Detection of orthologous genes with expression shifts linked to nickel hyperaccumulation across Eudicots

Mélina Gallopin[1*], Christine Drevet[1], Vanesa S. Garcia de la Torre[1], Sarah Jelassi[1], Marie Michel[1,4], Claire Ducos[1,5], Cédric Saule[1], Clarisse Majorel[2], Valérie Burtet-Sarramegna[2], Yohan Pillon[3], Paul Bastide[4], Olivier Lespinet[1], Sylvain Merlot[1*]

[1]*Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), 91198, Gif-sur-Yvette, France*

[2]*Institute of Exact and Applied Sciences (ISEA), Université de la Nouvelle-Calédonie, BP R4, Nouméa Cedex, 98851, New Caledonia*

[3] *Laboratoire des Symbioses Tropicales et Méditerranéennes (LSTM), IRD, INRAE, CIRAD, Institut Agro, Univ. Montpellier, Montpellier, France*

[4] *IMAG, Université de Montpellier, CNRS, 34000, Montpellier, France*

[5] *Present address: Theories and Approaches of Genomic Complexity (TAGC), INSERM U1090, Parc scientifique de Luminy, case 928, 13009 Marseille, France*

[6] *Present address: Hôpital Gustave Roussy, Bat. B2M, 114 rue Edouard Vaillant, 94805, Villejuif, France*

[*] For correspondence: melina.gallopin@i2bc.paris-saclay.fr; sylvain.merlot@i2bc.paris-saclay.fr

Keywords: differential expression analysis, RNA-seq, phylogenetic comparative methods, ortholog groups, nickel hyperaccumulation

Abstract

The remarkable capacity of plants to tolerate and accumulate tremendous amount of nickel is a complex adaptative trait that appeared independently in more than 700 species distributed in about fifty families. Nickel hyperaccumulation is thus proposed as a model to investigate the evolution of complex traits in plants. However, the mechanisms involved in nickel hyperaccumulation are still poorly understood in part because comparative transcriptomic analyses struggle to identify genes linked to this trait from a wide diversity of species. In this work, we have implemented a methodology based on the quantification of the expression of orthologous groups and phylogenetic comparative methods to identify genes which expression is correlated to the nickel hyperaccumulation trait. More precisely, we performed *de novo* transcriptome assembly and reads quantification for each species on its own transcriptome using available RNA-Seq datasets from 15 nickel hyperaccumulator and non-accumulator species. Assembled contigs were associated to orthologous groups built using proteomes predicted from completed plant genome sequences. We then analyzed the transcription profiles of 5953 orthologous groups from distant species using a phylogenetic ANOVA. We identified 31 orthologous groups with an expression shift associated with nickel hyperaccumulation. These orthologous groups correspond to genes that have been previously implicated in nickel accumulation, and to new candidates involved in this trait. We thus believe that this method can be successfully applied to identify genes linked to other complex traits from a wide diversity of species.

Introduction

Comparative biology is a fundamental strategy to study the evolution of complex developmental and physiological traits. The development of the RNA-Seq technology has opened the possibility to perform comparative studies at the molecular level in a wide diversity of plant species (Leebens-Mack *et al.*, 2019). In plants, RNA-Seq data have been used to compare distant species for phylogenetic tree inference and to study the evolution of complex traits such as the type of photosynthesis or the capacity to establish symbioses (Jiao *et al.*, 2011; Wickett *et al.*, 2014; Yang *et al.*, 2015; Heyduk *et al.*, 2019; Radhakrishnan *et al.*, 2020; Rich *et al.*, 2021). In most of these studies, RNA-Seq data are used to reveal gene loss, gene duplication and genetic variations linked to a specific trait in different phylae. However, because of the difficulty to use a unique sequence as reference, comparative studies rarely take full advantage of the quantitative information enclosed in RNA-Seq data to compare gene expression between distant species and identify genes which expression is linked to a particular trait (Roux *et al.*, 2015; Voelckel *et al.*, 2017; García de la Torre *et al.*, 2021; Rich *et al.*, 2021).

Metal hyperaccumulation represents an interesting case study to identify genes linked to a complex adaptative trait over a wide diversity of plant species (Manara *et al.*, 2020). Metal hyperaccumulation is defined as the capacity of plant species to accumulate in their leaves a high concentration of metal, such as nickel, manganese, zinc or cadmium, that is normally toxic for the vast majority of plants (van der Ent *et al.*, 2013). Today, about 700 plant species are known to hyperaccumulate metals but the large majority (*ie* 500 species) hyperaccumulates nickel (Reeves *et al.*, 2018). Nickel hyperaccumulators are distributed in about 50 families among more than 300 families of dicotyledon plants suggesting that the nickel hyperaccumulation trait appeared independently in several clades along plant evolution (Krämer, 2010; Cappa and Pilon-Smits,

2014). Comparative transcriptomic analysis of zinc and cadmium hyperaccumulators and non-accumulator species of the Brassicaceae family has first revealed that metal hyperaccumulation is linked to the high and constitutive expression of several genes involved in metal transport and homeostasis (Hammond *et al.*, 2006; Weber *et al.*, 2006; Hanikenne *et al.*, 2008; Halimaa *et al.*, 2014). Our knowledge of the molecular mechanisms involved in nickel hyperaccumulation is still limited but, as for the hyperaccumulation of zinc and cadmium, the hyperaccumulation of nickel likely evolved from the high and constitutive expression of genes involved in metal homeostasis. Halimaa *et al.* (2014) used SOLiD-based RNA-Seq approach to compare the expression of genes from three accessions of *Noccaea caerulescens* (Brassicaceae) with various abilities to tolerate and accumulate metals in order to identify genes linked to metal hyperaccumulation including nickel. In this study, the authors used the genome of the related model species *Arabidopsis thaliana* as a common reference to align RNA-Seq reads. Using an Illumina-based RNA-Seq approach, Meier *et al.* (2018) compared the expression of genes from various populations of *Senecio coronatus* (Asteraceae) hyperaccumulating (NiH) or not (NA) nickel. The authors generated a *S. coronatus* reference transcriptome by *de novo* assembly to quantify gene expression and identify differentially expressed genes in both type of populations. More recently, we used the same RNA-Seq technology to identify genes differentially expressed in pairs of NiH and closely related NA species from five distant plant families (García de la Torre *et al.*, 2021). Then, to identify convergent mechanisms involved in nickel accumulation, we used orthologous relationship between genes from these distant families and a multiple testing correction to identify orthologous groups (OG) containing genes differentially expressed between NiH and NA species in at least 3 plant families.

These comparative approaches rely on the possibility to have access to pairs of closely related species or populations with contrasting capacity to accumulate metals in order to use a common

reference sequence. However, the identification of such pairs of species is not possible in all plant clades. In addition, the output of these analyses strongly depends on the specific pair of species that have been selected for the study. Finally, none of these studies take into account the phylogenetic tree and the genetic drift associated with the selected species.

Phylogenetic relationships between species are known to induce correlations between trait measurements that can affect the analyses when ignored (Felsenstein, 1985). Phylogenetic Comparative Methods (PCMs) precisely aim at taking these relationships into account, and have been extensively studied over the last few decades (Harmon, 2019). In the context of gene expression, Bedford & Hartl (2009) studied several stochastic models, including the Ornstein-Uhlenbeck (OU) process. This process can be seen as modeling the evolution of a quantitative trait under stabilizing selection towards an optimal value (Hansen, 1997). Building on this process, Rohlfs and Nielsen (2015) proposed a phylogenetic ANOVA framework that takes into account both phylogenetic and individual variations. Individual variations represent both intra-specific variations and measurement errors, and ignoring them can lead to severe bias in PCMs (Silvestro *et al.*, 2015; Cooper *et al.*, 2016). This framework has been used to detect OG with significant mean expression shifts across groups of species from various clades of animal kingdom (Rohlfs and Nielsen, 2015; Stern and Crandall, 2018; Chen *et al.*, 2019; Catalán *et al.*, 2019).

In this work, we have implemented a methodology to identify orthologous groups (OG) with expression shifts linked to the nickel hyperaccumulation trait in a wide diversity of plant species. This method uses RNA-Seq datasets to produce reference transcriptomes by *de novo* assembly and quantify gene expression in each species. Genes from the different species are then associated to OG and the expression of OG identified in all species is then analyzed by PCM. Using this methodology, we have identified OGs previously associated with nickel hyperaccumulation as well

as new candidate genes involved in this trait. We believe that this methodology can be used more generally to identify genes associated to complex traits in a wide diversity of species.

Materials and Methods

Collection of RNA-Seq datasets

RNA-Seq datasets used in this study were collected from NCBI bioprojects PRJNA476917 (García de la Torre *et al.*, 2021), PRJNA312157 (Meier *et al.*, 2018) and PRJNA657163. The selected samples correspond to RNA extracted from leaves of nickel hyperaccumulator (NiH) and non-accumulator (NA) species or accessions from the families Asteraceae, Brassicaceae, Cunoniaceae, Phyllanthaceae, Rubiaceae and Salicaceae (representing five orders of Eudicots) and sequenced with the Illumina HiSeq2000 paired-end sequencing technology. Information on the different samples is summarized in **Table 1.**

Transcriptome assembly and expression quantification

Several *de novo* assembled transcriptomes used in this study were previously published and available from the bioproject PRJNA476917 (García de la Torre *et al.*, 2021). The transcriptomes of *Senecio coronatus* and *Microthlaspi perfoliatum* were assembled *de novo* with QIAGEN CLC Genomics Workbench v9 using the same assembly parameters (similarity ≥ 0.95; length fraction ≥ 0.75) as used in García de la Torre *et al.* (2021). For these assemblies, we used the paired-end Illumina reads SRX1901479 (*S. coronatus*), SRX8947157 and SRX8947158 (*M. perfoliatum*). For each species, the reads of each sample were mapped to the corresponding *de novo* transcriptome using CLC Genomics Workbench v9 (similarity ≥ 0.875; length fraction ≥ 0.75).

131

132 ***Construction of ortholog group seeds and annotation***

133 The sequence of predicted proteomes encoded by 12 plant genomes were downloaded from the
134 PLAZA 4.0 Dicots database (Van Bel *et al.*, 2018): *Arabidopsis thaliana* and *Brassica rapa*
135 (Brassicaceae), *Gossypium raimondii* and *Theobroma cacao* (Malvaceae), *Carica papaya*
136 (Caricaceae), *Prunus persica* (Rosaceae), *Cucumis melo* (Cucurbitaceae), *Glycine max* (Fabaceae),
137 *Ricinus communis* (Euphorbiaceae), *Populus trichocarpa* (Salicaceae), *Solanum lycopersicum*
138 (Solanaceae) and *Coffea canephora* (Rubiaceae). We used the meta-approach MARIO to build
139 ortholog group (OG) seeds using the 12 proteome sequences (Pereira *et al.*, 2014).
140 We performed the annotation of these OGs using the HMMER package (Eddy, 1998). For each
141 OG resulting from the MARIO output, we first performed a multiple alignment with MUSCLE
142 (Edgar, 2004) and created a HMM profile using hmmbuild. A profile database of the OG seeds
143 was created using hmmpress. To annotate the OGs, we searched for the closest homologs in the
144 SwissProt database with hmmsearch using the HMM profiles as queries. We extracted the function,
145 EC numbers, and GO terms from the hmmsearch hits with the lowest e-value (e-value ≤ 10e-45)
146 and transfered these annotations to the corresponding OGs.

147

148 ***Assignment of contigs to Orthologous Groups***

149 For each assembled transcriptome, we searched in each contig for the longest ORF on the forward
150 and reverse strands and translated this ORF to obtain the putative encoded protein. The assignment
151 of contigs to OGs was performed with the HMMER package (Eddy, 1998). We performed
152 hmmscan using the translation of the longest ORF of each contig as a query against the OG seeds

153 profile database. We assigned contigs to the OG profile having the lowest e-value (e-value ≤ 1e-
154 10, coverage ≥ 20%).
155
156 ***OG expression matrix construction***
157 We built an OG expression matrix associating each OG to its level of expression in each sample.
158 The expression level of an OG was calculated as the sum of the read counts corresponding to the
159 contigs assigned to this OG in each sample. We also summed the lengths of all contigs assigned
160 to a single OG to obtain an OG length matrix.
161
162 ***Data normalization and transformation***
163 We computed the normalization factors n_i for each sample i using the TMM method, implemented
164 in the function calcNormFactor of the edgeR package (Robinson and Oshlack, 2010). To take the
165 length of the OGs (sums of the lengths of each contigs within each OG) and the size of libraries
166 into account, we then computed log2 RPKM [reads per kilobase per million reads, (Mortazavi *et*
167 *al.*, 2008)] using the following formula:

168 $$y_{gi} = \log2\left(\frac{\frac{c_{gi} + 0.5}{l_{gi}}}{C_i n_i + 1} \times 10^9\right)$$

169 where $C_i = \sum_{g=1}^{G} c_{gi}$ is the library size of sample i, c_{gi} is the read count for OG g in sample i, and
170 l_{gi} is the length of the OG g in sample i. Note that the length can vary between samples for the
171 same OG, as samples are taken from different species. To ensure that the ratio inside the log is

strictly less than 1 and greater than zero, c_{gi} and $C_i n_i$ were offset away from zero by adding 0.5 and 1 respectively (Law *et al.*, 2014).

To perform PCA on the OG expression matrix, we used the plot.PCA function of the R package DESeq2. We used the rlog function to log-transform the data prior to the PCA.

Phylogenetic tree

A custom dated phylogenetic tree was built from the plant tree backbone at the family level proposed by Magallón *et al.* (2015). The tree topology and time divergence used in our study are based on (Barrabé *et al.*, 2014; Igea *et al.*, 2015; Razafimandimbison *et al.*, 2017) for Rubiaceae, (Pillon *et al.*, 2014) for Cunoniaceae and (Huang *et al.*, 2016) for Brassicaceae.

Phylogenetic ANOVA

We used the R package phylolm (Ho and Ané, 2014) to perform phylogenetic ANOVA, using the "OU with fixed root" model, with measurement error. In a phylogenetic regression using an OU with fixed root model, the residuals are assumed to be correlated with a correlation between two species that depends on their shared evolutionary time (*ie* the time between the root of the tree and the most recent common ancestor of the two species). In addition to the phylogenetic residuals, we included in the model independent identically distributed residuals to capture additional non-phylogenetic variance, that we fitted to the data. The factor of interest was the nickel hyperaccumulation capacity of the species or accession. The design matrix also included an intercept, and a factor representing the country of origin of the sample. For each OG, a p-value was computed, corresponding to a t-test (with correlated observations) on the coefficient associated to the hyperaccumulator factor. A Benjamini-Hochberg multiple testing correction was applied to the

195 vector of p-values. We used a threshold of 0.01, and selected OGs with a log2 fold change ≥ 1.5 or

196 ≤ -1.5.

Results

A methodology for detection of orthologous genes with mean expression shifts in nickel hyperaccumulators from distant plant families

In this work we wanted to develop a methodology to identify genes whose expression is linked to nickel hyperaccumulation across a wide diversity of plant species (**Figure 1**). We took advantage of RNA-Seq datasets previously generated from nickel hyperaccumulator (NiH) and related non-accumulator (NA) species or populations to generate *de novo* transcriptome assemblies for each species and then use these transcriptomes as references to quantify gene expression for each sample corresponding to these species. This methodology also uses the concept of orthologous groups (OG) to annotate genes putatively playing conserved functions in distant plant families (Altenhoff *et al.*, 2012). Finally, we quantified the expression of OG in NiH and NA species groups and analyzed the data with a Phylogenetic Comparative Method (PCM) to identify OG with an expression shift linked to the nickel hyperaccumulation trait.

De novo transcriptome assembly and quantification of contig expression from nickel hyperaccumulator and non-accumulator species

We used available RNA-Seq from nickel hyperaccumulator (NiH) and non-accumulator (NA) species or accessions that were generated using Illumina paired-end technology (**Table 1**). For most of the NiH and NA species used in this study, the transcriptomes were previously assembled *de novo* using the CLC Genomic Workbench software (García de la Torre *et al.*, 2021). We assembled the transcriptomes of *Senecio coronatus* and *Microthlaspi perfoliatum* using the same parameters. The number of contigs and the median size of the contigs for each species is given in

Table 2. The number of contigs obtained for *M. perfoliatum* is significantly higher than for the other species probably due to the tetraploid nature of this species and because we assembled reads from both roots and shoots samples.

The RNA-Seq reads from each replicate corresponding to the leaf samples of all NiH and NA species or populations (**Table 1**) were mapped to the corresponding *de novo* transcriptome. This generated a read count table for each contig in all sample replicate from each species.

Construction of orthologous group database and assignation of contigs

In the second step of our methodology, we wanted to establish orthologous relationships between genes expressed in the NiH and NA species in order to annotate genes potentially playing similar biological functions across these species (Altenhoff *et al.*, 2012). However, because the translation of *de novo* assembled transcriptomes generates a large number of truncated peptides that could affect the analysis of the orthologous relationships, we decided to first build up an orthologous group (OG) library using proteomes predicted from sequenced plant genomes. We selected 12 sequenced plant genomes chosen along the dicotyledon phylogenetic tree and including species belonging to the same families as the nickel hyperaccumulators (see Methods). These proteomes, containing from 27000 to 56000 peptides, were used to create Orthologous Group (OG) seeds using the MARIO meta-approach (Pereira *et al.*, 2014). MARIO combines the results of four methods: Best Reciprocal Hits, Inparanoid (O'Brien, 2004), OrthoFinder (Emms and Kelly, 2015) and Phylogeny (Lemoine *et al.*, 2007) to establish orthologous relationships between proteins and compute a consensus OG annotation. Using this method, we obtained 17830 OGs containing from 2 to 1466 proteins. We could attribute a function to 11301 OG (63 %) using the Swissprot database and 4486 OGs (25%) were associated to at least one EC number. 7779 OGs (43%) are represented

by at least one peptide in all model species. An HMM profile database was created from these OG seeds.

We then assigned each contig of the assembled transcriptomes from NiH and NA species to the OG seeds constructed with MARIO. For each contig, the longest Open Reading Frame (ORF) was identified and translated into a peptide. Each peptide was then associated to the closest OG using the HMM profile database (see Methods). Depending on the species, 44 to 69% of the contigs were assigned to an OG (**Table 2**). In total, among the 17830 OGs generated with the proteome of model plants, 15941 OGs are represented by at least one contig from at least one plant species of interest. It is important to note here that the RNA-Seq data used in this study only represent gene expressed in leaves. More importantly, for cross-species comparison, 5953 OGs are represented by at least one contig in each studied species.

Identification of differentially expressed Ortholog Groups between nickel hyperaccumulator and non-accumulator species

We focused our differential expression analysis on the 5953 OGs represented in all plant species. We built an OG expression matrix (see Methods), each row representing one of the 5953 OGs and each column representing an RNA-Seq sample from 9 nickel hyperaccumulator (NiH) and 9 non-accumulator (NA) species or populations (2 to 3 biological replicates per species). We first performed principal component analysis (PCA) of the dataset after normalization and a log transformation of the expression data to explore the correlation between samples (**Figure 2**). The representation of the two principal components having the most important effect on total variation indicates that the RNA-Seq samples do not cluster with respect to the NiH trait but rather with respect to the plant family to which they belong. This result further illustrates that it is important

266 to take into consideration the phylogenetic relationships between species in the transcriptomic
267 comparison between the NiH and NA groups.

268 To identify OGs differentially expressed between NiH and NA species from distant plant families,
269 we used a phylogenetic mixed model (Lynch, 1991; Housworth et al., 2004) implemented in the R
270 package phylolm (Ho and Ané, 2014).

271 Using this approach, we identified 31 OGs differentially expressed between NiH and NA species
272 with a log2FC threshold ≥ 1.5 or ≤ -1.5. and an FDR ≤ 0.01 (**Table 3**). We used the length
273 normalized unit RPKM to compare the expression of OG between samples. We also performed the
274 analysis with the TPM unit (transcripts per million) (Wagner *et al.*, 2012), but the results did not
275 differ significantly (data not shown). 17 OGs are more expressed in the group of NiH species. The
276 expression in NiH and NA species of OG 4147 and OG 7137, coding for a histidinol dehydrogenase
277 and a membrane transporter of the SLC40A family respectively, is presented in **Figure 3**. These
278 data illustrate the higher expression of these OG in several NiH species or populations compared
279 to NA species. In contrast, 14 OG are less expressed in NiH species including OGs coding for a
280 calcium-transporting ATPase (OG 10076) and a probable receptor-like serine/threonine-protein
281 kinase related to Lr10 (OG 13722).

Discussion

The comparison of the expression of genes playing conserved function over a wide diversity of plant species is still a challenging task in comparative and evolution biology. The goal of this work was to implement a method using the annotation of orthologous groups (OG) and a Phylogenetic Comparative Method (PCM) to compare the transcriptomes of evolutionary distant species to identify genes whose expression level is linked with the nickel hyperaccumulation trait.

Identification of orthologous genes with expression shifts linked to nickel hyperaccumulation

Genes linked to the nickel hyperaccumulation trait have been previously searched by comparing gene expression in pairs of closely related species or populations of the same species with contrasted capacity to accumulate nickel (Halimaa *et al.*, 2014; Meier *et al.*, 2018; García de la Torre *et al.*, 2021; Enomoto *et al.*, 2021). In this study, we have used available and comprehensive RNA-Seq datasets to identify OGs with expression shifts between nickel hyperaccumulator (NiH) and non-accumulator (NA) groups of species. Interestingly, we identified OG 7137 (**Table 3**, **Figure 3**) corresponding to the SLC40A membrane transporter family, also known as IREG or Ferroportin (FPN) transporters, as an OG more expressed in NiH compared to NA species. IREG/FPN transporters are able to transport divalent metal ions including nickel across membranes (Schaaf *et al.*, 2006; Morrissey *et al.*, 2009; Billesbølle *et al.*, 2020). In *Arabidopsis thaliana*, two transporters belonging to OG 7137, AtIREG1 and AtIREG2, have been shown to localize on the plasma membrane and the vacuolar membrane respectively (Schaaf *et al.*, 2006; Morrissey *et al.*, 2009). High expression of vacuolar localized IREG transporters in transgenic *A. thaliana* increases nickel tolerance (Schaaf *et al.*, 2006; Merlot *et al.*, 2014). The high expression of IREG/FPN genes

was previously shown to be linked to the nickel hyperaccumulation trait in species from several plant families (Meier *et al.*, 2018; García de la Torre *et al.*, 2021). The identification of OG 7137 thus validates the capacity of our methodology to identify OG linked to nickel hyperaccumulation. It also indicates that the high expression of IREG transporters in leaves is a robust characteristic of nickel hyperaccumulators that can be observed independently of the method used for cross species transcriptomic comparison. Our analysis also revealed a higher expression of OG 8462 and OG 6278, corresponding to the Ferric reductase FRO2 and the Cationic amino acid transporters CAT4 and CAT2 respectively, in the NiH species. These results confirm previous observations made in the Asteraceae species *S. coronatus* (Meier *et al.*, 2018) and further suggest that the high expression of these genes corresponds to convergent mechanisms implicated in metal hyperaccumulation. Recently, the cationic amino acid transporter CAT4, primarily localizing on the vacuolar membrane, was associated with the histidine level trait in *A. thaliana* (Angelovici *et al.*, 2017). Indeed, previous studies have highlighted the role of histidine as an important metal ligand involved in nickel hyperaccumulation in Brassicaceae species (Krämer *et al.*, 1996; Kozhevnikova *et al.*, 2014). Interestingly, our analysis also indicated that a high expression of Histidinol dehydrogenase (OG4147), the last step in histidine biosynthesis, is also associated with the nickel hyperaccumulation trait (**Table 3**, **Figure 3**). These results support a role for histidine synthesis and transport in nickel hyperaccumulation. The high expression of MATE and ABC transporters related to DTX27 (OG 6266) and ALS3 (OG 411), potentially involved in metal transport and tolerance (Larsen *et al.*, 2004; Liu *et al.*, 2009), was not previously associated with nickel hyperaccumulation. The validation of the contribution of these transporters to nickel hyperaccumulation will require further support.

On the contrary, our results suggest a lower expression of calcium-transporting ATPase (OG 10076) linked to the nickel hyperaccumulation trait. The rationale behind this result is not clear but it could be an indirect consequence of the edaphic conditions on which nickel hyperaccumulators are evolving. Indeed, NiH species are naturally growing on ultramafic soils that are rich in metals, including nickel, and with a strong calcium/magnesium imbalance affecting the development of most plant species (*aka* serpentine syndrome; Konečná *et al.*, 2020). Therefore, the lower expression of calcium-transporting ATPase in NiH species might be linked to the adaptation of these species to the serpentine syndrome.

Advantages and limitations of this methodology to compare distant species

One motivation of this work was to valorize the quantitative information contained in numerous RNA-Seq datasets already available from various plant species to identify gene functions associated with nickel hyperaccumulation. Most of these RNA-Seq datasets were produced with Illumina HiSeq paired-end technology. While the high number and quality of reads generated with this technology allow a good estimation of gene expression, their assembly frequently generates truncated ORF, thus affecting the annotation of OGs. To circumvent this limitation, we generated OG seeds using proteomes predicted from sequenced plant genomes. In a near future, the development of long-read sequencing technologies will undoubtably ease the assembly of full-length transcripts (Amarasinghe *et al.*, 2020). In this work, we produced the OG database using the Mario method (Pereira *et al.*, 2014). While this strategy allowed us to compare transcriptomes from distant species, the database is specific to this work and the results are thus difficult to compare with other studies. The use of more general OG databases for comparative and functional genomics

such as Bgee for animals (Bastian *et al.*, 2021) or PLAZA for plants (Van Bel *et al.*, 2022), would thus favor comparisons between studies.

One main advantage of our method is that gene expression is quantified in each species using its own transcriptome. Therefore, this method does not absolutely require the identification of pairs of closely related species or populations with contrasted traits to quantify and compare gene expression using a common reference sequence. The identification of such pairs of closely related species with contrasted traits may represent a limiting condition in some genera. For example, the nickel hyperaccumulator *Blepharidium guatemalense* represents a monotypic genus (Navarrete Gutiérrez *et al.*, 2021), and all species of the Cuban genus *Leucocroton* are able to hyperaccumulate nickel (Reeves *et al.*, 1996). In addition, the output of pairwise comparisons to identify genes linked to a specific trait such as metal hyperaccumulation, strongly depends on the particular pair of species chosen for the comparison (Halimaa *et al.*, 2014; Meier *et al.*, 2018; García de la Torre *et al.*, 2021). Our methodology based on the use of several species belonging to two contrasted phenotypic groups (*eg* NiH and NA) is less sensitive to the choice of species to identify conserved or convergent mechanisms linked in a complex trait.

The orthologous conjecture proposes that orthologous genes are functionally more similar than paralogous genes. While this conjecture is still disputed (Stamboulian *et al.*, 2020), the annotation of groups of orthologous genes or orthologous groups is a widely used strategy for comparative analysis over a wide diversity of species (Van Bel *et al.*, 2022). To compute the level of expression of OGs in each species, we decided to sum the counts of all contigs associated to the same OG as previously used (Lallemand *et al.*, 2019). This choice is consistent with the hypothesis that contigs belonging to the same OG more likely encode for proteins playing the same function. However, OG may also contain inparalog genes, resulting from gene duplication in some species lineages, that might have acquired specific function. This is the case for OG 7137 containing for example

373 *AtIREG1* and *AtIREG2* playing different roles in *A. thaliana* (see above). It is important to notice
374 that this OG does not appear to be more expressed in NiH species from several families. This might
375 be the consequence of the presence of inparalog genes in this OG with contrasted expression levels
376 in species from some families, thus affecting the calculation of the OG expression level.
377 Alternatively, the high expression of IREG in NiH species might not be a mechanism observed in
378 NiH species from all plant families. A more detailed analysis of the expression of contigs
379 composing this OG would be necessary to validate these hypotheses. However, the identification
380 of OG 7137 suggested that the method using the sum of read counts does not prevent the
381 identification of OG containing inparalog genes.
382 Altogether, our results suggest that this methodology based on the quantification of the expression
383 of orthologous groups allows the identification of genes with expression shifts linked to nickel
384 hyperaccumulation from distant plant species. Even though RNA-Seq sequencing technologies and
385 comparative genomics resources are evolving rapidly, we believe that the methodology presented
386 in this work could be used as a framework to identify genes linked to a specific complex trait in a
387 wide diversity of plant species or other organisms.
388

389 ## Data and code availability

390 The *Microthlaspi perfoliatum* Transcriptome Shotgun Assembly project has been deposited at
391 DDBJ/ENA/GenBank under the accession GITW00000000. The version described in this paper is
392 the first version, GITW01000000. The scripts and processed datasets are available at
393 https://github.com/i2bc/plant-nickel-accumulation
394

Contributions

M.G., O.L., S.M. conceived the study, V.S.GdlT., S.J., M.M., C.D., C.S., C.M, V.B. Y.P. collected the data, M.G., C.D., P.B., S.M performed the analyses and interpreted the results, M.G., O.L., S.M. wrote the manuscript. All authors have read and approved the final manuscript.

Acknowledgement / Funding

The authors would like to thank Claire Toffano-Nioche (I2BC) for valuable discussions and support, and Bruno Fogliani (IAC/UNC) for his expertise on New Caledonian species. This work was supported by the X-TreM grant from the CNRS MITI to SM and the special funding MODELCOG from I2BC to MG. RNA-Seq sequencing for *Microthlaspi* and *Phyllanthus* species was financed by the ANR grant ANR-13-ADAP-0004 to SM, VB and BF.

Bibliography

Altenhoff AM, Studer RA, Robinson-Rechavi M, Dessimoz C. 2012. Resolving the Ortholog Conjecture: Orthologs Tend to Be Weakly, but Significantly, More Similar in Function than Paralogs (JA Eisen, Ed.). PLoS Computational Biology **8**, e1002514.

Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. 2020. Opportunities and challenges in long-read sequencing data analysis. Genome Biology **21**, 30.

Angelovici R, Batushansky A, Deason N, Gonzalez-Jorge S, Gore MA, Fait A, DellaPenna D. 2017. Network-Guided GWAS Improves Identification of Genes Affecting Free Amino Acids.

Plant Physiology **173**, 872–886.

Barrabé L, Maggia L, Pillon Y, Rigault F, Mouly A, Davis AP, Buerki S. 2014. New Caledonian lineages of Psychotria (Rubiaceae) reveal different evolutionary histories and the largest documented plant radiation for the archipelago. Molecular Phylogenetics and Evolution **71**, 15–35.

Bastian FB, Roux J, Niknejad A, *et al.* 2021. The Bgee suite: integrated curated expression atlas and comparative transcriptomics in animals. Nucleic Acids Research **49**, D831–D847.

Bedford T, Hartl DL. 2009. Optimization of gene expression by natural selection. Proceedings of the National Academy of Sciences **106**, 1133–1138.

Van Bel M, Diels T, Vancaester E, Kreft L, Botzki A, Van de Peer Y, Coppens F, Vandepoele K. 2018. PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. Nucleic Acids Research **46**, D1190–D1196.

Van Bel M, Silvestri F, Weitz EM, Kreft L, Botzki A, Coppens F, Vandepoele K. 2022. PLAZA 5.0: extending the scope and power of comparative and functional genomics in plants. Nucleic Acids Research **50**, D1468–D1474.

Billesbølle CB, Azumaya CM, Kretsch RC, Powers AS, Gonen S, Schneider S, Arvedson T, Dror RO, Cheng Y, Manglik A. 2020. Structure of hepcidin-bound ferroportin reveals iron homeostatic mechanisms. Nature **586**, 807–811.

Cappa JJ, Pilon-Smits EAH. 2014. Evolutionary aspects of elemental hyperaccumulation. Planta **239**, 267–275.

Catalán A, Briscoe AD, Höhna S. 2019. Drift and Directional Selection Are the Evolutionary Forces Driving Gene Expression Divergence in Eye and Brain Tissue of Heliconius Butterflies.

Genetics **213**, 581–594.

Chen J, Swofford R, Johnson J, Cummings BB, Rogel N, Lindblad-Toh K, Haerty W, di Palma F, Regev A. 2019. A quantitative framework for characterizing the evolutionary history of mammalian gene expression. Genome Research **29**, 53–63.

Cooper N, Thomas GH, Venditti C, Meade A, Freckleton RP. 2016. A cautionary note on the use of Ornstein Uhlenbeck models in macroevolutionary studies. Biological Journal of the Linnean Society **118**, 64–77.

Eddy SR. 1998. Profile hidden Markov models. Bioinformatics **14**, 755–763.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research **32**, 1792–1797.

Emms DM, Kelly S. 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. Genome Biology **16**, 157.

Enomoto T, Yoshida J, Mizuno T, Watanabe T, Nishida S. 2021. Differences in mineral accumulation and gene expression profiles between two metal hyperaccumulators, Noccaea japonica and Noccaea caerulescens ecotype Ganges, under excess nickel condition. Plant Signaling & Behavior **16**, 1945212.

van der Ent A, Baker AJM, Reeves RD, Pollard AJ, Schat H. 2013. Hyperaccumulators of metal and metalloid trace elements: Facts and fiction. Plant and Soil **362**, 319–334.

Felsenstein J. 1985. Phylogenies and the Comparative Method. The American Naturalist **125**, 1–15.

García de la Torre VS, Majorel-Loulergue C, Rigaill GJ, *et al.* 2021. Wide cross-species RNA-Seq comparison reveals convergent molecular mechanisms involved in nickel

460 hyperaccumulation across dicotyledons. New Phytologist **229**, 994–1006.

461 **Halimaa P, Lin Y-F, Ahonen VH, *et al.*** 2014. Gene expression differences between Noccaea
462 caerulescens ecotypes help to identify candidate genes for metal phytoremediation.
463 Environmental science & technology **48**, 3344–53.

464 **Hammond JP, Bowen HC, White PJ, Mills V, Pyke KA, Baker AJM, Whiting SN, May ST,**
465 **Broadley MR**. 2006. A comparison of the Thlaspi caerulescens and Thlaspi arvense shoot
466 transcriptomes. New Phytologist **170**, 239–260.

467 **Hanikenne M, Talke IN, Haydon MJ, Lanz C, Nolte A, Motte P, Kroymann J, Weigel D,**
468 **Krämer U**. 2008. Evolution of metal hyperaccumulation required cis-regulatory changes and
469 triplication of HMA4. Nature **453**, 391–395.

470 **Hansen TF**. 1997. Stabilizing Selection and the Comparative Analysis of Adaptation. Evolution
471 **51**, 1341.

472 **Harmon LJ**. 2019. *Phylogenetic Comparative Methods: Learning From Trees* (LJ Harmon, Ed.).
473 Center for Open Science.

474 **Heyduk K, Ray JN, Ayyampalayam S, Moledina N, Borland A, Harding SA, Tsai C-J,**
475 **Leebens-Mack J**. 2019. Shared expression of crassulacean acid metabolism (CAM) genes pre-
476 dates the origin of CAM in the genus Yucca (J Cushman, Ed.). Journal of Experimental Botany
477 **70**, 6597–6609.

478 **Ho LST, Ané C**. 2014. A Linear-Time Algorithm for Gaussian and Non-Gaussian Trait
479 Evolution Models. Systematic Biology **63**, 397–408.

480 **Housworth EA, Martins EP, Lynch M**. 2004. The Phylogenetic Mixed Model. The American
481 Naturalist **163**, 84–96.

Huang C-H, Sun R, Hu Y, *et al.* 2016. Resolution of Brassicaceae Phylogeny Using Nuclear Genes Uncovers Nested Radiations and Supports Convergent Morphological Evolution. Molecular Biology and Evolution **33**, 394–412.

Igea J, Bogarín D, Papadopulos AST, Savolainen V. 2015. A comparative analysis of island floras challenges taxonomy-based biogeographical models of speciation. Evolution **69**, 482–491.

Jiao Y, Wickett NJ, Ayyampalayam S, *et al.* 2011. Ancestral polyploidy in seed plants and angiosperms. Nature **473**, 97–100.

Konečná V, Yant L, Kolář F. 2020. The Evolutionary Genomics of Serpentine Adaptation. Frontiers in Plant Science **11**.

Kozhevnikova AD, Seregin I V, Erlikh NT, Shevyreva TA, Andreev IM, Verweij R, Schat H. 2014. Histidine-mediated xylem loading of zinc is a species-wide character in Noccaea caerulescens. New Phytol **203**, 508–519.

Krämer U. 2010. Metal Hyperaccumulation in Plants. Annual Review of Plant Biology **61**, 517–534.

Krämer U, Cotter-Howells JD, Charnock JM, Baker AJM, Smith JAC. 1996. Free histidine as a metal chelator in plants that accumulate nickel. Nature **379**, 635–638.

Lallemand F, Martin-Magniette M, Gilard F, Gakière B, Launay-Avon A, Delannoy É, Selosse M. 2019. In situ transcriptomic and metabolomic study of the loss of photosynthesis in the leaves of mixotrophic plants exploiting fungi. The Plant Journal **98**, 826–841.

Larsen PB, Geisler MJB, Jones CA, Williams KM, Cancel JD. 2004. ALS3 encodes a phloem-localized ABC transporter-like protein that is required for aluminum tolerance in Arabidopsis. The Plant Journal **41**, 353–363.

Law CW, Chen Y, Shi W, Smyth GK. 2014. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology **15**, R29.

Leebens-Mack JH, Barker MS, Carpenter EJ, *et al.* 2019. One thousand plant transcriptomes and the phylogenomics of green plants. Nature **574**, 679–685.

Lemoine F, Lespinet O, Labedan B. 2007. Assessing the evolutionary rate of positional orthologous genes in prokaryotes using synteny data. BMC Evolutionary Biology **7**, 237.

Liu J, Magalhaes J V, Shaff J, Kochian L V. 2009. Aluminum-activated citrate and malate transporters from the MATE and ALMT families function independently to confer Arabidopsis aluminum tolerance. The Plant Journal **57**, 389–399.

Lynch M. 1991. Methods for the analysis of comparative data in evolutionary biology. Evolution **45**, 1065–1080.

Magallón S, Gómez-Acevedo S, Sánchez-Reyes LL, Hernández-Hernández T. 2015. A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. New Phytologist **207**, 437–453.

Manara A, Fasani E, Furini A, DalCorso G. 2020. Evolution of the metal hyperaccumulation and hypertolerance traits. Plant, Cell & Environment **43**, 2969–2986.

Meier SK, Adams N, Wolf M, Balkwill K, Muasya AM, Gehring CA, Bishop JM, Ingle RA. 2018. Comparative RNA-seq analysis of nickel hyperaccumulating and non-accumulating populations of *Senecio coronatus* (Asteraceae). The Plant Journal **95**, 1023–1038.

Merlot S, Hannibal L, Martins S, Martinelli L, Amir H, Lebrun M, Thomine S. 2014. The metal transporter PgIREG1 from the hyperaccumulator Psychotria gabriellae is a candidate gene for nickel tolerance and accumulation. J Exp Bot **65**, 1551–1564.

Morrissey J, Baxter IR, Lee J, Li L, Lahner B, Grotz N, Kaplan J, Salt DE, Guerinot M Lou. 2009. The Ferroportin Metal Efflux Proteins Function in Iron and Cobalt Homeostasis in Arabidopsis. The Plant Cell **21**, 3326–3338.

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods **5**, 621–628.

Navarrete Gutiérrez DM, Pollard AJ, van der Ent A, Cathelineau M, Pons M-N, Cuevas Sánchez JA, Echevarria G. 2021. Blepharidium guatemalense, an obligate nickel hyperaccumulator plant from non-ultramafic soils in Mexico. Chemoecology **31**, 169–187.

O'Brien KP. 2004. Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Research **33**, D476–D480.

Pereira C, Denise A, Lespinet O. 2014. A meta-approach for improving the prediction and the functional annotation of ortholog groups. BMC Genomics **15**, S16.

Pillon Y, Hopkins HCF, Rigault F, Jaffré T, Stacy EA. 2014. Cryptic adaptive radiation in tropical forest trees in New Caledonia. New Phytologist **202**, 521–530.

Radhakrishnan G V, Keller J, Rich MK, *et al.* 2020. An ancestral signalling pathway is conserved in intracellular symbioses-forming plant lineages. Nature Plants **6**, 280–289.

Razafimandimbison SG, Kainulainen K, Wikström N, Bremer B. 2017. Historical biogeography and phylogeny of the pantropical Psychotrieae alliance (Rubiaceae), with particular emphasis on the Western Indian Ocean Region. American Journal of Botany **104**, 1407–1423.

Reeves RD, Baker AJM, Borhidi A, Berazaín R. 1996. Nickel-accumulating plants from the ancient serpentine soils of Cuba. New Phytologist **133**, 217–224.

Reeves RD, Baker AJM, Jaffré T, Erskine PD, Echevarria G, Ent A. 2018. A global database

548 for plants that hyperaccumulate metal and metalloid trace elements. New Phytologist **218**, 407–
549 411.

550 **Rich MK, Vigneron N, Libourel C, *et al.*** 2021. Lipid exchanges drove the evolution of
551 mutualism during plant terrestrialization. Science **372**, 864–868.

552 **Robinson MD, Oshlack A**. 2010. A scaling normalization method for differential expression
553 analysis of RNA-seq data. Genome Biology **11**.

554 **Rohlfs R V, Nielsen R**. 2015. Phylogenetic ANOVA: The Expression Variance and Evolution
555 Model for Quantitative Trait Evolution. Systematic Biology **64**, 695–708.

556 **Roux J, Rosikiewicz M, Robinson-Rechavi M**. 2015. What to compare and how: Comparative
557 transcriptomics for Evo-Devo (M Robinson-Rechavi, Ed.). Journal of Experimental Zoology Part
558 B: Molecular and Developmental Evolution **324**, 372–382.

559 **Schaaf G, Honsbein A, Meda AR, Kirchner S, Wipf D, von Wirén N**. 2006. AtIREG2
560 Encodes a Tonoplast Transport Protein Involved in Iron-dependent Nickel Detoxification in
561 Arabidopsis thaliana Roots. Journal of Biological Chemistry **281**, 25532–25540.

562 **Silvestro D, Kostikova A, Litsios G, Pearman PB, Salamin N**. 2015. Measurement errors
563 should always be incorporated in phylogenetic comparative analysis (T Münkemüller, Ed.).
564 Methods in Ecology and Evolution **6**, 340–346.

565 **Stamboulian M, Guerrero RF, Hahn MW, Radivojac P**. 2020. The ortholog conjecture
566 revisited: the value of orthologs and paralogs in function prediction. Bioinformatics **36**, i219–
567 i226.

568 **Stern DB, Crandall KA**. 2018. The Evolution of Gene Expression Underlying Vision Loss in
569 Cave Animals (C Wilke, Ed.). Molecular Biology and Evolution **35**, 2005–2014.

570 **Voelckel C, Gruenheit N, Lockhart P**. 2017. Evolutionary Transcriptomics and Proteomics:
571 Insight into Plant Adaptation. Trends in Plant Science **22**, 462–471.

572 **Wagner GP, Kin K, Lynch VJ**. 2012. Measurement of mRNA abundance using RNA-seq data:
573 RPKM measure is inconsistent among samples. Theory in Biosciences **131**, 281–285.

574 **Weber M, Trampczynska A, Clemens S**. 2006. Comparative transcriptome analysis of toxic
575 metal responses in Arabidopsis thaliana and the Cd2+-hypertolerant facultative metallophyte
576 Arabidopsis halleri. Plant, Cell and Environment **29**, 950–963.

577 **Wickett NJ, Mirarab S, Nguyen N, *et al.*** 2014. Phylotranscriptomic analysis of the origin and
578 early diversification of land plants. Proceedings of the National Academy of Sciences **111**,
579 E4859–E4868.

580 **Yang Y, Moore MJ, Brockington SF, *et al.*** 2015. Dissecting Molecular Evolution in the Highly
581 Diverse Plant Clade Caryophyllales Using Transcriptome Sequencing. Molecular Biology and
582 Evolution **32**, 2001–2014.

583

584

585 **Table 1.** Description of RNA-Seq samples

Species	Families	Location	NiH	Short name	Bioprojects	SRA samples
Senecio coronatus Agnes mine	Asteraceae	South Africa	Yes	ScorA	PRJNA312157	SRX1901460 SRX1901471
Senecio coronatus Kaapsehoop	Asteraceae	South Africa	Yes	ScorC	PRJNA312157	SRX1901463 SRX1901464 SRX1901465
Senecio coronatus Galaxy mine	Asteraceae	South Africa	No	ScorB	PRJNA312157	SRX1901479 SRX1901480 SRX1901481
Senecio coronatus Pullen Farm	Asteraceae	South Africa	No	ScorD	PRJNA312157	SRX1901469 SRX1901470 SRX1901472
Noccaea caerulescens Firmiensis	Brassicaceae	France	Yes	Ncfi	PRJNA474900	SRX4174673 SRX4174674
Noccaea montana	Brassicaceae	France	No	Nmon	PRJNA474900	SRX4174677 SRX4174678
Microthlaspi perfoliatum	Brassicaceae	France	No	Mper	PRJNA657163	SRX8947159 SRX8947160 SRX8947161
Geissois pruinosa	Cunoniaceae	New Caledonia	Yes	Gpru	PRJNA476928	SRX4261243 SRX4261244 SRX4261245
Geissois racemosa	Cunoniaceae	New Caledonia	No	Grac	PRJNA476928	SRX4261246 SRX4261247 SRX4261248
Phyllanthus luciliae	Phyllanthaceae	New Caledonia	Yes	Phlu	PRJNA645979	SRX8723728 SRX8723729 SRX8723730
Phyllanthus conjugatus	Phyllanthaceae	New Caledonia	No	Phco	PRJNA645979	SRX8723731 SRX8723732 SRX8723733
Psychotria grandis	Rubiaceae	Cuba	Yes	Pgra	PRJNA476927	SRX4261234 SRX4261235
Psychotria costivenia	Rubiaceae	Cuba	Yes	Pcos	PRJNA476927	SRX4261236 SRX4261237
Psychotria revoluta	Rubiaceae	Cuba	No	Prev	PRJNA476927	SRX4261238 SRX4261239
Psychotria gabriellae	Rubiaceae	New Caledonia	Yes	Pgab	PRJNA476924	SRX4261225 SRX4261226 SRX4261227
Psychotria semperflorens	Rubiaceae	New Caledonia	No	Psem	PRJNA476924	SRX4261228 SRX4261229 SRX4261230
Homalium kanaliense	Salicaceae	New Caledonia	Yes	Hkan	PRJNA476925	SRX4261218 SRX4261219 SRX4261220
Homalium betulifolium	Salicaceae	New Caledonia	No	Hbet	PRJNA476925	SRX4261221 SRX4261222 SRX4261223

586

Table 2. *De novo* assemblies of contigs and assignment to OGs

Species	Number of contigs	Contig median length	ORF median length	Number of classified contigs (%)
Scor	46726	475	330	27423 (59)
Ncfi	41843	620	390	23608 (56)
Nmon	64367*	436	345	44460 (69)
Mper	144815	420	312	89067 (62)
Gpru	41188*	486	288	20998 (51)
Grac	37663*	649	366	20410 (54)
Phlu	42028	482	303	22136 (53)
Phco	45524	437	288	22414 (49)
Pgra	31362*	728	456	19506 (62)
Pcos	37309*	542	321	19158 (51)
Prev	45787*	534	285	20254 (44)
Pgab	46860*	558	273	20493 (44)
Psem	41854*	600	285	19379 (46)
Hkan	42325*	541	300	22500 (53)
Hbet	40774*	537	306	21890 (54)

*contigs with low expression (TPM<1) were filtered out.

Table 3. List of differentially expressed OG

OG id	log2FC	p-value	OG predicted function* (EC number)	Best Hit*
6278	3.33	9.3e-20	Cationic amino acid transporter 4, vacuolar	CAAT4_ARATH
6266	3.01	2.2e-08	Protein Detoxification 27	DTX27_ARATH
4147	2.79	6.6e-18	Histidinol dehydrogenase, chloroplastic (1.1.1.23)	HISX_BRAOC
411	2.17	6.0e-06	Protein Aluminum Sensitive 3	ALS3_ARATH
12179	2.09	1.3e-04	Purple acid phosphatase 3 (3.1.3.2)	PPA3_ARATH
1871	2.00	9.7e-08	-	-
1781	1.97	1.5e-14	Ferredoxin–NADP reductase, chloroplastic (1.18.1.2)	FENR2_PEA
8462	1.92	1.7e-06	Ferric reduction oxidase 2 (1.16.1.7)	FRO2_ARATH
1441	1.80	8.8e-14	Vegetative incompatibility protein HET-E-1 (2.7.11.1)	HETE1_PODAS
7137	1.79	1.2e-04	Solute carrier family 40 member 2	S40A2_ARATH
10886	1.77	9.5e-07	-	-
11472	1.74	1.4e-04	Phenolic glucoside malonyltransferase 2 (2.3.1.-)	PMAT2_ARATH
2692	1.65	4.8e-10	-	-
1778	1.60	1.3e-05	Probable glutamate carboxypeptidase 2 (3.4.17.21)	GCP2_ARATH
3868	1.54	9.6e-08	Probable folate-biopterin transporter 7	FBT7_ARATH
3738	1.51	4.3e-05	Shikimate O-hydroxycinnamoyltransferase (2.3.1.133)	HST_TOBAC
559	1.51	1.5e-05	Polynucleotide 5'-hydroxyl-kinase NOL9	NOL9_ARATH
11324	-1.63	3.7e-05	Protein Defective in Meristem Silencing 3	DMS3_ARATH
4494	-1.76	6.2e-05	DNA ligase 4 (6.5.1.1)	DNLI4_ARATH
10076	-1.76	2.6e-04	Calcium-transporting ATPase 12, plasma membrane-type (3.6.3.8)	ACA12_ARATH
3681	-1.77	1.4e-08	Dicer-like protein 4 (3.1.26.3)	DCL4_ARATH
2967	-1.93	3.4e-08	Uncharacterized protein HI_0077	Y077_HAEIN
693	-1.94	3.2e-10	-	-
8515	-1.96	2.8e-04	Putative axial regulator YABBY 2	YAB2_ARATH
3961	-2.07	2.4e-06	Probable GTP-binding protein OBGM, mitochondrial	OBGM_ARATH
3821	-2.09	1.8e-11	Probable protein phosphatase 2C 11 (3.1.3.16)	P2C11_ARATH
1916	-2.12	7.2e-05	Linoleate 13S-lipoxygenase 2-1, chloroplastic (1.13.11.12)	LOX2_ARATH
8093	-2.33	3.4e-05	Glycolipid transfer protein 2	GLTP2_ARATH
11720	-2.56	3.0e-05	Beta-fructofuranosidase, insoluble isoenzyme 1	INV1_DAUCA
5933	-3.17	4.1e-08	ACT domain-containing protein ACR4	ACR4_ARATH
13722	-3.19	1.1e-09	Rust resistance kinase Lr10 (2.7.11.1)	LRK10_WHEAT

590 *From UniProtKB/Swiss-Prot database

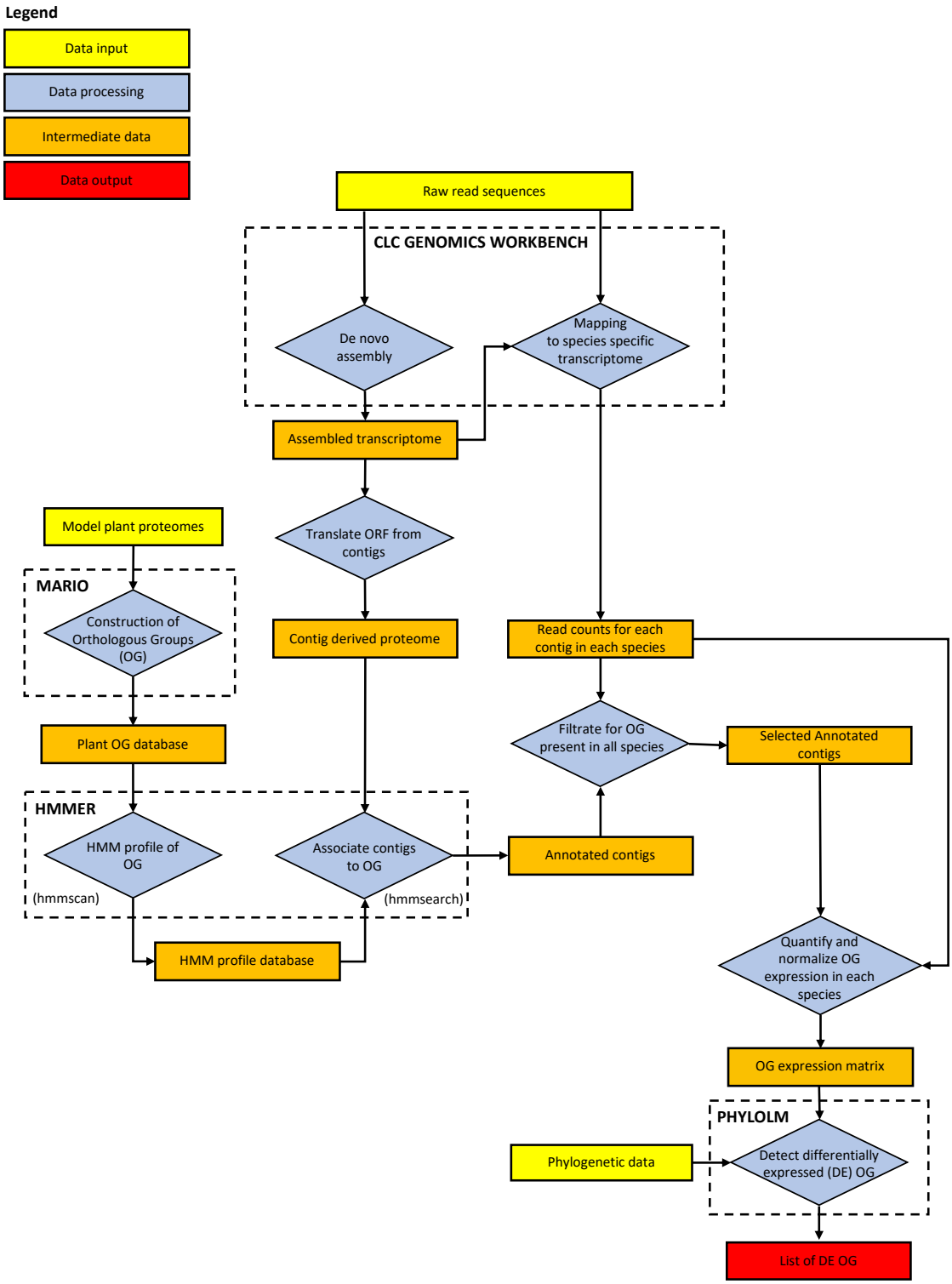

Figure 1: Workflow used to compare the expression of Orthologous groups in distant species (**a**) RNA-Seq datasets from different species were used to assemble transcriptomes *de novo* and to quantify gene (contig) expression in these transcriptomes using QIAGEN CLC Genomics Workbench. (**b**) Proteomes from sequenced plants species was used to generate a plant Orthologous group (OG) seed database using MARIO. (**c**) HMM profiles from the OG seed were produced by HMMER and used to annotate contigs from *de novo* transcriptomes. **(d)** The PHYLOLM package was used to detect differentially expressed OG using the normalized OG expression matrix.

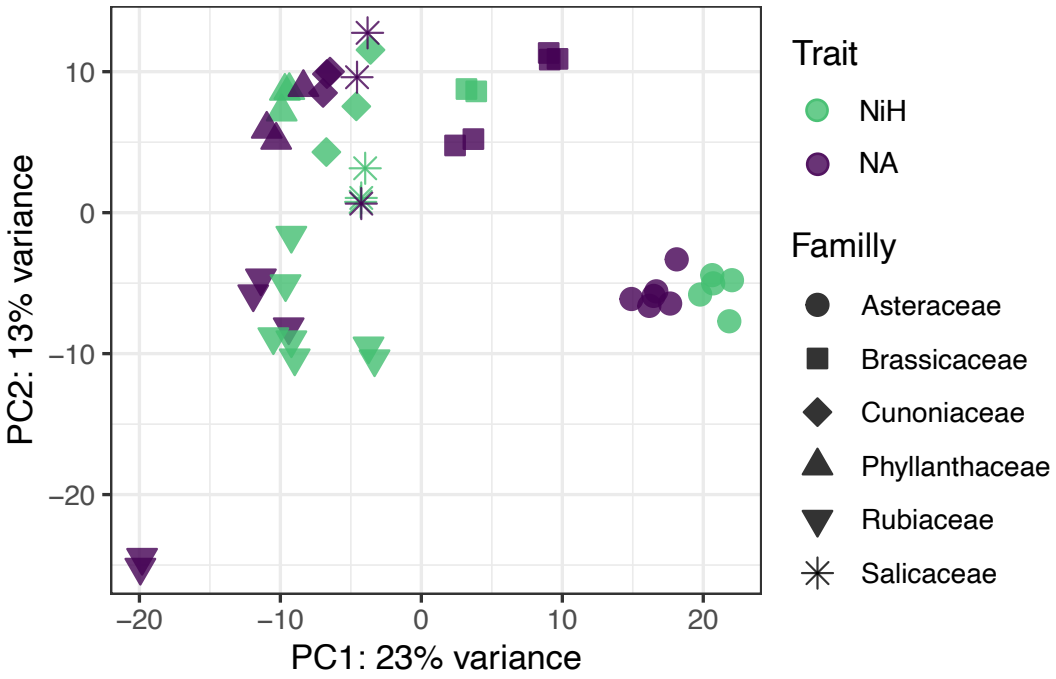

Figure 2. Principal Component Analysis performed on the OG expression table after normalization and log transformation. The shapes of the symbols correspond to the family of the species and the color to the nickel hyperaccumulation (NiH, green) or non-accumulating (NA, purple) trait.

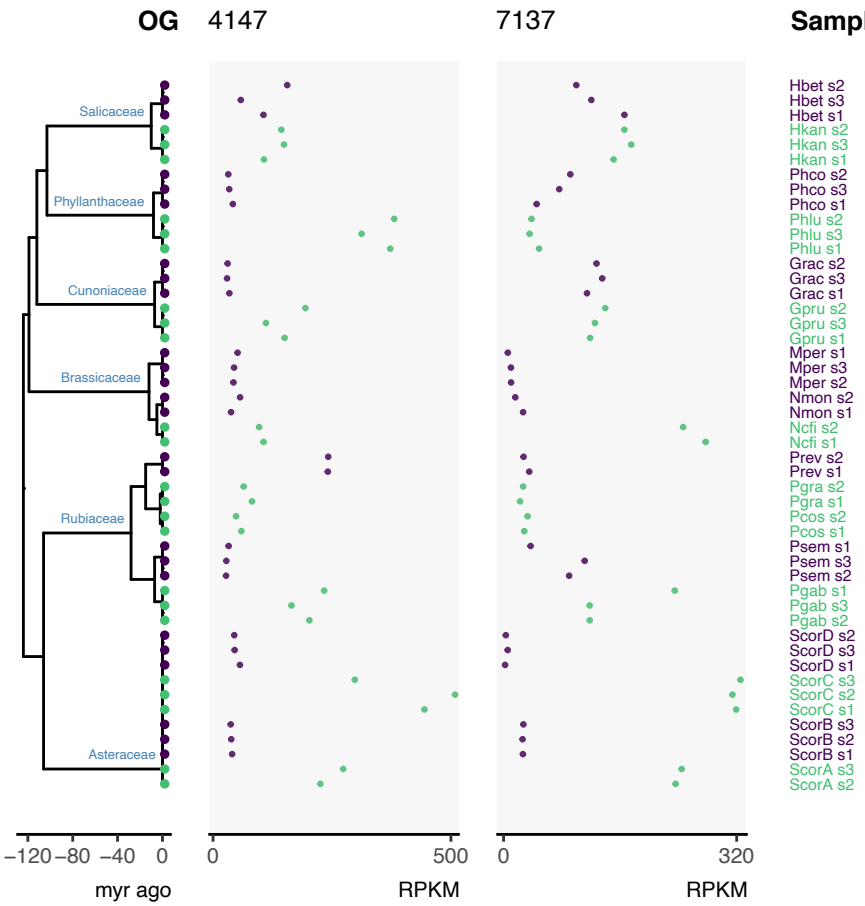

Figure 3. Representation of the expression of OG 4147 and OG 7137 along a plant phylogenetic tree. The expression of OG is expressed as Reads Per Kilobase of transcript, per Million mapped reads (RPKM). The phylogenetic tree is scaled in million years (myr), the six families considered are labeled on the tree. The name of the samples is given on the right. The color of the symbols correspond to NiH (green) or NA (purple) species or populations.