Network of orthologs and evolutionary context search unveil the landscape of gene transposition-duplication within the Brassicaceae family.

Dong-Ha Oh* and Maheshi Dassanayake*

Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803

Correspondence: Dong-Ha Oh (ohdongha@gmail.com), Maheshi Dassanayake (maheshid@lsu.edu)

Abstract

Gene duplication, deletion, and transposition erode co-linearity between genomes over time, providing a substantial source of genomic structural variations. We developed CLfinder-OrthNet pipeline to systematically trace gene duplication and transposition events by building networks based on co-linearity among orthologs (OrthNets) from multiple closely related genomes. Using this pipeline, we explored evolutionary events that result in erosion of co-linearity among six genomes in Brassicaceae, including the model plant *Arabidopsis thaliana* and two extremophyte wild-relatives of Brassica crops adapted to highly saline habitats. Out of 17,432 OrthNets derived from the six Brassicaceae genomes, 7,034 consisted of six co-linear orthologs, one from each genome, while the remaining included various degrees and combinations of gene duplication, deletion, and transposition. We focused on transposition-duplication (*tr-d*) events that lead to variations in both gene copy number and co-linearity. OrthNets identified a landscape of *tr-d* events including those with signatures of selective retention and putative sub-functionalization in all six genomes. These included lineage-specific *tr-d* of genes that may be critical for the local adaptation of extremophytes, such as orthologs of *SALT TOLERANCE 32* and *ZINC TRANSPORTER 3*. Among *tr-d* events uniquely shared by the two extremophytes, those resulted in duplication of putatively functional full-length ORFs were mostly derived from their common ancestor rather than from parallel independent events. Our approach enabled systematic detection and representation of a complex mosaic of evolutionary events among multiple genomes, as well as identification of candidate genes that may have played roles in the adaptation uniquely found in specific lineages.

Funding source

This work was supported by National Science Foundation (MCB 1616827) and the Next Generation BioGreen21 Program (PJ011379) of the Rural Development Administration, Republic of Korea.

Glossary

Best-hit : the gene locus in the target genome that is most likely orthologous to a given gene in the query genome. This can be the gene showing the highest sequence similarity score among all gene models in the target genome, for a query gene. CLfinder can also choose a best-hit based on co-linearity among multiple best-hit candidates showing sequence similarity.

Brassicaceae genomes: the six genomes used in this study are *Arabidopsis lyrata* (**Aly**), *A. thaliana* (**Ath**), *Capsella rubella* (**Cru**), *Eutrema salsugineum* (**Esa**), *Sisymbrium irio* (**Sir**), and *Schrenkiella parvula* (**Spa**), in the Brassicaceae (crucifer) family.

Co-linear (*cl*): the gene orders show a similarity that appears only explainable by a presence of a common ancestor, between genomes being compared.

CL copy in a *tr-d* event: among paralogs involved in a transposition-duplication (*tr-d*) event, the one showing co-linearity to the highest number of orthologs from other genomes being compared. This copy most likely represent the donor locus in the *tr-d* event. Multiple CL copies may exist in a *tr-d* event, if the donor locus underwent a tandem duplication.

Evolutionary context: patterns of evolutionary events either unique to a certain lineage or shared by multiple lineages. In this study, we aim to develop a method to search all orthologs from multiple closely-related genomes that shared a given evolutionary context.

Gene-only (go) duplication: a gene-only duplication includes sequence similarity only in the gene region but not in the adjacent intergenic regions, between the genomic loci being compared. A duplication that encompasses both gene and intergenic regions is indicated as a **"gene + intergenic (g+i)"** duplication.

Homologous Genome Segment (HGS): a stretch of genomic regions where sequence similarity can be detected between multiple paralogous or orthologous loci.

Lineage-specific (LS): a gene or evolutionary event found uniquely in a lineage. When used in the CLfinder module, lineage-specific (*ls*) indicates no best-hit found for the query gene among all genes in the target genome.

OrthNet: a network of orthologs, with gene loci as nodes, connected to their best-hits in other genomes with edges. Edge property includes co-linearity information between the two nodes being connected.

Independent parallel (Ind-parallel or Ind-par): a *tr-d* or other evolutionary event that occurred independently in multiple genomes in parallel. In an independent parallel *tr-d* event shared by multiple genomes, the Tr copies from different genomes do not show any co-linearity among each other.

Tr copies co-linear (Tr-*cl*): a *tr-d* event shared by multiple genomes, where Tr copies from different genomes show co-linearity among them. Such an event most likely originated from a common ancestor of the genomes sharing the event.

Tandem duplication (*td*): a gene duplication occurred locally, often resulting in a tandem array of duplicated genes.

Transposition (*tr*): an event where a gene moved to a different, unrelated locus in the genome, losing any trace of common ancestry except for the sequence similarity within the gene.

Transposition-duplication (*tr-d*): an event where a gene (donor) duplicates and gets inserted to a different, unrelated locus (acceptor) in the genome. A transposition event is most likely a result of transposition-duplication followed by the deletion at the donor locus.

Tr copy in a *tr-d* event: among paralogs involved in a transposition-duplication (*tr-d*) event, the one that is not a CL copy is a transposed (Tr) copy. Representing the acceptor locus of a *tr-d* event, it appears as transposed to orthologs from other genomes being compared. One CL copy can be associated with multiple Tr copies. A Tr copy may show co-linearity to Tr copies in other genomes (see "Tr-*cl*") and be subjected to tandem duplication.

Introduction

Co-linearity among closely related genomes erodes over time due to the accumulation of mutations including gene duplication, deletion, and transposition [1]. Such erosion is generally more rapid among land plants than in animals [2,3]. In addition to whole genome duplication (WGD) followed by fractionation found in most plant genomes during land plant evolution [4,5], smaller-scale lineage-specific gene duplications occur locally or between seemingly unrelated loci, often in different chromosomes, leading to tandem or transposition-duplications [6] in a continuum of structurally diverse genome space. Transposition-duplication followed by the deletion of the original donor copy results in transposition or "movement" of the gene locus, which is observed extensively in plant genomes not only for transposable elements (TEs) but also for protein-coding genes [2,7]. Gene duplication and transposition events provide a source of variation in copy numbers of genes and transcripts [8–10], as well as localization [11] and functions [12,13] of encoded proteins among orthologous genes.

Genomic structural variation that may contribute to the overall fitness of the organism can be naturally selected and fixed in the population. Although less frequently observed than single nucleotide polymorphsims and small indels, gene level duplications and transpositions are more significant in their contribution to overall fitness and speciation [14–19]. A large number of studies have reported

96 examples of gene level duplications and transpositions as key underlying sources for adaptations to
97 specific environments or speciation (reviewed in [20]) [21–26]. Therefore, as we gain access to more
98 recently sequenced genomes, it will be imperative to characterize gene structural organization unique
99 and shared in related lineages before attributing their relevance to adaptations specific to a subset of
100 lineages.

101 While the advent of next-generation and long read sequencing provides increasing numbers of
102 *de novo* assembled genomes [27–30], it also calls for novel methods for systematic comparative analysis
103 between multiple closely related genomes. There are two types of well-established computational tools
104 to identify and visualize orthologous groups among related genomes in comparative genomics. One set
105 of tools uses a gene or a gene family as the query to retrieve its orthologs in other genomes and
106 visualize synteny and other evolutionary events involving them, such as gene duplication and
107 transposition [31–35]. The other type of tools identify gene blocks or large genomic regions co-linear
108 among a pair or multiple genomes [36–38]. However, there is almost a void for methods that can
109 retrieve all orthologs within multiple genomes that have likely undergone the same evolutionary events
110 unique in a given lineage or shared by multiple lineages. Such methods will be especially useful to
111 explore signatures and mechanisms of shared, convergent, or parallel adaptive evolution. For example,
112 orthologs sharing the same structural variation independently in distantly related lineages that are
113 similar in a set of phenotypic adaptations may constitute mechanisms of convergent evolution. To
114 address this gap in computational tools, we introduce the CLfinder-OrthNet pipeline, which identifies co-
115 linearity (CL) in the arrangement of orthologous loci among multiple genomes and builds networks of
116 orthologs (OrthNets) to encode and visualize all evolutionary events, such as gene duplication, deletion,
117 and transposition, in each orthologs group. This pipeline essentially enables detection of all orthologs
118 that share a given evolutionary context, i.e. a pattern of evolutionary events uniquely found in a lineage
119 or multiple lineages, in a set of genomes.

120 In this study We applied the CLfinder-OrthNet pipeline to Brassicaceae genomes that share the
121 same whole genome duplication history [39]. Brassicaceae (crucifers) is one of the most highly
122 represented families for published genomes among land plants [28,39]. This includes the model species
123 *Arabidopsis thaliana* [40], extremophytes such as *Schrenkiella parvula* [41] and *Eutrema salsugineum*
124 [42,43], as well as a large number of food and biofuel crop species [44–50]. *S. parvula* and *E.*
125 *salsugineum*, the two most salt-tolerant crucifer species so far tested [51], are biogeographicaly
126 seperated and naturally adapted to multi-ion salts strsses in soils near a hypersaline lake in central
127 Anatolia [9] and combined salt and freezing stresses in salt pans of high latitude regions in the northern

128 hemisphere [52,53], respectively. These two extremophytes represent a framework for comparative
129 analysis to study mechanisms of plant adaptations to various environmental challenges [54,55],
130 facilitated by the presence of genome-wide synteny conserved among the Brassicaceae species [56]
131 including the model plant *A. thaliana.* Hence, a systematic identification of evolutionary events
132 modifying co-linearity and copy number of genes uniquely in these extremophytes, and not in their
133 stress-sensitive relatives, can provide insight to candidate genes and mechanisms that allow plant
134 adaptations to environmental challenges [54,55].
135 The CLfinder-OrthNet pipeline detects multiple forms of gene-level structural variation,
136 including tandem duplications, deletions, transpositions, and also combinations of them. For the
137 present application in this study, we focused on the relatively under-studied transposition-duplication
138 (*tr-d*) events [57] within a selected set of six Brassicaceae genomes. The pipeline enabled systematic
139 identification of *tr-d* events unique to a genome or shared by any possible combination of genomes, as
140 well as the original donor and duplicate loci in each *tr-d* event including loci with truncated coding
141 regions. Using this pipeline, we aim to identify the landscape of lineage-specific and shared *tr-d* events
142 among the target genomes; test whether there is a signature of selective retention among lineage-
143 specific *tr-d* events; and characterize *tr-d* events uniquely found in either or both extremophyte
144 genomes, which may have contributed to their adaptive evolution.
145
146 **Results**
147 **Patterns of co-linearity erosion within Brassicaceae genomes.**
148 We selected a set of six genomes with the same whole genome duplication history sampled
149 from the Brassicaceae Lineages I and II for the current study (Figure 1). This set includes the model plant
150 *A. thaliana* (Ath) and its relatives in Lineage I, *A. lyrata* (Aly) and *Capsella rubella* (Cru), as well as
151 *Sisymbrium irio* (Sir) and the two extremophytes, *E. salsugineum* (Esa) and *S. parvula* (Spa) in Lineage II.
152 Fig. 1 shows the phylogenetic relationships of the target species with other published genomes in
153 Brassicaceae based on amino acid sequence alignments of 14,614 homolog clusters (See Methods). In
154 the tree based on maximum likelihood, *A. lyrata* and *A. thaliana* were grouped together and sister to *C.*
155 *rubella* within the Lineage I. Similarly, *S. iro* and *S. parvula*, as well as various Brassica crop species were
156 placed in the same clade, while *E. salsugineum* was placed in a sister clade within the Lineage II.
157 We compared the gene orders in each of the Brassicaceae genomes except for *A. thaliana*
158 (query) with that of the *A. thaliana* genome (target) to analyze the degree of co-linearity erosion (Figure
159 2). We chose *A. thaliana* as the reference target because it has the best assembly among all plant

160 genomes. First, we assigned a numerical "Locus ID" to each non-transposable element (non-TE) protein-
161 coding gene locus according to its position in the chromosome or genome scaffold. Then, using the
162 Basic Local Alignment Search Tool for Nucleotides (BLASTN) [58], we aligned all loci in the query genome
163 with their best BLASTN ($e<10^{-5}$) hits ("best-hit") among *A. thaliana* protein-coding genes. We designated
164 the difference between Locus IDs of best-hits for query locus at position *n* and *n+1* as $d_{n,n+1}$ (Fig. 2A). The
165 $d_{n,n+1}$ for the query locus *n* is expected to be one if the gene order is perfectly co-linear, zero if the query
166 loci *n* and *n+1* had undergone tandem duplication (Fig. 2A, panel (1)), and larger than one in case of a
167 local insertion/deletion of gene loci or an inversion of genomic segments (Fig. 2A, panel (2)). A gene
168 transposition (*tr*) or transposition-duplication (*tr-d*) will result in the best-hit to move to a position
169 separated by X number of loci, which may be on a different chromosome (Fig. 2A, panel (3)). The
170 acceptor locus of a transposition or *tr-d* event (Fig. 2A, panel (3); Locus ID 3+x) may serve as the source
171 of insertion as in Fig. 2A, panel (2). We excluded query loci without a BLASTN hit in the *A. thaliana*
172 genome from the current analysis.

173 If the co-linearity eroded over time, we would expect that the proportion of query loci with $d_{n,n+1}$
174 $\neq 1$ will be larger in species diverged earlier from the target species. Indeed, $d_{n,n+1}$ distribution from the
175 *A. lyrata*-vs-*A. thaliana* comparison (Fig. 2B, "Aly vs Ath") showed a greater percentage of loci with $d_{n,n+1}$
176 = 1 than the *E. salsugineum*-vs-*A. thaliana* comparison (Fig. 2B, "Esa vs Ath"), reflecting the more recent
177 divergence of *A. thaliana* from *A. lyrata* than from *E. salsugineum*. However, as shown in column "Aly-
178 Ath" and "Esa-Ath" in Fig. 2C, while the proportion of query loci with lower $d_{n,n+1}$ (Fig. 2C, $d_{n,n+1}$ = 2~20)
179 was smaller in *A. lyrata* (10.58%) than in *E. salsugineum* (15.24%), those with higher $d_{n,n+1}$ (Fig. 2C, $d_{n,n+1}$ >
180 20 and "Diff Chr") were either comparable or more frequent in *A. lyrata* (7.98% + 19.34%) than *E.*
181 *salsugineum* (6.91% + 19.02%).

182 When we consider all five query species (Fig. 2D), the proportion of query loci with $d_{n,n+1}$ = 2~20
183 showed a positive correlation (R^2 = 0.85; Fig. 2D, upper left panel) with the median four degenerate site
184 (4d) substitution rate between all reciprocal best-hit pairs of the query and the target genomes. The
185 median 4d substituation rate reflects the divergence time between the two genomes. In contrast, the
186 proportion of query loci with $d_{n,n+1}$ > 20 or whose best-hits found in a target chromosome different from
187 their neighbors (Fig. 2C, $d_{n,n+1}$ > 20 and "Diff Chr") did not show any positive correlation with the median
188 4d substitution rate (see the negative linear fit slope in Fig. 2D, lower left panel). We found a moderate
189 positive correlation (R^2 = 0.59) between the proportion of query loci with $d_{n,n+1}$ > 20 and the overall
190 content of TE sequences (Fig. 2D, lower right panel), while the proportion of loci with $d_{n,n+1}$ = 2~20 was
191 not correlated with the TE content (R^2 = 0.028; Fig. 2D, upper right panel) in the respective genomes.

192 This result suggests that the gene order rearrangements have resulted from two different mechanisms:
193 proximal gene order rearrangements depend on mutations accumulated over time after the divergence
194 of species, while distal gene order rearrangements may have been initiated by the presence of repetitive
195 sequences and TE activities.
196
197 **Development of the CLfinder-OrthNet pipeline and application on crucifer genomes**
198 We aimed to identify co-linearity, and the erosion of it, among gene orders of multiple closely
199 related genomes systematically. We also wanted to retrieve groups of orthologs showing a specific
200 pattern of gene organization and evolutionary context, for example: all orthologous loci with a signature
201 of gene duplication and transposition found uniquely in a genome, within a lineage, or between groups
202 of genomes and lineages, in a high-throughput manner. To achieve these aims, we developed the
203 CLfinder-OrthNet pipeline (Figures 3, S1, and S2).
204 CLfinder compares all possible pairs of query and target genomes and identifies homologous
205 gene loci pairs either co-linear (*cl*), lineage-specific (*ls*), transposed (*tr*), or representing an end of a
206 segmental rearrangement of gene orders such as insertion, deletion, or inversion (*cl_end*) (Figures S1
207 and S2 and Supplementary Dataset S1). CLfinder accepts three inputs: representative gene models for
208 all loci in each genome, clusters of paralogs within each genome, and lists of best-hits between all
209 possible query-target genome pairs (Fig. 3). Users can select the methods and criteria for defining
210 paralog clusters and the list of best-hit pairs, as well as the sensitivity and stringency for the co-linearity
211 detection by controlling three parameters: window_size (*W*), num_CL_trshld (*N*), and gap_CL_trshld (*G*)
212 (Fig. S1). Based on these parameters, the CLfinder module searches both up- and down-stream of each
213 locus in the query genome for "loci-in-chain" in the target genome, to determine whether a query-
214 target best-hit pair is either co-linear (*cl*), transposed (*tr*), or not able to determine (*nd*) due to the query
215 genome assembly scaffold too fragmented (See Methods for detail). A query locus without a best-hit in
216 the target genome is marked as lineage-specific (*ls*). Figure S2 shows an example representing the
217 CLfinder process. The CLfinder program generates a summary of CLfinder analyses on all possible
218 reciprocal query-target pairs. As an example, we present the result for the CLfinder analyses on the six
219 Brassicaceae genomes in Table 1. CLfinder is the first module executed in the pipeline.
220 The second module of the pipeline, OrthNet, combines all pairwise comparisons generated by
221 CLfinder and encodes relationships among orthologs into a network with gene loci as nodes connected
222 by an edge to their best-hits in other genomes (Figures 3 and 4). Each edge has a property of either co-
223 linear (*cl*), transposed (*tr*), or tandem duplicated (*td*), or not determined (*nd*) if the query node is in a

224 genome scaffold containing less than *W* loci. The *cl* and *tr* edges can be either reciprocal or
225 unidirectional (Fig. 4A, "rc" and "uni", respectively). The OrthNet program also compare the ORF length
226 of a protein-coding node with the median length of ORFs for all other neighboring nodes and identify
227 nodes with truncated ORFs (Fig. 4). The module uses Markov clustering (MCL), based on edge weights
228 assigned according to edge properties (see Methods), aiming to divide large networks that are often a
229 result of expanded gene families with a large number of paralogs into smaller clusters derived from the
230 smallest distinguishable sets of co-linear loci. Each cluster of orthologs, separated by MCL, is given an
231 OrthNet ID and represented as an ortholog network or an OrthNet. Finally, the module can search with
232 user-defined patterns of ortholog copy numbers and co-linearity characteristics, to retrieve all OrthNets
233 sharing a given evolutionary context (see Methods, Fig. S1, and Table S1). A few selected examples of
234 OrthNets representing different evolutionary contexts are shown in Figure 4A.
235 We tested the CLfinder-OrthNet pipeline on the six Brassicaceae genomes using parameters and
236 input files as described in Methods and summarized the result in Table 1. For simplicity, we considered
237 *cl_end* loci pairs that could be a result of inversions, indels, and segmental duplications involving
238 multiple gene loci, as *cl* in this summary. All query-target genome pairs showed a comparable number
239 of *cl* loci pairs, ranging from 19,015 (Table 1, Sir as the query and Aly as the target) to 24,296 (Table 1,
240 Aly as the query and Ath as the target). The number of *cl* pairs follows the division of the Lineage I
241 (Table 1, Aly, Ath, and Cru) and II (Table 1, Esa, Sir, and Spa), with higher numbers observed between
242 query-target pairs within each Lineage. The number of *tr* pairs was proportional to the repeat contents
243 of the query genomes. For example, *A. lyrata* and *E. salsugineum* are the query genomes with the
244 highest content of *tr* pairs (Table 1, Aly and Esa), which correlated with the higher content of repeats in
245 these two genomes than in *A. thaliana*, *C. rubella*, or *S. parvula* genomes (Fig. 2D and Table S2, "TE(%)").
246 When *S. irio* was the query, the proportion of *nd* pairs was higher than all other genomes (Table 1, Sir),
247 because it had the most fragmented genome assembly among the six genomes. The entire output of
248 the CLfinder run for each query genome compared to the five other target genomes is in Supplementary
249 Dataset S1.
250 Combining all pairwise comparisons generated by the CLfinder, the OrthNet module developed
251 networks of orthologs we refer to as OrthNets, of various topologies, representing the evolutionary
252 history and context of each set of orthologous loci (Fig. 4). For an analysis of N genomes, a perfect
253 polygon (e.g. hexagon in the current study) with N nodes and N-1 bidirectional solid gray edges
254 represents a single-copy co-linear orthologous gene in all genomes. We identified a total of 7,034
255 OrthNets consisted of single-copy loci co-linear to each other in all genomes (Fig. 4A, panel (1)). Panel

256 (2) of Fig. 4A is an example from 50 OrthNets with co-linearity found within each of the Lineage I and II
257 while loci between the two Lineages, representing a transposition event right after the lineage
258 divergence. Panel (3) shows one of the nine OrthNets with only the locus in *S. parvula* transposed
259 compared to all other species. We found 44 OrthNets with the same evolutionary contexts depicted in
260 panel (4), i.e. *E. salsugineum*-specific tandem duplication, and 86 OrthNets for *S. parvula*-specific
261 transposition-duplication (*tr-d*) events shown in panel (5) of Fig. 4A. The OrthNet module also compares
262 the ORF size of a node with the median ORF size of all other orthologous nodes to which the node is
263 connected, to identify truncated ORFs. Examples of nodes with truncated ORFs are included in Fig. 4A,
264 panels (3), (4), and (5). We list all OrthNets identified by the OrthNet module as Supplementary Dataset
265 S1, together with the CLfinder results.
266 An OrthNet may include a disproportionately large number of duplicated gene loci in specific
267 genomes. For example, an OrthNet showing Aly-specific *tr-d* events included 82 nodes representing
268 additional transposed-duplicated copies only in Aly, while additionally preserving a complete hexagon of
269 single-copy loci that were reciprocally co-linear among all six genomes (Figure S3). Such duplication
270 events may result in an OrthNet with a large number of nodes. However, the majority of OrthNets
271 contain the same or less than 18 nodes per OrthNet (16,235 out of total 17,432 OrthNets), with a peak
272 at six nodes (8,119 OrthNets) and 7~12 nodes (5,641 OrthNets) (Fig. 4B). Larger OrthNets may represent
273 lineage-specific expansions of a certain gene as exemplified by Fig. S3 or a large gene family where exact
274 reciprocal orthologous loci pairs were hard to identify based on best-hit pairs (see Discussion).
275
276 **Characterization of lineage-specific (LS) transposition-duplication (*tr-d*) events**
277 While OrthNet and CLfinder can operate independently, when used in a pipeline, the OrthNet
278 module adds the results of its analysis to the output of the CLfinder module (Supplementary Dataset S1).
279 This includes a unique OrthNet ID, the number of nodes derived from each genome (i.e. gene copy
280 numbers), and the edge types connecting the locus to its best-hit nodes in other genomes. Users can
281 use a python script that searches a regular expression (Fig. S2, "search_OrthNet.py") to identify
282 OrthNets specific to putative lineage-specific or shared gene transposition and duplication events. We
283 used regular expressions in Table S1 to identify *tr* and *tr-d* events specific to each genome, shared by
284 any two pairs of genomes, and between Brassicaceae Lineage 1 and 2. In all genomes, the number of
285 OrthNets that showed lineage-specific *tr* events was smaller than those with lineage-specific *tr-d* events
286 (Table S2), in alignment with the model where a *tr* event is a result of a deletion of the original/donor
287 copy after a *tr-d* event.

288 For lineage-specific *tr-d* events, the OrthNet module enables effective identification of the
289 original donor or co-linear (CL) copy, or copies if the donor site included tandem duplications, and the
290 acceptor or transposed (Tr) copies (Figure 5A, "CL copy" and "Tr copies", respectively). For example, Fig.
291 5A represent OrthNets with *S. parvula* and *C. rubella* lineage-specific *tr-d* events for orthologs of the
292 *WRKY72* and *AGL87*, respectively. In each of these OrthNets, we could detect the CL copy (Fig. 5A, "CL
293 copy") as a part of the hexagon that were mostly reciprocally co-linear to the its ortholog nodes (Fig. 5A,
294 "Orthologs") from other genomes. In the same OrthNet, a Tr copy was connected to ortholog nodes
295 from other genomes in the hexagon through uni-directional *tr* edges (Fig. 5A, "Tr copies"). An OrthNet
296 may contain a single lineage-specific *tr-d* event as in the OrthNet for *WRKY72* (Fig. 5A, left) or multiple
297 lineage-specific *tr-d* events featuring one CL copy associated with multiple Tr copies. Also, Tr copies may
298 further undergo tandem duplications as shown in the OrthNet for *AGL87* (Fig. 5A, right). We compared
299 the ORF sizes between CL and Tr copies with the median ORF size of the orthologs from other genomes
300 in the hexagon for all OrthNets including lineage-specific tr-d events. In all genomes, we observed a
301 strong conservation of ORF sizes between most CL copies and their co-linear orthologs (Fig. 5B, blue
302 dots), while the majority of Tr copies had ORF sizes smaller than the median ORF sizes of orthologs from
303 other genomes (Fig. 5B, gray dots). Interestingly, we also found subsets of Tr copies which had ORFs
304 that were of similar size to their respective CL copy in all genomes (Fig. 5C, pink rectangles). Distribution
305 of the ORF size ratio between Tr and the CL copy showed peaks at 80~120% (Fig. 5D, pink rectangles).
306 These Tr copies that showed conservation in maintaining the original size of the ORFs were more
307 abundant in *A. thaliana*, *C. rubella*, *E. salsugineum*, and *S. parvula*, comprising more than 24% of all Tr
308 copies found in LS *tr-d* events in these genomes (Fig. 5D). We summarized OrthNets including lineage-
309 specific *tr-d* events for each of the six genomes in Supplementary Dataset S2.
310
311 **Characterization of shared transposition-duplication (*tr-d*) events**
312 For *tr-d* events shared between any pair within the six genomes, the OrthNet module identified
313 two categories with different evolutionary contexts: (1) parallel *tr-d* events independently occuring in
314 two genomes (Figure 6, "Ind-parallel") and (2) *tr-d* events where Tr copies from two genomes showing
315 co-linearity between them (Fig. 6, "Tr-*cl*"). Fig. 6A depicts examples of OrthNets including *tr-d* events in
316 the two categories. We found a total of seven and six OrthNets with *tr-d* events in "Ind-parallel" and
317 "Tr-*cl*" categories, respectively, shared between *E. salsugineum* and *S. parvula*. Genomes with higher TE
318 and repetitive sequence contents, such as *A. lyrata*, *S. irio*, and, to a lesser extent, *E. salsugineum*,
319 included more "ind-parallel" *tr-d* events shared with other genomes (Fig. 6B, left panel). Among Tr

320 copies in "ind-parallel" *tr-d* events, the proportion of complete ORFs (i.e. ORF size within ±20% of the
321 ORF of the corresponding CL copy) were comparable to Tr copies in LS *tr-d* events (Fig. 6C, left panel and
322 Fig. 5D).

323 The "Tr-*cl*" type *tr-d* events were mostly found between pairs of more recently diverged
324 genomes, e.g. *A. lyrata-A. thaliana* and *S. irio-S. parvula*. Contrastingly, the "Tr-*cl*" *tr-d* events were
325 detected at a very low frequency between Lineage I and II genomes (Fig. 6B, right panel). This
326 observation was consistent with the notion that such a rare event must involve a *tr-d* event before the
327 divergence of Lineages followed by deletions in all species and retention of the duplicated and original
328 copies selectively in the two genomes compared. The proportion of Tr copies that retained complete
329 ORFs compared to the CL copy in "Tr-*cl*" type *tr-d* events was higher (≥50%) than that found for "ind-
330 parallel" type *tr-d* events (Fig. 6C). We present the entire list of *tr-d* OrthNets shared by all genome
331 pairs, including information on CL and Tr copies, in Supplementary Dataset S2.

332 We hypothesized that selection has favored conservation of beneficial Tr copies to preserve the
333 ORF in additional gene copies (Fig. 5C and D, pink rectangles; Fig 6C), while majority of the Tr copies
334 were either duplicated incompletely or have undergone a high level of mutations over time that had led
335 to truncated ORFs. An alternative hypothesis is that these Tr copies may have been easier to duplicate
336 in their complete form by random chance due to their smaller gene size. Indeed, the gene lengths were
337 shorter for CL copies associated with Tr copies that had conserved complete ORF sizes, than in Tr copies
338 that had truncated ORFs (Figure S4).

339 To test our hypotheses, we shuffled all duplicated genomic regions and duplicated genes. Then,
340 we compared the occurrences of *tr-d* events showing complete duplication to the frequencies observed
341 in actual *tr-d* events (Figure 7). First, to detect duplicated genomic regions in a *tr-d* event, we compared
342 adjacent genomic regions, i.e. 5Kb up- and downstream regions, including the gene, of the CL copy and
343 each of Tr copies and searched for Homologous Genome Segments (HGSs) between the CL and Tr copy
344 loci. As depicted in Fig. 7A, an incomplete *tr-d* event results in the HGS including only a part of the CL
345 copy gene (HGS $\not\supset$ CL copy), while in a complete *tr-d*, the HGS encompasses the entire CL copy gene
346 (HGS $\supseteq$ CL copy). Interestingly, we found a subset of complete *tr-d* events where the start and end
347 positions of the HGS were similar to the start and end of the CL copy gene (i.e. HGS ≈ CL copy). We
348 named this subset "gene-only" *tr-d* (Fig. 7A) since the sequence homology was not detectable in the
349 intergenic region further from the CL copy coding regions by more than 20% of the CL copy gene size.
350 The length distribution of HGSs and CL copy genes in Fig. 7B matched the expected size of the CL copy
351 and the corresponding type of the tr ORF type shown in Fig. S4 for all six genomes. A total of 224

352 complete *tr-d* events showed a shift towards retaining longer HGS lengths corresponding to a shorter CL
353 copy gene length, compared to the 1,166 incomplete *tr-d* events (Fig. 7B).
354 Following random shuffling of all HSGs and CL copy genes as described in Methods, we counted
355 the occurrences of incomplete, complete, and gene-only *tr-d* events for each iteration. Fig. 7C shows
356 the comparison between the observed and expected occurrences of complete and gene-only *tr-d*
357 events, where expected values are the mean values from 10,000 iterations. Assuming a normal
358 distribution for the expected values, we estimated the P-value for the observed numbers of complete
359 and gene-only LS *tr-d* events for each genome. Both complete and gene-only *tr-d* events were much
360 more frequent than random chance. The gene-only *tr-d* events had smaller p-values than complete *tr-d*
361 events in all categories tested except in *A. thaliana* lineage-specific *tr-d* events (Fig. 6C, table in the
362 lower panel). We observe a smaller number of lineage-specific *tr-d* events in *A. thaliana* than in any
363 target genome. This may be a result of *A. thaliana* and *A. lyrata* sharing the same genus. Hence, we
364 included *Arabidopsis* genus-specific *tr-d* events into consideration, which led to comparable numbers
365 and enrichment of complete and gene-only *tr-d* events to other genomes (Fig. 6C, "Aly-Ath"). Hence,
366 random occurrence of duplications cannot explain the observed proportion of complete *tr-d* events,
367 which in >90% of the case resulted in Tr copies with their ORF sizes similar to that of the CL copy (e.g.
368 genes represented in pink shaded areas in Figs. 5C and 5D). Most likely, the observed proportion of
369 complete and gene-only duplications, which was higher than expected due to random chance, was the
370 sequential result of random duplications, purifying selection on desirable coding sequences, and the
371 intergenic spaces drifted in the absence of lack of selection pressure, leading to the loss of any
372 detectable sequence similarity beyond coding sequences. This explanation is consistent with 4d
373 substitution rates between complete ORFs of Tr copies and CL copies in all LS and shared complete *tr-d*
374 events. Higher 4d substitution rates were detected between ORFs of Tr and CL copy pairs in gene-only
375 *tr-d* events (Fig. 7D, "go"), compared to those in *tr-d* events where HGSs comprised both gene and
376 intergenic regions (Fig. 7D, "g+i"), for both lineage-specific (Fig. 7D, "LS") and indepedent parallel (Fig.
377 7D, "Ind-par") shared *tr-d* events. The 4d substitution rates associated with shared *tr-d* events with Tr
378 copies co-linear to each other (Fig. 7D, "Tr-*cl*") showed median values comparable or higher than the
379 median 4d substitution rates between all best-hit pairs of *A. thaliana* and Lineage II genomes (Fig. 2D).
380 This further agreed with the assumption that these *tr-d* events were derived from duplications dated
381 prior to the divergence between Lineage I and II followed by deletion in the genomes in which Tr copies
382 were not found. However, the 4d substitution rates between duplicates were not different between

383 gene-only *tr-d* events and those with sequence similarities detected in both gene and intergenic regions
384 among "Tr-*cl*" type shared *tr-d* events (Fig. 7D).
385 Complete *tr-d* events included a higher number of Tr copy genes that showed evidence of
386 expression compared to incomplete *tr-d* events (Fig. 7E). Expression evidence was obtained using RNA-
387 seq data for root and leaf/shoot tissues as described in Methods. Among complete *tr-d* events, there
388 was no noticeable difference in the proportion of expressed Tr copies between gene-only *tr-d* and *tr-d*
389 with sequence similarities extended to the adjacent intergenic regions (data not shown). Incomplete *tr-*
390 *d* was associated with most of the Tr copies with truncated ORFs, which comprised the majority of Tr
391 copies in both lineage-specific *tr-d* (Fig. 5C and 5D) and independent parallel *tr-d* events shared by a pair
392 of genomes (Fig. 6C). Out of total 1,706 Tr copies with truncated ORFs, only 135 were derived from
393 complete *tr-d* events, in which the ORFs were most likely truncated by null mutations after the
394 duplication (Fig. 7E). Finally, we compared the proportion of single exon genes among CL and Tr copy
395 genes to check if there was an enrichment of retrotransposon-derived *tr-d* events. In both lineage-
396 specific and shared *tr-d* events, proportions of single-exon genes were higher among Tr copies
397 compared to CL copies, consistent with the fact that the majority of Tr copies had shorter ORFs than CL
398 copies. However, even the higher proportion of single-exon genes in Tr copies was either less or
399 comparable to the proportion of single-exon genes among all protein-coding gene models in the six
400 target genomes, hence we didn't observe a noticeable enrichment of retrotransposon-derived
401 duplications among *tr-d* events (Figure S5).
402
403 **Genes associated with lineage-specific and shared *tr-d* events**
404 Table 2 presents a partial list of OrthNets associated with lineage-specific *tr-d* events for each of
405 the six Brassicaceae genomes, selected based on the most number of Tr copies with complete ORFs and
406 expression evidences, except for *S. irio*, for which RNAseq data was not available. The "Node count
407 pattern" column, generated by the OrthNet program, indicate the copy numbers in each genome, in the
408 alphabetic order of the species names, i.e. *A. lyrata* (Aly), *A. thaliana* (Ath), *C. rubella* (Cru), *E.*
409 *salsugineum* (Esa), *S. irio* (Sir), and *S. parvula* (Spa). For each OrthNet, we listed the number of Tr copies
410 with complete ORFs, with expression evidence, and both (Table 2, "exp", "cORF", and "cORF & exp",
411 respectively). OrthNets with a large number of Tr copies may contain Tr copies as tandem duplicates.
412 Therefore, we included the tandem duplicated numbers in parenthesis followed by the total number of
413 Tr copies. We also showed the number of gene-only and "g+i" types (as defined in Fig. 7D) among Tr
414 copies in complete *tr-d* events. The complete list of OrthNets including lineage-specific *tr-d* events is

415 available in Dataset S2, as in the same format given in Table 2. Among several notable high copy *tr-d*
416 events specific to *A. lyrata*, *C. rubella*, *E. salsugineum*, and *S. irio* were OrthNets that included Tr copy
417 paralogs encoding an unknown protein (ON_1227), a homeobox transcription factor AGAMOUS-like 87
418 (AGL87, ON_2442), an RNA-binding protein (ON_3071), a stress-signaling SALT-TOLERANCE32 (SAT32,
419 ON_2516), a cytochrome P450 (ON_3061), and an amino acid permease 3 (ON_3400) (Table 2). We did
420 not find lineage-specifc *tr-d* OrthNets with more than two Tr copies for either *A. thaliana* or *S. parvula*.
421 This may be due to their compact genomes with less repetitive elements and also large numbers of *tr-d*
422 events shared by *A. thaliana-A. lyrata* and *S. parvula-S. irio* pairs, based on their phylogenetic proximity
423 than other pairs in the target genomes (Fig. 6B and Dataset S2). The OrthNet "ON_1227", encoding an
424 unknown protein, included a mixture of tandem and transposition-duplications of eleven *A. lyrata* gene
425 loci, of which six were CL copies found as tandem duplicates, while five were Tr copies with three of
426 them in a tandem gene array (Table 2). The *E. salsugimeum* genome showed an expansion of loci
427 related to stress responses such as *SAT32* (ON_2516) and *NAC6* (ON_5701) through *tr-d*, resulting in
428 increased copy numbers with complete ORFs which were also expressed. In addition, *E. salsugimeum*
429 also had *tr-d* events leading to higher copy number of reproduction and development-related loci,
430 including *KOKOPELLI* (*KPL*) and *TERMINAL FLOWER 1* (*TFL1*). The *S. parvula* genome had increased copy
431 numbers with complete ORFs and expression for loci encoding a zinc transporter (ZIP3; Table 2,
432 ON_4815) and a WRKY transcription factor (WRKY72; Table 2, ON_6085).
433 We did not find a significantly over-represented gene ontology (GO) term among genes
434 associated with lineage-specific *tr-d* in each genome tested, most likely due to the too small number of
435 genes to be able to detect statistical enrichment. However, genes in lineage-specific *tr-d* events for all
436 genomes combined showed enrichment of several GO terms (Supplementary Dataset S3). GO branches
437 under "nitrogen compound metabolic process (adjusted P-value or P_{adj}=2.4 x10^{-4})" accounted for 172
438 out of total 812 lineage-specific *tr-d* genes with a GO annotation. Among them 121 were annotated
439 with the child GO term "nucleic acid metabolic process (P_{adj}=0.020)" including 18 "DNA repair
440 (P_{adj}=0.022)" genes. "Response to abiotic stimulus (P_{adj}=2.4x10^{-4})" included 93 genes, of which 37
441 belonged to "response to osmotic stress (P_{adj}=3.0x10^{-3})" while another 39 were "response to radiation
442 (P_{adj}=0.020)". Other notable GO branches included "reproduction" (P_{adj}=0.012) and "post-embryonic
443 development" (P_{adj}=0.014), which collectively included 88 genes. "Organelle organization (P_{adj}=2.9x10^{-}
444 ^{7})" accounted for 77 gene loci, of which 31 and 13 were in child GO terms "chromosome organization
445 (P_{adj}=1.1x10^{-4})" and "cytoskeleton organization (P_{adj}=0.031)", respectively. "Cell cycle (P_{adj}=1.1x10^{-4})"
446 included 32 genes. The entire list of enriched GO terms is available as Supplementary Dataset S3.

447 We selected the two largest OrthNets with *E. salsugineum*- and *C. rubella*-specific *tr-d* events
448 (Table 2. bold-faced) and independently visualized the extent of gene duplications using the GEvo tool in
449 the CoGE database [32] (Figure 8) to provide detailed representations of evolutionary events that lead
450 to copy number expansion. The *E. salsuginuem* genome included six copies of *SALT TOLERANCE 32*
451 (*SAT32*), which exists as single copies in all other Brassicaceae genomes so far sequenced (data not
452 shown). While one copy was reciprocally co-linear to its orthologous nodes in other genomes (Fig. 8A,
453 "CL copy"), five copies were transposed (Fig. 8A, "Tr copies") with three of them found as tandem
454 duplicates. Four of the Tr copies had ORF sizes similar to the CL copy and three of them showed
455 expression in either root or shoot tissues (Fig. 8A and Table 2). The GEvo plot illustrates extensive
456 sequence similarity among all loci and adjacent genomic regions that are reciprocally co-linear among
457 them (Fig. 8B, *AtSAT32*, *SpSAT32*, and *EsSAT32;1*). On one hand, the Tr copies (Fig. 8B, *EsSAT32;2/3/4/5*)
458 represented examples of gene-only *tr-d* events (Fig. 7A), where sequence similarities between the
459 genomic regions including the CL and Tr copies were restricted to the expected border regions of the
460 gene models (i.e. ±20% of the gene size). On the other hand, the CL copy (*EsSAT32;1*) showed sequence
461 similarity over both adjacent intergenic and neighboring genes, compared to its orthologs in other
462 genomes (Fig. 8B, *AtSAT32* and *SpSAT32*; similar patterns observed with the *S. iro*, *C. rubella*, and *A.*
463 *lyrata* co-linear orthologs). Interestingly, the Tr copies *EsSAT32;3/4/5* exhibited intron loss, resulting in
464 9, 4, and 1 exons, respectively, compared to the 13 exons in the CL copy *EsSAT32;1* (Fig. 7B) while
465 maintaining high deduced amino acid similarities over most of the coding regions (Figure S6). Among all
466 *EsSAT* paralogs, the highest average expression was observed for one of the Tr copies, *EsSAT32;2*
467 (Supplementary Dataset S1, OrthNet ID ON_2516 and gene ID 20186362).
468 The *AGL87* OrthNet (Fig. 5A, right panel) exhibited a *C. rubella*-specific increase in copy number
469 through multiple *tr-d* events. Similar to *SAT32*, the genomic landscape of the CL copy (Fig. 8C,
470 *CrAGL87;1*) showed extensive co-linear HGSs with other member nodes of the reciprocally co-linear
471 hexagon (Fig. 5A and 8C, *AtAGL87* and *SpAGL87*; similar patterns observed with the *A. lyrata*, *E.*
472 *salsugineum*, and *S. irio* co-linear orthologs). HGSs included the Tr copy genes (Fig. 8C,
473 *CrAGL87;2/3/4/5/6*) and additional ± ~2Kbp adjacent intergenic regions, exemplifying complete *tr-d*
474 events with sequence similarities found in both gene and adjacent intergenic regions (Fig. 5D, "g+i").
475 The two tandem duplication events, i.e. *CrAGL87;3/4* and *CrAGL87;5/6* were located closely in a genome
476 scaffold, about 133Kbp apart from each other. They were separated by eight loci, which was larger than
477 the maximum loci difference cutoff parameter used in this study, four, for detecting tandem duplicated

478 loci (i.e. max_TD_loci_dist (*T*) = 4; see Methods), hence identified as two separated tandem duplication
479 events.
480 For *tr-d* events shared by multiple genomes, we present the entire list of such OrthNets for all
481 pairs of genomes, as well as those associated with Lineage I and II-specific *tr-d* events, in Supplementary
482 Dataset S2. The *tr-d* events shared by *E. salsugineum* and *S. parvula* were of particular interest, because
483 they may indicate signatures of convergent evolution between these two species independently
484 adapted to high salinity [54,55]. Table 3 lists all OrthNets with "Ind-parallel" and "Tr-*cl*" type *tr-d* events
485 (as defined in Fig. 6A) shared between these two halophytes. In Table 3, we also included all Lineage II
486 (i.e. *E. salsugineum*, *S. iro*, and *S. parvula*)-specific *tr-d* events that had truncated Tr copy ORFs only for *S.*
487 *iro* (Table 2, marked by subscript "e"). Among "Ind-parallel" tr-d events, three out of seven events
488 detected were associated with stress signaling or response-related functions (Table 3, *CDPK1*, *5PTASE11*,
489 and *ABI1*). However, none of them included complete Tr copy ORFs in both *E. salsugineum* and *S.*
490 *parvula*. Interestingly, the "Tr-*cl*" category included more loci with Lineage II-specific *tr-d* followed by
491 truncation of the Tr copy ORF in *S. iro*, leaving complete ORFs in the Tr copy loci only for *E. salsugineum*
492 and *S. parvula*. Here, we found loci encoding orthologs of a putative basic helix-loop-helix (bHLH) type
493 transcription factor, a NAC transcription factor (NAC058), and a calcineurin B-like protein 10 (CBL10). All
494 Tr copies encoding these regulatory proteins showed expression evidence in both halophyte species.
495 Also, the *CRUMPLED LEAF/CRL* locus and a locus encoding a phosphorylcholine cytidylyltransferase
496 (*CCT1*) were associated with *tr-d* events that resulted in additional Tr copies with complete ORFs that
497 were also expressed (Table 3).
498
499 **Copy number variations between *S. parvula* and *A. thaliana* revisited using OrthNets**
500 In a previous study, we identified a number of genes known to function in stress responses that
501 showed increased copy numbers by gene duplication in *S. parvula*, compared to *A. thaliana* [9]. Using
502 the CLfinder-OrthNet pipeline, we revisited these genes to identify their evolutionary contexts in the six
503 crucifer genomes (Figure 9). The hexagons of reciprocally co-linear nodes often appeared disrupted in
504 OrthNets that included tandem duplication events, because each of the duplicated nodes found their
505 best-hit differently among duplicates in other genomes (Fig. 9A-D). The OrthNet for *HIGH-AFFINITY* K^+
506 *TRANSPORTER 1* (*HKT1*) locus, known for increased copy numbers due to tandem duplication in both *S.*
507 *parvula* and *E. salsugineum* compared to *A. thaliana* [9,42], showed tandem duplication events not only
508 in the two extremophytes but also in *C. rubella* and *S. irio* (Fig. 9A). In *E. salsugineum* and *S. iro*, one of
509 the three duplicates contained truncated ORF, while *C. rubella* and *S. parvula* contained two and four

510 tandem duplicates with complete ORFs, respectively. The *Na^+ H^+ EXCHANGER 8* (*NHX8*) locus, encoding
511 a Li^+ transporter in *A. thaliana* [59], showed tandem duplication in *S. parvula* [9] as well as *S. iro*.
512 However, the copy number was higher in *S. parvula* than *S. iro* (Fig. 9B). The OrthNet for *CALCINEURIN*
513 *B-LIKE PROTEIN 10* (*CBL10*) identified a Lineage II-specific *tr-d* event, where Tr copies in *S. parvula* were
514 further tandem duplicated (Fig. 9C). The *CBL10* Tr copy in *S. iro* was split into two gene models with
515 truncated ORFs and one of them was inverted and placed in the opposite strand (Supplementary
516 Dataset S1, scf890_22 and scf890_23). For simplicity, we collapsed the two truncated gene models as a
517 single node (Fig. 9C, Sir|scf890_22/23). The OrthNet containing the *K^+ EFFLUX ANTIPORTER 1* (*KEA 1*)
518 locus revealed an interesting difference between Lineage I and II species. A Lineage I-specific *tr-d* event
519 created an additional Tr copy locus (named *KEA2* in *A. thaliana*) that was co-linear among all three
520 Lineage I genomes. On the other hand, all three Lineage II genomes showed tandem duplications of the
521 *KEA1* locus, while the *KEA2* locus was absent in Lineage II genomes (Fig. 9D). Hence, both Lineage I and
522 II achieved the same copy number through different types of gene duplication events. Finally, previous
523 studies have reported the *ARABIDOPSIS THALIANA V-PPASE 1* (*AVP1*) locus underwent a *tr-d* in *S.*
524 *parvula* but not in *A. thaliana* [9,55]. Using OrthNet, we identified a more complex network that
525 included at least three loci (Figure S7). Fig. 9E presents a partial OrthNet, including a perfect hexagon of
526 reciprocally co-linear *AVP1* nodes and a *tr-d* event with the Tr copies in five genomes, except for *A.*
527 *thaliana*, all reciprocally co-linear. Hence, what appeared to be a *S. parvula*-specific *tr-d* event in *S.*
528 *parvula-A. thaliana* pairwise comparison was in fact an *A. thaliana*-specific deletion of the additional
529 *AVP1* copy. While all other nodes in Fig. 9E had complete ORFs with minimum 2,307bp, the *A. lyrata* Tr
530 copy (Aly|16052878) contained a truncated ORF of only 432bp, due to a deletion of a ~2.8Kbp genomic
531 region in the middle of the gene (data not shown). Notably, *A. thaliana* had a TE insertion (RC/Helitron)
532 at the same position [55], replacing the second *AVP1* locus.
533

534 ## Discussion

535 **A systematic identification of orthologs in a shared evolutionary context**

536 Conservation of gene order and positional context, or co-linearity, is only explainable by
537 investigating the shared ancestry of the genome loci involved. The study of genomic co-linearity has
538 advanced to incorporate recently sequenced genomes to understand evolutionary history of gene
539 families beyond the scope of model organism genomics. Comparative analysis of conservation and
540 variation of co-linearity enable identifying modes of gene duplications [6,60]; inferring genome
541 duplication and fractionation events especially among plant genomes [61,62]; and tracing the origin of

542 genes or gene families [63,64]. Among closely-related genomes, co-linearity or micro-synteny of
543 orthologous genes often coincide with conservation of intergenic non-coding regulatory regions [39,65].
544 Gene duplication and transposition events disrupt such conservation. Duplication further affects gene
545 dosage, which may eventually lead to divergence of expression and functions among duplicates [8,66].
546 Transposition also modifies strength and tissue-specificity of gene expression through changes in
547 promoter sequences and *cis*-elements [9,67], local epigenetic environment [68], and proximity to
548 enhancers and chromatin structural contexts [69–71]. In Arabidopsis populations, for example,
549 transposition or transposition-duplication (*tr-d*) events are associated with variations in growth
550 phenotypes [72], genetic incompatibilities [68], and various quantitative traits [73].
551 Inter-species comparative analysis can encompass a wider range of phenotypic variation than
552 population-level studies within a species. Lineage-specific (LS) duplication and transposition of genes
553 with putative functions derived from studies in a closely related model species, present candidate
554 targets to develop hypotheses explaining novel phenotypic variations. In plants, such hypotheses are
555 verifiable through studies using transgenic and genetic manipulations [63,74,75]. Multi-genome
556 comparison improves the resolution in identifying candidate genes for shared derived traits and also if
557 tested with appropriate genomes, enables detecting signatures of convergent evolution within distant
558 relatives that share analogous traits.
559 We developed the CLfinder-OrthNet pipeline that enables systematic detection and visualization
560 of clusters of orthologs that share an "evolutionary context." Here we describe an evolutionary context
561 as a distinct pattern of varying gene copy numbers due to co-linearity or loss of that in a lineage or a
562 certain set of lineages. In the current study, we used the CLfinder-OrthNet pipeline to a set of six
563 Brassicaceae genomes, including the leading model plant genome and two extremophyte species
564 [54,55]. We specifically focused on the gene transposition-duplication (*tr-d*) events that caused
565 variation in copy number and divergence in co-linearity of single gene loci in multiple combinations
566 within the target genomes. We characterized both lineage-specific *tr-d* events and those shared by
567 different lineages through either convergent, parallel evolution, or selective retention of duplicates
568 originated from common ancestors, among the six Brassicaceae genomes.
569
570 **Transposition-duplication as a major mechanism for erosion of co-linearity**
571 Before the application of the pipeline, we analyzed the extent of co-linearity erosion among the
572 six crucifer genomes, by comparing the *A. thaliana* genome to each of the remaining genomes. We used
573 the displacement of best-hit loci in the target *A. thaliana* genome for loci at position *n* and *n+1* in the

574 query genomes (Fig. 2A, $d_{n,n+1}$) as a proxy for co-linearity (Fig. 2A, $d_{n,n+1}$ = 1) or its erosion (Fig. 2A, $d_{n,n+1} \neq$
575 1) at the locus *n* of the each query genome. On average, 58.3% (±3.2%) of all genes with detectable
576 similarity to an *A. thaliana* gene showed undisturbed co-linearity (Fig. 2B and 2C, $d_{n,n+1}$ = 1). While the
577 distribution of query gene frequency with $d_{n,n+1}$ in the range from 2 to 20 showed a rapid decay from the
578 peak at $d_{n,n+1}$ = 1, the frequency of query genes stayed nearly constant as $d_{n,n+1}$ increased beyond 20 (Fig.
579 2B) and to very large $d_{n,n+1}$ values up to thousands (data not shown). Moreover, substantial proportion
580 of query genes found their best-hit in a target *A. thaliana* chromosome different from their neighbors
581 for all query genomes (Fig. 2C, "Diff Chr"). This pattern suggested that two distinct mechanisms were
582 responsible for eroding co-linearity. Proximal gene displacements (Fig. 2C, $d_{n,n+1}$ = 2~20) were likely
583 results of mutations and local indels and inversions accumulated over time, while distal gene
584 displacements (Fig. 2C, $d_{n,n+1}$>20 or "Diff Chr") coincide with TE expansion, which may happen
585 sporadically in each lineage regardless of the divergence time (Fig. 2D).
586 The distal displacements were mostly due to single-gene transposition and transposition-
587 duplication (*tr-d*) events, as illustrated in the proportion of best-hit pairs showing *tr* type in Table 1.
588 Translocations of large chromosome-level segments also contributed, however, the effect was small.
589 The six crucifer genomes consist of 24 large ancestral karyotype blocks reshuffled to five (*A. thaliana*),
590 seven (*E. salsugineum*, *S. irio*, and *S. parvula*), and eight (*A. lyrata* and *C. rubella*) chromosomes [76].
591 The CLfinder module identifies borders of such blocks, and other smaller scale genomic segments that
592 were either duplicated, deleted, inverted, or transposed, as *cl_end* types (see Methods), which account
593 for less than a few hundred loci in all pairwise comparisons combined (Supplementary DataSet S1). We
594 excluded all loci annotated as TE as well as lineage-specific gene models without detectable homology
595 to a gene in the target genome (*A. thaliana*), which may contain unannotated TE loci, from the analysis
596 of co-linearity erosion. As a result, the distal displacement counts did not include transposition of TE
597 genes.
598 For the transposition and transposition-duplication (*tr-d*) of non-TE gene loci, two types of
599 models, retrotransposon-associated and DNA repair or replication-associated models have been
600 suggested as the main mechanisms [77,78]. A *tr-d* event derived from retrotransposons often leads to
601 duplication of single exon genes (i.e. intronless compared to the parent copy) [8,79–82]. Transposition-
602 duplication may also arise during the non-homologous end-joining (NHEJ) repair process of DNA double-
603 strand breaks (DSB), where a short sequence motif may act as an anchor to a foreign sequence to fill-in a
604 gap [1]. In agreement with this model, a previous comparison of *A. lyrata* and *A. thaliana* found a
605 significant enrichment of flanking repeats, as short as 15bps, among transposed genes [83]. The

606 correlation between the proportion of query gene loci showing distal displacement (Fig. 2C, $d_{n,n+1}$>20 or
607 "Diff Chr") and overall TE contents of the query genome, rather than divergence time (Fig. 2D), supports
608 the DSB-repair model. Higher TE contents likely provide a higher frequency of short repeat anchors
609 required for the NHEJ DSB-repair, and TE activities themselves may also cause the DSB that lead to such
610 repairs [1]. The DSB-repair model can explain *tr-d* of multi-exon genes, which constitute the majority of
611 *tr-d* events we detected in the current study (Fig. S5). A transposition event likely starts with a *tr-d* and
612 requires an additional deletion of the donor copy locus over time, which corroborates the fewer
613 observed number of such events compared to *tr-d* events (Table S2)
614
615 **Markov clustering to resolve large OrthNets into representations of individual loci**
616 For all query-target pairs among six Brassicaceae genomes, the CLfinder module determined co-
617 linearity relationships for all gene loci in the query with their respective best-hits in the target genomes.
618 The OrthNet module combined the CLfinder results into networks where a gene locus node was
619 connected to their best-hits in all other genomes, with the co-linearity relationships between nodes as
620 the edge property. The basic network was then enhanced by adding tandem duplicated (*td*) paralogs
621 connected by undirected edges when present. To minimize getting distant multigene families into one
622 complex network, we further attempted to divide the network into the smallest possible units, using
623 MCL with higher edge weights given to tandem duplicated and reciprocally co-linear edges (See
624 Methods). Therefore, the resulting OrthNets represented groups of orthologs likely derived from a
625 single ancestral gene locus, with the addition of duplicated, transposed, and *tr-d* loci in varying
626 combinations among the six selected Brassicaceae genomes. If duplicates were found equally in all six
627 genomes, i.e. duplication dated before the divergence of the six genomes, they were likely represented
628 in two or more separate OrthNets, unless there was an incongruence between sequence homology and
629 co-linearity (see the "concluding remark" section below for further discussion).
630
631 **Signatures of selective retention among lineage-specific transposed-duplicated gene loci.**
632 In lineage-specific *tr-d* events retrieved by OrthNet, we found a subset of transposed-duplicated
633 gene loci (Fig. 5A, Tr copies) retaining similar ORF sizes compared to their respective donor locus (Fig.
634 5A, CL copy), as well as to orthologs in other species (Fig. 5B-D). The DSB repair model of *tr-d* suggests
635 that the duplicated region may start and end virtually at any random position in a genome, given that
636 the short sequence motif needed for the repair is likely ubiquitously available and can be as short as
637 several nucleotides [1,83]. However, our simulation reveled that both "complete" and "gene-only" *tr-d*

638 events were far more frequent than what expected from a random duplication model alone (Fig. 7C).
639 We are not aware of a gene duplication mechanism that preferably duplicates non-TE, protein-coding
640 gene sequences as entire units. Rather, our observation common to all six tested crucifer genomes is
641 likely a result of random *tr-d* events (e.g. through DSB repair), followed by accumulation of mutations
642 throughout the duplicated regions, except where the complete coding sequences were selectively
643 retained. These Tr copy genes with complete ORFs were also more likely to be expressed, and hence
644 less likely to be pseudogenes, compared to those with truncated ORFs, based on RNA-seq evidence (Fig.
645 7E).
646 Among such Tr copy loci that retained complete ORFs, a significant fraction had the duplication
647 start and end positions coincided with the boundaries of the coding region, resulting in sequence
648 similarity found exclusively for the entire gene model but not in adjacent intergenic regions between the
649 donor and duplicated loci. We described these as "gene-only" *tr-d* events (Fig. 7). Gene-only *tr-d* events
650 may have had more time to lose sequence similarity outside of the coding sequence than the ones that
651 retain similarity extended to intergenic regions. Indeed, we found gene-only *tr-d* events (Fig. 7D, "go")
652 older than *tr-d* events where sequence similarities could be detected in both gene and intergenic
653 regions (Fig. 7D, "g+i"), in all *tr-d* events either lineage-specific or independently occurred in parallel in
654 two genomes (Fig. 7D, "LS" and "Ind-par"). For shared "Tr-*cl*" type *tr-d* events , the 4d substitution
655 rates between duplicates (Fig. 7D, both "Tr-*cl*:g+i" and "Tr-*cl*:go") suggested that the majority of them
656 were older than the divergence of the Brassicaceae genomes (Fig. 2D), consistent with the notion that
657 these *tr-d* events were originated from the common ancestors of the genomes sharing them.
658 Considering that intergenic regions are generally more prone to mutations than genes, intergenic
659 sequence similarity among duplicated paralog loci in these old shared *tr-d* events (Fig. 7D, "Tr-*cl*;g+i")
660 may indicate purifying selection in intergenic regions suggestive of the presence of conserved non-
661 coding regulatory elements.
662 Overall, our analyses depicted the landscape of lineage-specific *tr-d* events among Brassicaceae
663 genomes, where the majority of *tr-d* events were incomplete, while *tr-d* events including Tr copies with
664 complete ORFs and gene-only *tr-d* were likely results of selective retention over time.
665
666 **Distribution of *tr-d* events among Brassicaceae lineages and TE contents.**
667 An intriguing pattern emerged when we considered the distribution of shared *tr-d* events in
668 relation to overall TE contents (Table S2). TE may facilitate *tr-d* of non-TE genes by providing short
669 sequence repeats that can act as anchors in DSB repair or TE activity may lead to initiation of DSB repair-

670 mediated *tr-d* events [1]. Hence, we expect genomes with higher TE contents to contain more recent *tr-*
671 *d* events. These recent *tr-d* will also have higher proportion of incomplete duplication and truncated
672 ORF in Tr copy loci, since the duplications will be more likely random before the functional duplicates
673 get selectively retained over time (Fig. 2C).
674 Indeed, the three more TE-rich (>30% TE contents) genomes, *A. lyrata*, *E. salsugineum*, and *S.*
675 *irio*, contained more lineage-specific *tr* and *tr-d* events, compared to the remaining three TE-depleted
676 (<20% TE contents) genomes (Table S2). Interestingly, the highest number of lineage-specific *tr-d*
677 events, majority of which occured recently after the species' divergence from the closest relative (Fig.
678 7C-D), was observed in *A. lyrata* rather than in *E. salsugineum*, although the latter contained the highest
679 proportion of the genome annotated as TE (Table. S2). *A. lyrata* and *S. irio* also included the smallest
680 proportions of complete ORFs in Tr copy loci of lineage-specific *tr-d* events at 12.4% and 14.4%,
681 respectively. Contrastingly, *E. salsugineum* showed a proportion of complete ORFs in Tr copy loci at
682 28%, comparable to other TE-depleted genomes (Fig. 2D).
683 One possible explanation is that *A. lyrata* and *S. irio* genomes may contain TEs that had
684 expanded more recently than in the *E. salsugineum* genome. Both *A. lyrata* and *S. irio* have close
685 relatives with low TE contents, *A. thaliana* and *S. parvula*, respectively. Between *A. lyrata* and *A.*
686 *thaliana* genomes, TEs seem to have expanded in *A. lyrata* rather than being lost in *A. thaliana*. The TE
687 expansion in *A. lyrata*, as inferred by the expansion of recent long terminal repeat (LTR)
688 retrotransposons starting at 1 million years ago (MYA) [84], is indeed more recent than the estimated *A.*
689 *lyrata-A. thaliana* divergence at 3~5MYA [85]. While beyond the scope of this study, it will be
690 interesting to test whether the same is true for *S. irio* and *S. parvula* genomes, as well as comparing the
691 age of LTRs in *E. salsugineum* with those in *A. lyrata* and *S. irio*. The percentage of loci showing distal
692 best-hit displacement (Fig. 2C, $d_{n,n+1}$ >20 or "Diff Chr"), which reflect *tr-d* events, may show a better
693 correlation with the content of more recently expanded TEs, than with total contents of sequences
694 annotated as TE/repeats (Fig. 2D).
695
696 **Search for extremophyte lineage-specific *tr-d* events using OrthNets**
697 One possible application of the CLfinder-OrthNet pipeline is to retrieve orthologs sharing
698 evolutionary events uniquely found in a lineage or multiple lineages exhibiting a convergent trait. For
699 example, the two extremophytes included in the current study, *S. parvula* and *E. salsugineum*, share the
700 capacity to survive and reproduce under highly saline environments [51]. The habitats of these two
701 extremophytes also provide additional challanges dintinct to each species, multi-ion salt stresses for *S.*

702 *parvula* [9] and combined salt and freezing stresses for *E. salsugineum* [52,53]. Genomes of these two
703 Lineage II extremophytes may contain signatures of adaptation both shared and unique to each species.
704 Comparisons of the genomes and transcriptomes of *E. salsugineum* and *S. parvula* with those of
705 the model plant *A. thaliana* have identified a number of stress response-related orthologs showing
706 differences in copy number, promoter sequences, and basal level expression strengths, as well as
707 signatures of sub- and neo-functionalization among duplicates [9,10,42,86,87]. Functions of some of
708 these genes in stress adaptation have been tested and confirmed by transgenic studies in the
709 extremophyte species [87–89]. However, pairwise comparison to a model species often has limited
710 resolution in identifying candidate variations of interest. For example, comparison of *S. parvula* with the
711 model *A. thaliana* genomes detected thousands of orthologous genes that show different duplication or
712 gene transposition events (Table 1) and hundreds even after filtering for those showing variation in
713 expression strengths [9]. Given the high number of these candidates, often the hypothesis generation
714 step has to focus on a subset of candidates with stronger prior functional evidence related to plant
715 stress responses. The majority of these variations may reflect the more ancient divergence between the
716 Brassicaceae Lineage I and II, rather than more recent local adaptations specific to an extremophyte.
717 The CLfinder-OrthNet pipeline enables identifying lineage (monophyletic) or multiple lineages
718 (polyphyletic or paraphyletic)-specific variations in a multi-genome comparative framework. Our
719 pipeline identified 63, 26, and 14 ortholog groups showing gene copy number increases through *tr-d*
720 events specific to *E. salsugineum*, *S. parvula*, and both, respectively (Fig. 5D and Table 3). These
721 numbers are orders of magnitude fewer than previous searches from a pairwise comparison only to *A.*
722 *thaliana* [9], signifying the vastly improved resolutions in finding lineage-specific events.
723 *E. salsugineum* contained more lineage-specific *tr-d* events than *S. parvula*, in alignment with
724 the higher TE content of the genome (Fig. 5D and Table S2). The largest OrthNet with *E. salsugineum*-
725 specific *tr-d* events included orthologs of the *SALT-TOLERANCE 32* (*SAT32*) locus (Figure 8A and Table 2,
726 ON_2516), that encoded a transcription regulator, of which the expression level positively correlated
727 with the survival rate of the model plant *A. thaliana* under a salt stress condition [90]. Among the *SAT32*
728 orthologs in *E. salsugineum* (*EsSAT32s*), three of the four Tr copy loci with complete ORFs exhibited
729 intron losses, without any loss of deduced amino acid sequence homology for entire coding regions
730 (Figs. 8B and S6). Intron losses and smaller transcript sizes are thought to enable rapid responses and
731 regulation of expression timing in Drosophila and mouse [91,92] as well as in yeast [93]. It may also
732 indicate variations in DNA methylation landscapes among duplicated paralogs, which facilitate intron
733 losses [94]. It is not clear whether the lack of sequence homology in the adjacent intergenic regions

734 among *EsSAT32* paralogs (i.e. "gene-only" *tr-d*; Fig. 8B) is indicative of reverse transcriptase-mediated
735 duplication and intron losses [95] or different rate of mutation between gene and intergenic regions.
736 Either way, such variations in intergenic regions including promoters may lead to sub-functionalization
737 [8] of *EsSAT32* paralogs, at least three of which exhibited tissue-specific expressions with different basal
738 expression strengths in root and shoot tissues (data not shown).
739 Except for *SAT32*, OrthNets showing *E. salsugineum*- or *S. parvula*-specific *tr-d* did not include
740 genes with known functions directly related to salt stress responses in *A. thaliana*. This is rather
741 unsurprising considering that *A. thaliana* is naturally a salt-sensitive species and its response to salinity is
742 mostly a symptom than a selective and strategic response aiming for tolerance or avoidance and,
743 eventually, survival in extremely saline environments [96]. OrthNets with *E. salsugineum*- or *S. parvula*-
744 specific *tr-d* included loci encoding proteins annotated based on functional evidences in the model *A.*
745 *thaliana*, involved in defense (Table 2, ON_4608, ON_6085, and ON_1764-1, encoding MAPKKK5,
746 WRKY72, and DGK5, respectively; see the table legend for full names of proteins), regulation of growth
747 and development (Table 2, ON_4071, ON_4867, and ON_5499, encoding CYCA2;4, TFL1, and XCP1,
748 respectively), regulation of organelle functions and senescence (Table 2, ON_5701 and ON_4729 for
749 encoding and a UQCRX-like family protein), and embryogenesis (Table 2, ON_4633/KPL) [97]. A notable
750 example of *S. parvula*-specific increases in gene copy numbers through *tr-d* events includes the *ZRT/IRT-*
751 *LIKE PROTEIN 3* (*ZIP3*) locus encoding a zinc transporter protein precursor (Table 2). This may be a
752 possible adaptation strategy of the species in soils that are highly saline and also depleted in
753 micronutrients such as zinc and iron in central Anatolia [98,99]. We suggest that these genes, based on
754 their lineage-specific *tr-d* events leading to additional gene copies that are expressed and also encode
755 complete ORFs, may have functions in salt adaptation yet unknown in the salt-sensitive model *A.*
756 *thaliana*, or their functions, either conserved or modified, may regulate a strategic decision-making
757 process in growth and development in avoiding or alleviating salt stresses [100] in these two
758 extremophyte species. Additional tissue- and stress response-specific expression data and functional
759 studies using genetic and transgenic methods will be able to refine and confirm these hypotheses.
760
761 **Shared *tr-d* events and signatures of a convergent evolution**
762 Gene copy number variations through *tr-d* events shared exclusively by the two extremophytes
763 may represent a signature of convergent evolution in two separate lineages that show salt-adaptive
764 traits [51]. The OrthNet pipeline identified two subsets of *tr-d* events shared by the two extremophytes,
765 of which the complete list is presented in Table 3. Three parallel *tr-d* events independently occurred in

766 both *E. salsugineum* and *S. parvula* (Table 3, "Ind-parallel") involved loci encoding well-known salt stress
767 signal transducers in the model plant *A. thaliana* (Table 3, ABI1, CDPK1, and 5PTASE11, see the table
768 legend for full names) [97]. However, none of these events resulted in additional gene copies with
769 complete ORFs and expression evidence in both genomes (Table 3). The other subset of shared *tr-d*
770 events most likely took place in a Lineage II ancestor before the divergence of *E. salsugineum* and *S.*
771 *parvula*, based on the fact that the Tr copies in both genomes showed co-linearity as depicted in Fig. 6A.
772 Subsequent deletion or truncation of all duplicated Tr copies in *S. irio* made these tr-d events unique to
773 the two extremophytes (Table 3, "Tr-*cl*"). This subset included two OrthNets with *tr-d*, encoding
774 transcription factors of unknown functions (Table 3, ON_170-3 and ON_3847) [97]. In both cases, Tr
775 copy loci included complete ORFs with expression evidence in both extremophyte species. Additionally,
776 a well-known salt stress signaling gene locus encoding the Calcineurin B-Like protein 10 [101,102]
777 showed *tr-d* in both *E. salsugineum* and *S. parvula* with complete ORFs in Tr copy loci. However,
778 expression was detected only for the two *S. parvula* Tr copies (Table 3, ON_2430/CBL10).
779 The paucity of copy number enhancement with complete ORFs through shared parallel *tr-d*
780 events suggests that the adaptation path was distinct and unique to each extremophyte species. While
781 being the two most salt-tolerant among tested Brassicaceae species [51], the natural habitats of these
782 extremophytes do present distinct combinations of challenges for each species [9,52,103]. The
783 increased copy numbers of the two transcription factors found among the "*Tr-cl*" subset however,
784 present a good set of candidates for further genetic and transgenic studies, to determine their possible
785 contribution to the extremophytes' responses to high salinity, the common stress component of their
786 habitats.
787
788 **OrthNets reveal complex evolutionary histories of stress-related gene loci**
789 In a previous study using *A. thaliana-S. parvula* pairwise comparison, we found several genes
790 with stress responsive functions in the model *A. thaliana* that showed increased copy numbers through
791 gene duplication in *S. parvula* [9]. These included loci for ion transporters (*HKT1*, *NHX8*, *KEA 1*, and
792 *AVP1*), and a stress signal mediator (*CBL10*), with functions associated with ion homeostasis in the
793 model plant *A. thaliana* in response to salt stresses [97]. OrthNet visualization on the comparative
794 framework expanded to six species (Fig. 9) found that their evolutionary histories were more complex
795 than previously thought.
796 For example, the increased copy numbers of *HKT1* loci were not unique to extremophytes as
797 previously thought [9,42]. They were also found in *C. rubella* and *S. iro* genomes (Fig. 9A). Recent

798 studies have demonstrated that *E. salsugineum* and *S. parvula* share a copy of HKT1 that included two
799 amino acid substitutions unique to the two extremophytes, that were responsible for differentially
800 transporting Na^+ and K^+ [87,89]. None of the HKT1 orthologs from *C. rubella*, or any other crucifers
801 tested in this study, included this amino acid substitution (data not shown). Interestingly, OrthNet
802 showed that the HKT1 copy containing the amino acid substitutions was truncated in *E. salsugineum*
803 (Fig. 9A, Esa|20197527). This may be an error in the gene model annotation in the *E. salsugineum* draft
804 genome [43] used in this study. Previous studies mentioned above have successfully cloned a full length
805 cDNA from this particular locus using an alternative *E. salsugineum* genome annotation [42], which
806 contained a complete ORF in the same locus. In addition to the amino acid substitutions, the Tr copy of
807 *HKT1* in *S. parvula* (Fig. 9A, Spa|Sp6g07110) showed a tissue-specific expression pattern distinct from its
808 orthologs in both *A. thaliana* and *E. salsugineum* [9,42]. This may explain the unique adaptation of *S.*
809 *parvula* to its habitat, where soils contain toxic levels of both Na^+ and K^+ ions [103]. Overall, this
810 exemplifies the complexity of candidate gene mining and hypothesis development for functional studies
811 in non-model species, where the validity of gene model annotations and possibilities of neo- and sub-
812 functionalization need to be considered.
813 For *NHX8* and *CBL10* loci, OrthNets confirmed *S. parvula*-specific copy number increases
814 compared to the other five crucifers, previously speculated through *A. thalina-S. parvula* pairwise
815 comparison [9], although duplications were found in species other than *S. parvula* (Fig. 9B and 9C). The
816 increased *NHX8* copies in *S. parvula* may account for the the extreme Li^+ toxicity tolerance of the species
817 [9], based on an ectopic over-expression of the gene conferring tolerance to Li^+ toxicity in the model *A.*
818 *thaliana* [59], as well as expression patterns of the three *S. parvula* paralogs that indicate both increased
819 basal-level expression strengths and tissue-specific sub-functionalization [9]. The *NHX8* OrthNet
820 revealed copy number increases also in *A. lyrata* and *S. iro* through segmental duplication and tandem
821 duplication, respectively, although *S. parvula* maintained the highest copy numbers among them (Fig.
822 9B). We confirmed that *S. iro* was extremely sensitive to Li^+ levels *S. parvula* is tolerant of (Figure S8)
823 and it will be interesting to test the Li^+ toxicity of *A. lyrata*, similarly. The contribution of increased *NHX8*
824 copy numbers in *S. parvula* adaptation to high soil [Li^+], a characteristic of its natural habitat [103], is
825 currently under investigation through transgenic approaches. The OrthNet for *CBL10*, which encodes a
826 stress signal transducer interacting with a number of Na^+ and K^+ transporters [101,102], includes a *tr-d*
827 event of the locus shared by all three Lineage II genomes tested in this study and likely originated from a
828 Lineage II ancestor, followed by a further copy number increase via tandem duplication unique to *S.*

829 *parvula* (Fig. 9C). Due to the disruption of the transposed copy in *S. iro*, the copy number increase was
830 in effect specific to the extremophytes, with the highest copy number found in *S. parvula*.
831 For *KEA1* and *AVP1*, OrthNet revealed that the copy number increases were not unique to *S.*
832 *parvula* (Fig. 9D and 9E), unknown at the time of the previous study [9]. Both Lineage I and Lineage II
833 species acquired the second *KEA1* copy through *tr-d* (Fig. 9D, annotated as *KEA2* in the model *A. thaliana*
834 [97]) and tandem duplication, respectively. The *tr-d* of *AVP1* likely dates to an ancestor before the
835 division of Lineage I and II, while one of the duplicates got deleted in *A. thaliana* and truncated in *A.*
836 *lyrata* (Fig. 9E). Overall, these examples illustrates the improved resolution of the OrthNet approach
837 over the previous approach relied on a pairwise comparison to the model plant *A. thaliana* [9], in
838 identifying lineage-specific gene duplication events.
839
840 **Concluding remarks: potential and challenges of evolutionary context-based searches using the**
841 **CLfinder-OrthNet pipeline**
842 The CLfinder-OrthNet pipeline provides a flexible tool to identify co-linearity among orthologs in
843 multiple closely related genomes and organize them into networks enabling a search based on
844 evolutionary contexts for lineage- or lineages-specific gene duplication and transposition events. The
845 input for paralog clusters, used for identification of tandem duplication events in each genome (Fig. S1,
846 "Input 2"), can be derived from any gene clustering method, with a user-defined max_TD_loci_dist
847 parameter. Likewise, users can choose the method and criteria for inter-species comparisons to
848 determine best-hit pairs (Fig. S1, "Input 3"). In this study, we aimed to trace the *tr-d* events including
849 truncated ORFs. Hence we determined best-hits using BLASTN [58] with an e-value cutoff ($e<10^{-5}$)
850 among coding sequences. Users interested more on identifying conserved protein functions can apply
851 additional filters, such as conserved ORF lengths, deduced amino acid similarities, and conserved
852 peptide domain structures, to ensure best-hit pairs guided by functional ORFs. Finally, the genome
853 annotation input (Fig. S1, "Input 1") can include non-protein-coding features such as putative enhancers
854 identified as intergenic DNase Hypersensitive regions [70], miRNA genes, or any other conserved
855 regulatory feature that can be traced among multiple closely related genomes which often share
856 substantial sequence similarity over intergenic regions [39].
857 A challenge to the CLfinder-OrthNet approaches, especially for plant genomes, is that often the
858 co-linear loci do not necessarily include the most similar sequences, especially when each genome being
859 compared include multiple paralogs. Such incongruence between co-linearity and sequence similarity
860 results in a best-hit of a query locus found in a transposed location in the target genome, even when the

861 co-linear location in the target genome also contained a gene showing homology with the query locus.
862 To solve this problem, we designed the CLfinder module to accept multiple best-hit candidates for each
863 query locus and scan for an alternative best-hit that give a reciprocal best-hit pair (see Methods). In
864 addition, the OrthNet module employed MCL to separate large OrthNets into smallest possible units, so
865 that majority of OrthNets included a single locus in each genome plus additional duplicates in subsets of
866 genomes. Using this approach, we identified 86% of the total OrthNets including the same or less than
867 12 nodes in the current study (Fig. 4B). MCL has been used to identify orthologous groups based on a
868 network graph with edges weighted for the sequence similarity (e.g. OrthoMCL) [31]. Using different
869 edge weights based on co-linearity relationships, the OrthNet module essentially attempts identification
870 of orthologous gene groups based on genomic locations, instead of sequence similarities.
871 Even after clustering with MCL, some OrthNets appeared to carry multiple loci, likely derived
872 from the same origin. For example, the complete OrthNet for *AVP1* (Fig. S7 and Supplementary Dataset
873 S1, ON_0-0398) appeared to include orthologs derived from three loci. The first two included the
874 reciprocally co-linear hexagon including *AVP1* (Fig. S7, *AVP1*) and a transposed group of five loci
875 indicating an ancient *tr-d* event with the *A. thaliana* locus deleted, which agreed with previous studies
876 [9,55]. The third group showed the *S. iro* locus (Fig. S7, Sir|scf2049_71) connected to the *AVP2* locus in
877 *A. thaliana* as a reciprocal best-hit pair with a *tr* edge, rather than with the *AVP1* locus. While lacking
878 the additional duplicated copy of the *AVP1* locus, *A. thaliana* included an additional locus *VHP2;2*,
879 through a segmental duplication of the *AVP2* locus, which was unique to *A. thaliana* (Fig. S7). Overall,
880 the OrthNet revealed a complex evolutionary history involving three loci showing high sequence identity
881 (minimum 84% among ORF nucleotide sequences) for this vacuolar membrane proton pump, essential
882 for maintaining a proton gradient across the membrane required for ion transport to plant vacuoles and
883 other organelles [104,105].
884 Such complex patterns are expected to be more commonplace when the analysis include
885 species with different WGD histories, common among plant genomes. Crop wild relatives (CWRs) often
886 possess traits desirable for developing crops more resilient to environmental challenges, either lost
887 during the domestication of the crop species or acquired uniquely in CWRs through adaptive evolution
888 [106]. The two extremophytes included in this study are CWRs of a large number of food and biofuel
889 crops, Brassica species [44,107] (Fig. 1). Many crop species, including Brassica crops, have undergone
890 recent polyploidization events. Our next goal includes application of the CLfinder-OrthNet pipeline for
891 comparison of crop and CWRs with different WGD histories, which may require additional layers of
892 analyses on duplication depth and fractionations among sub-genomes [108,109].

893

894 ## Materials and Methods

895 **Genome and gene models**

896 We obtained genome annotations of *A. lyrata* (Aly, version 1.0), *A. thaliana* (Ath, v. "TAIR10"), *C.*
897 *rubella* (Cru, v. 1.0), and *E. salsugineum* (Esa, v. 1.0) from Phytozome v. 11 (http://genome.jgi.doe.gov/),
898 while genomes of *S. irio* (Sir, v. 0.2; CoGE genome id 19579) and *S. parvula* (v. 2.0) were downloaded
899 from CoGE (https://genomevolution.org/coge/) and thellungiella.org (http://thellungiella.org/data/),
900 respectively. For *S. irio* annotation, we used a combination of RepeatMasker, a BLASTN search versus
901 Repbase v. 20170127 (http://www.girinst.org/repbase/) and OrthoMCL [110] to further filter out
902 lineage-specific clusters of genes encoding components of transposons. To generate a species tree of
903 crucifer genomes including the six target species, we used Agalma [111] which built a maximum
904 likelihood tree based on 14,614 alignments of homologous protein-coding gene clusters. Each cluster
905 contained sequences from more than four crucifer genomes.

906 For the analysis of co-linearity erosion, we compared all protein-coding gene loci in five crucifer
907 genomes to those in *A. thaliana* using BLASTN. Four degenerate site (4d) substitution rate distributions
908 were analyzed for all reciprocal BLASTN pairs of protein-coding genes between the five crucifer genomes
909 and *A. thaliana*, using codeml [112] and custom scripts. Transposable element (TE) contents were
910 determined using RepeatMasker v. 4.0.7 and RepeatModeler v. 1.0.8 (http://www.repeatmasker.org/),
911 for all six target species.

912

913 **Identification of co-linearity among crucifer genomes using CLfinder**

914 We developed the CLfinder process to scan the order of homologous gene loci between a pair of
915 genomes. The CLfinder module perform this process automatically for all pairs among multiple closely-
916 related genomes. The CLfinder process uses three inputs (Fig. 3 and S1). We used only primary
917 transcript models for protein-coding gene loci for all three inputs. **Input 1**: A custom script
918 ("parse_gtf.py") parsed genome annotation in .gtf format. **Input 2:** Intra-species sequence comparison
919 detected paralogous loci by clustering primary protein-coding gene model sequences within each single
920 species using OrthoMCL [110], as described previously [9]. **Input 3:** For inter-species sequence
921 comparison, we performed pairwise reciprocal BLASTN ($e<10^{-5}$) between all primary protein-coding gene
922 model ORFs to obtain ten most similar BLASTN hits ("best-hits") in the target species, for each gene
923 model in the query species.

924 The sensitivity and stringency of the CLfinder process are adjustable based on three user-
925 defined parameters, window_size (*W*), num_CL_trshld (*N*), and gap_CL_trshld (*G*) (Fig. S2). For each
926 query locus (from Input 1), CLfinder scans and counts number of 'loci-in-chain', whose best-hits (from
927 Input 3) are separated by the same or less than *G* from the best-hit of their immediate neighbors. The
928 scan starts from the query loci and moves towards both up- and downstream of the query genome. If
929 more than *N* loci-in-chain, including the query locus, are found within the window of size 2 x *W* (i.e. *W* to
930 the up- and downstream), the query locus is declared co-linear (*cl*) with its best-hit locus in the target
931 species. We excluded lineage-specific (*ls*) loci that do not have a best-hit in the target species from the
932 scan and counting process for loci-in-chain. Tandem duplicated (*td*) loci, defined as adjacent loci
933 separated by the same or less than max_TD_loci_dist (*T*) that are in the same paralog cluster defined in
934 Input 2, are counted as one locus during the scan and counting. Query locus neither co-linear (*cl*) nor
935 lineage-specific (*ls*) is transposed (*tr*). If query locus is in a genome assembly scaffold that contains less
936 than *W* loci, the pipeline mark the query locus as not-determined (*nd*) instead of transposed (*tr*).
937 Segmental rearrangements such as local inversion are identified based on loci-in-chain detected
938 exclusively in either up- or down-stream. In such cases, CLfinder declares the query and its best-hit as
939 *cl_end*.
940 For the analysis of six crucifer genomes, CLfinder parameters *W* and *G* were decided based on
941 the distribution of protein-coding gene locus content in the scaffolds of the most fragmented genome
942 (*S. irio*) and the results from the analysis of co-linearity erosion. We used 20, 3, 20, and 4 for *W*, *N*, *G*,
943 and *T*, respectively. The CLfinder process (Fig. S2) was performed for all possible query-target pairs for
944 the six crucifer species using a wrapper script in the CLfinder module (Fig. S1, "CLfinder_multi.py").
945
946 **Construction and visualization of OrthNets**
947 The OrthNet module combines the output of the CLfinder module and produces networks of
948 orthologous loci (OrthNet). The network contains primary protein-coding gene loci from all species
949 (nodes) connected with their best-hits with directional edges, and the co-linearity relationship (i.e.
950 either *cl*, *tr*, or *nd*) between them as the edge property. Additionally, the OrthNet module connects
951 tandem duplicated paralogs with undirected edges (*td*), and compares the ORF size of a node to the
952 median ORF size of all neighboring nodes to detect truncated ORFs (i.e. of sizes less than 40 and 80% of
953 the median ORF size).
954 The OrthNet module first clusters all nodes from all species connected with an edge (Fig. S1,
955 "create_OrthNet.py"). This process often results in an aggregation of multiple loci connected with

956 unidirectional edges in a single large OrthNet, due to duplicated paralogs, lineage-specific deletions, and
957 partial gene models that lead to non-reciprocal best-hit pairs. We employed two methods to alleviate
958 this issue.
959 First, for all unidirectional edges from node A to node B, the OrthNet module searches for an
960 alternative best-hit for B, among the list of the ten best-hit candidates, that makes A and B a reciprocal
961 best-hit pairs (Fig. S1, "update_BestHitPairs.py"). The best-hit candidates were obtained by running the
962 BLASTN with the "-max_target_seqs 10" option when creating the Input 2 for the CLfinder module. The
963 OrthNet module records how many best-hit candidates it tested before identifying a reciprocal best-hit
964 pair (Supplmentary Dataset S1, legend). Only when a reciprocal best-hit does not exist among all ten
965 alternatives, the OrthNet will have an unidirectional edge. Second, each OrthNet is subjected to further
966 clustering to finer OrthNets using Markov Chain Clustering (MCL) [113], with a higher edge weight given
967 to the order of *td*, reciprocal *cl*, reciprocal *tr*, unidirectional *cl*, and unidirectional *tr* edges. This process
968 effectively removed spurious unidirectional edges that connect nodes from different loci (Fig S1.
969 "update_OrthNet_afterMCL.py"). The resulted OrthNets are converted to .sif file format for
970 visualization using Cytoscape (http://cytoscape.org/).
971
972 **Analysis of lineage-specific transposition-duplication events using OrthNet**
973 The OrthNet module codifies each OrthNet based on the number of nodes from each genome as
974 well as edge properties connecting them, enabling effective search for patterns of interest (Fig. S1,
975 "search_OrthNet.py"). The module assigned all nodes in an OrthNet with an "OrthNet_code", which
976 included copy numbers in all six genomes used in this study, plus edge properties to their best-hit
977 ortholog nodes and sorted by the alphabetical order of genome names (Supplementary DataSet S1,
978 column "OrthNet_code"). We used the python regular expression patterns shown in Table S1, to search
979 for all *tr-d* events unique to one or multiple lineages among the six Brassicaceae genomes.
980 We used this functionality to search for OrthNets including lineage-specific (LS) transposition-
981 duplication (*tr-d*) events. Within a *tr-d* event, the *tr-d* donor or "CL copy" was the node connected to
982 nodes from other species with the most *cl* edges, while the remaining were *tr-d* acceptors or "Tr copies."
983 When multiple CL copies existed due to tandem duplication, we used the one with the longest ORF as
984 the representing CL copy. We compared ORF sizes between nodes from all other species and both CL
985 and Tr copies, as well as between the CL copy and Tr copies. Homologous Genome Segments (HGSs)
986 were detected between the gene and ±5Kb intergenic regions of the CL copy and each of Tr copies, using
987 LASTZ with chaining as previously described [9]. A *tr-d* event was "complete" if the HGS included the

988 entire CL copy gene. A "gene-only" *tr-d* was defined as a complete *tr-d* event with the size of HGS less
989 than 120% of the CL copy gene. We used a custom script to determine the expected occurrences of
990 complete and gene-only *tr-d* by random shuffling and overlapping of HGSs and CL copy genes. The
991 distribution of such occurrences from 10,000 iterations was fitted to a normal distribution to calculate
992 the P-value of the observed occurrence, using the fitdistr function in R MASS package (https://cran.r-
993 project.org/web/packages/MASS).
994 The 4d substation rates were calculated for all CL and Tr copy pairs where the Tr copy ORF size
995 was comparable (with ±20% in proportion) to that of the representative CL copy (i.e. Tr copy with a
996 complete ORF), using codeml and a custom script (Fig. S2, "pairwiseKs_by_codeml.py"). To determine
997 Tr copies with expression evidence, we used RNA-seq data for leaf and root tissues obtained from
998 studies by Wang et al. (2016)[13] for *A. lyrata*, *A. thaliana*, and *C. rubella*, and Oh et al. (2014) for *A.*
999 *thaliana* and *S. parvula* [9] as well as samples prepared for the current study (for *E. salsugineum*,
1000 BioProject ID PRJNA63667) essentially as previously described [9]. FPKM values of representative
1001 transcript models were estimated using Stringtie (v. 1.3.1c) with the '-e' option [114], after RNAseq
1002 reads were aligned to the genome using HISAT2 (v. 2.0.5) [114].
1003
1004 **Availability of software**
1005 All custom scripts and the CLfinder-OrthNet pipeline are available on GitHub
1006 (https://github.com/ohdongha/CL_finder), together with a tutorial with a sample dataset and
1007 parameters used for the current study.
1008
1009 **Figure legends**
1010 **Figure 1.** A comparative genomics framework including the two extremophyte/halophyte crucifers, *S.*
1011 *parvula* and *E. salsugineum*. Boxes and stars indicate the six Brassicaceae species selected for this work
1012 and halophytes, respectively. The tree was based on an alignment of 14,614 homologous gene clusters,
1013 as described in Methods.
1014
1015 **Figure 2.** Erosion of co-linearity between Brassicaceae (crucifer family) genomes
1016 **(A)** Each locus in the query genome (orange) is paired with an arrow to its most homologous
1017 counterpart locus ("best-hit") in the target genome (blue), after loci in both genomes numbered based
1018 on their positions. The difference in best-hit loci positions between the query locus at position *n* and its
1019 neighbor at position *n+1* ($d_{n,n+1}$) is 1 when the query locus and its best-hit in the target are co-linear,

1020 while a tandem duplication unique to the query genome results in $d_{n,n+1}$ = 0 (1). Co-linearity is disrupted
1021 by mutations and indels (2), as well as a gene transposition event (3).
1022 **(B, C)** The distribution of $d_{n,n+1}$ with either *A. lyrata* (Aly), *C. rubella* (Cru), *E. salsugineum* (Esa), *S. irio*
1023 (Sir), and *S. parvula* (Spa) as the query and the *A. thaliana* (Ath) genome as the target. Aly-Ath and Esa-
1024 Ath pairs were shown as exemplary histograms in B. Results for all pairs were summarized in C.
1025 **(D)** The proportion of gene loci with $d_{n,n+1}$ ranging from 2 to 20 showed a higher correlation with the
1026 median four-degenerated (4d) substitution rate (upper panels). Contrastingly, the proportion of those
1027 with $d_{n,n+1}$ larger than 20 or their best-hits found in a different target chromosome was better correlated
1028 with the overall transposable element (TE) contents (lower panels).
1029 [a] The query locus involved in a tandem duplication event unique to the query genome.
1030 [b] The query locus co-linear with its best-hit without any modification in gene order.
1031 [c] The best-hit of the query locus *n* is located in a different chromosome ("Diff Chr") in Ath, compared to
1032 the best-hit of its immediate neighbor locus *n+1*.
1033 [d] Includes loci with $d_{n,n+1}$ > 20 as well as the "Diff Chr" indicated in (*C*)
1034
1035 **Figure 3.** The CLfinder-OrthNet pipeline
1036 [a] For the detailed method to determine co-linearity (CL) relationship between the query loci and their
1037 most homologous counterpart ("best-hits") in the target genome, see Figures S2 and S3. An example
1038 CLfinder output for the six crucifer species highlighted in Fig. 1 are summarized in Table 1, with the full
1039 results given as Supplementary Dataset S1.
1040 [b] See Figures 4-6, for examples of OrthNets with different evolutionary contexts, e.g. transposition (*tr*)
1041 and transposition-duplication (*tr-d*) unique to each species or a group of species.
1042
1043 **Figure 4.** The OrthNet module encodes the evolutionary context of an orthologous gene group into a
1044 network.
1045 **(A)** OrthNet examples representing five different evolutionary contexts. Nodes are color-coded
1046 according to the species. Transparent nodes with dashed borders indicate loci with truncated ORFs, i.e.
1047 ORF sizes smaller than either 80% or 50% compared to the median ORF size of nodes they are connected
1048 to. Edges show properties either co-linear (*cl*) or transposed (*tr*), reciprocally (rc) or uni-directionally
1049 (uni). Tandem duplicated (*td*) paralogs are also connected by an edge. The lower panel includes a table

1050 with the ID and annotation for representative OrthNets, as well as the number of OrthNets with the
1051 same evolutionary context among total 17,432 identified for the six genomes.
1052 **(B)** A histogram showing the size distribution of OrthNets.
1053
1054 **Figure 5.** Characterization of lineage-specific transposition-duplication (*tr-d*) events among the six
1055 Brassicaceae genomes
1056 **(A)** Examples of OrthNets with *tr-d* events specific for *S. parvula* (*WRKY DNA-BINDING PROTEIN*
1057 *72*/*WRKY72*) and *C. rubella* (*AGAMOUS-LIKE 87*/*AGL87*). Within a *tr-d* event, OrthNet identifies the
1058 original donor copy (CL copy), which is co-linear to orthologs in other genomes (Orthologs), and the
1059 transposed and duplicated copies (Tr copies). Node and edges are as described in Fig. 4A.
1060 **(B)** ORF size comparison between all loci involved in a *tr-d* event (i.e. both CL and Tr copies) with the
1061 median of Orthologs. Blue dots indicate CL copies.
1062 **(C and D)** ORF size comparison between Tr copies and their corresponding CL copy within each of the *tr-*
1063 *d* events, as a scatterplot (C) and a histogram of ORF size ratio (D). Pink shades indicate Tr copies with
1064 complete ORFs whose sizes are comparable (±20% in proportion) to that of the CL copy. Panel D also
1065 shows numbers and percentages (in parentheses) of Tr copies with complete ORFs below the species
1066 labels. The entire list of OrthNets showing lineage-specific *tr-d*, including CL and Tr copies, is in
1067 Supplementary Dataset S2.
1068
1069 **Figure 6.** Transposition-duplication (*tr-d*) events shared by a pair of Brassicaceae genomes
1070 **(A)** Example OrthNets with *tr-d* events shared by *E. salsugineum* and *S. parvula*, representing two
1071 categories: (1) independent-parallel ("Ind-parallel") *tr-d* events and (2) *tr-d* events with Tr copies co-
1072 linear to each other ("Tr-*cl*"). Node and edges are as described in Fig. 4A.
1073 **(B)** Number of OrthNets shared by pairs of genomes in "Ind-parallel" and "Tr-*cl*" categories, with the
1074 cladogram of the six crucifer genomes on the top. Heatmap colors visualize the rank of each cell based
1075 on OrthNet numbers in each category.
1076 **(C)** Proportion of Tr copies with complete ORFs (i.e. ORF size ±20% of the CL copy in proportion) within
1077 OrthNets presented in (B). The genome 1 (row)-genome 2 (column) position shows the number of Tr
1078 copies with complete ORF / total Tr copies in genome 1, for all OrthNets with *tr-d* shared by genome 1
1079 and 2. For example, the Aly-Ath and Ath-Aly positions in "Ind-parallel" category indicate 6 out of 29 *A.*
1080 *lyrata* (Aly) Tr copies and 6 out of 22 *A. thaliana* (Ath) Tr copies, respectively, have complete ORFs in the

1081 17 “Ind-parallel” *tr-d* events shared by *A. lyrata* and *A. thaliana*. Heatmap colors indicate the
1082 percentage of Tr copies with complete ORFs, for cells with >5 total Tr copies.
1083
1084 **Figure 7.** Characterization of duplicated genomic regions in transposition-duplication (*tr-d*) events.
1085 **(A)** We Identified Homologous Genome Segments (HGSs) between the CL and Tr copy genes and
1086 adjacent genomic regions (±5Kbp) in a *tr-d* event as described in Methods. A *tr-d* event is either
1087 complete or incomplete, based on whether the HGS included the full CL copy gene or not. A subset of
1088 complete *tr-d* events had HGS coinciding with the start and end of the CL copy gene without extending
1089 to the intergenic regions (“gene-only” *tr-d*).
1090 **(B)** Histograms and box-and-whisker plots (inlets) showing size distributions of HGSs and CL copy genes
1091 for complete and incomplete *tr-d* events.
1092 **(C)** Comparison of observed occurrences (Obs.) and expected occurrences (Exp.) of complete and gene-
1093 only *tr-d* events. Upper panel shows the distribution of expected occurrences from 10,000 random
1094 shuffling of HGSs and CL copy genes for *A. lyrata* (Aly)-specific complete and gene-only *tr-d* events.
1095 Fitting the random shuffling results to normal distributions (upper panel, blue curves) generated P-
1096 values of observed occurrences for *tr-d* events unique to each genome and the genus Arabidopsis (lower
1097 panel).
1098 **(D)** Four degenerate site (4d) substitution rates between ORFs of CL and Tr copy genes in different types
1099 of complete *tr-d* events. Complete *tr-d* events were either gene-only (“go”) or with HGSs detected in
1100 both gene and intergenic regions (“g+I”). We compared lineage-specific (LS) and shared *tr-d* events that
1101 are either independent-parallel (“Ind-par”) or with Tr copies co-linear to each other (“Tr-*cl*”). Lines and
1102 “x” marks in the box indicate medians and means, respectively.
1103 (*E*) Proportion of Tr copy genes with expression evidences (RNA-seq FPKM>0) in all *tr-d* events either
1104 lineage-specific or shared by a pair of genomes. The *tr-d* type is as described in (A) and Tr copy ORF type
1105 is as in Fig. 5C and D (pink shade) and Fig. 7C. *S. irio* genes were excluded due to the lack of RNA-seq
1106 data.
1107
1108 **Figure 8.** Examples of lineage-specific *tr-d* events.
1109 **(A)** OrthNet for the *SALT-TOLERANCEE 32* (*SAT32*). Nodes and edges are as described in Fig. 4A. The
1110 OrthNet showed *E. salsugineum*-specific *tr-d*, with three of the five Tr copies tandem duplicated.
1111 **(B)** Comparison of *SAT32* loci and adjacent ±30Kbp genomic regions between *A. thaliana*, *S. parvula*, and
1112 *E. salsugineum* as a GEvo plot (https://genomevolution.org/r/maxx). Pink shades connect Homologous

1113 Genomic Segments (HGSs) between genomes, while gene models, mRNAs, and coding sequences are
1114 depicted in gray, blue, and green, respectively (see https://genomevolution.org/wiki/index.php/GEvo
1115 for detailed GEvo legends). *EsSAT32;1* is the CL copy (Esa|20186362), while *EsSAT32;2/3/4/5* indicate
1116 the four Tr copies with complete ORFs. The yellow arrow marks the position of Esa|20181274, which
1117 contains a truncated ORF. *EsSAT32;3/4/5* showed intron losses without compromising gene products
1118 (see text and Figure S6 for details).
1119 **(C)** GEvo plot visualizing *AGAMOUS-LIKE 87* (*AGL87*) loci of *S. parvula*, *E. salsugineum*, *A. thaliana*, and *C.*
1120 *rubella* (https://genomevolution.org/r/olzb). The OrthNet for *AGL87* in the six crucifer species is in Fig.
1121 5A. *CrAGL87;1* is the CL copy (Cru|20890904) and *CrAGL87;2/3/4/5/6* represent Tr copies. The genomic
1122 regions including tandem duplicates *CrAGL87;3/4* and *CrAGL87;5/6* are located only 133Kbp apart from
1123 each other in the same genome scaffold and contained a large number of short HGSs between them.
1124 Only HGSs including the *CrAGL87* loci are connected with pink shades in these genome regions.
1125
1126 **Figure 9.** OrthNets associated with stress-related gene loci known for copy number variations between
1127 *S. parvula* and *A. thaliana*.
1128 OrthNets for **(A)** *HIGH AFFINITY K^+ TRANSPORTER 1* (*HKT1*; ON_1313), **(B)** *Na^+ H^+ EXCHANGER 8* (*NHX8*;
1129 ON_1214-1), **(C)** *CALCINEURIN B-LIKE PROTEIN 10* (*CBL10*; ON_2430), **(D)** *K^+ EFFLUX ANTIPORTER 1* (*KEA*
1130 *1*; ON_0-0500), and **(E)** *ARABIDOPSIS THALIANA V-PPASE 1* (*AVP1*; ON_0-0398, part) are presented, with
1131 nodes and edges as described in Fig. 4A. (E) shows a part of the OrthNet including orthologs derived
1132 from the *AVP1* loci, while the complete OrthNet is in Figure S7 (see the Discussion section for detail).
1133

1134 **List of Supplementary Information**

1135 **Table S1.** Regular expressions to retrieve OrthNets containing transposition (*tr*) and transposition-
1136 dupcliation (*tr-d*) events unique to a genome and shared by two genomes or within Lineage I and II,
1137 among the six Brassicaceae genomes.
1138 **Table S2.** Number of OrthNets showing lineage-specific (LS) transposition (*tr*) and transposition-
1139 duplication (*tr-d*)
1140
1141 **Figure S1.** A detailed flowchart of the CL_finder-OrthNet pipeline.
1142 **Fig. S2.** An example of the CLfinder process, comparing *S. parvula* (Spa) and *A. thaliana* (Ath) as the
1143 query and target genome, respectively.

1144 **Fig. S3.** An example of *A. lyrata* (Aly)-specific *tr-d* events with a large number of Tr copies (OrthNet ID:
1145 On_40)
1146 **Fig. S4.** Comparison of CL copy gene lengths between Tr copies with complete (-c) and truncated (-t)
1147 ORFs in six crucifer genomes
1148 **Fig. S5** Proportion of single-exon genes among CL and Tr copies in lineage-specific (LS) and shared *tr-d*
1149 events, compared to all genes in six crucifer genomes.
1150 **Fig. S6.** An alignment of deduced amino acid sequences for crucifer *SALT TOLERANCE 32* (*SAT32*) genes.
1151 **Fig. S7** Full OrthNet including the *AVP1* loci (ON_0-0398).
1152
1153 **Dataset S1.** Combined output of the CLfinder-OrthNet including gene annotation and RNAseq
1154 expression values for the six Brassicaceae genomes.
1155 **Dataset S2.** List of OrthNets that include transposition and transposition-duplication events unique to a
1156 genome or shared by groups of genomes among the six Brassicaceae genomes.
1157 **Dataset S3.** List of Gene Ontology (GO) terms enriched in genes associated with lineage-specific
1158 transposition-duplication events among the six Brassicaceae genomes.
1159

1160 ## References

1161 1. Wicker TM, Buchmann JP, Keller B. Patching gaps in plant genomes results in gene movement and
1162 erosion of colinearity. Genome Res [Internet]. 2010;20:1229–37. Available from:
1163 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2928501&tool=pmcentrez&rendertype=ab
1164 stract
1165 2. Freeling M, Lyons E, Pedersen B, Alam M, Ming R, Lisch D. Many or most genes in Arabidopsis
1166 transposed after the origin of the order Brassicales. Genome Res [Internet]. 2008 [cited 2011 Jul
1167 25];18:1924–37. Available from:
1168 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2593585&tool=pmcentrez&rendertype=ab
1169 stract
1170 3. Salse J, Abrouk M, Bolot S, Guilhot N, Courcelle E, Faraut T, et al. Reconstruction of
1171 monocotelydoneous proto-chromosomes reveals faster evolution in plants than in animals. Proc Natl
1172 Acad Sci USA [Internet]. 2009;106:14908–13. Available from:
1173 http://www.pnas.org/cgi/doi/10.1073/pnas.0902350106
1174 4. Van de Peer Y, Mizrachi E, Marchal K. The evolutionary significance of polyploidy. Nat Rev Genet
1175 [Internet]. 2017 [cited 2017 Jun 5]; Available from: http://www.ncbi.nlm.nih.gov/pubmed/28502977

1176 5. Li Z, Defoort J, Tasdighian S, Maere S, Van de Peer Y, De Smet R. Gene Duplicability of Core Genes Is
1177 Highly Consistent across All Angiosperms. Plant Cell [Internet]. 2016;28:326–44. Available from:
1178 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4790876&tool=pmcentrez&rendertype=ab
1179 stract
1180 6. Freeling M. Bias in plant gene content following different sorts of duplication: tandem, whole-
1181 genome, segmental, or by transposition. Annu Rev Plant Biol [Internet]. 2009 [cited 2011 Jul 27];60:433–
1182 53. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19575588
1183 7. Woodhouse MR, Tang H, Freeling M. Different Gene Families in Arabidopsis thaliana Transposed in
1184 Different Epochs and at Different Frequencies throughout the Rosids. Plant Cell [Internet].
1185 2011;23:4241–53. Available from: http://www.plantcell.org/cgi/doi/10.1105/tpc.111.093567
1186 8. Wang Y, Wang X, Paterson AH. Genome and gene duplications and gene expression divergence: A
1187 view from plants. Ann N Y Acad Sci. 2012;1256:1–14.
1188 9. Oh D-H, Hong H, Lee SY, Yun D-J, Bohnert HJ, Dassanayake M. Genome Structures and Transcriptomes
1189 Signify Niche Adaptation for the Multiple-Ion-Tolerant Extremophyte Schrenkiella parvula. Plant Physiol
1190 [Internet]. 2014 [cited 2014 Apr 8];164:2123–38. Available from:
1191 http://www.plantphysiol.org/cgi/doi/10.1104/pp.113.233551
1192 10. Oh D-H, Dassanayake M, Haas JS, Kropornika A, Wright C, D'Urzo MP, et al. Genome structures and
1193 halophyte-specific gene expression of the extremophile *Thellungiella parvula* in comparison with
1194 *Thellungiella salsuginea* (*Thellungiella halophila*) and Arabidopsis. Plant Physiol [Internet]. 2010 [cited
1195 2010 Nov 27];154:1040–52. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20833729
1196 11. Liu S-L, Pan AQ, Adams KL. Protein subcellular relocalization of duplicated genes in Arabidopsis.
1197 Genome Biol Evol [Internet]. 2014;6:2501–15. Available from:
1198 http://gbe.oxfordjournals.org/content/early/2014/09/04/gbe.evu191.abstract
1199 12. Panchy N, Lehti-Shiu MD, Shiu S-H. Evolution of gene duplication in plants. Plant Physiol [Internet].
1200 2016;171:pp.00523.2016. Available from:
1201 http://www.plantphysiol.org/lookup/doi/10.1104/pp.16.00523
1202 13. Wang J, Tao F, Marowsky NC, Fan C. Evolutionary Fates and Dynamic Functionalization of Young
1203 Duplicate Genes in Arabidopsis Genomes. Plant Physiol [Internet]. 2016;172:427–40. Available from:
1204 http://www.plantphysiol.org/lookup/doi/10.1104/pp.16.01177
1205 14. Ossowski S, Schneeberger K, Lucas-Lledo JI, Warthmann N, Clark RM, Shaw RG, et al. The Rate and
1206 Molecular Spectrum of Spontaneous Mutations in Arabidopsis thaliana. Science (80-.). [Internet].
1207 2010;327:92–4. Available from: http://www.sciencemag.org/cgi/doi/10.1126/science.1180677

1208 15. Campbell CD, Eichler EE. Properties and rates of germline mutations in humans. Trends Genet
1209 [Internet]. Elsevier Ltd; 2013;29:575–84. Available from: http://dx.doi.org/10.1016/j.tig.2013.04.005
1210 16. Kondrashov FA. Gene duplication as a mechanism of genomic adaptation to a changing environment.
1211 Proc R Soc B [Internet]. 2012 [cited 2013 Sep 19];279:5048–57. Available from:
1212 http://rspb.royalsocietypublishing.org/content/279/1749/5048
1213 17. Qian W, Zhang J. Genomic evidence for adaptation by gene duplication. Genome Res. 2014;24:1356–
1214 62.
1215 18. Hardigan MA, Crisovan E, Hamilton JP, Kim J, Laimbeer P, Leisner CP, et al. Genome Reduction
1216 Uncovers a Large Dispensable Genome and Adaptive Role for Copy Number Variation in Asexually
1217 Propagated Solanum tuberosum. Plant Cell [Internet]. 2016;28:388–405. Available from:
1218 http://www.plantcell.org/lookup/doi/10.1105/tpc.15.00538
1219 19. Panchy N, Lehti-Shiu MD, Shiu S-H. Evolution of gene duplication in plants. Plant Physiol [Internet].
1220 2016;171:pp.00523.2016. Available from:
1221 http://www.plantphysiol.org/lookup/doi/10.1104/pp.16.00523
1222 20. Kondrashov FA. Gene duplication as a mechanism of genomic adaptation to a changing environment.
1223 Proc R Soc B [Internet]. 2012;279:5048–57. Available from:
1224 http://rspb.royalsocietypublishing.org/cgi/doi/10.1098/rspb.2012.1108
1225 21. Cook DE, Lee TG, Guo X, Melito S, Wang K, Bayless A, et al. Copy number variation of multiple genes
1226 at *Rhg1* mediates nematode resistance in soybean. Science (80-.). [Internet]. 2012 [cited 2012 Nov
1227 19];338:1206–9. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23065905
1228 22. Gan X, Hay A, Kwantes M, Haberer G, Hallab A, Ioio R Dello, et al. The Cardamine hirsuta genome
1229 offers insight into the evolution of morphological diversity. Nat Plant [Internet]. 2016;2:16167. Available
1230 from: http://www.nature.com/articles/nplants2016167
1231 23. Shirai K, Matsuda F, Nakabayashi R, Okamoto M, Tanaka M, Fujimoto A, et al. A Highly Specific
1232 Genome-Wide Association Study Integrated with Transcriptome Data Reveals the Contribution of Copy
1233 Number Variations to Specialized Metabolites in Arabidopsis thaliana Accessions. Mol Biol Evol
1234 [Internet]. 2017;1–12. Available from:
1235 http://academic.oup.com/mbe/article/doi/10.1093/molbev/msx234/4110007/A-Highly-Specific-
1236 GenomeWide-Association-Study
1237 24. Simon S, Ruhl M, De Montaigu A, Wotzel S, Coupland G. Evolution of CONSTANS regulation and
1238 function after gene duplication produced a photoperiodic flowering switch in the brassicaceae. Mol Biol
1239 Evol. 2015;32:2284–301.

1240 25. Grandaubert J, Lowe RGT, Soyer JL, Schoch CL, Van de Wouw AP, Fudal I, et al. Transposable
1241 element-assisted evolution and adaptation to host plant within the Leptosphaeria maculans-
1242 Leptosphaeria biglobosa species complex of fungal pathogens. BMC Genomics [Internet]. 2014;15:1–27.
1243 Available from: http://www.biomedcentral.com/1471-
1244 2164/15/891%5Cnhttp://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-15-891
1245 26. Boore JL, Lavrov D V, Brown WM. Gene translocation links insects and crustaceans. Nature.
1246 1998;392:667–8.
1247 27. Wang J, Sun P, Li Y, Liu Y, Yu J, Ma X, et al. Hierarchically Aligning 10 Legume Genomes Establishes a
1248 Family-Level Genomics Platform [Internet]. Plant Physiol. 2017. Available from:
1249 http://www.plantphysiol.org/lookup/doi/10.1104/pp.16.01981
1250 28. Koenig D, Weigel D. Beyond the thale: comparative genomics and genetics of Arabidopsis relatives.
1251 Nat Rev Genet [Internet]. Nature Publishing Group; 2015;16:285–98. Available from:
1252 http://dx.doi.org/10.1038/nrg3883
1253 29. Wang M, Yu Y, Haberer G, Marri PR, Fan C, Goicoechea JL, et al. The genome sequence of African rice
1254 (Oryza glaberrima) and evidence for independent domestication. Nat Genet [Internet]. Nature
1255 Publishing Group; 2014;46:982–8. Available from: http://www.nature.com/doifinder/10.1038/ng.3044
1256 30. Du H, Yu Y, Ma Y, Gao Q, Cao Y, Chen Z, et al. Sequencing and de novo assembly of a near complete
1257 indica rice genome. Nat Commun [Internet]. Nature Publishing Group; 2017;8:15324. Available from:
1258 http://www.nature.com/doifinder/10.1038/ncomms15324
1259 31. Chen F, Mackey AJ, Stoeckert CJ, Roos DS. OrthoMCL-DB: querying a comprehensive multi-species
1260 collection of ortholog groups. Nucleic Acids Res [Internet]. 2006 [cited 2017 Jun 4];34:D363–8. Available
1261 from: http://www.ncbi.nlm.nih.gov/pubmed/16381887
1262 32. Lyons E, Freeling M. How to usefully compare homologous plant genes and chromosomes as DNA
1263 sequences. Plant J [Internet]. 2008 [cited 2011 Feb 8];53:661–73. Available from:
1264 http://www.ncbi.nlm.nih.gov/pubmed/18269575
1265 33. Vandepoele K. A Guide to the PLAZA 3.0 Plant Comparative Genomic Database. Methods Mol Biol
1266 [Internet]. 2017 [cited 2017 Jun 4]. p. 183–200. Available from:
1267 http://www.ncbi.nlm.nih.gov/pubmed/27987171
1268 34. Grin I, Linke D. GCView: The genomic context viewer for protein homology searches. Nucleic Acids
1269 Res. 2011;39:353–6.
1270 35. Wall PK, Leebens-Mack J, Müller KF, Field D, Altman NS, Depamphilis CW. PlantTribes: A gene and
1271 gene family resource for comparative genomics in plants. Nucleic Acids Res. 2008;36:970–6.

1272 36. Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, et al. MCScanX: A toolkit for detection and
1273 evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 2012;40:1–14.
1274 37. Tang H, Bomhoff MD, Briones E, Zhang L, Schnable JC, Lyons E. SynFind: compiling syntenic regions
1275 across any set of genomes on demand. Genome Biol Evol [Internet]. 2015;7:3286–98. Available from:
1276 http://gbe.oxfordjournals.org/content/early/2015/11/11/gbe.evv219.abstract
1277 38. Proost S, Fostier J, De Witte D, Dhoedt B, Demeester P, Van De Peer Y, et al. i-ADHoRe 3.0-fast and
1278 sensitive detection of genomic homology in extremely large data sets. Nucleic Acids Res. 2012;40.
1279 39. Haudry A, Platts AAE, Vello E, Hoen DRD, Leclercq M, Williamson RJ, et al. An atlas of over 90,000
1280 conserved noncoding sequences provides insight into crucifer regulatory regions. Nat Genet [Internet].
1281 2013 [cited 2013 Oct 9];45:891–8. Available from: http://www.nature.com/doifinder/10.1038/ng.2684
1282 40. Cheng C-Y, Krishnakumar V, Chan A, Schobel S, Town CD. Araport11: a complete reannotation of the
1283 Arabidopsis thaliana reference genome. bioRxiv [Internet]. 2016;47308. Available from:
1284 http://biorxiv.org/content/early/2016/04/05/047308.abstract
1285 41. Dassanayake M, Oh D-H, Haas JSJS, Hernandez A, Hong H, Ali S, et al. The genome of the
1286 extremophile crucifer *Thellungiella parvula*. Nat Genet [Internet]. Nature Publishing Group; 2011 [cited
1287 2011 Nov 10];43:913–918. Available from: http://www.nature.com/doifinder/10.1038/ng.889
1288 42. Wu H-J, Zhang Z, Wang J-Y, Oh D-H, Dassanayake M, Liu B, et al. Insights into salt tolerance from the
1289 genome of *Thellungiella salsuginea*. Proc Natl Acad Sci USA [Internet]. 2012 [cited 2013 Sep
1290 19];109:12219–24. Available from: http://www.pnas.org/content/109/30/12219.short
1291 43. Yang R, Jarvis DE, Chen H, Beilstein M a, Grimwood J, Jenkins J, et al. The Reference Genome of the
1292 Halophytic Plant Eutrema salsugineum. Front Plant Sci [Internet]. 2013 [cited 2013 Mar 28];4:46.
1293 Available from:
1294 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3604812&tool=pmcentrez&rendertype=ab
1295 stract
1296 44. Cheng F, Mandáková T, Wu J, Xie Q, Lysak M a., Wang X, et al. Deciphering the diploid ancestral
1297 genome of the mesohexaploid Brassica rapa. Plant Cell [Internet]. 2013 [cited 2013 May 8];25:1541–54.
1298 Available from: http://www.plantcell.org/cgi/doi/10.1105/tpc.113.110486
1299 45. Kagale S, Koh C, Nixon J, Bollina V, Clarke WE, Tuteja R, et al. The emerging biofuel crop Camelina
1300 sativa retains a highly undifferentiated hexaploid genome structure. Nat Commun [Internet]. Nature
1301 Publishing Group; 2014;5:3706. Available from:
1302 http://www.nature.com/doifinder/10.1038/ncomms4706
1303 46. Chalhoub B, Denoeud F, Liu S, Parkin IAP, Tang H, Wang X, et al. Early allopolyploid evolution in the

1304 post-Neolithic *Brassica napus* oilseed genome. Science (80-.). 2014;345:950 LP-953.
1305 47. Liu S, Liu Y, Yang X, Tong C, Edwards D, Parkin I a P, et al. The Brassica oleracea genome reveals the
1306 asymmetrical evolution of polyploid genomes. Nat Commun [Internet]. Nature Publishing Group; 2014
1307 [cited 2014 May 25];5:3930. Available from: http://www.ncbi.nlm.nih.gov/pubmed/24852848
1308 48. Yang J, Liu D, Wang X, Ji C, Cheng F. The genome sequence of allopolyploid Brassica juncea and
1309 analysis of differential homoeolog gene expression influencing selection. Nat Genet. 2016;
1310 49. Dorn KM, Fankhauser JD, Wyse DL, Marks MD. A draft genome of field pennycress (Thlaspi arvense)
1311 provides tools for the domestication of a new winter biofuel crop. DNA Res [Internet]. 2015;22:121–31.
1312 Available from: http://dnaresearch.oxfordjournals.org/cgi/doi/10.1093/dnares/dsu045
1313 50. Kitashiba H, Li F, Hirakawa H, Kawanabe T, Zou Z, Hasegawa Y, et al. Draft Sequences of the Radish
1314 (Raphanus sativus L.) Genome. DNA Res [Internet]. 2014;1–10. Available from:
1315 http://www.ncbi.nlm.nih.gov/pubmed/24848699
1316 51. Orsini F, D'Urzo MP, Inan G, Serra S, Oh DH, Mickelbart M V., et al. A comparative study of salt
1317 tolerance parameters in 11 wild relatives of Arabidopsis thaliana. J Exp Bot [Internet]. Oxford University
1318 Press; 2010;61:3787–98. Available from: http://www.ncbi.nlm.nih.gov/pubmed/20595237
1319 52. Inan G, Zhang Q, Li P, Wang Z, Cao Z, Zhang H, et al. Salt cress. A halophyte and cryophyte
1320 *Arabidopsis* relative model system and its applicability to molecular genetic analyses of growth and
1321 development of extremophiles. Plant Physiol [Internet]. 2004 [cited 2011 Mar 14];135:1718–37.
1322 Available from: http://www.plantphysiol.org/cgi/content/abstract/135/3/1718
1323 53. Amtmann A. Learning from evolution: *Thellungiella* generates new knowledge on essential and
1324 critical components of abiotic stress tolerance in plants. Mol Plant [Internet]. 2009 [cited 2010 Dec
1325 3];2:3–12. Available from:
1326 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2639741&tool=pmcentrez&rendertype=ab
1327 stract
1328 54. Dittami SM, Tonon T. Genomes of extremophile crucifers: new platforms for comparative genomics
1329 and beyond. Genome Biol [Internet]. 2012;13:166. Available from:
1330 http://genomebiology.com/2012/13/8/166
1331 55. Oh D-H, Dassanayake M, Bohnert HJ, Cheeseman JM. Life at the extreme: lessons from the genome.
1332 Genome Biol [Internet]. 2012;13:241. Available from: http://genomebiology.com/2012/13/3/241
1333 56. Mandáková T, Lysak MA. Chromosomal phylogeny and karyotype evolution in x=7 crucifer species
1334 (Brassicaceae). Plant Cell [Internet]. 2008 [cited 2012 Mar 13];20:2559–70. Available from:
1335 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2590746&tool=pmcentrez&rendertype=ab

1336 stract
1337 57. Wang Y, Li J, Paterson AH. MCScanX-transposed: Detecting transposed gene duplications based on
1338 multiple colinearity scans. Bioinformatics. 2013;29:1458–60.
1339 58. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and
1340 applications. BMC Bioinformatics [Internet]. 2009 [cited 2017 Dec 14];10:421. Available from:
1341 http://www.ncbi.nlm.nih.gov/pubmed/20003500
1342 59. An R, Chen Q-J, Chai M-F, Lu P-L, Su Z, Qin Z-X, et al. AtNHX8, a member of the monovalent cation:
1343 proton antiporter-1 family in Arabidopsis thaliana, encodes a putative Li+/H+ antiporter. Plant J
1344 [Internet]. 2007 [cited 2010 Nov 29];49:718–28. Available from:
1345 http://www.ncbi.nlm.nih.gov/pubmed/17270011
1346 60. Emms DM, Covshoff S, Hibberd JM, Kelly S. Independent and Parallel Evolution of New Genes by
1347 Gene Duplication in Two Origins of C4 Photosynthesis Provides New Insight into the Mechanism of
1348 Phloem Loading in C4 Species. Mol Biol Evol. 2016;33:1796–806.
1349 61. Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH. Synteny and Collinearity in Plant
1350 Genomes. Science (80-.). [Internet]. 2008;320:486–9. Available from:
1351 http://www.sciencemag.org/content/320/5875/486.full.pdf
1352 62. Jiao Y, Paterson AH. Polyploidy-associated genome modifications during land plant evolution. Phil
1353 Trans R Soc B [Internet]. 2014;369:20130355. Available from:
1354 http://rstb.royalsocietypublishing.org/lookup/doi/10.1098/rstb.2013.0355
1355 63. Vlad D, Kierzkowski D, Rast MI, Vuolo F, Dello Ioio R, Galinha C, et al. Leaf Shape Evolution Through
1356 Duplication, Regulatory Diversification, and Loss of a Homeobox Gene. Science (80-.). [Internet].
1357 2014;343:780–3. Available from: http://www.sciencemag.org/cgi/doi/10.1126/science.1248384
1358 64. Zhao T, Holmer R, Bruijn S de, Angenent GC, van den Burg HA, Schranz ME. Phylogenomic Synteny
1359 Network Analysis of MADS-Box Transcription Factor Genes Reveals Lineage-Specific Transpositions,
1360 Ancient Tandem Duplications, and Deep Positional Conservation. Plant Cell [Internet]. American Society
1361 of Plant Biologists; 2017 [cited 2017 Jun 8];tpc.00312.2017. Available from:
1362 http://www.ncbi.nlm.nih.gov/pubmed/28584165
1363 65. Engström PG, Sui SJH, Drivenes Ø, Becker TS, Lenhard B. Genomic regulatory blocks underlie
1364 extensive microsynteny conservation in insects. Genome Res. 2007;17:1898–908.
1365 66. Assis R, Bachtrog D. Neofunctionalization of young duplicate genes in Drosophila. Proc Natl Acad Sci
1366 USA [Internet]. 2013;110:17409–14. Available from:
1367 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3808614&tool=pmcentrez&rendertype=ab

1368 stract
1369 67. Arsovski AA, Pradinuk J, Guo XQ, Wang S, Adams KL. Evolution of cis-regulatory elements and
1370 regulatory networks in duplicated genes of Arabidopsis thaliana. Plant Physiol [Internet].
1371 2015;169:pp.00717.2015. Available from:
1372 http://www.ncbi.nlm.nih.gov/pubmed/26474639%5Cnhttp://www.plantphysiol.org/lookup/doi/10.110
1373 4/pp.15.00717
1374 68. Durand S, Bouché N, Perez Strand E, Loudet O, Camilleri C. Rapid establishment of genetic
1375 incompatibility through natural epigenetic variation. Curr Biol. 2012;22:326–31.
1376 69. Feuerborn A, Cook PR. Why the activity of a gene depends on its neighbors. Trends Genet [Internet].
1377 Elsevier Ltd; 2015;31:483–90. Available from: http://dx.doi.org/10.1016/j.tig.2015.07.001
1378 70. Zhu B, Zhang W, Zhang T, Liu B, Jiang J. Genome-Wide Prediction and Validation of Intergenic
1379 Enhancers in Arabidopsis Using Open Chromatin Signatures. Plant Cell [Internet]. 2015;27:2415–26.
1380 Available from: http://www.ncbi.nlm.nih.gov/pubmed/26373455
1381 71. Hu Z, Tee WW. Enhancers and chromatin structures: regulatory hubs in gene expression and
1382 diseases. Biosci Rep [Internet]. 2017;0:BSR20160183. Available from:
1383 http://bioscirep.org/lookup/doi/10.1042/BSR20160183
1384 72. Vlad D, Rappaport F, Simon M, Loudet O. Gene transposition causing natural variation for growth in
1385 Arabidopsis thaliana. PLoS Genet. 2010;6:21.
1386 73. Imprialou M, Kahles A, Steffen JG, Osborne EJ, Gan X, Lempe J, et al. Genomic Rearrangements in
1387 Arabidopsis Considered as Quantitative Traits. Genetics [Internet]. 2017 [cited 2017 Jun 7];205:1425–41.
1388 Available from: http://www.ncbi.nlm.nih.gov/pubmed/28179367
1389 74. Dun X, Shen W, Hu K, Zhou Z, Xia S, Wen J, et al. Neofunctionalization of Duplicated Tic40 Genes
1390 Caused a Gain-of-Function Variation Related to Male Fertility in Brassica oleracea Lineages. Plant Physiol
1391 [Internet]. 2014;166:1403–19. Available from:
1392 http://www.plantphysiol.org/cgi/doi/10.1104/pp.114.246470
1393 75. Das M, Haberer G, Panda A, Laha S Das, Ghosh TC, Schäffner AR. Expression pattern similarities
1394 support the prediction of orthologs retaining common functions after gene duplication events. Plant
1395 Physiol [Internet]. 2016;171:pp.01207.2015. Available from:
1396 http://www.plantphysiol.org/lookup/doi/10.1104/pp.15.01207
1397 76. Lysak MA, Berr A, Pecinka A, Schmidt R, McBreen K, Schubert I. Mechanisms of chromosome number
1398 reduction in Arabidopsis thaliana and related Brassicaceae species. Proc Natl Acad Sci USA [Internet].
1399 2006;103:5224–9. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.0510791103

1400 77. Hastings PJ, Lupski JR, Rosenberg SM, Ira G. Mechanisms of change in gene copy number. Nat Rev
1401 Genet [Internet]. 2009;10:551–64. Available from: http://www.nature.com/doifinder/10.1038/nrg2593
1402 78. Robberecht C, Voet T, Esteki MZ, Nowakowska BA, Vermeesch JR. Nonallelic homologous
1403 recombination between retrotransposable elements is a driver of de novo unbalanced translocations.
1404 Genome Res. 2013;23:411–8.
1405 79. Cusack BP, Wolfe KH. Not born equal: Increased rate asymmetry in relocated and retrotransposed
1406 rodent gene duplicates. Mol Biol Evol. 2007;24:679–86.
1407 80. Moran J V. Exon Shuffling by L1 Retrotransposition. Science (80-.). [Internet]. 1999;283:1530–4.
1408 Available from: http://www.sciencemag.org/cgi/doi/10.1126/science.283.5407.1530
1409 81. Abdelkarim BTM, Maranda V, Drouin G. The fate of retrotransposed processed genes in Arabidopsis
1410 thaliana. Gene [Internet]. The Author(s); 2017;609:1–8. Available from:
1411 http://dx.doi.org/10.1016/j.gene.2017.01.029
1412 82. Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A. Gene duplication and exon shuffling
1413 by helitron-like transposons generate intraspecies diversity in maize. Nat Genet [Internet]. 2005;37:997–
1414 1002. Available from: http://www.nature.com/doifinder/10.1038/ng1615
1415 83. Woodhouse MR, Pedersen B, Freeling M. Transposed genes in Arabidopsis are often associated with
1416 flanking repeats. PLoS Genet. 2010;6:26.
1417 84. Baidouri M El, Panaud O. Comparative genomic paleontology across plant kingdom reveals the
1418 dynamics of TE-driven genome evolution. Genome Biol Evol. 2013;5:954–65.
1419 85. Hu TTTT, Pattyn P, Bakker EGEG, Cao J, Cheng JJF, Clark RM, et al. The Arabidopsis lyrata genome
1420 sequence and the basis of rapid genome size change. Nat Genet [Internet]. Nature Publishing Group;
1421 2011 [cited 2011 Oct 16];43:476–481. Available from: http://dx.doi.org/10.1038/ng.807
1422 86. Dassanayake M, Oh D-H, Hong H, Bohnert HJ, Cheeseman JM. Transcription strength and halophytic
1423 lifestyle. Trends Plant Sci [Internet]. 2011 [cited 2010 Nov 27];16:1–3. Available from:
1424 http://dx.doi.org/10.1016/j.tplants.2010.10.006
1425 87. Ali Z, Park HC, Ali A, Oh D-HD-H, Aman R, Kropornicka a., et al. TsHKT1;2, a HKT1 Homolog from the
1426 Extremophile Arabidopsis Relative Thellungiella salsuginea, Shows K+ Specificity in the Presence of NaCl.
1427 Plant Physiol [Internet]. Am Soc Plant Biol; 2012 [cited 2012 Mar 6];158:1463–74. Available from:
1428 http://www.ncbi.nlm.nih.gov/pubmed/22238420
1429 88. Oh D-H, Leidi E, Zhang Q, Hwang S-M, Li Y, Quintero FJ, et al. Loss of halophytism by interference
1430 with *SOS1* expression. Plant Physiol [Internet]. 2009 [cited 2010 Aug 25];151:210–22. Available from:
1431 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2735974&tool=pmcentrez&rendertype=ab

1432 stract
1433 89. Ali A, Raddatz N, Aman R, Kim S, Park HC, Jan M, et al. A Single Amino Acid Substitution in the
1434 Sodium Transporter HKT1 Associated with Plant Salt Tolerance. Plant Physiol [Internet].
1435 2016;171:pp.00569.2016. Available from:
1436 http://www.plantphysiol.org/lookup/doi/10.1104/pp.16.00569
1437 90. Park MY, Chung MS, Koh HS, Lee DJ, Ahn SJ, Kim CS. Isolation and functional characterization of the
1438 Arabidopsis salt-tolerance 32 (AtSAT32) gene associated with salt tolerance and ABA signaling. Physiol
1439 Plant. 2009;135:426–35.
1440 91. Hao S, Baltimore D. RNA splicing regulates the temporal order of TNF-induced gene expression. Proc
1441 Natl Acad Sci USA [Internet]. 2013;110:11934–9. Available from:
1442 http://www.pnas.org/cgi/doi/10.1073/pnas.1309990110
1443 92. Jiang L, Li XN, Niu DK. Higher frequency of intron loss from the promoter proximally paused genes of
1444 Drosophila melanogaster: Evidence consistent with delays in intron splicing as a selective force. Fly
1445 (Austin). 2014;8:120–5.
1446 93. Jeffares DC, Penkett CJ, Bähler J. Rapidly regulated genes are intron poor. Trends Genet.
1447 2008;24:375–8.
1448 94. Wang H, Devos KM, Bennetzen JL. Recurrent Loss of Specific Introns during Angiosperm Evolution.
1449 PLoS Genet. 2014;10.
1450 95. William Roy S, Gilbert W. The evolution of spliceosomal introns: patterns, puzzles and progress. Nat
1451 Rev Genet [Internet]. 2006;7:211–21. Available from:
1452 http://www.nature.com/doifinder/10.1038/nrg1807
1453 96. Gong Q, Li P, Ma S, Indu Rupassara S, Bohnert HJ. Salinity stress adaptation competence in the
1454 extremophile Thellungiella halophila in comparison with its relative Arabidopsis thaliana. Plant J
1455 [Internet]. 2005;44:826–39. Available from:
1456 http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=1
1457 6297073
1458 97. Cheng C-Y, Krishnakumar V, Chan AP, Thibaud-Nissen F, Schobel S, Town CD. Araport11: a complete
1459 reannotation of the Arabidopsis thaliana reference genome. Plant J [Internet]. 2017 [cited 2017 Mar
1460 8];89:789–804. Available from: http://doi.wiley.com/10.1111/tpj.13415
1461 98. Eide D. The Zip Family of Zinc Transporters. Plant Cell. 2005;1465:190–8.
1462 99. Cakmak I, Kalayci M, Ekiz H, Braun HJ, Kilinç Y, Yilmaz A. Zinc deficiency as a practical problem in
1463 plant and human nutrition in Turkey: A NATO-science for stability project. F. Crop Res. 1999;60:175–88.

1464 100. Julkowska MM, Testerink C. Tuning plant signaling and growth to survive salt. Trends Plant Sci
1465 [Internet]. Elsevier Ltd; 2015;1–9. Available from:
1466 http://linkinghub.elsevier.com/retrieve/pii/S1360138515001727
1467 101. Quan R, Lin H, Mendoza I, Zhang Y, Cao W, Yang Y, et al. SCABP8/CBL10, a putative calcium sensor,
1468 interacts with the protein kinase SOS2 to protect Arabidopsis shoots from salt stress. Plant Cell
1469 [Internet]. 2007 [cited 2010 Nov 29];19:1415–31. Available from:
1470 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1913747&tool=pmcentrez&rendertype=ab
1471 stract
1472 102. Ren XL, Qi GN, Feng HQ, Zhao S, Zhao SS, Wang Y, et al. Calcineurin B-like protein CBL10 directly
1473 interacts with AKT1 and modulates K+ homeostasis in Arabidopsis. Plant J [Internet]. 2013 [cited 2015
1474 Jan 1];74:258–66. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23331977
1475 103. Helvaci C, Mordogan H, Çolak M, Gündoğan I. Presence and Distribution of Lithium in Borate
1476 Deposits and Some Recent Lake Waters of West-Central Turkey. Int. Geol. Rev. [Internet]. 2004;46:177–
1477 90. Available from: http://www.informaworld.com/openurl?genre=article&doi=10.2747/0020-
1478 6814.46.2.177&magic=crossref%7C%7CD404A21C5BB053405B1A640AFFD44AE3
1479 104. Schilling RK, Tester M, Marschner P, Plett DC, Roy SJ. AVP1: One Protein, Many Roles. Trends Plant
1480 Sci [Internet]. Elsevier Ltd; 2016;22:154–62. Available from:
1481 http://linkinghub.elsevier.com/retrieve/pii/S1360138516301972
1482 105. Kronzucker HJ, Britto DT. Sodium transport in plants: a critical review. New Phytol [Internet]. 2011
1483 [cited 2010 Dec 8];189:54–81. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21118256
1484 106. Dassanayake M, Yun DOD, Bressan RA, Cheeseman JM, Bohnert JH. The scope of things to come:
1485 New paradigms in biotechnology. Plant Biotechnol. Agric. 2009.
1486 107. Cheng F, Liang J, Cai C, Cai X, Wu J, Wang X. Genome sequencing supports a multi-vertex model for
1487 Brassiceae species. Curr Opi Plant Biol [Internet]. Elsevier Ltd; 2017;36:79–87. Available from:
1488 http://dx.doi.org/10.1016/j.pbi.2017.01.006
1489 108. Tasdighian S, Van Bel M, Li Z, Van de Peer Y, Carretero-Paulet L, Maere S. Reciprocally Retained
1490 Genes in the Angiosperm Lineage Show the Hallmarks of Dosage Balance Sensitivity. Plant Cell
1491 [Internet]. American Society of Plant Biologists; 2017 [cited 2017 Nov 25];tpc.00313.2017. Available
1492 from: http://www.ncbi.nlm.nih.gov/pubmed/29061868
1493 109. Renny-Byfield S, Rodgers-Melnick E, Ross-Ibarra J. Gene Fractionation and Function in the Ancient
1494 Subgenomes of Maize. Mol Biol Evol [Internet]. Oxford University Press; 2017 [cited 2017 Nov
1495 25];34:1825–32. Available from: http://academic.oup.com/mbe/article/34/8/1825/3742530/Gene-

1496 Fractionation-and-Function-in-the-Ancient
1497 110. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes.
1498 Genome Res [Internet]. 2003 [cited 2012 Mar 2];13:2178–89. Available from:
1499 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=403725&tool=pmcentrez&rendertype=abst
1500 ract
1501 111. Dunn CW, Howison M, Zapata F. Agalma: an automated phylogenomics workflow. BMC
1502 Bioinformatics [Internet]. 2013;14:330. Available from:
1503 http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3840672&tool=pmcentrez&rendertype=ab
1504 stract
1505 112. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol [Internet]. 2007 [cited
1506 2012 Feb 29];24:1586–91. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17483113
1507 113. van Dongen S, Abreu-Goodger C. Using MCL to extract clusters from networks. Methods Mol Biol
1508 [Internet]. 2012 [cited 2016 Nov 26];804:281–95. Available from:
1509 http://www.ncbi.nlm.nih.gov/pubmed/22144159
1510 114. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq
1511 experiments with HISAT, StringTie and Ballgown. Nat Protoc. [Internet]. Nature Publishing Group;
1512 2016;11:1650–67. Available from: http://dx.doi.org/10.1038/nprot.2016.095
1513

Table 1. Summary of CLfinder results showing pairwise comparisons among 6 crucifer species

Query species	# protein-coding genes	CL type[a]	Target species						# *td* [b] events (# *td* genes)
			Aly	Ath	Cru	Esa	Sir	Spa	
Aly	32,657	*cl*		24,296	23,055	21,416	19,988	21,032	2,163 (5,733)
		tr		4,881	5,375	6,668	8,104	6,478	
		ls		2,876	3,611	3,954	3,902	4,530	
		nd		604	616	619	663	617	
Ath	27,206	*cl*	23,436		22,683	21,187	19,821	20,851	1,747 (4,770)
		tr	2,431		2,804	4,032	5,355	4,064	
		ls	1,339		1,719	1,987	2,030	2,291	
		nd	0		0	0	0	0	
Cru	26,521	*cl*	22,371	22,836		20,906	19,350	20,436	1,752 (4,996)
		tr	3,036	2,836		4,338	5,817	4,267	
		ls	950	666		1,112	1,154	1,646	
		nd	164	183		165	200	172	
Esa	26,351	*cl*	20,384	20,884	20,460		19,699	20,612	1,646 (4,461)
		tr	4,465	4,137	4,460		5,431	4,046	
		ls	1,452	1,274	1,377		1,146	1,631	
		nd	50	56	54		75	62	
Sir	32,524	*cl*	19,015	19,538	19,068	19,728		19,766	1,795 (4,586)
		tr	3,062	2,860	2,998	2,722		2,697	
		ls	5,520	5,148	5,496	5,054		4,225	
		nd	4,927	4,978	4,962	5,020		5,836	
Spa	26,847	*cl*	19,849	20,358	19,934	20,380	19,546		1,242 (3,049)
		tr	2,830	2,452	2,718	2,534	4,097		
		ls	3,649	3,541	3,688	3,432	2,526		
		nd	519	496	507	501	678		

[a] Co-linear (*cl*), transposed (*tr*), lineage-specific (*ls*), or not determined due to too small genome scaffold (*nd*)

[a] Using CLfinder parameters {window_size, num_CL_trshld, gap_CL_trshld} = { 20, 3, 20 }

[b] Tandem duplication (*td*) detected using the parameter max_TD_loci_dist = 4

Table 2. A partial list of OrthNets associated with lineage-specific (LS) transposition-duplication (*tr-d*) events [a]

LS *tr-d*	OrthNet ID (ON_)	Node count pattern [b]	Annotation	CL copy [c] ORF length (nt) / # exons	Number of Tr copies: Total (*td*), cORF, exp, cORF & exp [d]	Number of Tr copies: Duplication mode: g+i, go [e]
Aly	1227	11.1.1.1.1.1	unknown protein	630 / 5	5 (3), 4, 2, 2	5, 0
	3580	4.1.1.1.1.1	cytochrome c oxidase 15 (COX15)	1374 / 7	2, 2, 2, 2	2, 0
	4486	3.1.1.1.1.1	carbamoyl phosphate synthetase A (CARA)	1293 / 10	2, 2, 2, 2	0, 2
	3527	4.1.1.1.1.1	RNA-dependent RNA polymerase 1 (RDR1)	3324 / 3	3, 1, 1, 1	1, 0
	3557	4.1.1.1.1.1	beta-amylase 3 (BMY3)	1605 / 3	3, 1, 2, 1	1, 0
	4003	3.1.1.1.1.1	peroxin4 (PEX4)	474 / 5	1, 1, 1, 1	0, 1
	200-3	3.1.1.1.1.1	homeobox protein 5 (HB5)	942 / 3	2, 1, 2, 1	1, 0
Ath	4411	1.3.1.1.1.1	PAI3*	828 / 5	2, 2, 2, 2	1, 1
	4622	1.2.1.1.1.1	HPA1*	1254 / 7	1, 1, 1, 1	1, 0
	5135	1.2.1.1.1.1	Ribosomal protein S5 family	2964 / 10	1, 1, 1, 1	1, 0
	5699	1.2.1.1.1.1	UGT72E3*	1446 / 1	1, 1, 1, 1	1, 0
	5831	1.2.1.1.1.1	PM* choline transporter family	1449 / 3	1, 1, 1, 1	1, 0
	0-1050	1.2.1.1.1.1	zinc finger protein-related	5328 / 4	1, 1, 1, 1	1, 0
Cru	**2442**	**1.1.6.1.1.1**	**AGAMOUS-like 87 (AGL87)**	**492 / 1**	**5 (4), 5, 1, 1**	**5, 0**
	3071	1.1.5.1.1.1	RNA-binding (RRM/RBD/RNP motifs) family	546 / 4	4 (2), 3, 4, 3	3, 0
	4286	1.1.3.1.1.1	NMNAT*	639 / 8	2, 2, 2, 2	0, 2
	4551	1.1.3.1.1.1	LOB domain-containing protein 41 (LBD41)	795 / 2	2, 2, 2, 2	0, 2
	4140	1.1.3.1.1.1	myb domain protein 95 (MYB95)	852 / 3	2, 2, 1, 1	0, 2
	4421	1.1.3.1.1.1	hydroxycinnamoyl transferase (HCT)	1296 / 2	2 (2), 2, 1, 1	0, 2
	5691	1.1.2.1.1.1	SRP-related	1821 / 13	1, 1, 1, 1	0, 1
	5251	1.1.2.1.1.1	Calcium-binding EF-hand family	639 / 1	1, 1, 1, 1	1, 0
Esa	**2516**	**1.1.1.6.1.1**	**SALT-TOLERANCE 32 (SAT32)**	**1317 / 13**	**5 (3), 4, 3, 3**	**0, 4**
	4071	1.1.1.3.1.1	Cyclin A2;4 (CYCA2;4)	1401 / 10	2, 2, 2, 2	0, 2
	3945	1.1.1.3.1.1	prenylated RAB acceptor 1.E (PRA1.E)	621 / 1	2, 2, 1, 1	2, 0
	4237	1.1.1.3.1.1	RNA-DIRECTED DNA METHYLATION 1 (RDM1)	480 / 2	2, 2, 1, 1	2, 0
	4314	1.1.1.3.1.1	SPFH/Band 7/PHB domain-containing	1077 / 9	2 (2), 2, 1, 1	0, 2
	4608	1.1.1.2.1.1	MAPKKK5*	2154 / 11	1, 1, 1, 1	0, 1
	4633	1.1.1.2.1.1	KOKOPELLI (KPL)	1473 / 3	1, 1, 1, 1	0, 1
	4867	1.1.1.2.1.1	TERMINAL FLOWER 1 (TFL1)	534 / 4	1, 1, 1, 1	1, 0
	5499	1.1.1.2.1.1	xylem cysteine peptidase 1 (XCP1)	1068 / 4	1, 1, 1, 1	1, 0
	5701	1.1.1.2.1.1	NAC domain containing protein 6 (NAC6)	870 / 3	1, 1, 1, 1	1, 0
Sir	3061	1.1.1.1.5.1	cytochrome P450 (CYP78A10)	1620 / 2	4, 3, - , -	3, 0
	3400	1.1.1.1.4.1	amino acid permease 3 (AAP3)	1431 / 7	3, 2, - , -	0, 2
	3964	1.1.1.1.3.1	PALE CRESS (PAC)	603 / 5	2, 1, - , -	1, 1
	5051	1.1.1.1.2.1	GLOBULAR ARREST1 (GLA1)	1707 / 8	1, 1, - , -	0, 1
	5943	1.1.1.1.2.1	DROUGHT SENSITIVE 1 (DRS1)	1662 / 4	1, 1, - , -	0, 1
	6060	1.1.1.1.2.1	YUCCA5 (YUC5);	1281 / 3	1, 1, - , -	1, 0
Spa	3944	1.1.1.1.1.3	Endomembrane protein 70 ;	1779 / 7	2, 2, 2, 2	1, 1
	4729	1.1.1.1.1.2	UQCRX* -like family	204 / 3	1, 1, 1, 1	0, 1
	4815	1.1.1.1.1.2	zinc transporter 3 precursor (ZIP3)	1029 / 3	1, 1, 1, 1	1, 0
	5521	1.1.1.1.1.2	20S proteasome alpha subunit G1 (PAG1)	750 / 10	1, 1, 1, 1	0, 1
	6085	1.1.1.1.1.2	WRKY DNA-binding protein 72 (WRKY72)	1659 / 4	1, 1, 1, 1	0, 1
	1764-1	1.1.1.1.1.2	diacylglycerol kinase 5 (DGK5)	1482 / 13	1, 1, 1, 1	0, 1
	403-2	1.1.1.1.1.2	Alba DNA/RNA-binding protein	396 / 5	1, 1, 1, 1	0, 1

[a] A partial list of OrthNets including the most Tr copies with complete ORFs (i.e. ORF size within ±20% in proportion compare to the CL copy) and with RNA-seq expression evidence. The complete list is in Dataset S2.
[b] Number of nodes for each genome, sorted alphabetically, i.e. Number of nodes for Aly.Ath.Cru.Esa.Sir.Spa
[c] If multiple CL copies exist, mostly due to tandem duplication (TD), we chose the one with the longest ORF.
[d] Number of Total Tr copies with Tr copies tandem duplicated in parentheses, Tr copies with complete ORF (cORF), Tr copies with RNA-seq expression evidence (exp), and Tr copies with both cORF and exp (cORF & exp)
[e] g+i, the homologous genome segment (HGS) includes both gene and intergenic regions; go, gene-only *tr-d*, same as in Fig. 6D. The sum of g+i and go sets is equal to the "Complete duplication" set in Fig. 6.

* Abbreviations in annotation: PAI3, phosphoribosylanthranilate isomerase 3; HPA1, histidinol phosphate aminotransferase 1; UGT72E3, an UDPG:coniferyl alcohol glucosyltransferase; PM, plasma membrane; NMNAT, nicotinate/nicotinamide mononucleotide adenyltransferase; MAPKKK5, mitogen-activated protein kinase kinase kinase 5; UQCRX, ubiquinol-cytochrome C reductase

Table 3. List of OrthNets including *tr-d* events shared by *E. salsugineum* (Esa) and *S. parvula* (Spa)[a]

Category	OrthNet ID (ON_)	Node count pattern [b]	Annotation	Genomes	CL copy [c] ORF length (nt) / # exons	Total, cORF, exp, cORF & exp [d]
Ind-parallel	3530	1.1.1.3.1.2	Nucleotidylyl transferase superfamily	Esa	1200 / 12	2, 2, 2, 2
				Spa	1209 / 12	1, 1, 0, 0
	0-0850	1.1.1.2.1.3	Zinc-finger domain, monoamine-oxidase A repressor R1	Esa	1410 / 10	1, 1, 0, 0
				Spa	1656 / 13	1, 1, 0, 0
	3452	1.1.1.3.1.2	novel cap-binding protein (NCBP)	Esa	678 / 6	2, 0, 2, 0
				Spa	678 / 6	1, 1, 1, 1
	3602	1.1.1.2.1.3	calcium-dependent protein kinase 1 (CDPK1)	Esa	1638 / 7	1, 0, 1, 0
				Spa	1464 / 8	2, 1, 0, 0
	4248	1.1.1.2.1.2	general regulatory factor 2 (GRF2)	Esa	783 / 4	1, 0, 1, 0
				Spa	783 / 4	1, 1, 1, 1
	4216	1.1.1.2.1.2	ABA INSENSITIVE 1 (ABI1)	Esa	1320 / 4	1, 0, 1, 0
				Spa	975 / 4	1, 0, 0, 0
	4253	1.1.1.2.1.2	inositol polyphosphate 5-phosphatase 11 (5PTASE11)	Esa	933 / 6	1, 0, 1, 0
				Spa	1008 / 7	1, 0, 0, 0
	3427 [e]	1.1.1.2.2.2	embryo defective 2191 (emb2191)	Esa	654 / 5	1, 1, 1, 1
				Spa	561 / 7	1, 1, 1, 1
Tr-*cl*	170-3	1.1.1.2.1.2	basic helix-loop-helix (bHLH) DNA-binding superfamily	Esa	1092 / 7	1, 1, 1, 1
				Spa	1077 / 7	1, 1, 1, 1
	8-07	1.1.1.2.1.2	polygalacturonase 4 (PGA4)	Esa	1266 / 4	1, 1, 1, 1
				Spa	1266 / 4	1, 1, 0, 0
	3919	1.1.1.2.1.2	unknown protein	Esa	360 / 1	1, 1, 0, 0
				Spa	357 / 1	1, 1, 1, 1
	3913	1.1.1.2.1.2	SNARE-like superfamily	Esa	456 / 6	1, 1, 1, 1
				Spa	408 / 5	1, 1, 0, 0
	4390	1.1.1.2.1.2	RING/U-box superfamily	Esa	426 / 1	1, 0, 1, 0
				Spa	375 / 2	1, 1, 0, 0
	4526	1.1.1.2.1.2	RING/FYVE/PHD zinc finger superfamily	Esa	1209 / 4	1, 0, 0, 0
				Spa	1188 / 5	1, 1, 1, 1
	2430 [e]	1.1.1.2.3.3	calcineurin B-like protein 10 (CBL10)	Esa	780 / 9	1, 1, 0, 0
				Spa	747 / 9	2, 2, 2, 2
	3847 [e]	1.1.1.2.2.2	NAC domain containing protein 58 (NAC058)	Esa	921 / 3	1, 1, 1, 1
				Spa	927 / 3	1, 1, 1, 1
	3474 [e]	1.1.1.2.2.2	P-loop nucleoside triphosphate hydrolases superfamily	Esa	3465 / 17	1, 1, 1, 1
				Spa	3492 / 17	1, 1, 0, 0
	240-2 [e]	1.1.1.2.2.2	Ribosomal protein L22p/L17e family	Esa	528 / 7	1, 1, 0, 0
				Spa	528 / 7	1, 1, 0, 0
	3310 [e]	1.1.1.2.3.2	CRUMPLED LEAF (CRL)	Esa	783 / 10	1, 1, 1, 1
				Spa	810 / 9	1, 1, 1, 1
	0-0834 [e]	1.1.1.2.2.2	photosystem II reaction center W (PSBW)	Esa	402 / 2	1, 1, 1, 1
				Spa	402 / 2	1, 1, 0, 0
	1066-1 [e]	1.1.1.2.3.3	phosphorylcholine cytidylyltransferase (CCT1)	Esa	996 / 8	1, 1, 1, 1
				Spa	996 / 8	2, 1, 1, 1

[a] The complete list for all pairs of genomes is in Dataset S2.

[b] Number of nodes for each genome, sorted alphabetically, i.e. Number of nodes for Aly.Ath.Cru.Esa.Sir.Spa

[c] If multiple CL copies exist, mostly due to tandem duplication (TD), we chose the one with the longest ORF

[d] Number of Total Tr copies with Tr copies tandem duplicated in parentheses, Tr copies with complete ORF (cORF), Tr copies with RNAseq expression evidence (exp), and Tr copies with bot cORF and exp (cORF & exp)

[e] These OrthNets showed Lineage II-specific *tr-d* events, with all additional Tr copies in Sir having incomplete ORFs, essentially resulting in copy number increments specific to Esa and Spa only.

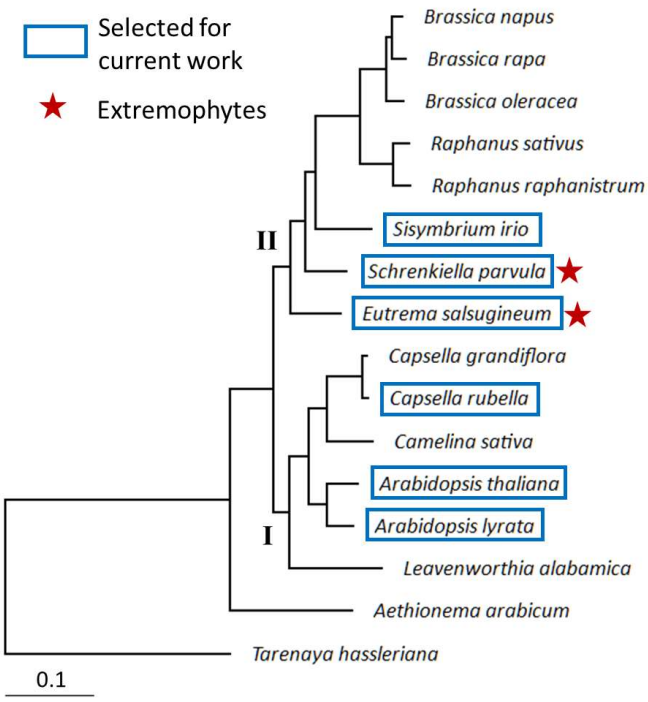
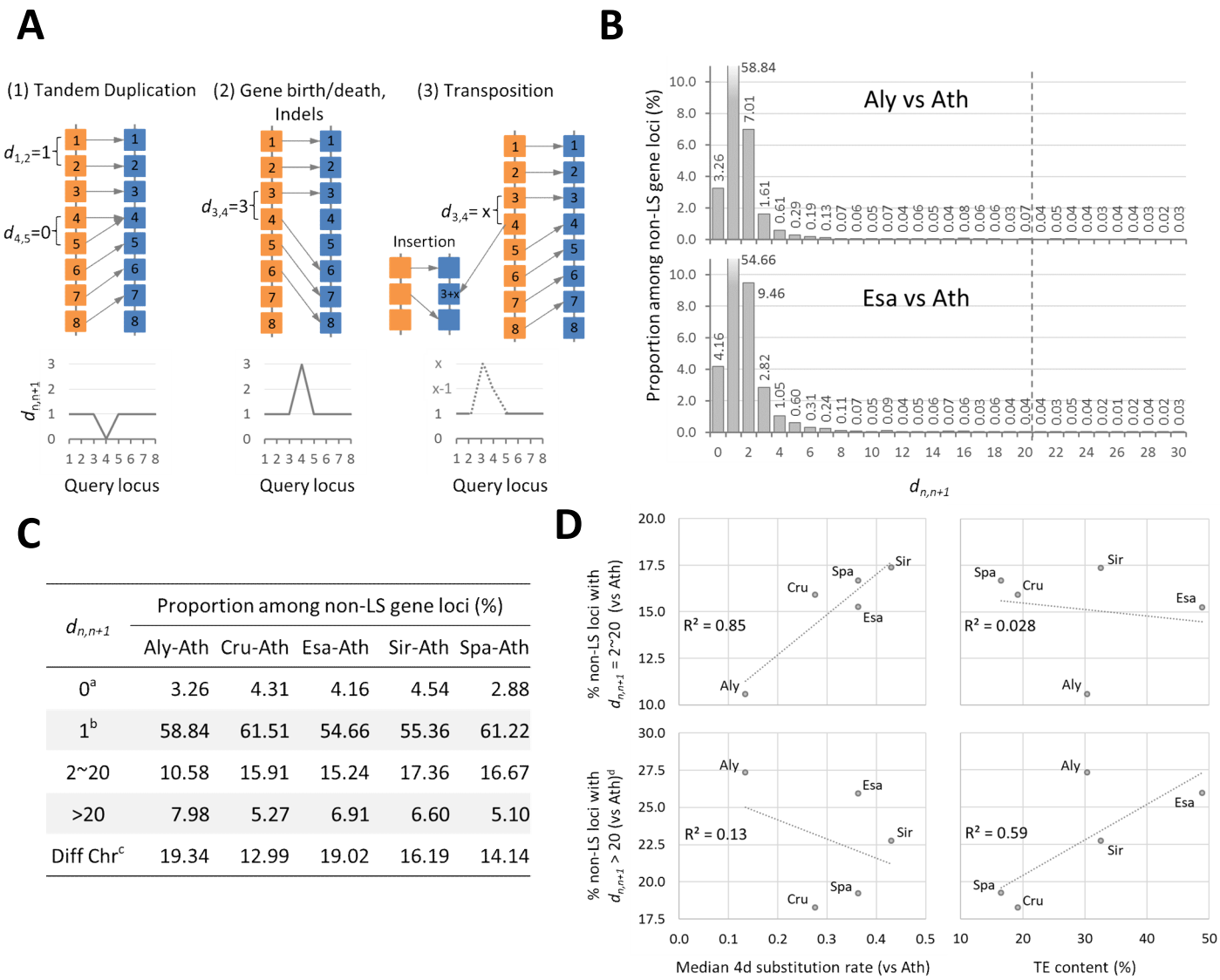


Figure 1. A comparative genomics framework including the two extremophyte/halophyte crucifers, *S. parvula* and *E. salsugineum*.

Boxes and stars indicate the six Brassicaceae species selected for this work and halophytes, respectively. The tree was based on an alignment of 14,614 homologous gene clusters, as described in Methods.



$d_{n,n+1}$	Proportion among non-LS gene loci (%)				
	Aly-Ath	Cru-Ath	Esa-Ath	Sir-Ath	Spa-Ath
0[a]	3.26	4.31	4.16	4.54	2.88
1[b]	58.84	61.51	54.66	55.36	61.22
2~20	10.58	15.91	15.24	17.36	16.67
>20	7.98	5.27	6.91	6.60	5.10
Diff Chr[c]	19.34	12.99	19.02	16.19	14.14

Figure 2. Erosion of co-linearity between Brassicaceae (crucifer family) genomes

(A) Each locus in the query genome (orange) is paired with an arrow to its most homologous counterpart locus ("best-hit") in the target genome (blue), after loci in both genomes numbered based on their positions. The difference in best-hit loci positions between the query locus at position *n* and its neighbor at position *n+1* ($d_{n,n+1}$) is 1 when the query locus and its best-hit in the target are co-linear, while a tandem duplication unique to the query genome results in $d_{n,n+1}$ = 0 (1). Co-linearity is disrupted by mutations and indels (2), as well as a gene transposition event (3).
(B, C) The distribution of $d_{n,n+1}$ with either *A. lyrata* (Aly), *C. rubella* (Cru), *E. salsugineum* (Esa), *S. irio* (Sir), and *S. parvula* (Spa) as the query and the *A. thaliana* (Ath) genome as the target. Aly-Ath and Esa-Ath pairs were shown as exemplary histograms in B. Results for all pairs were summarized in C.
(D) The proportion of gene loci with $d_{n,n+1}$ ranging from 2 to 20 showed a higher correlation with the median four-degenerated (4d) substitution rate (upper panels). Contrastingly, the proportion of those with $d_{n,n+1}$ larger than 20 or their best-hits found in a different target chromosome was better correlated with the overall transposable element (TE) contents (lower panels).
[a] The query locus involved in a tandem duplication event unique to the query genome.
[b] The query locus co-linear with its best-hit without any modification in gene order.
[c] The best-hit of the query locus *n* is located in a different chromosome ("Diff Chr") in Ath, compared to the best-hit of its immediate neighbor locus *n+1*.
[d] Includes loci with $d_{n,n+1}$ > 20 as well as the "Diff Chr" indicated in (*C*)

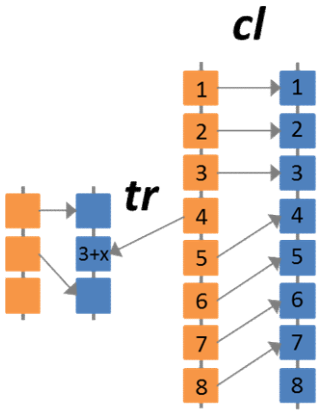
Input

1. Gene models for multiple closely related genomes (.gtf)
2. Paralog clusters within each genome (e.g. OrthoMCL)
3. Homolog pairs ("best-hits") between genomes (e.g. BLASTN)

CLfinder

For all query-target genome pairs, the CLfinder module:

- assigns co-linearity relationship between each locus in the query genome and its best-hit in the target genome;
- searches an window adjacent to the query locus for co-linear "loci-in-chains" [a], to determine whether the query and its best-hit are co-linear (*cl*) or transposed (*tr*); identifies also tandem duplicated (*td*) and lineage-specific (*ls*) loci.



OrthNet

- combines CLfinder results into Ortholog Networks (OrthNets) connecting gene loci based on their co-linearity relationships;
- clusters the OrthNets using Markov Clustering (MCL);
- codifies OrthNets based on gene copy numbers in each genome and edge attributes, enabling retrieval of all OrthNets matching user-inquired evolutionary context patterns [b].

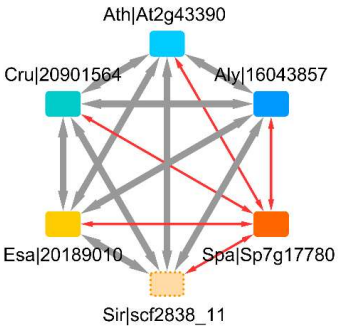
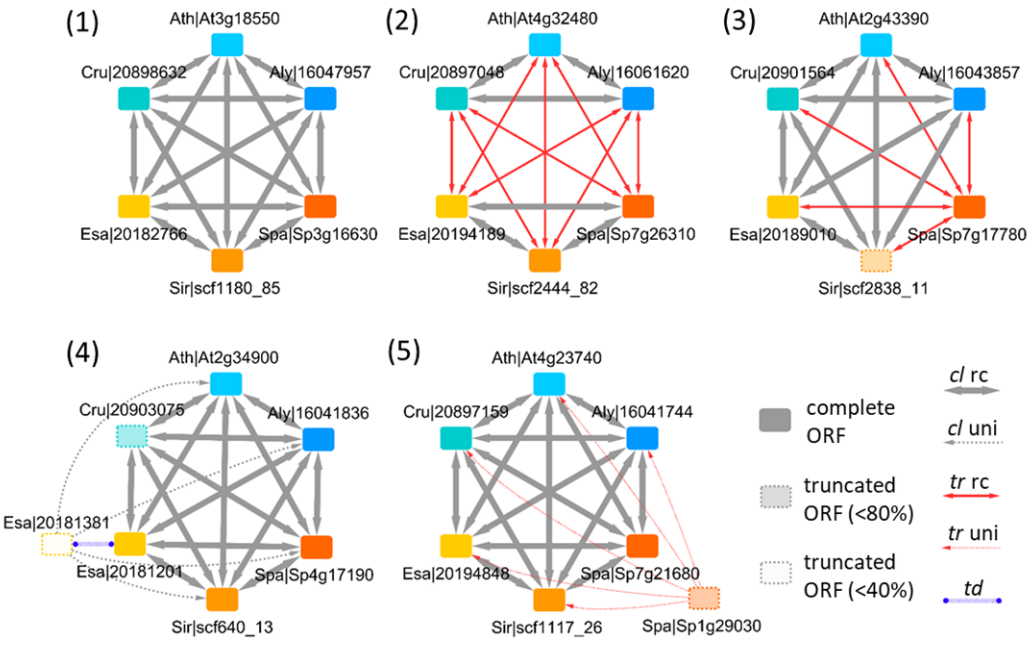


Figure 3. The CLfinder-OrthNet pipeline

[a] For the detailed method to determine co-linearity (CL) relationship between the query loci and their most homologous counterpart ("best-hits") in the target genome, see Figures S2 and S3. An example CLfinder output for the six crucifer species highlighted in Fig. 1 are summarized in Table 1, with the full results given as Supplementary Dataset S1.

[b] See Figures 4-6, for examples of OrthNets with different evolutionary contexts, e.g. transposition (*tr*) and transposition-duplication (*tr-d*) unique to each species or a group of species.

A



Evolutionary context	Number of OrthNets	Example OrthNet ID (Annotation)
(1) All co-linear, single copy	7,034	On_6361 (*BRANCHED 1*)
(2) Transposition (*tr*) between Lineage I and II	50	On_8867 (unknown function)
(3) Spa-specific transposition	9	On_12904 (unknown function)
(4) Esa-specific tandem duplication (*td*)	44	On_4974 (*IMBIBITION-INDUCIBLE 1*)
(5) Spa-specific transposition-duplication (*tr-d*)	86	On_4629 (LRR kinase family)

B

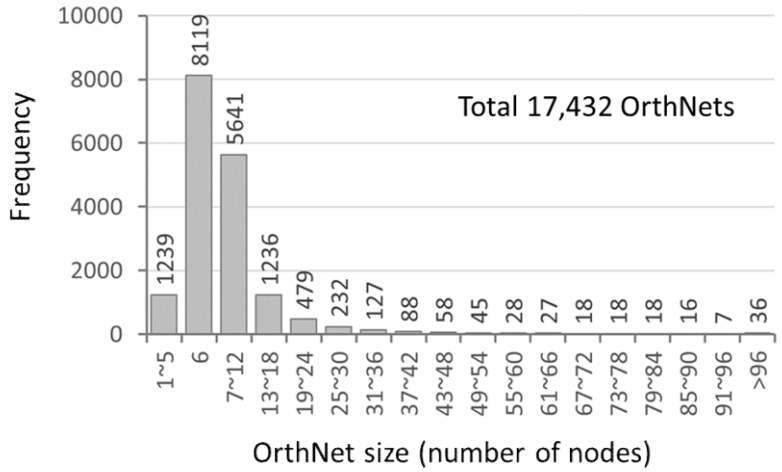


Figure 4. The OrthNet module encodes the evolutionary context of an orthologous gene group into a network.

(A) OrthNet examples representing five different evolutionary contexts. Nodes are color-coded according to the species. Transparent nodes with dashed borders indicate loci with truncated ORFs, i.e. ORF sizes smaller than either 80% or 50% compared to the median ORF size of nodes they are connected to. Edges show properties either co-linear (*cl*) or transposed (*tr*), reciprocally (rc) or uni-directionally (uni). Tandem duplicated (*td*) paralogs are also connected by an edge. The lower panel includes a table with the ID and annotation for representative OrthNets, as well as the number of OrthNets with the same evolutionary context among total 17,432 identified for the six genomes.
(B) A histogram showing the size distribution of OrthNets.

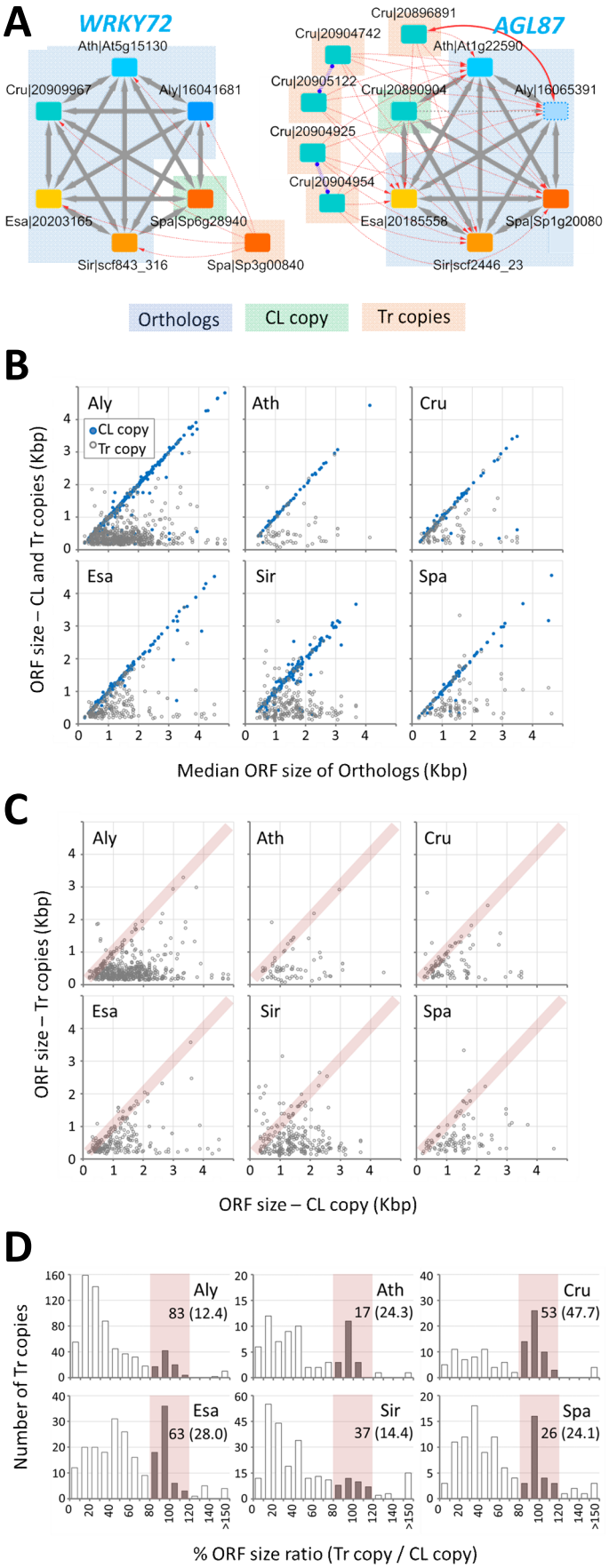
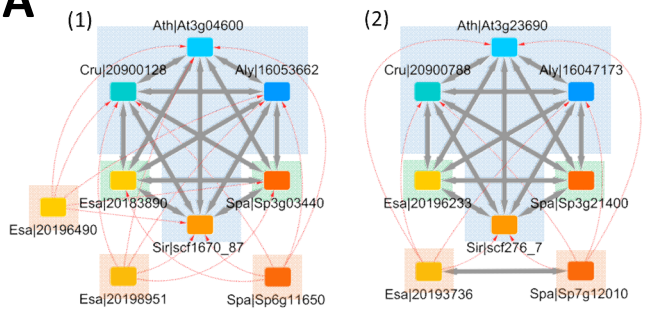


Figure 5. Characterization of lineage-specific transposition-duplication (*tr-d*) events among the six Brassicaceae genomes

(A) Examples of OrthNets with *tr-d* events specific for *S. parvula* (*WRKY DNA-BINDING PROTEIN 72*/*WRKY72*) and *C. rubella* (*AGAMOUS-LIKE 87*/*AGL87*). Within a *tr-d* event, OrthNet identifies the original donor copy (CL copy), which is co-linear to orthologs in other genomes (Orthologs), and the transposed and duplicated copies (Tr copies). Node and edges are as described in Fig. 4A.
(B) ORF size comparison between all loci involved in a *tr-d* event (i.e. both CL and Tr copies) with the median of Orthologs. Blue dots indicate CL copies.
(C and D) ORF size comparison between Tr copies and their corresponding CL copy within each of the *tr-d* events, as a scatterplot (C) and a histogram of ORF size ratio (D). Pink shades indicate Tr copies with complete ORFs whose sizes are comparable (±20% in proportion) to that of the CL copy. Panel D also shows numbers and percentages (in parentheses) of Tr copies with complete ORFs below the species labels. The entire list of OrthNets showing lineage-specific *tr-d*, including CL and Tr copies, is in Supplementary Dataset S2.

A



Orthologs | CL copy | Tr copies

	Evolutionary context	Number of OrthNets	Example OrthNet ID (Annotation)
(1)	Esa-Spa, independent parallel *tr-d* (ind-parallel)	7	On_3530 (Nucleotidylyl transferase)
(2)	Esa-Spa, *tr-d* with Tr copies co-linear (Tr-*cl*)	6	On_170-3 (bHLH transcription factor)

B

Number of OrthNets

	Aly	Ath	Cru	Esa	Sir	Spa
Aly		17	10	17	47	8
Ath	17		0	3	8	3
Cru	10	0		3	9	4
Esa	17	3	3		23	7
Sir	47	8	9	23		29
Spa	8	3	4	7	29	

Ind-parallel

	Aly	Ath	Cru	Esa	Sir	Spa
Aly		24	7	1	1	1
Ath	24		1	0	0	0
Cru	7	1		0	1	0
Esa	1	0	0		9	6
Sir	1	0	1	9		22
Spa	1	0	0	6	22	

Tr-*cl*

C

Proportion of Tr copies with complete ORFs

10 20 30 40 50 60 70 80 90 100 (%)

1 \ 2	Aly	Ath	Cru	Esa	Sir	Spa
Aly		6/29	4/15	2/27	8/84	3/13
Ath	6/22		-	1/3	2/10	1/3
Cru	5/13	-		1/5	4/11	2/4
Esa	8/19	0/3	2/3		5/27	3/9
Sir	11/73	1/10	4/11	5/34		6/61
Spa	2/8	0/5	3/17	5/8	13/38	

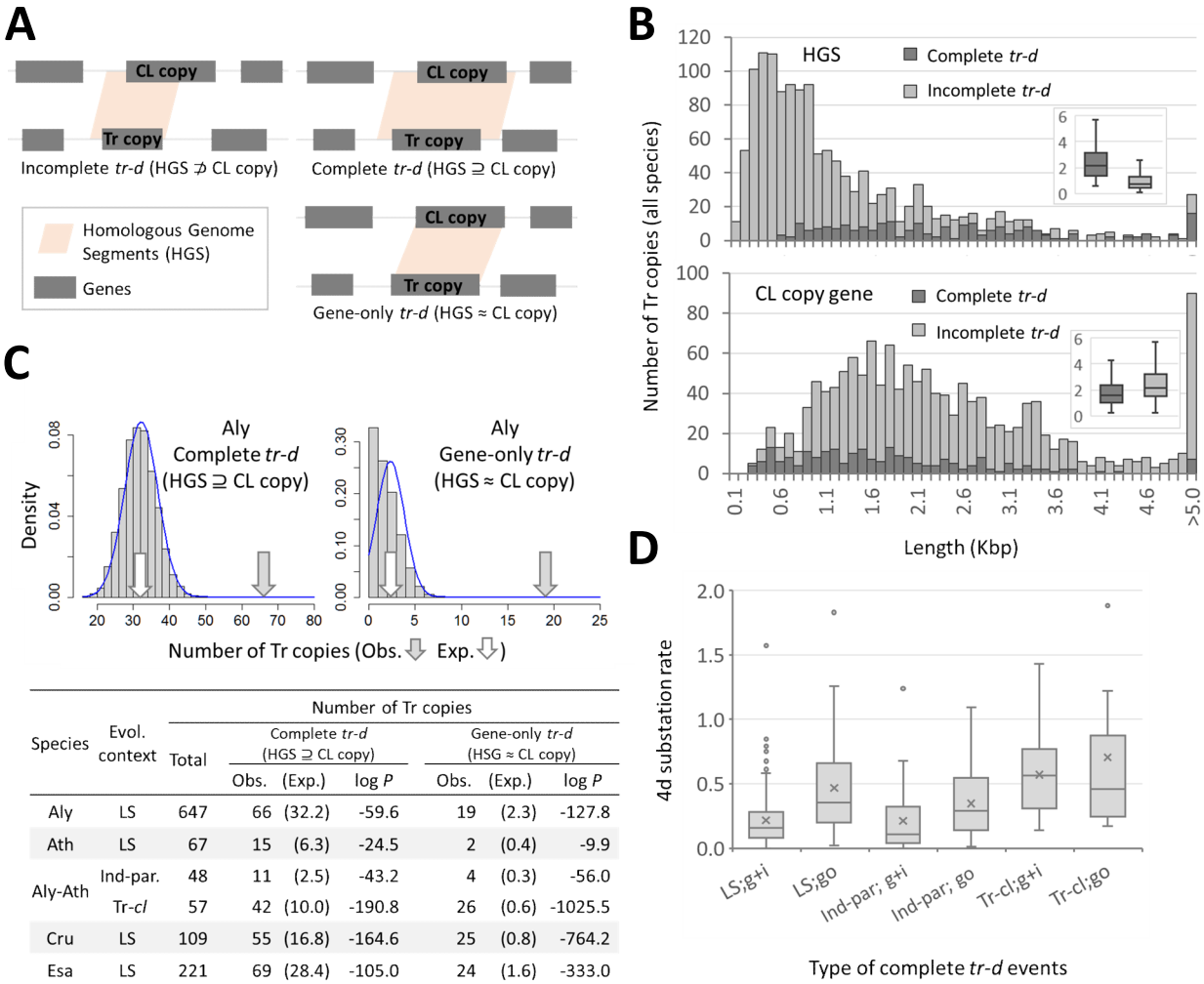
Ind-parallel

1 \ 2	Aly	Ath	Cru	Esa	Sir	Spa
Aly		22/32	8/10	0/1	1/1	1/1
Ath	22/25		1/1	-	-	-
Cru	8/8	1/1		-	0/1	-
Esa	1/1	-	-		7/9	4/6
Sir	1/1	-	0/2	7/9		13/26
Spa	0/1	-	-	6/6	14/28	

Tr-*cl*

Figure 6. Transposition-duplication (*tr-d*) events shared by a pair of Brassicaceae genomes

(A) Example OrthNets with *tr-d* events shared by *E. salsugineum* and *S. parvula*, representing two categories: (1) independent-parallel ("Ind-parallel") *tr-d* events and (2) *tr-d* events with Tr copies co-linear to each other ("Tr-*cl*"). Node and edges are as described in Fig. 4A.
(B) Number of OrthNets shared by pairs of genomes in "Ind-parallel" and "Tr-*cl*" categories, with the cladogram of the six crucifer genomes on the top. Heatmap colors visualize the rank of each cell based on OrthNet numbers in each category.
(C) Proportion of Tr copies with complete ORFs (i.e. ORF size ±20% of the CL copy in proportion) within OrthNets presented in (B). The genome 1 (row)-genome 2 (column) position shows the number of Tr copies with complete ORF / total Tr copies in genome 1, for all OrthNets with *tr-d* shared by genome 1 and 2. For example, the Aly-Ath and Ath-Aly positions in "Ind-parallel" category indicate 6 out of 29 *A. lyrata* (Aly) Tr copies and 6 out of 22 *A. thaliana* (Ath) Tr copies, respectively, have complete ORFs in the 17 "Ind-parallel" *tr-d* events shared by *A. lyrata* and *A. thaliana*. Heatmap colors indicate the percentage of Tr copies with complete ORFs, for cells with >5 total Tr copies.



Species	Evol. context	Number of Tr copies						
		Total	Complete *tr-d* (HGS ⊇ CL copy)			Gene-only *tr-d* (HSG ≈ CL copy)		
			Obs.	(Exp.)	log *P*	Obs.	(Exp.)	log *P*
Aly	LS	647	66	(32.2)	-59.6	19	(2.3)	-127.8
Ath	LS	67	15	(6.3)	-24.5	2	(0.4)	-9.9
Aly-Ath	Ind-par.	48	11	(2.5)	-43.2	4	(0.3)	-56.0
	Tr-*cl*	57	42	(10.0)	-190.8	26	(0.6)	-1025.5
Cru	LS	109	55	(16.8)	-164.6	25	(0.8)	-764.2
Esa	LS	221	69	(28.4)	-105.0	24	(1.6)	-333.0
Sir	LS	241	39	(12.6)	-86.9	16	(1.1)	-206.1
Spa	LS	105	24	(9.5)	-153.2	11	(0.6)	-392.6

E

tr-d type	Tr copy ORF type	# Tr copies	# Tr copies in RNA-seq	# Tr copies (%) expressed
Incomplete	Truncated	1,571	1,206	479 (39.7)
Complete	Truncated	135	103	67 (65.0)
Complete	Complete	368	315	231 (73.3)

Figure 7. Characterization of duplicated genomic regions in transposition-duplication (*tr-d*) events.

Figure legends in the next page.

Figure 7. Characterization of duplicated genomic regions in transposition-duplication (*tr-d*) events.

(A) We Identified Homologous Genome Segments (HGSs) between the CL and Tr copy genes and adjacent genomic regions (±5Kbp) in a *tr-d* event as described in Methods. A *tr-d* event is either complete or incomplete, based on whether the HGS included the full CL copy gene or not. A subset of complete *tr-d* events had HGS coinciding with the start and end of the CL copy gene without extending to the intergenic regions ("gene-only" *tr-d*).
(B) Histograms and box-and-whisker plots (inlets) showing size distributions of HGSs and CL copy genes for complete and incomplete *tr-d* events.
(C) Comparison of observed occurrences (Obs.) and expected occurrences (Exp.) of complete and gene-only *tr-d* events. Upper panel shows the distribution of expected occurrences from 10,000 random shuffling of HGSs and CL copy genes for *A. lyrata* (Aly)-specific complete and gene-only *tr-d* events. Fitting the random shuffling results to normal distributions (upper panel, blue curves) generated P-values of observed occurrences for *tr-d* events unique to each genome and the genus Arabidopsis (lower panel).
(D) Four degenerate site (4d) substitution rates between ORFs of CL and Tr copy genes in different types of complete *tr-d* events. Complete *tr-d* events were either gene-only ("go") or with HGSs detected in both gene and intergenic regions ("g+I"). We compared lineage-specific (LS) and shared *tr-d* events that are either independent-parallel ("Ind-par") or with Tr copies co-linear to each other ("Tr-*cl*"). Lines and "x" marks in the box indicate medians and means, respectively.
(*E*) Proportion of Tr copy genes with expression evidences (RNA-seq FPKM>0) in all *tr-d* events either lineage-specific or shared by a pair of genomes. The *tr-d* type is as described in (A) and Tr copy ORF type is as in Fig. 5C and D (pink shade) and Fig. 7C. *S. irio* genes were excluded due to the lack of RNA-seq data.

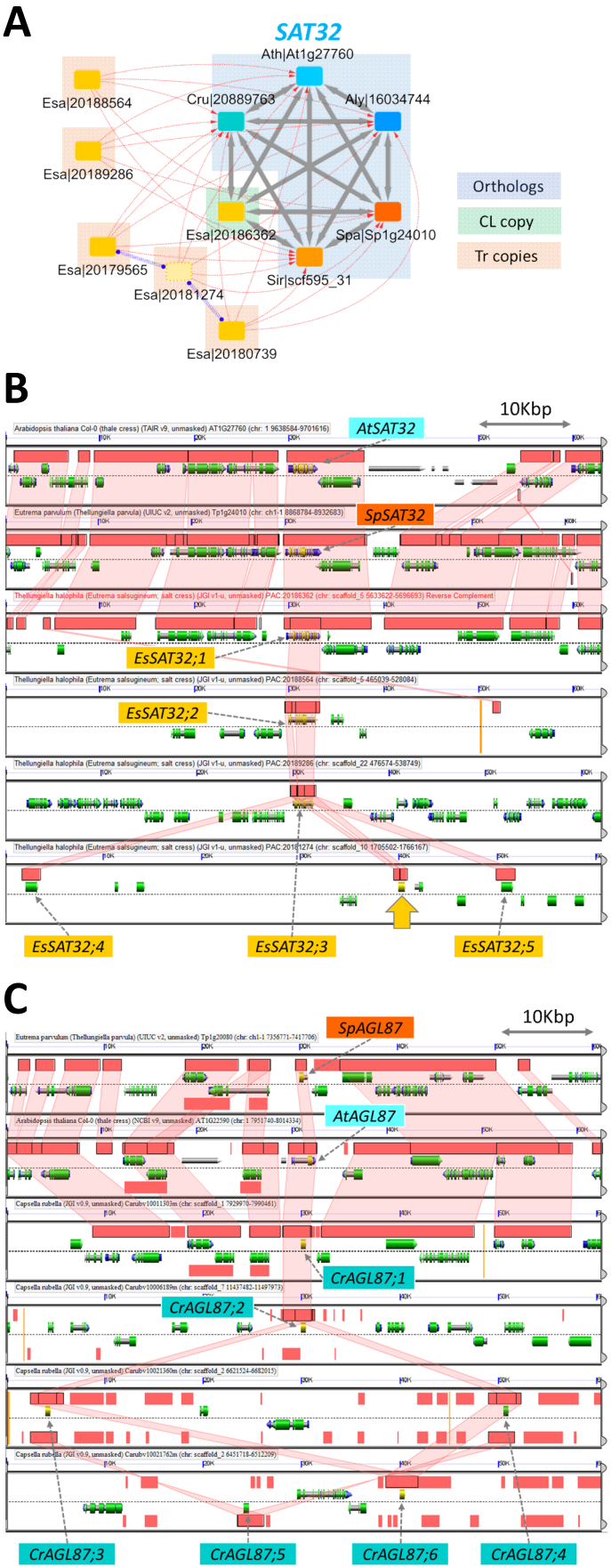


Figure 8. Examples of lineage-specific *tr-d* events.

(A) OrthNet for the *SALT-TOLERANCE 32* (*SAT32*). Nodes and edges are as described in Fig. 4A. The OrthNet showed *E. salsugineum*-specific *tr-d*, with three of the five Tr copies tandem duplicated.
(B) Comparison of *SAT32* loci and adjacent ±30Kbp genomic regions between *A. thaliana*, *S. parvula*, and *E. salsugineum* as a GEvo plot (https://genomevolution.org/r/maxx). Pink shades connect Homologous Genomic Segments (HGSs) between genomes, while gene models, mRNAs, and coding sequences are depicted in gray, blue, and green, respectively (see https://genomevolution.org/wiki/index.php/GEvo for detailed GEvo legends). *EsSAT32;1* is the CL copy (Esa|20186362), while *EsSAT32;2/3/4/5* indicate the four Tr copies with complete ORFs. The yellow arrow marks the position of Esa|20181274, which contains a truncated ORF. *EsSAT32;3/4/5* showed intron losses without compromising gene products (see text and Figure S6 for details).
(C) GEvo plot visualizing *AGAMOUS-LIKE 87* (*AGL87*) loci of *S. parvula*, *E. salsugineum*, *A. thaliana*, and *C. rubella* (https://genomevolution.org/r/olzb). The OrthNet for *AGL87* in the six crucifer species is in Fig. 5A. *CrAGL87;1* is the CL copy (Cru|20890904) and *CrAGL87;2/3/4/5/6* represent Tr copies. The genomic regions including tandem duplicates *CrAGL87;3/4* and *CrAGL87;5/6* are located only 133Kbp apart from each other in the same genome scaffold and contained a large number of short HGSs between them. Only HGSs including the *CrAGL87* loci are connected with pink shades in these genome regions.

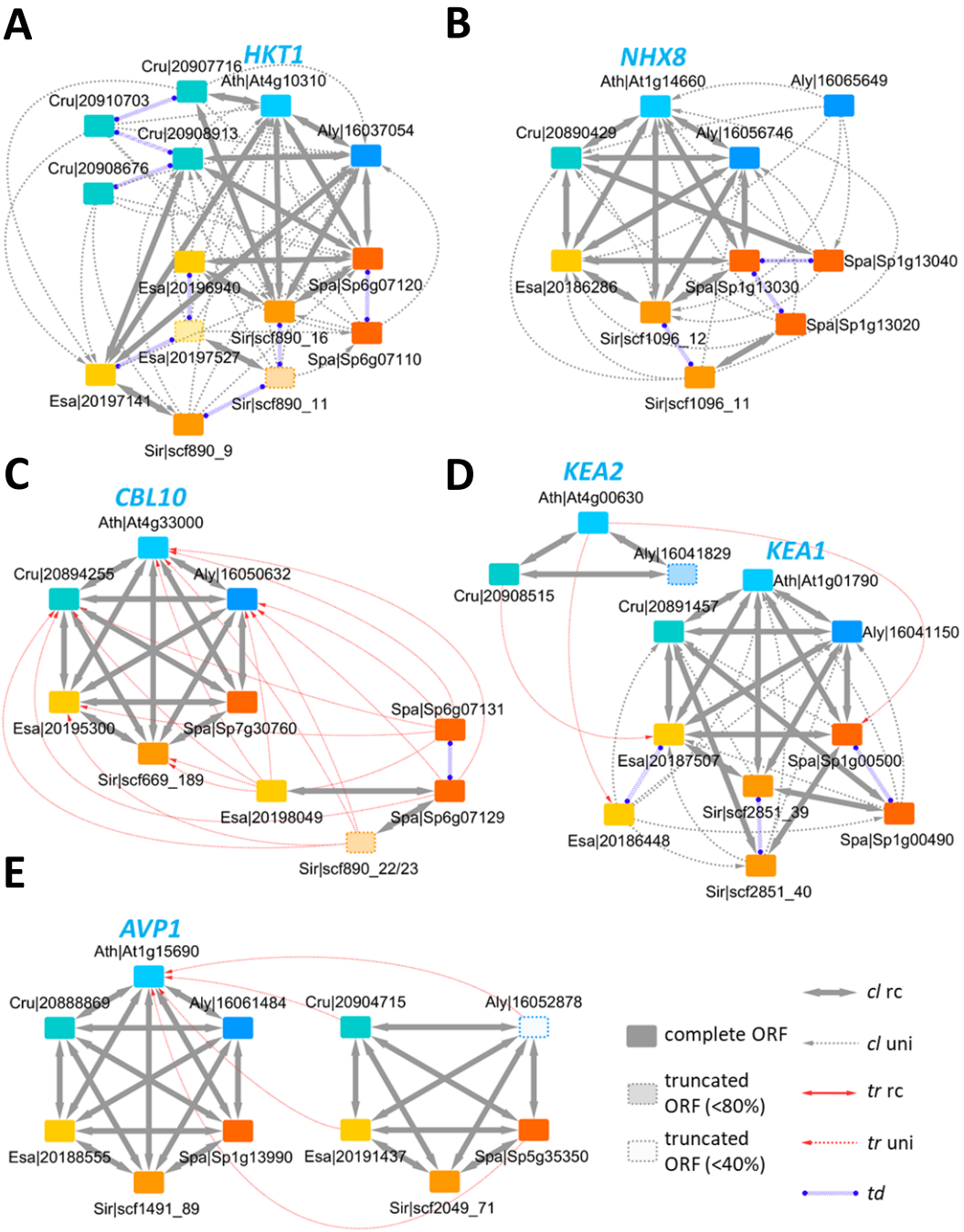


Figure 9. OrthNets associated with stress-related gene loci known for copy number variations between *S. parvula* and *A. thaliana*.

OrthNets for **(A)** *HIGH AFFINITY K⁺ TRANSPORTER 1* (*HKT1*; ON_1313), **(B)** *Na⁺ H⁺ EXCHANGER 8* (*NHX8*; ON_1214-1), **(C)** *CALCINEURIN B-LIKE PROTEIN 10* (*CBL10*; ON_2430), **(D)** *K⁺ EFFLUX ANTIPORTER 1* (*KEA 1*; ON_0-0500), and **(E)** *ARABIDOPSIS THALIANA V-PPASE 1* (*AVP1*; ON_0-0398, part) are presented, with nodes and edges as described in Fig. 4A. (E) shows a part of the OrthNet including orthologs derived from the *AVP1* loci, while the complete OrthNet is in Fig. S7 (see the Discussion section for detail).