

1 **A web-based platform of nucleotide sequence alignments of plants**

2

3

4 Chiara Santos^{1,¶}, João Carneiro^{1,¶} and Filipe Pereira^{2*}

5

6

7 ¹ Interdisciplinary Centre of Marine and Environmental Research (CIIMAR), University of Porto,

8 Terminal de Cruzeiros do Porto de Leixões; Avenida General Norton de Matos, S/N 4450-208

9 Matosinhos – Portugal

10 ² University of Porto, Praça Gomes Teixeira 4099-002 Porto, Portugal

11 [¶] These authors contributed equally to this work.

12

13

14

15 *Corresponding author:

16 F. Pereira; E-mail fpereirapt@gmail.com

17 ORCID Filipe Pereira 0000-0001-8950-1036

18 **Abstract**

19 In recent years, a large number of nucleotide sequences have become available for plant
20 species by the advent of massive parallel sequencing. The use of genomic data has been
21 important for agriculture, food science, medicine or ecology. Despite the increasing amount of
22 data, nucleotide sequences are usually available in public databases as isolated records with
23 some descriptive information. Researchers interested in studying a wide range of specific
24 plant families are forced to do multiple searches, sequence downloads, data curation and
25 sequence alignments. In order to help researches overcoming these problems, we have built a
26 comprehensive on-line resource of curated nucleotide sequence alignments for plant research,
27 named PlantAligDB (available at <http://plantaligdb.portugene.com>). The latest release
28 incorporates 514 alignments with a total of 66,052 sequences from six important genomic
29 regions: *atpF-atpH*, *psbA-trnH*, *trnL*, *rbcL*, *matK* and ITS. The alignments represent 223
30 plant families from a variety of taxonomic groups. The users can quickly search the database,
31 download and visualize the curated alignments and phylogenetic trees using dynamic
32 browser-based applications. Different measures of genetic diversity are also available for each
33 plant family. We also provide the workflow script that allows the user to do the curation
34 process, explaining the steps involved. Overall, the PlantAligDB provides a complete, quality
35 checked and regularly updated collection of alignments that can be used in taxonomic, DNA
36 barcoding, molecular genetics, phylogenetic and evolutionary studies.

37

38 **Keywords**

39 DNA sequences, Multiple sequence alignments, Plant families

40 **Introduction**

41 The recent development of high-throughput sequencing technologies has significantly
42 increased the number of nucleotide sequences available in public databases (Egan *et al.* 2012;
43 Feuillet *et al.* 2011). Complete genome sequences are now accessible in public databases
44 (e.g., EnsemblPlants) (<https://plants.ensembl.org/index.html>) for the analysis and visualisation
45 of genomic data for an ever-growing number of plants, such as *Beta vulgaris*, *Prunus persica*
46 and *Citrus sinensis*, among many others. Sequences from individual genes or gene regions
47 have also been deposit in public databases as a result of international initiatives. For instance,
48 the DNA barcoding project has released thousands of sequences aiming at species
49 identification and taxonomic classification of plants, mostly from the chloroplast DNA
50 (cpDNA) protein-coding genes *rbcL* and *matK* [CBOL Plant Working Group' - (Group *et al.*
51 2009; Hollingsworth *et al.* 2011)]. The plastid *trnL* (UAA) intron is another good example of
52 a cpDNA region highly represented in sequence databanks (Taberlet *et al.* 2007).

53 Several web-based databases are available for plant genome sequences, usually
54 dedicated to a single species or a genomic feature [e.g., (Lai *et al.* 2012; Meyer *et al.* 2005;
55 Numa & Itoh 2014; Sakai *et al.* 2013)]. However, most nucleotide sequences are accessible in
56 public databases as isolated records with simple descriptive information (taxonomy,
57 geography, publications, etc.). For instance, the NCBI Entrez Nucleotide database
58 (<http://www.ncbi.nlm.nih.gov>) and the BOLD - The Barcode of Life Data System
59 (www.barcodinglife.org) (Ratnasingham & Hebert 2007) are useful repositories with
60 descriptive information for sequence or species. The TreeBASE (<https://www.treebase.org>) is
61 a repository of phylogenetic information with user-submitted phylogenetic trees and the data
62 used to generate them. Nevertheless, researchers interested in studying a large number of
63 plant families are forced to do multiple searches and sequence downloads of genetic data for
64 their investigations. Moreover, the available sequences are not aligned and curated. The

65 multiple sequence alignment step is critical because it determines the accuracy of the
66 subsequence analyses, such as phylogenetic inference, identification of conserved motifs,
67 function prediction, etc. Building accurate sequence alignments involves many steps,
68 including sequence file conversion, run of alignment algorithms in local computers or
69 webservers, selection of best alignment parameters, and manual fine-tuning of the alignment.
70 This process is laborious and requires costly computational resources, which are not always
71 available.

72 We describe an on-line database (PlantAligDB, available at
73 <http://plantaligdb.portugene.com>) with a comprehensive, automatically curated and regularly
74 updated collection of alignments from diverse plant families (Figure 1), and respective
75 phylogenetic inference of each alignment. The PlantAligDB is a consistent repository of
76 curated alignments and phylogenetic trees that are generated by the same workflow. For
77 example, the Orchidaceae family is frequently referred to as a critical group, whose species
78 are difficult to identify. Because of their ecological importance, ongoing studies continue to
79 be made to reduce their extinction risk and maintain their diversity (Li *et al.* 2018). The
80 PlantAligDB can help researchers designing accurate methods for plant identification (*matK*
81 and *rbcL* are used in DNA barcoding projects) whether by identification of conserved motives
82 that enable the design of primers. The PlantAligDB alignments can be used as a reference
83 database for phylogenetic studies, allowing the construction of reference phylogenetic trees of
84 the different regions [e.g., genomic regions *atpF-atpH* (Domenech *et al.* 2014), *psbA-trnH*
85 (Dong *et al.* 2012) and ITS (Karehed *et al.* 2008)]. Moreover, it provides useful data to
86 understand the genetic diversity of the selected genomic regions.

87

88 **Materials and Methods**

89

90 **Data curation**

91 We retrieved all nucleotide sequences of different genomic regions from the NCBI
92 Entrez Nucleotide database (<http://www.ncbi.nlm.nih.gov>) using the Geneious software
93 (Drummond 2009). Different combinations of search terms (e.g. ‘*gene name*’; ‘viridiplantae’;
94 ‘chloroplast’; ‘gene’; ‘complete’) and a maximum limit of 5000 bp as sequence length were
95 used in searches to retrieve the largest number of sequences. The selection was done by
96 eliminating sequences showing one or more of the following features: a) ambiguous name
97 description; b) sequences with high number of nucleotide ambiguities (>50%); c) sequences
98 with large stretches of the region missing (>50%). After a preliminary curation of the data, we
99 selected five cpDNA regions and one nuclear DNA region, which were the most represented
100 in the NCBI database, commonly used in phylogenetic studies for being relevant and
101 informative. The six genomic regions were named according to the gene regions where they
102 are located: *atpF-atpH* (*ATPase I subunit – ATPase III subunit*), *psbA-trnH* [*Photosystem II*
103 *32 kDa protein – tRNA-His (GUG)*], *trnL* [*tRNA-Leu (UAA)*], *rbcL* (*rubisco large subunit*),
104 *matK* (*maturase K*) and ITS (internal transcribed spacer). We then removed from the datasets
105 all redundant sequences belonging to the same species and sequences without a clear species
106 assignment. We also reverse complement the sequences that were found in the opposite
107 direction. The sequence orientation for each region is that of the most commonly found in the
108 NCBI database. Therefore, the orientation of the *trnL (UAA)*, *atpF-atpH* and *rbcL* regions are
109 the same of that used in the reference cpDNA sequence of *Nicotiana tabacum*
110 (NC_001879.2), while the opposite orientation is used for regions *psbA-trnH* and *matK*. The
111 target region named ITS in our database includes the internal transcribed spacer 1, 5.8S rRNA
112 and internal transcribed spacer 2 section of the nuclear ribosomal DNA.

113 Because a high number of sequences were detected for the *trnL (UAA)* region (more
114 than 50,000 hits), we used the external regions named “C” and “D” and the internal regions

115 named “G” and “H” by (Taberlet *et al.* 2007) as queries in the NCBI Basic Local Alignment
116 Search Tool (BLAST; <http://blast.ncbi.nlm.nih.gov/>). The search was made against the
117 nucleotide collection (nr/nt) of Tracheophyta (vascular plants) using the Biopython package
118 (www.biopython.org) with an expected threshold of 1000 and a minimum word size of 16.
119 Therefore, our database includes two datasets for the *trnL* (UAA) genomic region: the ‘*trnL*
120 CD’ target region with a length of 577 bp in *N. tabacum*, and the ‘*trnL* GH’ with a length of
121 78 bp in *N. tabacum*, located inside the *trnL* CD region. All information regarding the selected
122 target regions can be found in the *Genomic Regions* section of the PlantAligDB.

123 The nucleotide sequences of the six regions were organized by family according to the
124 NCBI taxonomy and were aligned (each region and family in separated alignments) using the
125 default parameters of the MUSCLE software (Edgar 2004) running in the Geneious software.
126 The alignment was repeated in some families after excluding sequences that do not cover the
127 entire region of interest and that had large stretches of nucleotide ambiguities. We only used
128 alignments with ten or more species per family to build the PlantAligDB. Some species
129 sequences were lost in this process filter. The neighbor joining phylogenetic tree of each
130 region-family were calculated using Tamura-Nei model using Geneious Tree Builder. The
131 methodology was built in Armadillo Workflow (<http://www.bioinfo.uqam.ca/armadillo/>) to
132 automate the update process of the database. The latest release update of June 2018
133 incorporates 514 alignments and phylogenetic trees, from 223 plant families.

134

135 Conservation measures

136 The database includes two measures of sequence conservation for each alignment:
137 *percentage of identical sites* (PIS), calculated by dividing the number of identical positions in
138 the alignment for an oligonucleotide by its length and the *percentage of pairwise identity*
139 (PPI), calculated by counting the average number of pairwise matches across the positions of

140 the alignment, divided by the total number of pairwise comparisons.

141

142 **Results and Discussion**

143

144 **Database organization**

145

146 **Basic structure**

147 The PlantAligDB database is divided in nine sections (Figure 1): 1) *Home*, provides a
148 brief description of what can be done in the database; 2) *Genomic regions*, describes the
149 regions used in the database; 3) *Taxonomic groups*, the table containing the plant
150 family/region alignments and phylogenetic trees; 4) *BLAST*, search the database using a query
151 sequence by means of the BLASTN algorithm (Altschul *et al.* 1990); 5) *Genetic diversity*,
152 describes the percentage of identical sites (PIS) and pairwise identity values (PPI) for each
153 alignment; 6) *Download*, provides hyperlinks to download the curated alignments; 7)
154 *Tutorials*, contains information about how the database was built, and how to use it; 8)
155 *Citations*; 9) *Contacts*.

156

157 **Taxonomic groups**

158 The database is being regularly updated by our team and currently includes 514
159 alignments and phylogenetic trees from seven target regions: *atpF-atpH*, *psbA-trnH*, *trnL* CD,
160 *trnL* GH, *rbcL*, *matK* and ITS (Table 1). Sequence alignments are provided for 223 different
161 plant families. Currently, the *trnL* GH region has the largest number of sequences ($n =$
162 34,674). The *Fabaceae* family has the largest number of aligned species in a target region,
163 with 2599 sequences for the *trnL* GH region. When considering all regions together, the
164 *Fabaceae* ($n = 4714$), *Poaceae* ($n = 4494$) and *Asteraceae* ($n = 4459$) families are those with

165 the highest number of sequences. The alignments for each plant family can be accessed
166 through a dynamic table in the *Taxonomic groups* section of the database by following a
167 hyperlink with the number of species included in each alignment
168 (http://plantaligdb.portugene.com/cgi-bin/PlantAligDB_taxonomicgroups.cgi). The users are
169 able to quickly search and locate a queried feature, order each column using the ascendant or
170 descendent mode, filter the information, download the curated datasets, among other features.
171 The multiple sequence alignment and phylogenetic tree can be visualized by clicking in the
172 number of species present in the alignment.

173

174 **Genetic diversity**

175 The PIS values in our current dataset vary from 0.16% to 99.07% (Table 2).
176 Nevertheless, the PIS and PPI sequence conservation measures are not intended to be used for
177 comparison of different families and/or regions, since the number of sequences in each
178 alignment can be very different. The results must be interpreted at the light of the number of
179 sequences and the representativeness of the sequences included in the alignment file. The
180 *matK* was the region with the lowest PIS value (0.16%) [Figure 2. f)], while the *trnL* GH was
181 the region with the highest PIS value (99.07%), as can be seen in Figure 2 d) and Table 2. The
182 *rbcL* was the most conserved region [Figure 2 e)] with an average of 78.48%, while the ITS
183 was less conserved with an average of 28.84% (Table 2). Our results are in accordance with
184 earlier studies where *atpF-atpH* and *psbA-trnH* were found to be more variables than *matK*
185 (Lahaye 2008). The *trnL* CD regions showed values slightly more conserved than *atpF-atpH*
186 [Figure 2 c) and a)]. The lowest PPI value (69.96%) was found in *psbA-trnH* region [Figure 2
187 b)] and the highest was 100% in *trnL* GH, as shown in Figure 2 d) and Table 2. The ITS
188 region was less conserved with an average of 87.21% [Table 3 and Figure 2 g)], while the
189 *trnL* GH was the most conserved with an average of 97.09% [Table 3 d)].

190

191 **Sequence alignments and phylogenetic trees**

192 The alignments stored in the database can be visualized using a dynamic browser-based
193 application named Wasabi (<http://wasabiapp.org/>) (Veidenberg *et al.* 2016) with multiple
194 options for the visualization and analysis of sequence data and phylogenetic trees. This
195 resource is particularly useful to help researchers selecting the most appropriated genomic
196 regions for their investigations. The users can zoom in and out the selected regions of the
197 alignments, collapse regions with gaps, alternate between column and row selection, remove
198 or add sequences, realign sequences, and export the sequence data in the FASTA format. If an
199 Wasabi account is created, the user can re-align specific alignments with PAGAN (Loytynoja
200 *et al.* 2012) and PRANK (Loytynoja 2014). The user can merge different alignments from
201 identical regions (Hollingsworth *et al.* 2009) and different families using the PAGAN
202 application. The download of the complete database of curated alignments is accessible in the
203 *Download* section of the database.

204

205

206 **How to use the PlantAligDB**

207 To explore a genomic region, a researcher must start by accessing the ‘Genomic
208 Regions’ section in the menu bar. For example, by selecting *psbA-trnH*, the user will find a
209 brief description of the genomic region and the name, size and position of that region in the
210 reference genome, and the families with available alignments. The user should select the
211 ‘Taxonomic Groups’ section in the menu bar to search for a specific plant family. Then,
212 through the search tool on the right side of the page, the user can type the name of the family
213 and a dynamic table will be displayed with the number of sequences for each region. The user
214 can also access a hyperlink with the description of the family and taxonomic tree using the

215 resource Tree of Life Web Project (<http://tolweb.org>). Clicking on the number of sequences,
216 the database is redirected to the Wasabi tool. The user can create a Wasabi account by
217 providing an e-mail or choosing a temporary account, which allows to realign sequences with
218 PAGAN or PRANK. The user can merge a PAGAN realignment with alignments of other
219 families on the same region, by selecting a file on his local computer in the “alignment
220 extension” option.

221

222 **Availability and design**

223 The PlantAligDB is freely available at <http://plantaligdb.portugene.com> and is optimized for
224 the major web browsers (Internet Explorer, Firefox, Safari, and Chrome). The SQLite local
225 database is used for data storage and runs on an Apache web server. The dynamic HTML
226 pages were implemented using CGI-Perl and JavaScript and the dataset table views were
227 generated using the JQuery plugin DataTables v1.9.4 (<http://datatables.net/>). The PlantAligDB
228 visualization tables are generated automatically. The process of database update is optimized
229 for large datasets. There are no access restrictions for academic and commercial use.

230

231 **Acknowledgments**

232 This work was supported by the Portuguese Foundation for Science and Technology (FCT),
233 European Regional Development Fund (ERDF) and Operational Human Potential Program
234 [IF/01356/2012 to F.P. and SFRH/BPD/100912/2014 to J.C.]. J.C. also acknowledges the FCT
235 funding for his research contract at CIIMAR, established under the transitional rule of Decree
236 Law 57/2016, amended by Law 57/2017. C.S. was supported by the “Programa Ciências Sem
237 Fronteiras” from the Conselho Nacional de Desenvolvimento Científico e Tecnológico
238 (CNPq) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil.

239 CIIMAR was partially supported by the Strategic Funding UID/Multi/04423/2019 through
240 national funds provided by FCT and ERDF in the framework of the programme PT2020.

241

242 **Author Contributions**

243 CS, JC and FP contributed equally to designed and performed research, analysed data and wrote
244 the paper.

245 **References**

246

- 247 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol*
248 *Biol* **215**, 403-410.
- 249 Domenech B, Asmussen-Lange CB, Baker WJ, *et al.* (2014) A phylogenetic analysis of palm subtribe
250 Archontophoenicinae (Arecaceae) based on 14 DNA regions. *Botanical Journal of the*
251 *Linnean Society* **175**, 469-481.
- 252 Dong W, Liu J, Yu J, Wang L, Zhou S (2012) Highly variable chloroplast markers for evaluating plant
253 phylogeny at low taxonomic levels and for DNA barcoding. *PLoS One* **7**, e35071.
- 254 Drummond AB, Cheung M, Heled J, Kearse M, Moir R, Stones-Havas S, Thierer T, Wilson A (2009)
255 Geneious Pro 4.8.2.
- 256 Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput.
257 *Nucleic Acids Res* **32**, 1792-1797.
- 258 Egan AN, Schlueter J, Spooner DM (2012) Applications of next-generation sequencing in plant
259 biology. *Am J Bot* **99**, 175-185.
- 260 Feuillet C, Leach JE, Rogers J, Schnable PS, Eversole K (2011) Crop genome sequencing: lessons and
261 rationales. *Trends Plant Sci* **16**, 77-88.
- 262 Group CPW (2009) A DNA barcode for land plants. *Proc Natl Acad Sci U S A* **106**, 12794-12797.
- 263 Group CPW, Hollingsworth PM, Forrest LL, *et al.* (2009) A DNA barcode for land plants. *Proceedings*
264 *of the National Academy of Sciences* **106**, 12794-12797.
- 265 Hollingsworth ML, Andra Clark A, Forrest LL, *et al.* (2009) Selecting barcoding loci for plants:
266 evaluation of seven candidate loci with species-level sampling in three divergent groups of
267 land plants. *Mol Ecol Resour* **9**, 439-457.
- 268 Hollingsworth PM, Graham SW, Little DP (2011) Choosing and using a plant DNA barcode. *PLoS*
269 *One* **6**, e19254.
- 270 Karehed J, Groeninckx I, Dessein S, Motley TJ, Bremer B (2008) The phylogenetic utility of
271 chloroplast and nuclear DNA markers and the phylogeny of the Rubiaceae tribe
272 Spermaceae. *Mol Phylogenet Evol* **49**, 843-866.
- 273 Lahaye RS, V.; Duthoit, S.; Maurin, O. and van der Bank, M. (2008) A test of psbK-psbI and atpF-
274 atpH as potential plant DNA barcodes using the flora of the Kruger National Park as a model
275 system (South Africa). *Nature Precedings*, 21.
- 276 Lai K, Berkman PJ, Lorenc MT, *et al.* (2012) WheatGenome.info: an integrated database and portal
277 for wheat genome information. *Plant Cell Physiol* **53**, e2.
- 278 Li J, Gale SW, Kumar P, Zhang J, Fischer G (2018) Prioritizing the orchids of a biodiversity hotspot
279 for conservation based on phylogenetic history and extinction risk. *Botanical Journal of the*
280 *Linnean Society*, box084-box084.
- 281 Loytynoja A (2014) Phylogeny-aware alignment with PRANK. *Methods Mol Biol* **1079**, 155-170.
- 282 Loytynoja A, Vilella AJ, Goldman N (2012) Accurate extension of multiple sequence alignments using
283 a phylogeny-aware graph algorithm. *Bioinformatics* **28**, 1684-1691.

- 284 Meyer S, Nagel A, Gebhardt C (2005) PoMaMo--a comprehensive database for potato genome data.
285 *Nucleic Acids Res* **33**, D666-670.
- 286 Numa H, Itoh T (2014) MEGANTE: a web-based system for integrated plant genome annotation.
287 *Plant Cell Physiol* **55**, e2.
- 288 Ratnasingham S, Hebert PD (2007) bold: The Barcode of Life Data System
289 (<http://www.barcodinglife.org>). *Mol Ecol Notes* **7**, 355-364.
- 290 Sakai H, Lee SS, Tanaka T, *et al.* (2013) Rice Annotation Project Database (RAP-DB): an integrative
291 and interactive database for rice genomics. *Plant Cell Physiol* **54**, e6.
- 292 Taberlet P, Coissac E, Pompanon F, *et al.* (2007) Power and limitations of the chloroplast trnL (UAA)
293 intron for plant DNA barcoding. *Nucleic Acids Res* **35**, e14.
- 294 Veidenberg A, Medlar A, Loytynoja A (2016) Wasabi: An Integrated Platform for Evolutionary
295 Sequence Analysis and Data Visualization. *Mol Biol Evol* **33**, 1126-1130.
- 296

297 **Tables**

298

299

300

Table 1: Summary of data currently available in the PlantAligDB.

Target region	Genome	Type	Length (bp) in <i>Nicotiana tabacum</i>	Number of alignments	Number of sequences
<i>atpF-atpH</i>	cpDNA	Inter-genic spacer	502	31	1025
<i>psbA-trnH</i>	cpDNA	Inter-genic spacer	509	79	4852
<i>trnL CD</i>	cpDNA	Intron	577	44	2527
<i>trnL GH</i>	cpDNA	Intron	78	173	34674
<i>rbcL</i>	cpDNA	Protein-coding gene	1434	39	1748
<i>matK</i>	cpDNA	Protein-coding gene	1530	113	11341
ITS	nuDNA	Transcribed spacers and 5.8S gene	678	35	9885
<i>Total</i>				514	66052

301

302

303

304

305

306

Table 2: Average percentage of identical sites (PIS) values in all plant families organized by genomic region.

PIS	<i>atpF-atpH</i>	<i>psbA-trnH</i>	<i>trnL CD</i>	<i>trnL GH</i>	<i>rbcL</i>	<i>matK</i>	ITS
Mean	55.05	39.1	61.13	58.62	78.48	65	28.84
Max	85.95	97.42	96.46	99.07	95.24	97.3	78.45
Min	15.19	2.3	21	6.43	27.47	0.16	1.13

307

308

309

310

311

312

Table 3: Average percentage of pairwise identity (PPI) values in all plant families organized by genomic region.

PPI	<i>atpF-atpH</i>	<i>psbA-trnH</i>	<i>trnL CD</i>	<i>trnL GH</i>	<i>rbcL</i>	<i>matK</i>	ITS
Mean	95.49	94.02	96.34	97.09	96.94	95.23	87.21
Max	99.37	99.94	99.6	100	99.48	99.62	97.75
Min	87.59	69.96	90.54	88.54	81.18	78.72	70.79

313

314

315 **Legends to figures**

316

317 Figure 1: Workflow used to generate the curated alignments, phylogenetic trees, and genetic
318 conservation values stored in PlantAligDB.

319

320 Figure 2: Graphic representation of the measures of sequence conservation PPI and PIS for each
321 region-family alignment: a) *atpF-atpH* region b), *psbA-trnH* region, c) *trnL* CD region, d) *trnL* GH
322 region, e) *rbcL* region, f) *matK* region and g) ITS region.

Selection of target genomic regions



DNA sequences obtained in public databases



Manual curation

removal of duplicated sequences from the same species
removal of sequences without a clear species assignment
reverse complement of sequences found in the opposite direction



Organization of sequences by family according to the NCBI taxonomy



Selection of families represented by ten or more species



Alignment of sequences in each family



Manual curation

removal of sequences that do not cover the entire target region
removal of sequences with large stretches of nucleotide ambiguities



Final sequence alignment

Figure 1

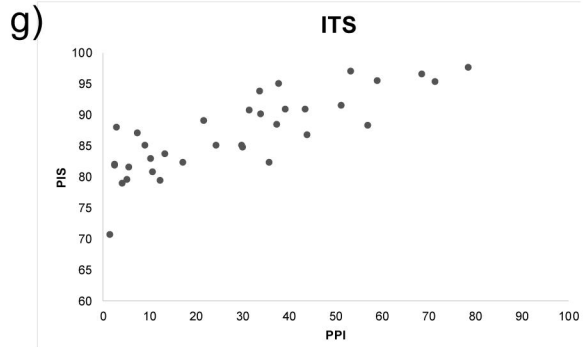
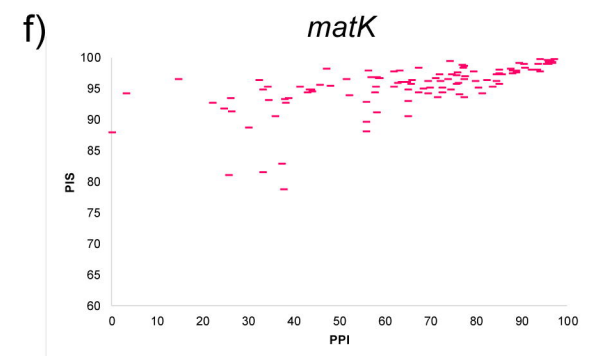
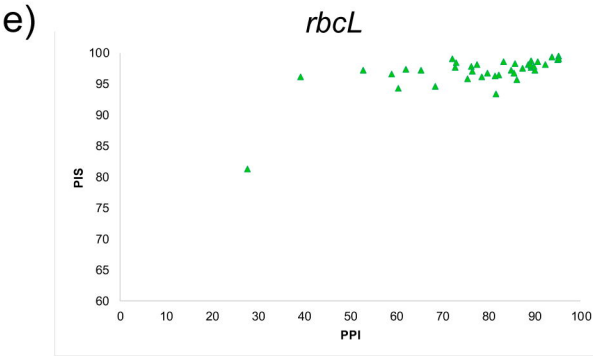
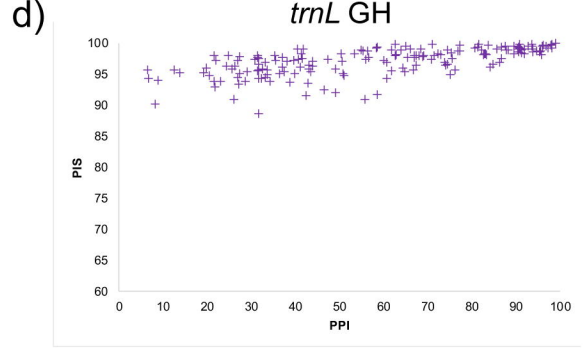
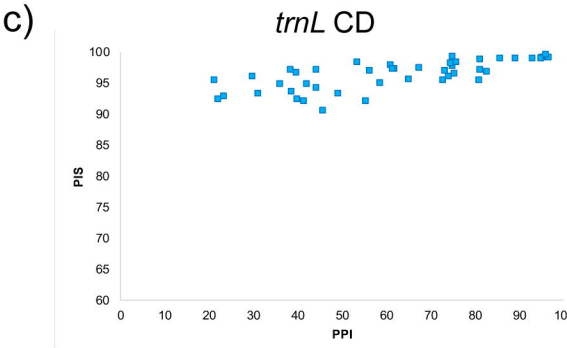
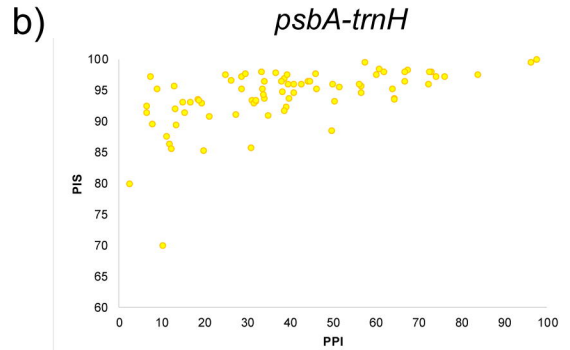
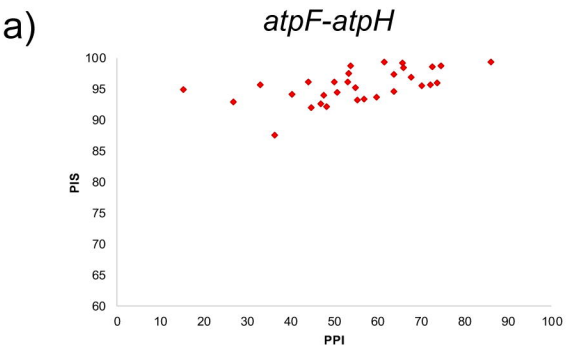


Figure 2