow).

Figure 4- UPGMA phenogram of the 101 GLS biosynthetic genes. Bootstrap values above 50 are represented on the branches. Detailed UPGMA tree is available at Supplementary Figure 1.

Figure 5- Phylogenetic tree from the Maximum Likelihood analysis of CYP79F1-F2 and CYP81F1-F4 genes in Brassicaceae with *A. thaliana* CYP79F genes as outgroups. Acronyms are present as the first letter of the genus and the second to species, e.g. At for *Arabidopsis thaliana*, and gene identification when possible. Upon lack of complete CYP annotation, accession numbers were used. Different copies of the same gene are identified by an "X" following sequential numbering, e.g. *A. thaliana* X1, *A. thaliana* X2. Only bootstrap values above 50 are presented. Accession numbers of the sequences analyzed are provided in Supplementary Table 3.

Figure 6– Chemodiversity profiling of GLS in *Brassica* and rocket species (*Diplotaxis* and *Eruca*). Data matrix of GLS chemodiversity is provided in Supplementary Table 4. Colors indicate presence (red) and absence (yellow) of a glucosinolate compound. Bootstraps values above 50 are presented in the clustering phenogram resulted from the Eucledian distances method.

Tables

Table 1- GLS genes information according to sub-pathways of indolic, aliphatic and aromatic. Number of genes - total of genes annotated in each sub-pathway; Number of specific genes - genes exclusive to a given sub-pathway; Number of shared genes - genes shared in at least two sub-pathways.

Supplementary Material

Supplementary Figure 1- Detailed UPGMA phenogram of the 101 GLS biosynthetic genes. Bootstrap values above 70 are represented on the branches.

Supplementary Figure 2- PCA analysis using the 101 GLS genes (**A**) and PCAs loading plots (**B**) of PCA 1 (above) and PCA 2 (below).

Supplementary Table 1- Gene compilation of GLS biosynthetic pathway using a multi-databasing approach. For each gene, accession numbers annotated for *A. thaliana* are provided, alongside the number of sequences available at NCBI database, restricted to Brassicaceae.

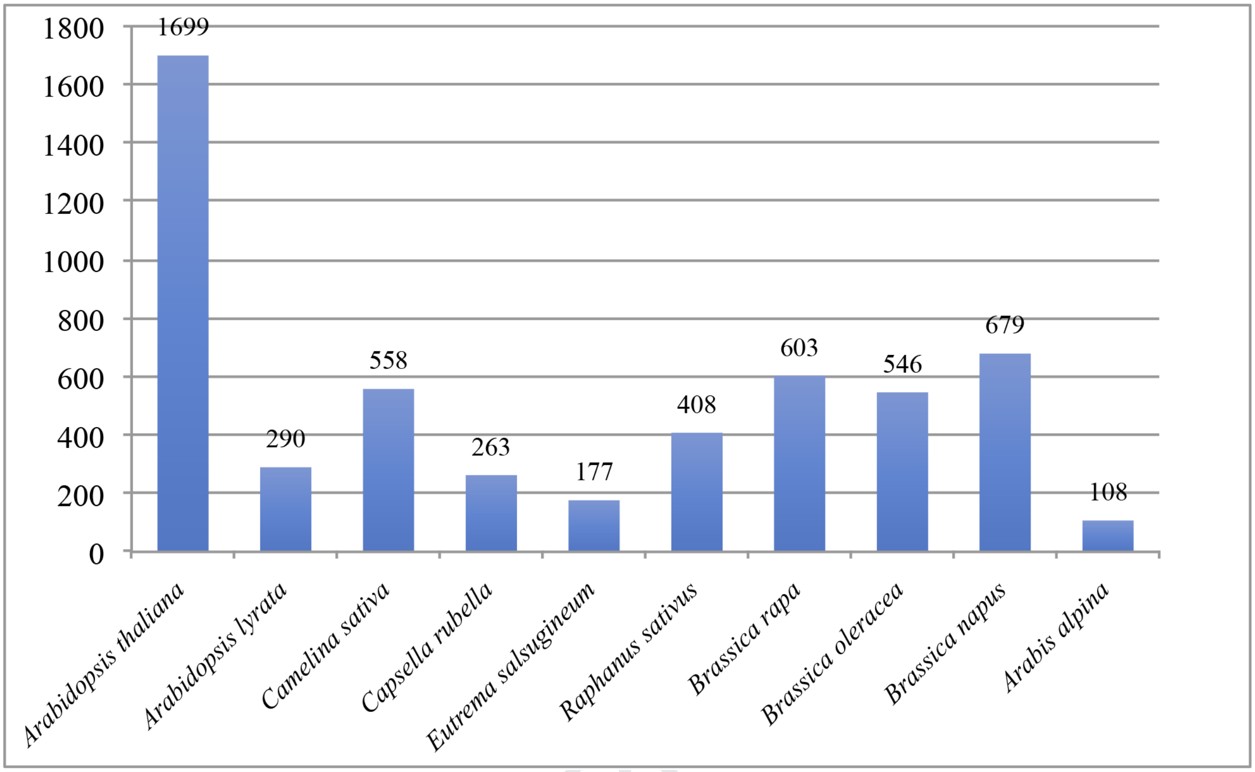
Supplementary Table 2- GLS gene classification according to each sub-pathway (indolic, aliphatic and aromatic).

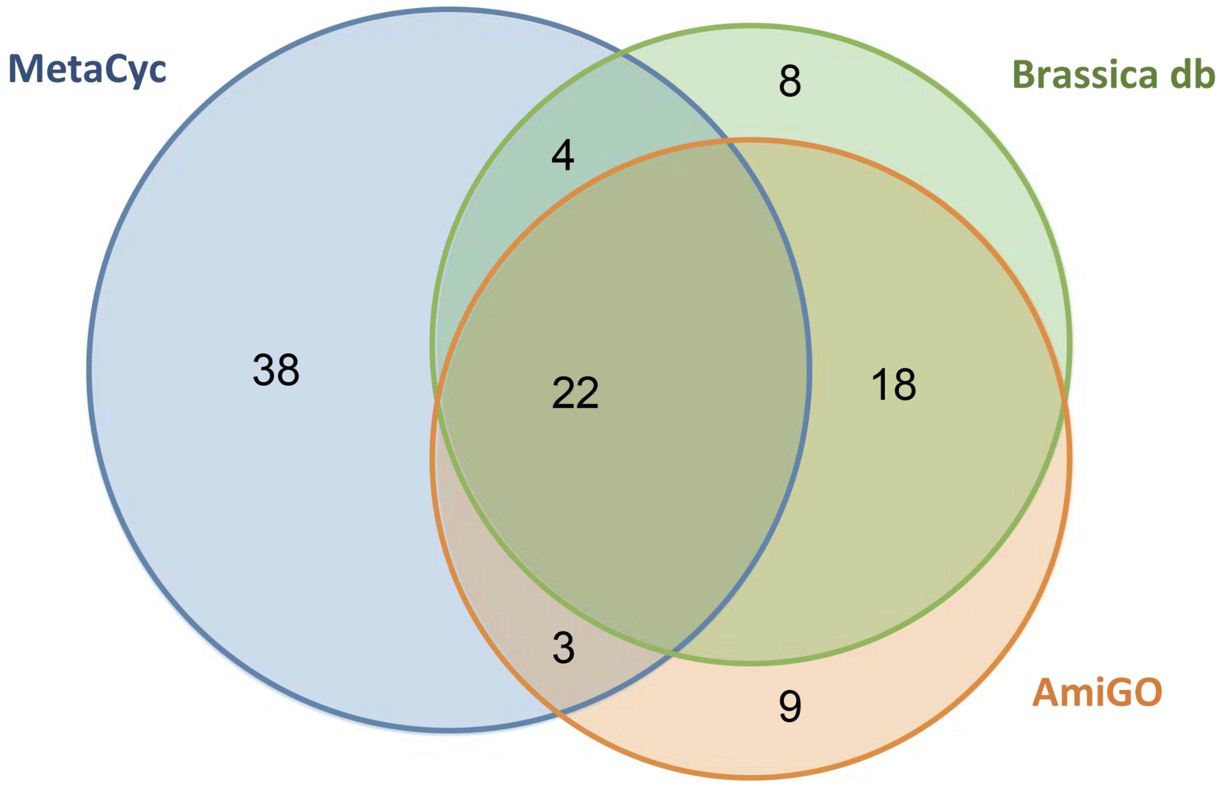
Supplementary Table 3- *CYP79F1-F2* and *CYP81F1-F4* sequences retrieved from NCBI database to perform phylogenetic analysis in Brassicaceae available species. Accession numbers, species and number of sequences are provided, together with code identification used in the Maximum Likelihood tree.

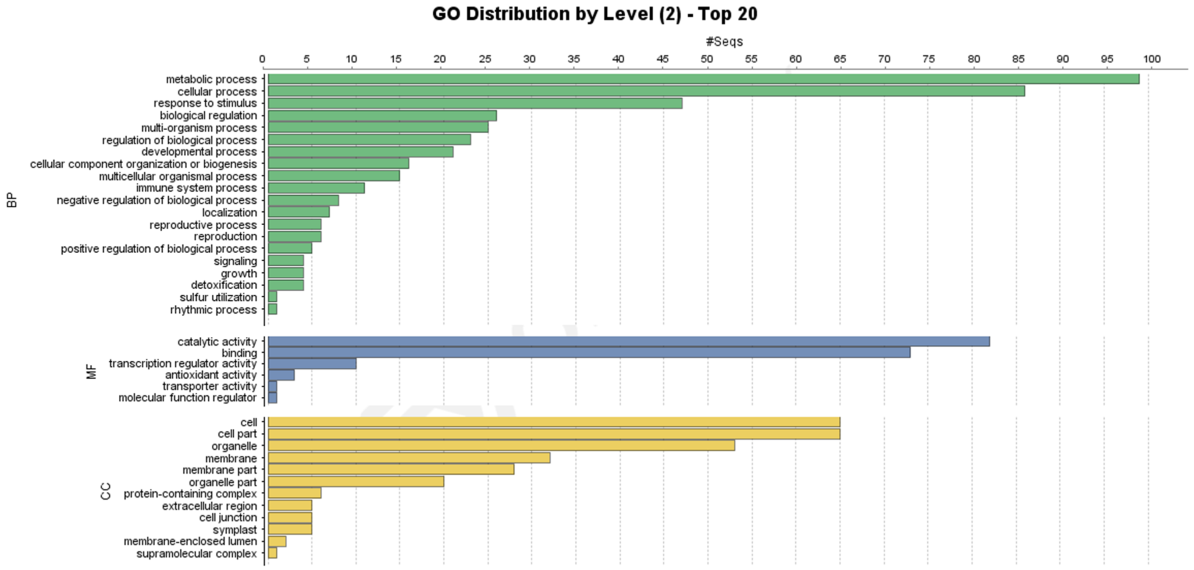
Supplementary Table 4- Data matrix of GLS used for a chemodiversity snapshot on *Brassica* species (*B. napus*, *B. olearacea*, *B. rapa*, *B. juncea*), *Eruca vesicaria* and several wild rocket *Diplotaxis* species.

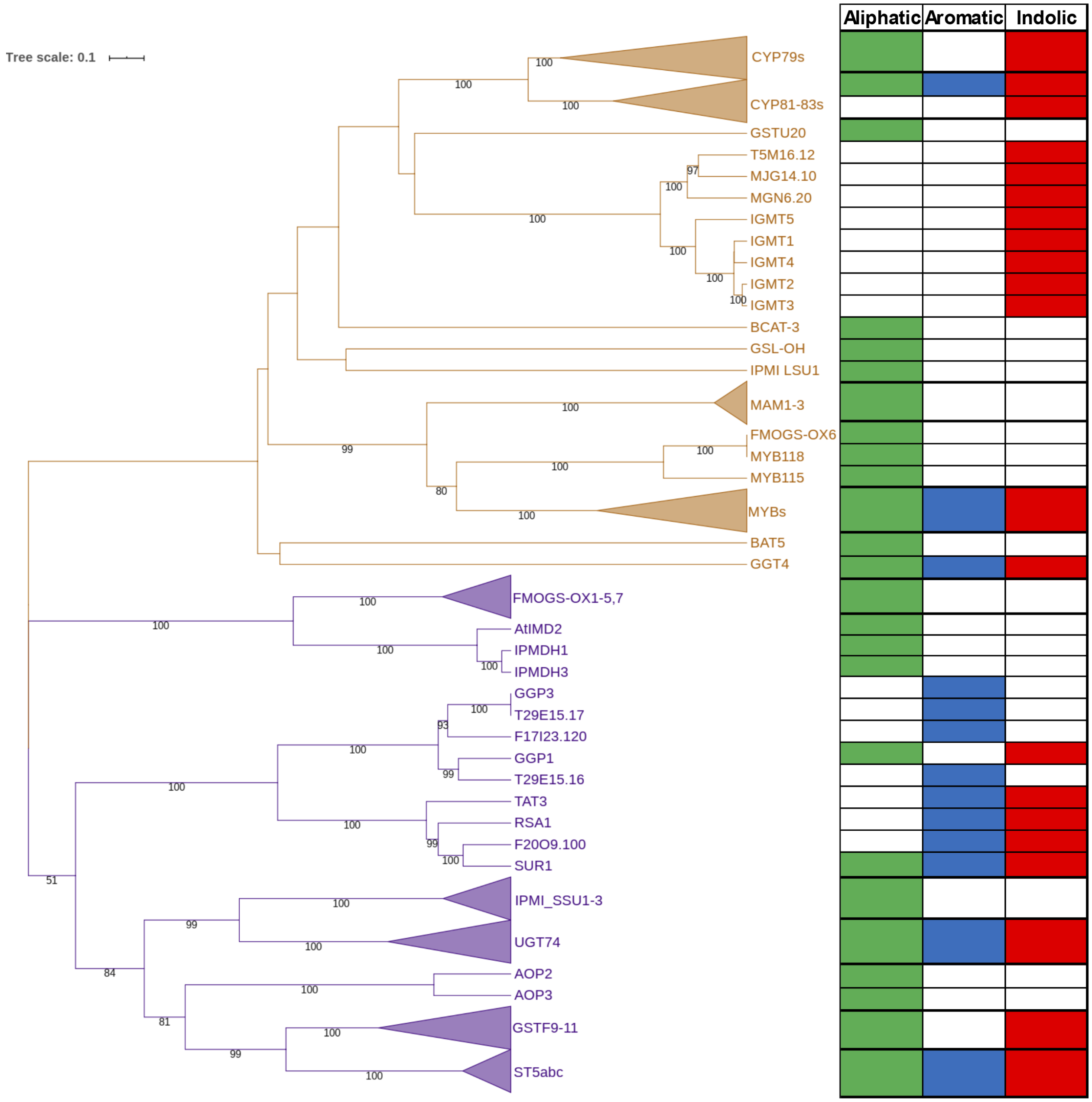
Table 1- GLS genes information according to sub-pathways of indolic, aliphatic and aromatic. Number of genes - total of genes annotated in each sub-pathway; Number of specific genes - genes exclusive to a given sub-pathway; Number of shared genes - genes shared in at least two sub-pathways.

	Aliphatic	**Aromatic**	**Indolic**	**Combined unigenes of the 3 pathways**
Nº. Genes	40	20	41	78
Nº. Specific Genes	31	6	26	-
Nº. Shared Genes	9	14	14	8

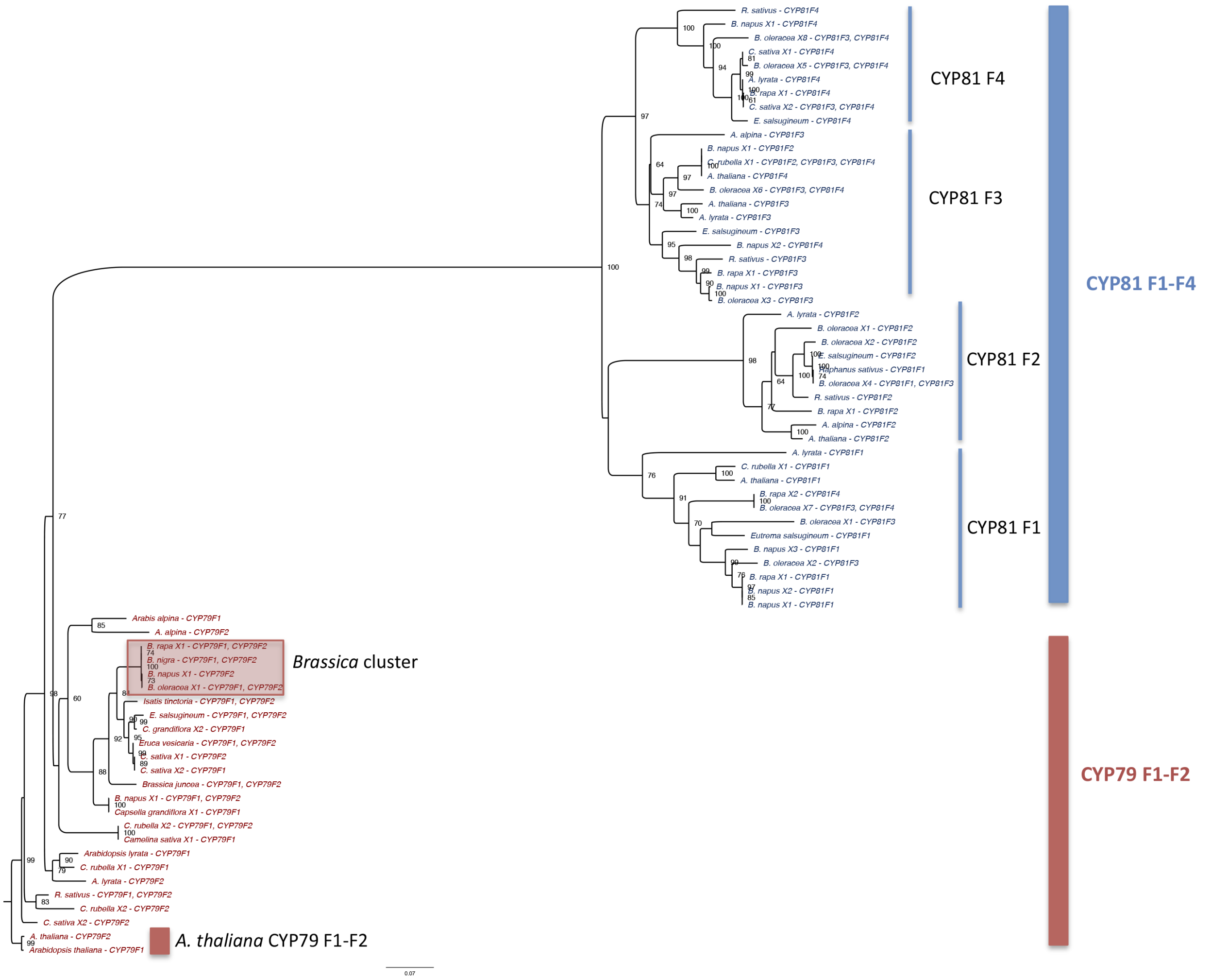

1800
1600
1400
1200
1000
800
600
400
200
0
1699
290
558
263
177
408
603
546
679
108
Arabidopsis thaliana
Arabidopsis lyrata
Camelina sativa
Capsella rubella
Eutrema salsugineum
Raphanus sativus
Brassica rapa
Brassica oleracea
Brassica napus
Arabis alpina

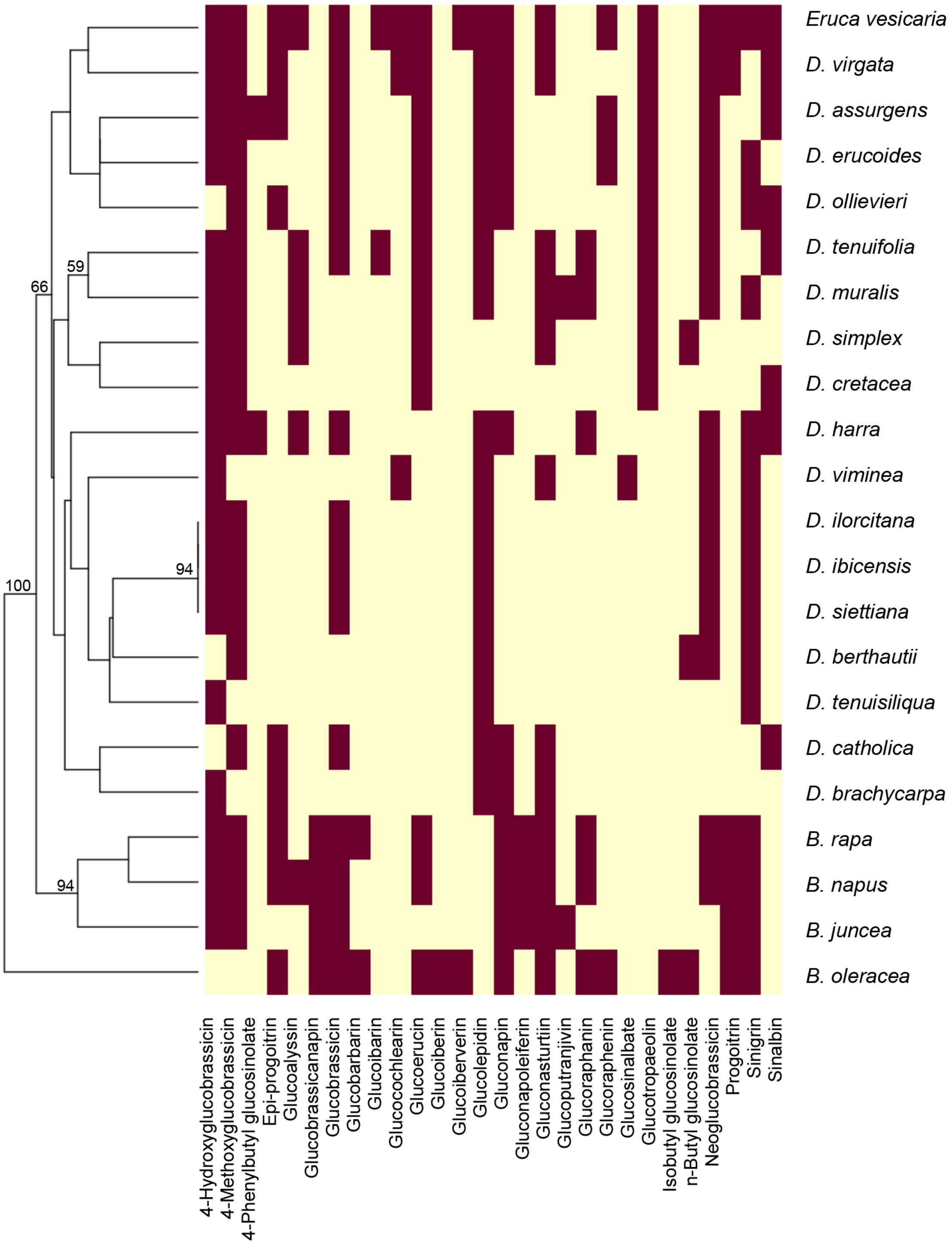

MetaCyc
Brassica db
AmiGO
8
4
38
22
18
3
9


GO Distribution by Level (2) - Top 20
#Seqs
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
BP
metabolic process
cellular process
response to stimulus
biological regulation
multi-organism process
regulation of biological process
developmental process
cellular component organization or biogenesis
multicellular organismal process
immune system process
negative regulation of biological process
localization
reproductive process
reproduction
positive regulation of biological process
signaling
growth
detoxification
sulfur utilization
rhythmic process
MF
catalytic activity
binding
transcription regulator activity
antioxidant activity
transporter activity
molecular function regulator
CC
cell
cell part
organelle
membrane
membrane part
organelle part
protein-containing complex
extracellular region
cell junction
symplast
membrane-enclosed lumen
supramolecular complex



	Aliphatic	Aromatic	Indolic
CYP79s	🟩	⬜	🟥
CYP81-83s	🟩	🟦	🟥
	⬜	⬜	🟥
GSTU20	🟩	⬜	⬜
T5M16.12	⬜	⬜	🟥
MJG14.10	⬜	⬜	🟥
MGN6.20	⬜	⬜	🟥
IGMT5	⬜	⬜	🟥
IGMT1	⬜	⬜	🟥
IGMT4	⬜	⬜	🟥
IGMT2	⬜	⬜	🟥
IGMT3	⬜	⬜	🟥
BCAT-3	🟩	⬜	⬜
GSL-OH	🟩	⬜	⬜
IPMI LSU1	🟩	⬜	⬜
MAM1-3	🟩	⬜	⬜
FMOGS-OX6	🟩	⬜	⬜
MYB118	🟩	⬜	⬜
MYB115	🟩	⬜	⬜
MYBs	🟩	🟦	🟥
BAT5	🟩	⬜	⬜
GGT4	🟩	🟦	🟥
FMOGS-OX1-5,7	🟩	⬜	⬜
AtIMD2	🟩	⬜	⬜
IPMDH1	🟩	⬜	⬜
IPMDH3	🟩	⬜	⬜
GGP3	⬜	🟦	⬜
T29E15.17	⬜	🟦	⬜
F17I23.120	⬜	🟦	⬜
GGP1	🟩	⬜	🟥
T29E15.16	⬜	🟦	⬜
TAT3	⬜	🟦	🟥
RSA1	⬜	🟦	🟥
F20O9.100	⬜	🟦	🟥
SUR1	🟩	🟦	🟥
IPMI_SSU1-3	🟩	⬜	⬜
UGT74	🟩	🟦	🟥
AOP2	🟩	⬜	⬜
AOP3	🟩	⬜	⬜
GSTF9-11	🟩	⬜	🟥
ST5abc	🟩	🟦	🟥


R. sativus - CYP81F4
100
B. napus X1 - CYP81F4
B. oleracea X8 - CYP81F3, CYP81F4
100
C. sativa X1 - CYP81F4
81
B. oleracea X5 - CYP81F3, CYP81F4
94
99
A. lyrata - CYP81F4
100
B. rapa X1 - CYP81F4
100
61
C. sativa X2 - CYP81F3, CYP81F4
E. salsugineum - CYP81F4
CYP81 F4
97
A. alpina - CYP81F3
B. napus X1 - CYP81F2
64
C. rubella X1 - CYP81F2, CYP81F3, CYP81F4
100
97
A. thaliana - CYP81F4
B. oleracea X6 - CYP81F3, CYP81F4
97
74
A. thaliana - CYP81F3
100
A. lyrata - CYP81F3
CYP81 F3
E. salsugineum - CYP81F3
95
B. napus X2 - CYP81F4
98
R. sativus - CYP81F3
100
99
B. rapa X1 - CYP81F3
90
B. napus X1 - CYP81F3
100
B. oleracea X3 - CYP81F3
CYP81 F1-F4
A. lyrata - CYP81F2
B. oleracea X1 - CYP81F2
B. oleracea X2 - CYP81F2
100
E. salsugineum - CYP81F2
98
100
Raphanus sativus - CYP81F1
100
74
B. oleracea X4 - CYP81F1, CYP81F3
CYP81 F2
64
R. sativus - CYP81F2
77
B. rapa X1 - CYP81F2
100
A. alpina - CYP81F2
A. thaliana - CYP81F2
A. lyrata - CYP81F1
C. rubella X1 - CYP81F1
76
100
A. thaliana - CYP81F1
91
B. rapa X2 - CYP81F4
100
B. oleracea X7 - CYP81F3, CYP81F4
70
B. oleracea X1 - CYP81F3
Eutrema salsugineum - CYP81F1
CYP81 F1
B. napus X3 - CYP81F1
99
B. oleracea X2 - CYP81F3
76
B. rapa X1 - CYP81F1
97
B. napus X2 - CYP81F1
85
B. napus X1 - CYP81F1
77
Arabis alpina - CYP79F1
85
A. alpina - CYP79F2
B. rapa X1 - CYP79F1, CYP79F2
74
B. nigra - CYP79F1, CYP79F2
100
Brassica cluster
B. napus X1 - CYP79F2
73
B. oleracea X1 - CYP79F1, CYP79F2
98
60
84
Isatis tinctoria - CYP79F1, CYP79F2
E. salsugineum - CYP79F1, CYP79F2
90
99
C. grandiflora X2 - CYP79F1
92
95
Eruca vesicaria - CYP79F1, CYP79F2
99
C. sativa X1 - CYP79F2
89
C. sativa X2 - CYP79F1
88
CYP79 F1-F2
Brassica juncea - CYP79F1, CYP79F2
B. napus X1 - CYP79F1, CYP79F2
100
Capsella grandiflora X1 - CYP79F1
C. rubella X2 - CYP79F1, CYP79F2
100
Camelina sativa X1 - CYP79F1
Arabidopsis lyrata - CYP79F1
99
90
C. rubella X1 - CYP79F1
79
A. lyrata - CYP79F2
R. sativus - CYP79F1, CYP79F2
83
C. rubella X2 - CYP79F2
C. sativa X2 - CYP79F2
A. thaliana - CYP79F2
99
Arabidopsis thaliana - CYP79F1
A. thaliana CYP79 F1-F2
0.07


59
66
94
100
94
Eruca vesicaria
D. virgata
D. assurgens
D. erucoides
D. ollievieri
D. tenuifolia
D. muralis
D. simplex
D. cretacea
D. harra
D. viminea
D. ilorcitana
D. ibicensis
D. siettiana
D. berthautii
D. tenuisiliqua
D. catholica
D. brachycarpa
B. rapa
B. napus
B. juncea
B. oleracea
4-Hydroxyglucobrassicin
4-Methoxyglucobrassicin
4-Phenylbutyl glucosinolate
Epi-progoitrin
Glucoalyssin
Glucobrassicanapin
Glucobrassicin
Glucobarbarin
Glucoibarin
Glucocochlearin
Glucoerucin
Glucoiberin
Glucoiberverin
Glucolepidin
Gluconapin
Gluconapoleiferin
Gluconasturtiin
Glucoputranjivin
Glucoraphanin
Glucoraphenin
Glucosinalbate
Glucotropaeolin
Isobutyl glucosinolate
n-Butyl glucosinolate
Neoglucobrassicin
Progoitrin
Sinigrin
Sinalbin

Highlights

- Brassicaceae genes involved in GLS biosynthesis were identified using a multi-database approach
- UPGMA and PCA separation between genes in GLS core structure and CYP450/MYB gene families.
- Phylogenetics revealed a recent diversification of aliphatic genes and an earliest for indolic.
- Distinct GLS chemo-profiles between *Brassica* crops and *Diplotaxis* species, wild relatives.
- GLS-rich species as a new source of taxa with great agronomic potential for abiotic stress tolerance.

CONTRIBUTION

The Brassicaceae family is one of the world's most economically important plant groups. They include important crop species (e.g., *Brassica* spp.), weeds (e.g., *Capsella*, *Lepidium*, *Sisymbrium*, and *Thlaspi*), ornamentals (e.g., *Hesperis*, *Lobularia*, and *Matthiola*), and the model organism for flowering plants *Arabidopsis thaliana*. Among the most important chemical compounds produced by Brassicaceae species, are glucosinolates (GLS) with proven and widely documented health promoting effects. Glucosinolates have been the subject of several studies in Brassicaceae as important chemical compounds, particularly in chemical assessment in commercial crops, and also on the characterization of its biochemical pathway reconstruction. However, an integrated approach covering genomic, phylogenetic and chemical analysis in GLS pathway in Brassicaceae remains limited. There are several novel and important aspects to our paper, namely it is the first time where a taxa approach is performed on GLS pathway genes in Brassicaceae species, while in *A. thaliana* its assessment has been extensively studied.

In our paper, we looked through gene clusters involved in the biosynthesis of GLS, by combining genome analysis with biochemical pathways and chemical diversity assessment. Considering the high diversity in GLS content in Brassicaceae species, an integrated approach was performed by assessing a global GLS gene inventory in Brassicaceae and its diversity, analysing a potential genetic basis for GLS divergence using 6 CYP genes (CYP79F1-F2 and CYP81F1-F4), known to be key genes of indolic and aliphatic GLS biosynthetic pathways, linked to a chemical diversity evaluation of GLS compounds in major *Brassica* crops compared to the wild relative genus (*Diplotaxis*).

Our results point to a recent diversification of the aliphatic CYP79F1 and F2 genes in *Brassica* crops, while for indolic genes a clear separation is observed for CYP81F1-F4 genes, revealing an earliest divergence on this GLS sub-pathway. Chemical diversity snapshot allowed recognizing that *Brassica* and *Diplotaxis* have distinct GLS chemo-profiles, highlighting that the latter genus appears as an extremely GLS-rich species. Given the importance of GLS in abiotic stress tolerance, we have explored *Diplotaxis* species, the closest wild relatives of *Brassica* crops, as a new source of taxa with great agronomic potential. Understanding the genomic diversity responsible for the corresponding GLS biosynthetic pathways linked to the chemical diversity could bring insights for exploring new opportunities for using GLS-rich species, yet unexplored.

In summary, this work provides an integrated framework to analyse the chemical diversity of GLS in Brassicaceae, and provides data that complement current state of the art studies performed in GLS within Brassicaceae to answer a wide range of scientific questions in the fields of the genomic basis of chemical diversity and on species diversity assessment using an integrative approach.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.