Accurate identification of de novo genes in plant genomes using machine learning algorithms

Claudio Casola[1,2], Adekola Owoyemi[1], Alan E. Pepper[2,3] and Thomas R. Ioerger[4]

[1] Department of Ecology and Conservation Biology, Texas A&M University, College Station, USA

[2] Interdisciplinary Doctoral Degree Program in Ecology and Evolutionary Biology, Texas A&M University, College Station, USA

[3] Department of Biology, Texas A&M University, College Station, USA

[4] Department of Computer Science and Engineering, Texas A&M University, College Station, USA

*Author for Correspondence: Claudio Casola, Department of Ecology and Conservation Biology, Texas A&M University, College Station, TX, 77843, 979-845-8803, ccasola@tamu.edu

Abstract

De novo gene birth—the evolution of new protein-coding genes from ancestrally noncoding DNA—is increasingly appreciated as an important source of genetic and phenotypic innovation. However, the frequency and overall biological impact of de novo genes (DNGs) remain controversial. Large-scale surveys of de novo genes are critical to address these issues, but DNG identification represents a persistent challenge due to the lack of standardized protocols and the laborious analyses traditionally used to detect DNGs. Here, we introduced novel approaches to identify de novo genes that rely on Machine Learning Algorithms (MLAs) and are poised to accelerate DNG discovery. We specifically investigated if MLAs developed in one species using known DNGs can accurately predict de novo genes in other genomes. To maximize the applicability of these methods across species, we relied only on DNA and protein sequence features that can be easily obtained from annotation data. Using hundreds of published and newly annotated DNGs from three angiosperms, we trained and tested both Decision Tree (DT) and Neural Network (NN) algorithms. Both MLAs showed high levels of accuracy and recall within-genomes. Although accuracies and recall decreased in cross-species analyses, they remained elevated between evolutionary closely related species. A few training features, including presence of a protein domain and coding probability, held most of the MLAs predictive power. In analyses of all genes from a genome, recall was still elevated. Although false positive rates were relatively high, MLA screenings of whole-genome datasets reduced by up to ten-fold the number of genes to be examined by conventional comparative genomic methods. Thus, a combination of MLAs and traditional strategies can significantly accelerate the accurate discovery of DNG and the annotation in angiosperm genomes.

Introduction

Novel genes are major drivers of adaptation and evolutionary innovation. A large body of work suggests that new protein-coding genes form at a high rate and represent major contributors to genome evolution and phenotypic variation in plants (1-8). As a result of this rapid evolutionary gene turnover, all species contain hundreds to thousands young, lineage-specific protein-coding sequences that are absent from other taxa (9-12). Both small-scale and whole-genome duplications are responsible for the formation of many novel genes, which tend to share the biological function of their parent genes (1, 3, 4). Occasionally, plants acquired novel genes throughout non-duplicative mechanisms, including horizontal DNA transfer from other species (13, 14) and via the "exaptation" of the coding regions of transposable elements (15-17). Although fundamentally different in nature, all these processes generate new protein-coding sequences from pre-existing genes.

Evolutionary genomic analyses have revealed that new genes can also emerge from ancestrally noncoding DNA sequences wherein an open reading frame (ORF) and new regulatory sequences originate through 'enabler substitutions', typically nucleotide substitutions (18-20). Long considered unlikely evolutionary accidents (Jacob 1977), these so-called de novo genes (hereafter, DNGs) encode for evolutionary novel protein sequences that share no homology with genes from other species. Some DNGs have been shown to bear significant phenotypic impacts, from regulating the mating pathway in budding yeast (21, 22), to producing antifreeze proteins in some fish (23, 24) and affecting human health (25-27).

69 First unequivocally discovered in *Drosophila* (28, 29), de novo genes have been identified across
70 several other groups of animals (23, 30-33) and are well characterized in some model organisms,
71 particularly *Saccharomyces cerevisiae* (21, 34-38). In plants, thousands of potential DNGs have
72 been retrieved through computational surveys in *Arabidopsis thaliana*, *Brassica rapa*, poplar,
73 rice, sweet orange and *Triticeae* (8, 10-12, 39-47).
74
75 Despite the growing number of species and genomes analyzed, the current understanding of de
76 novo gene birth and evolution in plants remain severely limited for several reasons. For instance,
77 the identification of DNGs has traditionally relied on comparative genomic strategies that can be
78 computationally demanding, remain difficult to implement on a large scale, and tend to produce
79 many false positives (48). DNG surveys typically include an initial step wherein all genes (or
80 proteins) in a focal species are queried against genes from other species through sequence
81 homology searches, for example using Blast algorithms (49). The pattern of presence/absence of
82 homologous coding sequences of a given gene along a phylogeny of species allows to estimate
83 its approximate time of origin, a procedure known as 'phylostratigraphy' (50). Genes that share
84 no homology outside of a given genome according to the phylostratigraphic method should thus
85 be considered species-specific (taxonomically restricted). However, there are significant caveats
86 associated with phylostratigraphy.
87
88 First, homology searches can only generate catalogs of all genes that lack homology, also known
89 as 'orphan genes', of which DNGs represent only a subset. Rapidly evolving ancestral genes (51,
90 52), genes derived from exapted transposable elements (11, 53), horizontally transferred genes
91 (53) and genes with alternative coding frame (11, 53) also contribute to the pool of orphan genes.

Although some of these processes are thought to occur at much lower rates than de novo gene birth, the proportion of DNGs among orphan genes might be low (54, 55). Discriminating DNGs from other types of orphan genes requires accurate investigation of synteny conservation across species to identify the enabler substitutions that are uniquely associated with de novo gene birth. Substitutions that enable longer ORFs are especially useful but can be observed only by comparing the DNG coding region with the syntenic genomic regions from several other species (18, 19). While critical to the correct identification of DNGs, the search for enabler substitutions is rarely implemented.

Second, it has been shown that homology searches can underestimate the age of some types of genes, i.e. rapidly evolving genes (56, 57), which can directly affect estimates of DNG rate of formation (48).

Third, homology searches against large datasets require extensive computing resources. Although strategies exist to accelerate this analysis (58), a thorough search of the known sequence space remains challenging and is not reproducible over time, given that sequences databases are expanding exponentially.

Additionally, a wide spectrum of strategies and bioinformatic protocols have been applied to the search of de novo genes in plants and in other organisms. The combination of these issues can produce significant discrepancies in estimates of DNGs even within the same species. For instance, the number of DNGs discovered in *A. thaliana* ranges from 364 (11) to 782 (41). The lack of standardized accurate approaches to assess the number of DNGs represents a major challenge to estimates the rate of de novo gene birth and is diminishing our ability to characterize the biological impact of DNG across plant species.

Machine learning algorithms (MLAs) offer a set of approaches with the potential to mitigate the limitations in DNG detection outlined above. MLAs have proven to be powerful methods for learning models from non-linear datasets in a variety of domains, including many applications in genomics and bioinformatics. In fact, MLAs have been developed to annotate genomic features that include protein-coding genes, RNA genes, enhancers, transcription start sites, splice sites and gene function (59-61). To the best of our knowledge, MLAs applied to the detection of de novo genes have not been developed yet. Interestingly, a few studies have explored the ability of MLAs to identify the broader category of orphan genes in plants (62, 63). A variety of MLAs trained and tested on 1,784 orphan genes from *A. thaliana* showed up to 92% accuracy and 95% sensitivity (62), whereas hybrid deep-learning algorithms applied to 1,544 moso bamboo (*Phyllostachys edulis*) orphan genes reached up to 87% balanced accuracy (63). These results suggest that MLAs can achieve high levels of accuracy for orphan gene prediction. However, as discussed above, DNGs represent only a fraction of orphan genes, and are evolutionarily distinct from other types of genes that lack homology across species. Thus, the ability of MLAs to accurately predict DNGs remains untested.

One of the challenges in applying MLAs to learning classifiers for DNGs is the (typically) small number of positive examples compared to the size of the rest of the genome. While current annotations of plant genomes usually contain tens of thousands of genes, only a few hundred genes can be confidently identified as de novo genes for training. Some MLAs can be sensitive to this asymmetry, outputting models with low information content that appear accurate only because the majority of genes are ancestral genes (hereafter, AGs), while being very inaccurate

for DNGs. There are various methods that have been proposed for handling this significant class imbalance (64). We show that sub-sampling of AGs as negative examples can be effective in training accurate models for DNGs. However, although the accuracy on balanced testing sets can be high, even a moderate false positive rate (FPR) can lead to many false positive predictions when the classifier is applied to tens-of-thousands of genes in a whole genome. We show that the FPR can be reduced somewhat by adjusting the selection of examples during training, though at the cost of increasing the false negative rate (FNR). However, false positives can also be removed through traditional comparative genomic analyses that allows to detect signatures of de novo gene birth, i.e. lack of homology in other species and presence of enabler substitutions.

The ability to detect de novo genes through machine learning classifiers depends on the presence of features that show different distributions of values between DNGs and AGs, defined here as genes with no recent de novo origin. Studies across eukaryotes indicate that DNGs and AGs exhibit different distributions in multiple features associated with gene and protein sequences. For instance, DNGs tend to be shorter and with fewer exons than most AGs (8, 11, 35, 36). Proteins encoded by DNGs typically have fewer annotated domains and possess more structurally disordered regions than ancestral proteins in some eukaryotes (18, 20, 51, 65).

Leveraging these observations, we sought to develop and test MLAs aimed at discriminating DNGs from ancestral genes using sequence-derived information. MLA were trained using DNA and protein sequence attributes from DNGs and AGs obtained from three plant species. The de novo gene catalogs consisted of 331 putative DNGs from *Arabidopsis thaliana* (41), 175 and 343 DNGs from *Oryza sativa* (8) and 754 novel DNGs from *Brassica rapa*. These species represent

161 evolutionary lineages with different levels of divergence, as *A. thaliana* and *B. rapa* are
162 relatively closely related species belonging to the Brassicaceae family within the dicotyledon
163 (dicots) clade, whereas rice is a much more evolutionary distant species in the monocotyledon
164 (monocots) clade.
165
166 Our investigation had the following goals: (1) Assessing the accuracy and recall of different
167 MLA approaches, including decision trees and neural networks, in discriminating DNGs and
168 AGs; (2) Identifying DNA and protein features with high predictive power for detecting DNGs;
169 (3) Assessing the accuracy and recall of MLAs based exclusively on sequence features compared
170 to those incorporating both sequence features and functional genomic data (gene expression
171 levels, translation level, protein-protein interactions, etc.); (4) Determining the predictive ability
172 of MLAs built on data from one species across other taxa; (5) Determining the predictive ability
173 of MLA approaches in detecting DNGs using whole-genome sequence data that include all genes
174 in a species/accession.
175
176
177 ## Results and Discussion
178
179 **A set of high confidence *A. thaliana*-specific DNGs**
180 A dataset 782 of putative *A. thaliana*-specific DNGs was recently generated by Li et al. (41).
181 These genes were identified using sequence homology searches on a limited set of databases and
182 without validating the de novo status of each gene through synteny conservation with closely
183 related species. To produce a set of high confidence DNGs from the Li et al. (2016) catalogue,

we conducted additional homology searches and retained only genes that passed a series of stringent criteria, including conserved synteny with other Brassicaceae and lack of homology with genomes and transcriptomes deposited on NCBI, as described in **Methods**. This approach follows a robust computational framework developed to identify DNGs (58). A total of 298 putative DNGs were excluded as they shared homology with genes in other species (**S1 Table**), leaving 331 high confidence *A. thaliana*-specific DNGs, similarly to the number of DNGs reported by Donoghue et al. (11); however, we could not directly compare our list of DNGs to those identified in this paper, as de novo gene IDs were not provided by Donoghue et al. (11). The remaining 153 putative *A. thaliana* DNGs shared no conserved synteny with other species and were removed from the catalog as the de novo birth pathway could not be determined.

Identification of high confidence DNGs in *Brassica*

We first analyzed all available *Brassica* genome assemblies to determine gene annotation completeness. Most assemblies showed a high level of completeness according to BUSCO scores (**S3 Table**). We selected *B. rapa* as the primary focal species for a de novo gene survey due to its agricultural importance and due to the extensive genomic and functional data available for this species (66-71). As the main focal genome to identify de novo genes, we selected the recently improved *B. rapa* v3.0 genome assembly from the Chiifu-401-42 genotype (71), which contained 46,248 protein coding annotated genes and showed a slightly higher annotation completeness than other available assemblies. We performed extensive sequence similarity searches to identify putative *Brassica*-specific genes (see **Methods**) and identified 754 candidate *B. rapa* DNGs.

Sequence and structural features of de novo genes, ancestral genes and their proteins

On average, de novo gene and protein sequences are known to diverge from ancestral genes in several features. To determine if these qualities can be used to discriminate between DNGs and AGs, we compiled 22 sequence features from the 331 DNGs from *A. thaliana*, the 754 DNGs from *B. rapa*, and the recently published catalogs of 175 and 343 DNGs from rice, *Oryza sativa* (8) (**S1-3 Datasets**). These sequence features are straightforward to calculate for any annotated genome, and do not require any additional data collection (such as gene expression by RNAseq or ribosome profiling). The rice DNGs were obtained integrating sequence homology searches with synteny analyses and were therefore considered well-curated datasets comparable to those of *A. thaliana* and *B. rapa*. To the best of our knowledge, these datasets represent the largest comparative catalog of plant de novo genes to date. The same features were also retrieved by all AGs in these three species (**Table 1**).

Table 1. Primary sequence features and mean values of de novo and ancestral genes and proteins.

	AT DNGs	AT AGs	BR DNGs	BR AGs	OS DNGs-1	OS DNGs-2	OS AGs
# Genes	**331**	26,423	**754**	34,354	**175**	**343**	38,405
Gene length (bp)	**376**	2,269	**630**	1,988	**4,612**	**3,788**	3,729
CDS length (bp)	**218**	1,251	**350**	1,169	**418**	**365**	1157
CDS #exons	**1.4**	5.3	**1.9**	5.4	**3.0**	**2.7**	4.7
GC-content	**43**	44	**49**	47	**58**	**58**	58
SSRs in CDS (bp)	**22**	18	**16**	11	**16**	**10**	16
%SSRs in CDS	**4.1**	1.3	**4.2**	0.9	**2.8**	**1.9**	1.8
Gene overlap	**0.0**	0.1	**0.5**	0.1	**13.1**	**8.7**	7.9
CAI	**0.76**	0.77	**0.77**	0.79	**0.79**	**0.79**	0.81
%Coding	**31**	98	**39**	92	**44**	**38**	84

Coding probability	**0.20**	0.93	**0.49**	0.91	**0.50**	**0.43**	0.85
Protein domains	**0.02**	0.85	**0.02**	1.15	**0**	**0.01**	0.95
%ISD	**0.28**	0.31	**0.33**	0.30	**0.46**	**0.44**	0.34
pI	**8.7**	7.5	**8.6**	7.5	**9.7**	**9.6**	8.0
#TM helices	**0.15**	0.78	**0.09**	0.60	**0.06**	**0.05**	0.51
%w/TM helices	**13**	25	**7**	15	**5**	**4**	14
%w/SP	**15**	15	**7**	11	**5**	**5**	12
%w/mTP	**5**	4	**3**	3	**1**	**1**	2
%w/cTP	**1**	6	**2**	5	**2**	**2**	5
%w/luTP	**0**	0.5	**0**	0	**0**	**0**	0
%w/NoTP	**78**	74	**88**	80	**92**	**92**	81
Kozak score	**-5.44**	-4.99	**-5.40**	-5.06	**-5.15**	**-5.37**	-4.93

AT: *A. thaliana.*
BR: *B. rapa.*
OS: *O. sativa.*
CDS: coding sequence.
SSRs: simple sequence repeats (microsatellites).
CAI: codon adaptation index.
Coding probability according to CPC2 (72).
%Coding: proportion of genes predicted by CPC2 to be coding.
%ISD: proportion of protein sequence with intrinsic structural disorder.
pI: Isoelectric point.
#TM helices: average of transmembrane helices per gene.
%w/TM helices: proportion of genes with transmembrane helices.
%w/SP: proportion of genes with signal peptide.
%w/mTP: proportion of genes with Mitochondrial transit peptide.
%w/cTP: proportion of genes with Chloroplast transit peptide.
%w/luTP: proportion of genes with Luminal transit peptide.
%w/NoTP: proportion of genes with No-targeting peptide.

Along the lines of previous studies (8, 41), all DNGs components tend to be shorter (except for introns in the *O. sativa* DNG sets), especially at the level of the coding region, and contain fewer exons compared to AGs (**Table 1**; **S2A-C Fig**). The GC-content varied more significantly across species than between DNGs and AGs (**Table 1; S2D Fig**). Interestingly, the GC-content distribution peaks at higher values for AGs in *A. thaliana* and for DNGs in *B. rapa* (**S2D Fig**). In rice, the GC-content distribution is bimodal in AG, as previously described (73), with DNG GC values peaking in between (**S2D Fig**). This pattern in the rice genome is mirrored at the level of

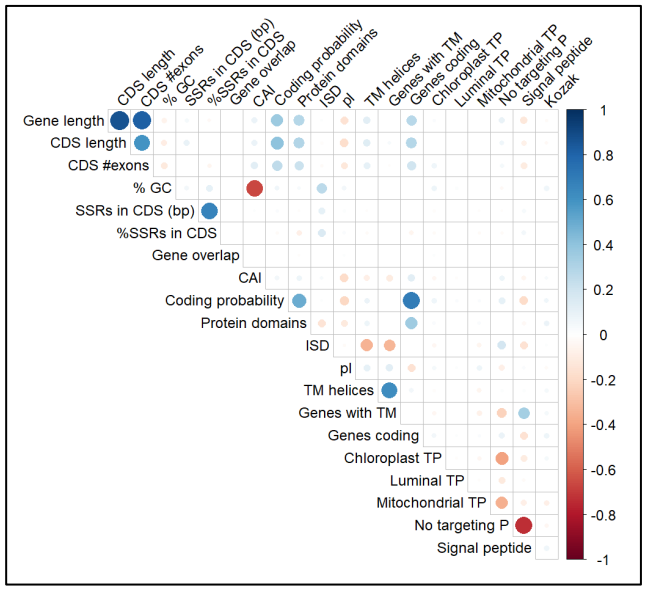
the codon adaptation index (**S2G Fig**). Additionally, DNG coding regions contain a higher proportion of DNA derived from microsatellites identified by Tandem Repeat Finder. Interestingly, the microsatellite content is lower in the large rice DNG dataset, which includes older DNGs, suggesting that simple repeat content may decrease with time due to substitutions. As expected, the predicted coding potential of DNGs is much lower than in AGs, with much fewer DNGs being labeled 'coding' according to the coding potential calculator 2. In agreement with this, the coding adaptation index of DNGs is on average significantly lower than in AGs (**Table 1; S2E-G Fig**). Furthermore, Kozak scores were lower in DNGs compared to AGs, although their distributions largely overlapped between the two types of genes (**Table 1; S2H Fig**).

Protein features included the presence of conserved domains, predicted proportion of intrinsic structural disorder (ISD) residues, isoelectric point, predicted number of transmembrane helices and subcellular localization peptides (**Table 1; S2I-K Fig**). As expected, given the small size of de novo proteins and their recent origin, very few of them contained conserved domains. Similarly, we found a much lower number of transmembrane motifs (TMs) and genes with TMs in de novo proteins than ancestral proteins. This is in contrast with the observation that *S. cerevisiae* de novo ORFs with adaptive potential are enriched for TM domains (74).

Across all species, DNG proteins showed consistently higher isoelectric point (pI), indicating a higher proportion of basic residues, and very few transmembrane domains compared to AG proteins (**Table 1**). Elevated pIs due to a depletion of acid residues have been also found in mammalian orphan proteins (75) and in S. *cerevisiae*-specific translated ORFs (37), but no

272 explanations for these trends have been put forward. We also found that the distribution of pI
273 values is very different between proteins encoded by DNGs and AGs. While in AG proteins the
274 pI values follow an approximate trimodal distribution, isoelectric points in DNG proteins cluster
275 around two peaks around low (~4) and high (~11) values (**S2I Fig**).
276
277 Additionally, we observed higher ISD levels in DNGs than AGs only in *B. rapa* and rice (**Table**
278 **1; S2K Fig**). This is in agreement with what observed in *Drosophila melanogaster* DNGs (76)
279 and in orphan genes in *D. melanogaster* (77) and *Leishmania* (78). A similar trend was also
280 reported in rodents young genes (65), although follow up studies suggest that this pattern is
281 likely to be an artifact (51, 79). Some authors have shown that high ISD levels are associated
282 with elevated GC content in orphan genes or young genes (80), possibly a result of some young
283 genes overlapping with ancestral genes (51). It is unclear if the modest difference in GC content
284 between DNGs and AGs in *B. rapa* and rice drives their elevated ISD levels.
285
286 Overall, DNG proteins also contained fewer localization peptides compared to AG proteins, with
287 the notable exception in *A. thaliana* of both a higher proportion of mitochondrial transit peptides
288 in DNGs vs. AGs, and a comparable number of DNG and AG proteins with a signal peptide
289 (**Table 1**).
290
291 We further investigated possible correlation among features in each species (**Fig 1**). As expected,
292 gene length, CDS length and CDS #exons (the number of coding exons) were positively
293 correlated. Similarly, the coding probability and the presence of protein domains increased with
294 gene length and CDS length and, to a smaller extent, with CDS #exons. Longer coding regions

295 are more likely to be predicted as coding by the coding potential calculator (72). Longer genes
296 are also more likely to encode protein domains. In agreement with previous studies (51, 80), we
297 found that GC-content and ISD were positively correlated. Another expected pattern is the
298 negative correlation between ISD and the presence of transmembrane helices, which cannot form
299 in disordered protein regions. The anticorrelation between GC-content and codon adaption index
300 may be due to the paucity of GC-rich codons among the most-used codons in angiosperms.
301 Similar correlations were observed in *B. rapa* and rice (**S3 Fig**).
302
303



312 **Fig 1**. Correlation matrix of *A. thaliana* sequence features used to train ML classifiers. Feature
313 names follow the nomenclature of Table 1.
314
315
316

317 **Both Decision Tree and Neural Network classifiers detect the majority of de novo genes in**
318 **test gene sets**

319 As a preliminary attempt to develop a predictive model for DNGs, we trained both a Decision
320 Tree (DT) classifier and a Neural Network (NN) classifier on the *A. thaliana* gene datasets. The
321 set of positive examples was formed by the 331 DNGs we identified in this species. Because
322 using all 26,423 AGs as negative examples led to degenerate tree where every gene was
323 classified as negative due to class imbalance (64), the AGs were sub-sampled (81) by choosing
324 an equal number of 331 genes at random for the negative examples. The NN classifier consisted
325 of a fully-connected network with a hidden layer of 20 units (see **Methods**). We also evaluated
326 the effect of a different number of hidden units (from 10 to 100 in intervals of 10 units) and 2
327 hidden layers, but these did not significantly improve the accuracy of the NN. When the 662
328 selected examples were divided randomly into 70% for training and 30% for testing, the DT and
329 NN models were found to have 91-92.0% accuracy in predicting DNGs in the test set. Thus,
330 even though Decision Trees and Neural Networks represent completely different methods for
331 capturing patterns in training data, they are both able to learn how to discriminate DNGs from
332 AGs using sequenced-based features. The confusion matrix for the 100 genes from the 30% test
333 sets shows that the errors are evenly distributed between false positives and false negatives, with
334 recall values above 90% (**Table 2**).
335
336
337
338

Table 2. Confusion matrix of DT and NN classifiers, trained on *A. thaliana* genes, applied to an independent balanced test set. Values represent counts of test examples.

			predicted class labels		
			DNG	**AG**	**Recall**
DT classifier	**actual class labels**	**DNG**	91	9	91%
		AG	7	93	
NN classifier	**actual class labels**	**DNG**	92	8	92%
		AG	8	92	

In order to determine a more general estimate of the predictive accuracy (since the single tree above is dependent on the specific AGs chosen as negative examples), 10-fold cross-validation was carried out (where a decision tree was generated based on 90% of the data and tested on the remaining 10%, repeated 10 times in a rotated manner), which resulted in a performance estimate of 89.7%, with a 95%-confidence interval (CI95) of 87.3-92.1% (**Table 3**). The same protocol was used to train classifiers on species-specific sequence features retrieved in DNGs and AGs of *B. rapa* and *O. sativa* (using the larger dataset of 343 de novo genes in the latter, see **Table 1**). In all species, NN models performed slightly better than DT models, significantly so in *A. thaliana* and *B. rapa* but not in rice (**Table 3**). DT and NN classifiers showed significantly lower accuracies in *B. rapa* and *O. sativa* compared to *A. thaliana* (**Table 3**; DT *A. thaliana-B. rapa* P=0.0038; DT *B. rapa-O. sativa* P=0.0003; NN *A. thaliana-B. rapa* P=0.0005; NN *B. rapa-O. sativa* P=0.0001, unpaired T-test). Overall, these results indicate that MLAs trained on species-specific datasets can successfully retrieve the vast majority of DNGs. Confusion matrices for both classifiers indicate that NN models achieve substantially higher recall than DT models in all

358 species (**S4 Table**). Recall is comparably high (~92%) in NN models of Brassicaceae, but
359 decreased to ~83% in rice, where DT models achieved only ~76% of recall (**S4 Table**).
360
361 **Table 3. Accuracy of DT and NN classifiers in the three angiosperms**.

Species	Decision Tree balanced accuracy† (95% C.I.)	Neural Network balanced accuracy† (95%C.I.)
A. thaliana	89.7% (87.3-92.1%)	93.2% (91.5-94.8)*
B. rapa	85.1% (84.3-86.0%)	88.9% (88.1-89.8)*
O. sativa	76.5% (73.1-80.0%)	80.2% (77.0-83.3)

362 †Averaged over 10-fold cross-validation
363 *NN model vs. DT model within species, *P*-value<0.05, unpaired T-test
364
365
366 The variation in accuracy and recall across species may be due to several factors. A higher
367 quality of gene annotation in *A. thaliana* may explain the increased accuracy of MLAs in this
368 species. The lower accuracy in rice could in part depend on the slightly older age of the larger
369 dataset of 343 DNGs used for this species (8), as DNGs should acquire features that are more of
370 typical genes through time (35). Differences in age of DNGs could also explain the significant
371 overlap in the distribution of continuous features between DNGs and AGs (**S2 Fig**). We also
372 observed that some features show a varying degree of predictive importance between
373 Brassicaceae and rice, which could further contribute to differences in accuracy (see below).
374
375
376 **Training DT classifiers with functional genomic features**
377 We evaluated whether adding functional genomic data could improve the accuracy of the
378 decision tree classifier using datasets available for *A. thaliana*. The classifier was re-trained
379 using 28 additional features which are not systematically available in all genomes, including

transcription data (RNAseq), translation levels estimated through Ribosomal profiling (RiboSeq), proximity to transposable elements, selective constraint, and phenotype data for gene knockout mutants (**S5 Table**). The 10-fold cross-validated accuracy of models extended with these functional features was 91.4%, (89.7-93.1%), which is not significantly greater than models without these functional features (P>0.05, unpaired T-test). Some of the functional features were occasionally used as decision criteria in lower branches of some of the decision trees; the functional feature with highest importance (0.04) was "AVG RiboP RPKM 25 samples", which suggests that lack of expression evidence can be an important discriminator for DNGs.

Feature Importance in Decision Trees

To better assess the contribution of each feature to DT classifiers in the three species, we calculated feature importances, which are based on the relative contribution of each feature to splitting the data in the tree and range between 0 and 1 (see **Methods** and **Fig 2**). Feature importances averaged over 10 runs for each species are shown in **S6 Table**. The DT classifier developed in *A. thaliana* consisted of 70 nodes (including 24 leaves) with a depth of 11. The attribute tested for splitting at the root of the tree, considered the most important based on reduction of Gini impurity, was "Protein domains". This turns out to be a highly discriminating feature: most AGs (22458/26423=85.0%) have at least one domain (recognized as a known fold family by Pfam based on amino acid sequence), whereas most DNGs do not (only 7/331=2.1% had a recognized domain) (**Fig 2**).

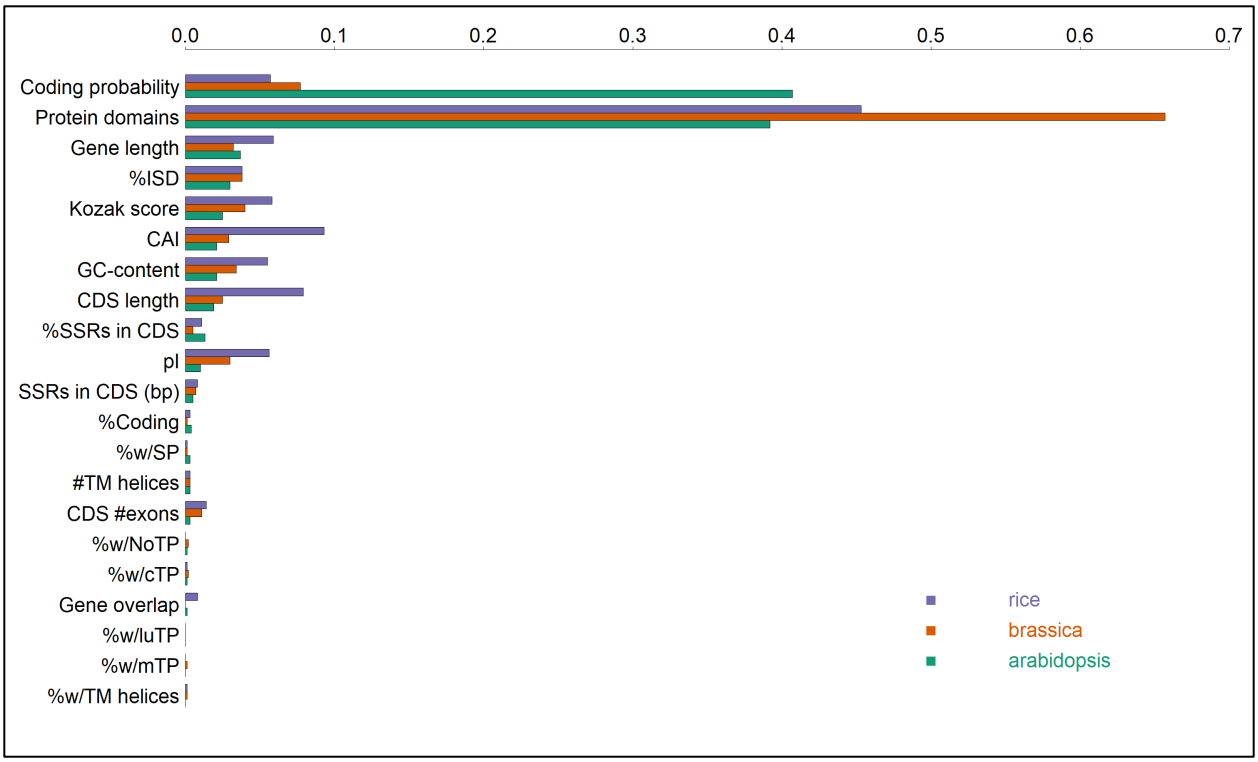


Fig 2. Top ten features in the DT classifier ranked by importance.

Overall, the top ten features by importance are largely the same across all species. "Protein domains" is the most prominent feature across species (**Fig 2**; **S6 Table**). Several other top features that are known to be significantly different between DNGs and AGs, including "Coding probability", "Gene length", "CDS length", "Codon adaptation index (CAI)", "%GC" and the proportion of "intrinsic structurally disordered (ISD)" regions in proteins, show high importance. Although the presence of a conserved Kozak motif has not been investigated before in DNGs, the "Kozak score" feature showed a relatively high importance, in agreement with other findings suggesting that de novo genes might acquire more 'gene-like' regulatory sequences by natural selection after their emergence (31).

Furthermore, some features exhibited substantially higher importance in rice than Brassicaceae. For instance, "Codon adaptation index (CAI)" represents the second most important feature in rice while ranking sixth and eight in *A. thaliana* and *B. rapa*, respectively (**Fig 2; S6 Table**). This is interesting as CAI is only slightly higher in AGs than DNGs in rice (**Table 1; S2G Fig**). "Gene length", "CDS length", "Kozak score" are also more prominent features in the monocot.

Interestingly, the top 2 features alone in *A. thaliana* ("Coding probability" and "Protein domains") can be used to construct decision trees with nearly equivalent accuracy of the ones trained on all features. For example, in *A. thaliana*, the balanced accuracy with such trees (averaged over 10-fold CV) is 91.4% (95%CI: 89.1-93.7%). The decision trees still have multiple nodes in them, typically around 30; they just include splits on multiple threshold values (i.e. sub-ranges) of "Coding probability". The average feature importances are 0.407 for "Coding probability" and 0.392 for "Protein domains". In fact, the neural network performs even better, probably due to the reduction in parameters (weights in the network) with just two inputs: 94.0% (95%CI: 92.2-95.7%). Further analyses will be necessary to determine if this applies to other plant genomes.

Accuracy is not limited by small training set size

The size of the training dataset can affect accuracy in MLA predictions. We tested if this is the case for the DT classifiers using randomly selected subsets of *A. thaliana* de novo genes and ancestral genes equal to 10-100% of the original training set of 464 genes. This was repeated 30 times for each set size to obtain a DT classifier learning curve. The testing dataset for each

analysis was carried out on a group of DNGs that did not overlap with the training set. Additionally, during training and testing, the number of negative examples (AGs) was always balanced with an equal number of positive examples. We found that the accuracy of the DT remains elevated (>85%) even for training with one-tenth of the full training-set size of 464 genes and is nearly equal to the highest accuracy with only 30% of the full training size, corresponding to 138 genes (**Fig 3**). The observed trend also suggests that the accuracy would not be significantly improved by using a larger training set with more known DNGs.

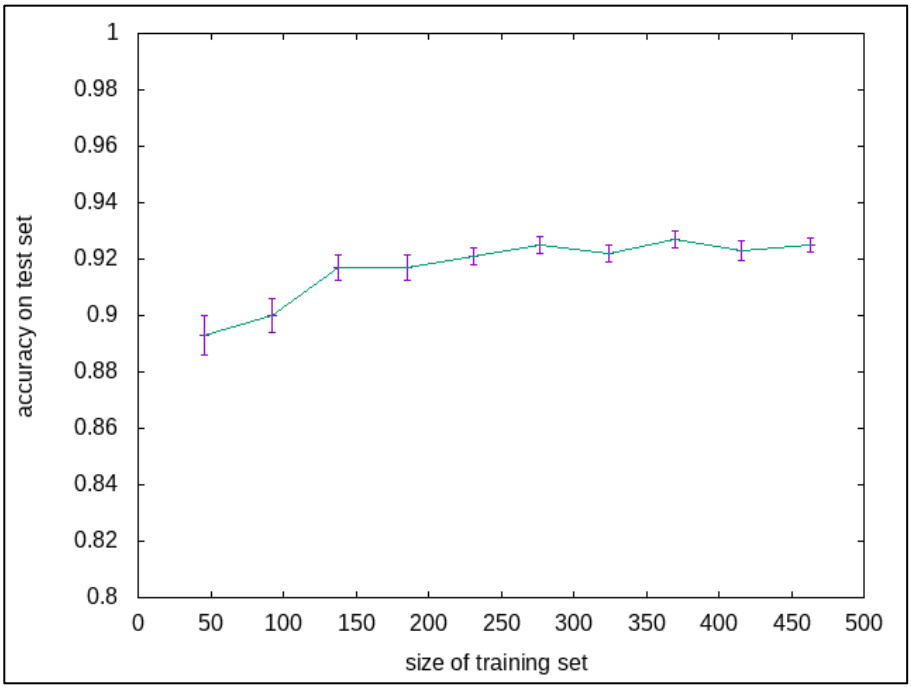


Fig 3. Learning curve of the *A. thaliana* DT classifier on 10-100% of the original training dataset. Standard errors are shown.

Cross-species models for de novo gene prediction are nearly as accurate as species-specific models

An important question is whether the patterns extracted by the MLAs for discriminating DNGs are species-specific, or whether the MLAs are capturing general properties of DNGs that extend across multiple plant species. To assess this, we trained a DT and NN classifiers using data on genes from *A. thaliana*, and then applied these models to the *B. rapa* and *O. sativa* datasets. In 10-fold cross-validation analyses, the *A. thaliana* DT model achieved 83.1% and 67.2% accuracy in *B. rapa* and rice, respectively (**Table 4**). The *A. thaliana* NN model reached slightly higher accuracy in both species (**Table 4**). Overall, species-specific models (**Table 3**) achieved significantly higher accuracy than cross-species models (P-value<0.05, unpaired T-test), except for the *A. thaliana* DT model applied to *B. rapa* datasets (P-value=0.0855, unpaired T-test). Taken together, these results indicate that models trained on the *A. thaliana* genome, which has been more carefully and thoroughly annotated, can be applied with nearly equal accuracy on *B. rapa*, and does not require an MLA to be re-trained on each new gene set. Conversely, *A. thaliana* classifiers achieved substantially lower accuracies in *O. sativa* compared to the rice-specific models. It appears that the features associated with DNGs in *A. thaliana*, such as coding probability, lacking recognizable protein domains, and having lower Kozak score, generalize across species, and are associated with DNGs in other plant genomes.

Table 4. Comparison of performance of DT and NN models using a common model trained on *A. thaliana* versus species-specific models.

Species	Decision Tree balanced accuracy† (95% C.I.)	Neural Network balanced accuracy† (95%C.I.)
B. rapa	83.1% (81.2-85.0)	85.1% (83.5-86.6)
O. sativa	67.2% (63.9-70.5)	69.4% (66.5-72.2)

†Averaged over 10-fold cross-validation

The higher predictive ability of *A. thaliana* MLA classifiers in *B. rapa* compared to rice suggests that cross-species DNG identification with MLAs tend to be more accurate in closely related genomes. Features that differ significantly between Brassicaceae and rice genes, including gene length, %GC and simple repeat content in the coding region, may drive the lower sensitivity of the *A. thaliana* DT classifier in rice. A broader taxonomic sampling at varying phylogenetic distances from *A. thaliana* will be required to test this hypothesis more thoroughly. We noticed that the recall decreased in cross-species prediction, with a limited difference in *B. rapa* (from ~84-92% to ~82-84%) and a significant drop in in rice *(*from ~76-83% to ~55-59%) (**S7 Table**). This indicates that cross-species MLA predictions of DNGs may achieve acceptable levels of sensitivity within taxonomic families but fail to detect a substantial fraction of DNGs in more distantly related species. Thus, broad MLA-based de novo gene surveys in plants may require the training of classifiers using DNGs detected with comparative genomic approaches in at least one species per family.

Whole-genome predictions of DNGs using DT and NN classifiers

We next assessed the accuracy of species-specific DT and NN classifiers to predict DNGs in whole-genome gene sets. We calculated the balanced accuracy, which is better suited to assess the performance of classifiers when classes are imbalanced. The overall balanced accuracy ranged from 92.3% in *A. thaliana* to 76.9% in *O. sativa*, with slightly higher accuracy in DT vs. NN models (**Table 5**). A high recall of ~94-99% was found across species and classifiers, although DT models also achieved lower FPRs compared to NN models (**S8 Table**). Given the much higher number of AGs than DNGs, the total number of false positives reached ~2,055 in *A. thaliana* and a maximum of ~8,948 genes in *O. sativa* (**S8 Table**). Given the high FPRs, MLAs alone may achieve the level of accuracy required to entirely replace traditional comparative genomic analyses in DNG surveys; however, the application of MLAs as a first step would decrease by up to 10-fold the number of genes that need to be investigated with homology searches and other time-consuming approaches in order to remove false positives.

Table 5. Comparison of performance of DT and NN species-specific models applied to whole genomes.

species	Decision Tree genome-wide accuracy† (95% C.I.)	Neural Network genome-wide accuracy† (95%C.I.)
A. thaliana	92.3% (91.9-92.7)	91.1% (90.4-91.9)
B. rapa	87.4% (87.1-87.7)	86.0% (85.6-86.4)
O. sativa	79.3% (78.7-79.9)	76.9% (74.9-78.8)

†Averaged over 10-fold cross-validation

We further examined if the performance of models trained on *A. thaliana* data but applied to the two other species (in this analysis, there are no confidence intervals because a single input model trained on one species was tested for accuracy on the whole genome of another). The *A. thaliana* classifiers, particularly NN models, showed relatively high balanced accuracy in *B. rapa* and in rice (**Table 6**). However, recall values dropped significantly in both species, from ~97-98% to ~82-84% in *B. rapa* and from ~94-98% to ~55-59% in rice (**S9 Table**). Interestingly, the *A. thaliana* NN model resulted in lower FPRs than the within-species models while the opposite was true for the *A. thaliana* DT model (**S9 Table**).

Table 6. Comparison of performance of DT and NN models using a common model trained on *A. thaliana* versus species-specific models applied to whole genomes.

species	Decision Tree genome-wide accuracy	Neural Network genome-wide accuracy
B. rapa	84.9%	86.2%
O. sativa	74.6%	81.6%

The DT classifier specificity can be adjusted by increasing the proportion of negative examples during training

Given the high false positive rates in classifiers, we investigated if including negative examples during training could increase the model specificity using *A. thaliana* datasets and the DT model. We scaled up the number of negative examples up to 10 times the original set of 232 AGs while maintaining a balanced test set with equal numbers of DNGs and AGs. Each iteration at different training sizes was repeated 30 times. We found that for increasing numbers of negative examples, the balanced accuracy steadily decreases from ~91% to ~85%, primarily due to the

increasing false negative rate from <10% up to 25% (**Fig 4**). Concomitantly, the false positive rate decreased from ~8% to <4% (**Fig 4**). As the main goal of the MLA approach is to detect DNGs, the loss of sensitivity associated with the higher number of negative examples might be not worthwhile. However, we noticed that this tradeoff between false positive and false negative rates might be acceptable for limited increases in the size of negative examples.

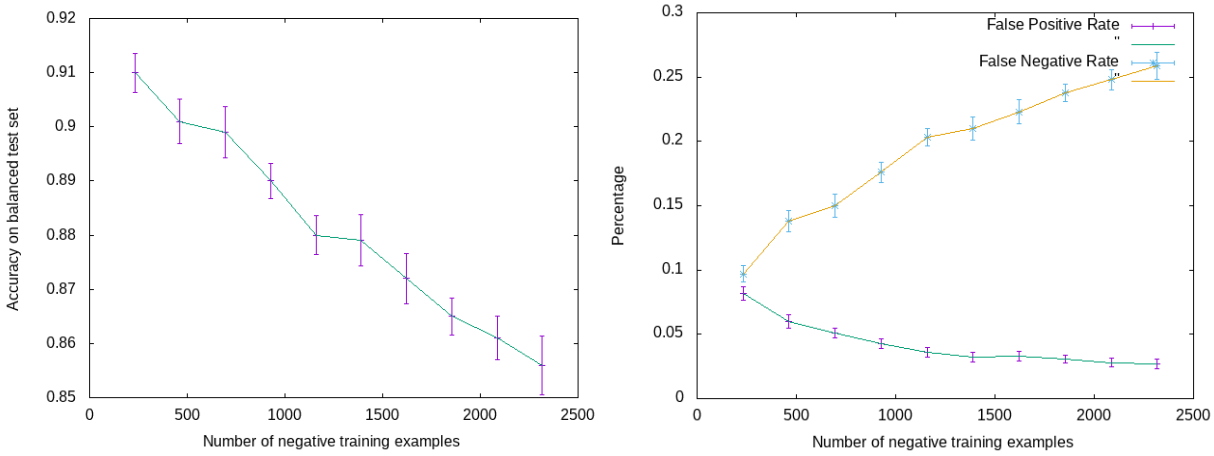


Fig 4. Accuracy (left) and false positive and false negative rates (right) for increasing number of negative examples used in the training of the DT classifier in *A. thaliana*. Standard errors are shown.

Conclusions

In this study, we have developed and assessed the first machine learning framework to identify de novo genes. In order to make these approaches readily applicable in species with limited functional genomic data, we have specifically selected basic sequence features that can be obtained from DNA and protein sequences available in annotated genomes. Using DNG datasets

558 from three plant species, including an updated gene set from *A. thaliana* and the first group of de
559 novo genes in *B. rapa*, we have found that both decision tree (DT) and neural network (NN)
560 classifiers achieve high levels of accuracy and recall in predicting DNGs. Using DT algorithms
561 applied to sub-sampled sets of DNGs and AGs, we identified a few features with significant
562 predictive power for DNGs. This is in line with performance ability of MLAs to discover orphan
563 genes in *A. thaliana* based on six DNA sequence features (62). Importantly, orphan genes are not
564 equivalent to de novo genes, as the former appear to be mostly constituted by rapidly evolving
565 genes (54). Training MLA models with additional features derived from functional genomic data
566 (transcription and translation data) and information from phenotypic assays that are not readily
567 available in most sequenced genomes does not lead to a substantial increase in accuracy or
568 reduction in the number false positives.
569
570 A major advantage offered by MLAs is the significant decrease in computational time compared
571 to traditional genomic approaches to find DNGs—essentially, a time contraction from weeks or
572 months to minutes. For MLAs to be successfully applied in DNG surveys across hundreds to
573 thousands of species, it is critical to train models using datasets of known DNGs from a few key
574 species, and for these models to obtain high accuracy and recall in other species. We conducted
575 initial tests to explore this possibility using DT and NN models trained on *A. thaliana* to predict
576 DNGs in *B. rapa* and rice. We found that these cross-species predictions achieved comparable
577 accuracy of species-specific models in *B. rapa,* and somewhat lower accuracy in rice. This
578 suggests that MLAs trained in one species can likely be used to infer DNGs in closely related
579 species, as in the case of the two Brassicaceae, *A. thaliana* and *B. rapa*. Given that many
580 angiosperm families now contain sequenced genomes from multiple species, and considering

581 both the rapid increased of the number and quality of new genome assemblies, de novo gene
582 discovery based on MLAs could likely be applied to a large number of flowering plant taxa.
583 Future work on more taxa should help better determine how cross-species MLAs accuracy
584 decreases when the evolutionary distance between taxa increases, as this study indicates.
585
586 Genome-wide analyses showed that species-specific models predicted well above 90% of known
587 DNGs, although the much higher number of ancestral genes lead to several thousand false
588 positive cases. In cross-species genome-wide analyses, *A. thaliana* models identified 82-84% of
589 true DNGs in *B. rapa*, but achieved less than 60% recall in rice. Notably, FPRs were
590 substantially lower in cross-species NN models compared to species-specific models. The
591 combination of high recall, at least in some cross-species tests, and high FPRs suggests that a
592 three-step pathway can be employed to accelerate DNG discovery in angiosperms. First, NN
593 models are developed in one or two species with the best gene annotation quality in each
594 angiosperm family. Second, these NN models are applied to an array of target species in the
595 same family. Third, the candidate DNGs predicted from each target species, comprising only a
596 few thousand genes, are analyzed post-hoc with traditional comparative genomic approaches to
597 remove false positives.
598
599 As this represents the first systematic study to assess machine learning approaches in de novo
600 gene discovery, we expect that further developments of in this area could significantly increase
601 accuracy and recall while reducing the false positive rate in DNG detection. Along these lines,
602 alternative machine learning approaches, including deep neural networks, and methods to

603 address the class imbalance between DNGs and AGs different from sub-sampling, such as
604 synthetic minority over-sampling algorithms, or SMOTE (62, 82), warrant future investigations.
605
606
607

608 # Methods

609 ## Validation of *Arabidopsis thaliana* de novo genes

610 The TAIRv10 (TAIR10) DNA and protein sequences of *A. thaliana* were obtained from the
611 folder "TAIR10 blastsets" in the TAIR repository (83). Data from the files
612 "TAIR10_pep_20101214", "TAIR10_cds_20101214" and "TAIR10_exon_20101028" were
613 used in sequence similarity searches and sequence feature analyses. *A. thaliana* de novo genes
614 were retrieved from a set of 782 putative DNGs recently described by Li et al (2016) using
615 sequence homology searches. These genes were screened to identify high-confidence DNGs
616 supported by further comparative genomic data, particularly synteny information. Specifically,
617 we performed Blast v2.11.0 (49) searches of the corresponding protein sequences against several
618 NCBI databases. We used Blastp to search the "nr" and "tsa_nr" databases, and tBlastn to search
619 the "nr/nt", "refseq_rna", "est" and "TSA" databases with the following modified parameters:
620 num_descriptions 100 -num_alignments 100 -max_hsps 5 -evalue 0.001 -seg yes. The seg filter
621 was turned on in order to remove spurious hits due to nonhomologous stretches of similar amino
622 acids. These searches were carried out against increasingly broader taxonomic units that include
623 *Arabidopsis* species excluding *A. thaliana*, Brassicaceae (excluding *Arabidopsis*), rosidae
624 (excluding Brassicaceae), angiosperms (excluding rosidae), green plants (excluding
625 angiosperms). We also searched for homologous sequences of the 782 DNGs in fungi, bacteria

and archaea to identify and remove possible horizontal transfer cases (**S1 Table**). Subject sequences were screened using unix scripts to remove truncated proteins. We excluded from the catalog of DNGs all cases with any homology with sequences from a non-focal species.

To determine synteny conservation of DNG coding regions we searched 45 Brassicaceae genome assemblies obtained from Phytozome (https://phytozome-next.jgi.doe.gov) using the translated DNA sequence of each exon of the 782 putative DNGs, which allowed us to detect conserved coding regions for genes with one or multiple exons, in tBlastx run with the following modified parameters: num_descriptions 100 -num_alignments 100 -max_hsps 5 -evalue 0.001 -seg yes. A list of the 45 species investigated in available in **S2 Table**.

First, we used these alignments to identify Brassicaceae genomic regions that were syntenic with DNGs and with the potential to encode proteins. Although these regions did not include annotated genes, they maintained long coding regions and were thus considered *bona fide* homologs to DNGs. We based this selection on two criteria. We selected for hits with a conserved methionine within the first five amino acids of the *A. thaliana* DNG protein, in order to account for alternative first codon positions (84). Additionally, we included only hits wherein the total alignment length from the first codon to the first stop codon equal to at least 75% of the query protein. This threshold was selected to include hits that are likely to encode a protein, given the lack of disabling mutations along most the of the coding, while allowing for slightly shorter loci, as stop codons may also vary slightly between orthologs. Using this strategy, we identified 44 DNGs with putative homologous genes in non-*A. thaliana* Brassicaceae, which were thus discarded.

649 Second, we used the same alignment data for the remaining DNGs to identify those that maintain
650 synteny conservation *in noncoding regions* with other Brassicaceae. To this end, we applied a
651 minimum threshold of 30% coverage between *A. thaliana* DNGs and Brassicaceae genomes as
652 corresponding to conserved synteny. This length threshold is lower than those used in previous
653 DNG analyses in animals (31, 85, 86), in order to account for the decreased overall synteny
654 conservation among Brassicaceae. Furthermore, the syntenic alignments were screened for the
655 presence of enabler substitutions, represented by a novel start codon, the removal of stop codons
656 and/or frameshifts in the DNG coding region compared to the syntenic DNA of the outgroup
657 species. Overall, we found 604 *A. thaliana* putative DNGs with apparent synteny conservation
658 with at least one other Brassicaceae genome.
659
660 **Identification of *Brassica rapa* de novo genes**
661 We retrieved all *Brassica rapa* genome assemblies and gene sets available as of October 2021
662 and screened each assembly for completeness using BUSCO v3.0 (87) (**S3 Table**). The *B. rapa*
663 coding regions, protein and genome assembly fasta files were downloaded from:
664 https://ngdc.cncb.ac.cn/search/?dbId=gwh&q=GWHAAES00000000.
665 *B. rapa* proteins containing stops ("Xs") within their amino acid sequences were removed,
666 leaving 45,912 proteins. We further screened the remaining *B. rapa* proteins to identify and
667 remove sequences mostly formed by transposable elements (TEs), as they likely represent
668 misannotated TEs. To this aim, we first downloaded the sequences of 39,197 TE families from
669 31 Brassicaceae reported in PlantRep (88). Proteins containing TE sequences were retrieved by
670 performing a tBlastn search with the following modified parameters: -evalue 1e-10 -
671 max_target_seqs 10 -max_hsps 5.

672 A total of 9,791 *B. rapa* proteins shared sequence similarity with TEs over at least 50% of their
673 length and were removed from the dataset. Proteins with Blast matches uniquely with unknown
674 repeats were not removed as those repeats might represent microsatellites, which could
675 potentially form a portion of de novo gene coding regions.
676
677 To search for DNGs among the remaining 36,121 *B. rapa* proteins, we carried out a
678 multipronged homology search strategy to identify proteins with homology in non-*Brassica rapa*
679 genomes. First, the *B. rapa* proteins were searched against the plant NCBI refseq protein set
680 obtained on August 31, 2021 throughout a Blastp run with the following modified parameters: -
681 num_descriptions 5 -num_alignments 5 -evalue 0.001 -seg yes. A total of 18,413 proteins
682 showed no sequence homology to NCBI refseq proteins with the exception of *Brassica*
683 sequences, thus representing candidate *Brassica* orphan proteins. A further Blastp search was
684 performed against the NCBI non-random, tsa_nr, refseq_rna and est databases of all
685 Brassicaceae proteins with the following modified parameters: -max_target_seqs 50 -max_hsps
686 5. All hits containing premature stop codons were removed as they could represent expressed
687 sequences of noncoding genes or truncated and thus non-functional proteins.
688
689 Similarly to the procedure applied to *A. thaliana* putative DNGs, we carried out Blast searches
690 against increasingly broad taxonomic units containing *B. rapa*, starting from *Brassica* but
691 excluding *B. rapa*, then other *Brassicaceae*, rosidae, angiosperms, green plants, excluding the
692 previous taxon at each step, and using the following modified parameters: -max_target_seqs 50 -
693 max_hsps 5-evalue 0.001 -seg yes. Fungal proteomes were also screened, whereas Archaea and
694 Bacteria sequences were not included as our analyses in *A. thaliana* DNGs showed that

prokaryotic databases contributed marginally to the detection of homologs. We found 2,089 *B. rapa* proteins sharing no sequenced homology with two or more non-*Brassica* proteins, thus representing *B. rapa* orphan genes. A cut-off of at least two non-*Brassica* proteins was implemented to take into account possible contamination from *A. thaliana* into other Brassicaceae genome datasets (89, 90). This number of orphan genes in *B. rapa* is similar to but higher than the 1,540 *B. rapa* orphan genes recently described (45), probably because of differences in the homology search criteria and in the gene annotation version. We further removed from the list of orphan genes 35 genes with annotated protein domains in eggNOG Mapper (91), as they likely represent fast evolving non-DNGs.

In order to identify enabler changes uniquely associated with de novo gene birth, we inspected the alignments from the Blast searches between *B. rapa* orphan genes and the genome of 45 Brassicaceae (**S2 Table**). After applying the same approach described to filter *A. thaliana* putative DNGs, we obtained 754 *B. rapa*-specific de novo genes.

Examination of DNG coding and protein sequence features

DNA sequence features. Length of genes, predicted coding sequences (CDS) and (where available) UTRs were retrieved from gff files of the assemblies of each species. The GC-content of each coding region was calculated using an in-house perl script. Transposable element (TE) genome coordinates were obtained from the TAIR10 gff3 file in the *A. thaliana* TAIR repository and from the Brapa_genome_v3.0_TE.gff file downloaded from http://brassicadb.cn/#/Download/ for *B. rapa*.

718 *O. sativa* de novo gene IDs were retrieved from supplementary information in Zhang et al. (8).
719 *O. sativa* release v3 coding region fasta sequences were downloaded from the Gramene
720 repository (http://ftp.gramene.org/oge/release-3/fasta/oryza_sativa/dna/). Rice protein fasta
721 sequences were obtained translating the CDS sequences using the ORF finder program in the
722 SMS suite (92). The initial set of 50,556 genes was parsed to retain only the longest isoform of
723 each locus and to remove genes with premature stop codons, leaving 38,748 genes.
724
725 Microsatellites were retrieved from the coding regions of each gene using Tandem Repeat Finder
726 v4.09 (93) with default options. Overlap of TEs and microsatellites with the coding region and
727 gene distance from TEs were obtained using bedtools (94). The coding potential was estimated
728 using the Coding Potential Calculator (72). The codon adaptation index was calculated using the
729 'cai' tool in the EMBOSS suite (95).
730
731 Kozak scores were computed as the sum of the logs of nucleotide probabilities within a window
732 of 12 bp around the ATG start codon (96), based on probabilities extracted from all genes in each
733 species genome:

734 $$s = \sum_{i=-6}^{5} \log\left(p(n_i)\right)$$

735
736 where n_i is the observed nucleotide at position *i* relative to the start of the ATG. The scores are
737 generally negative with a mean around -5.1, but the closer to zero, the more like the consensus
738 sequence (AAAAAAATGGCG) they are. This is similar, but not identical, to the
739 acAACAATGGC consensus sequence for terrestrial plants (97), reflecting biases that promote

740 translation initiation by the ribosome. The nucleotide probability profiles surrounding start
741 codons in *A. thaliana, B. rapa,* and *O. sativa*, as well as other diverse plant genomes (*Petunia*
742 *inflata*: 36,489 genes; *Quercus robur*: 25,808 genes) are highly similar, although genes in *O.*
743 *sativa* appear to have a relaxed constraint in the nucleotide following the ATG, whereas it is
744 guanine over 50% of the time in the other two species (see **S1 Fig**), which could be related to
745 fact that rice is a monocot and thus distantly related to the dicot family Brassicaceae. The
746 preference for adenines upstream of the ATG in plants is much less pronounced than in
747 nucleotide profiles of Kozak sequences in other eukaryotes (e.g. human, *Drosophila*) (see **S1**
748 **Fig**).
749
750 *Protein sequence features*. Protein domains were obtained from the NCBI Conserved Domain
751 Database (98). Protein structural disorder was calculated using IUPred2 (99) after removing
752 cysteines from the protein sequences in order to account for the possible presence of the disulfide
753 bonds, which can strongly affect ISD estimates (100). Transmembrane helices were estimated
754 using TMHMM Server v. 2.0 (101). The identified helices may represent either transmembrane
755 structures or signal peptides, which tend to occur within the first 60 amino acid in the N-terminal
756 of the protein. Proteins were conservatively assigned putative transmembrane helices if they
757 contained one or more predicted helices and at least 18 amino acids in a helix past the first 60
758 amino acids. Signal peptides were predicted using TargetP2.0 (102). The isoelectric point of each
759 protein was calculated using the Sequence Manipulation Suite Protein Isoelectric Point tool (92).
760 All programs above were run using default settings.
761

762 *Functional genomic features in* A. thaliana. The *A. thaliana* TAIR10 gff3 file
763 "TAIR10_GFF3_genes.gff" in the TAIR repository was used to obtain the length of "5'UTR
764 length" and "3'UTR length" and the number of coding exons ("#Exons"). The proportion of the
765 coding region overlapping with transposable elements (TEs; "TEs in CDS (bp)" and "%TEs in
766 CDS") and the gene distance to the nearest TE ("TEdist") were calculated using bedtools (94)
767 and the genome coordinate of TAIR10 coding exons and TEs. Genome coordinate of TEs were
768 obtained from the gff3 file "TAIR10_GFF3_genes_transposons.gff" available in the TAIR
769 repository. Possible regulatory motifs ("#Motifs promoter") in the promoter regions of DNGs
770 and AGs were identified using the MEME suite (103). The DNA sequences corresponding to the
771 300bp upstream of the transcription start site of each TAIR10 gene were retrieved using bedtools
772 and screened using the MEME Streme tools (104).
773
774 Transcription factor binding site (TFBS) information was retrieved from the Plant *cis*-Map
775 genome browser (http://ucsc.gao-lab.org/index.html). The Conserved TFBS dataset included
776 binding sites deposited in the PlantRegMap database (105, 106). Conservation of TFBSs was
777 assessed using multiple genome alignments across species from the Plant *cis*-Map genome
778 browser conservation track. Binding sites with at least 50% of their sequence falling within
779 conserved elements were considered conserved ("#Conserved TFBSs"). The pipeline to identify
780 putative functional transcription factor binding sites ("#FUN TFBSs") is described in Tian et al.
781 (105). The AtRegNet confirmed TFBSs dataset ("#AtRegNet TFBSs") was downloaded from the
782 AGRIS database (107). The bedtools suite was used to extract the DNA sequences of the 200bp
783 upstream of transcription start site of TAIR10 genes, corresponding to the putative promoter
784 regions, and intersected with TFBS genome coordinates.

Transcription quantification features were obtained from a study of 18 natural *A. thaliana* accessions (108) study. The average ("RNAseq AVG") and maximum ("RNAseq MAX") expression across 48 samples, reported as log(rpkm) values, were calculated for each gene across 45 samples. Expression in only 1 out of 45 samples was also added ("RNAseq <2 samples"). Average ("RP AVG") and maximum ("RP MAX") ribosome profiling (RP) expression data, reported as log(rpkm) values, were also calculated for root tissues including control and deficient phosphorous nutrition conditions in a total of 25 samples (109). The maximum RP expression was calculated only for genes expressed in at least two samples.

The "Missense variant" feature represents the number of missense (nonsynonymous) substitutions divided by the length of the coding in each gene ("Missense variation"). Missense substitutions were obtained from the 1001 *A. thaliana* Genomes portal (https://1001genomes.org/index.html).

Protein domains and protein-protein interactions (PPIs) were obtained from the files TAIR10_all.domains ("#HMMPfam Domain") and TairProteinInteraction 20090527.txt ("#PPIs", "#PPIs (w/predicted)"), respectively, from the TAIR repository. The PPI data contain interaction annotations extracted from the literature by TAIR and BIOGRID (110). The frequency of three different categories of amino acids, "Tiny", "Aromatic" and "Acidic", were obtained using the pepstats program in the EMBOSS suite (95).

We gathered phenotypic information using data deposited in the TAIR repository containing phenotypic data extracted from the literature by TAIR. Data files names: TAIR_Phenotypes_9-2019.txt ("TAIR_Phenotypes_9-2019.txt"), Locus_Germplasm_Phenotype_20190630.txt ("LGP_20190630.txt"), Locus_Germplasm_Phenotype_20130122 ("LGP_20130122"). Additionally, data from manually curated meta-analysis of loss-of-function phenotypes (111) ("Lloyd and Meinke 2012") and from a high-throughput phenotype screening of annotated genes with unknown function (112) ("Luhua, et al. 2013") were included.

Machine learning training and testing

Balanced sets of DNGs and AGs were selected for ML training and testing within each species. For instance, in *A. thaliana* 331 AGs were randomly chosen among the 26,423 available AGs. The Decision Tree (DT) classifier was trained using the *scikit-learn* package in Python (113). The feed-forward fully-connected neural network (NN) with a single hidden layer with 20 hidden nodes was also trained with the MLPClassifier implementation in *scikit-learn*, using *tanh* activation functions and the 'adam' solver. Log transformations were applied to length-based features (Gene length, CDS length, Distance from TEs) and to functional genomic features with RPKM in *A. thaliana*.

Feature importance (normalized decrease in Gini impurity index at each node where a feature is used, weighted by the fraction of training examples represented at those nodes) was calculated using the 'feature_importance' attribute of the DecisionTreeClassifier generated by *scikit-learn*. In some runs, "At-least-one-domain" was the feature at the root of the tree, and in other runs,

"Coding potential (cpc2)" represented the splitting feature at the root. Thus, we estimated the importance of features by averaging them over multiple runs.

Scripts and data from this study are available at https://github.com/ioerger2/DNG. This repository contains instructions on how to run Decision Tree and Neural Network training and testing. A python script generates accuracies, confusion matrices, decision trees and feature importances for each dataset. The *A. thaliana* features are available in the repository.

Statistical analyses

All statistical analyses were performed in R (cit). In MLA testing, accuracy corresponds to the sum of true positives (TPs) and true negatives (TNs), divided by all genes: (TP+TN)/(TP+FP+TN+FN). Sensitivity and specificity are represented by true positive rate (TPR=TP/(TP+FN)) and true negative rate (TNR=TN/(TN+FP)). Test sets and training sets are always kept disjoint (with no overlap of genes) in all runs.

Acknowledgements

The authors would like to thank Michael Dickens and the Texas A&M University High Performance Research Computing facility for assistance with performing homology search analyses on HPC clusters.

Author Contributions

CC, AEP and TRI conceived and designed the study. CC, AO and TRI analyzed the data. TRI.

CC, AEP and TRI wrote the paper.

REFERENCES

1. Cui L, Wall PK, Leebens-Mack JH, Lindsay BG, Soltis DE, Doyle JJ, et al. Widespread genome duplications throughout the history of flowering plants. Genome research. 2006;16(6):738-49.
2. Lin R, Ding L, Casola C, Ripoll DR, Feschotte C, Wang H. Transposase-derived transcription factors regulate light signaling in Arabidopsis. Science (New York, NY). 2007;318(5854):1302-5.
3. Flagel LE, Wendel JF. Gene duplication and evolutionary novelty in plants. New Phytologist. 2009;183(3):557-64.
4. Panchy N, Lehti-Shiu M, Shiu S-H. Evolution of gene duplication in plants. Plant physiology. 2016;171(4):2294-316.
5. Wendel JF, Jackson SA, Meyers BC, Wing RA. Evolution of plant genome architecture. Genome biology. 2016;17(1):37.
6. Qiao X, Li Q, Yin H, Qi K, Li L, Wang R, et al. Gene duplication and evolution in recurring polyploidization–diploidization cycles in plants. Genome biology. 2019;20(1):1-23.
7. Li L, Wurtele ES. The QQS orphan gene of Arabidopsis modulates carbon and nitrogen allocation in soybean. Plant Biotechnol J. 2015;13(2):177-87.
8. Zhang L, Ren Y, Yang T, Li G, Chen J, Gschwend AR, et al. Rapid evolution of protein diversity by de novo origination in Oryza. Nat Ecol Evol. 2019;3(4):679-90.
9. Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TCG. More than just orphans: are taxonomically-restricted genes important in evolution? Trends in Genetics. 2009;25(9):404-13.
10. Yang X, Jawdy S, Tschaplinski TJ, Tuskan GA. Genome-wide identification of lineage-specific genes in Arabidopsis, Oryza and Populus. Genomics. 2009;93(5):473-80.
11. Donoghue MT, Keshavaiah C, Swamidatta SH, Spillane C. Evolutionary origins of Brassicaceae specific genes in Arabidopsis thaliana. BMC Evol Biol. 2011;11:47.
12. Arendsee ZW, Li L, Wurtele ES. Coming of age: orphan genes in plants. Trends Plant Sci. 2014;19(11):698-708.
13. Richardson AO, Palmer JD. Horizontal gene transfer in plants. Journal of experimental botany. 2007;58(1):1-9.

14. Wickell DA, Li FW. On the evolutionary significance of horizontal gene transfers in plants. New Phytologist. 2020;225(1):113-7.
15. Joly-Lopez Z, Forczek E, Hoen DR, Juretic N, Bureau TE. A gene family derived from transposable elements during early angiosperm evolution has reproductive fitness benefits in Arabidopsis thaliana. PLoS Genet. 2012;8(9):e1002931.
16. Lisch D. How important are transposons for plant evolution? Nat Rev Genet. 2013;14(1):49-61.
17. Gould SJ, Vrba ES. Exaptation - a Missing Term in the Science of Form. Paleobiology. 1982;8(1):4-15.
18. Van Oss SB, Carvunis AR. De novo gene birth. PLoS Genet. 2019;15(5):e1008160.
19. McLysaght A, Hurst LD. Open questions in the study of de novo genes: what, how and why. Nat Rev Genet. 2016;17(9):567-78.
20. Bornberg-Bauer E, Hlouchova K, Lange A. Structure and function of naturally evolved de novo proteins. Curr Opin Struct Biol. 2021;68:175-83.
21. Li D, Dong Y, Jiang Y, Jiang H, Cai J, Wang W. A de novo originated gene depresses budding yeast mating pathway and is repressed by the protein encoded by its antisense strand. Cell Res. 2010;20(4):408-20.
22. Bungard D, Copple JS, Yan J, Chhun JJ, Kumirov VK, Foy SG, et al. Foldability of a Natural De Novo Evolved Protein. Structure. 2017;25(11):1687-96 e4.
23. Baalsrud HT, Torresen OK, Solbakken MH, Salzburger W, Hanel R, Jakobsen KS, et al. De Novo Gene Evolution of Antifreeze Glycoproteins in Codfishes Revealed by Whole Genome Sequence Data. Mol Biol Evol. 2018;35(3):593-606.
24. Zhuang X, Yang C, Murphy KR, Cheng CC. Molecular mechanism and history of non-sense to sense evolution of antifreeze glycoprotein gene in northern gadids. Proc Natl Acad Sci U S A. 2019;116(10):4400-5.
25. Li CY, Zhang Y, Wang Z, Zhang Y, Cao C, Zhang PW, et al. A human-specific de novo protein-coding gene associated with human brain functions. PLoS Comput Biol. 2010;6(3):e1000734.
26. Samusik N, Krukovskaya L, Meln I, Shilov E, Kozlov AP. PBOV1 is a human de novo gene with tumor-specific expression that is associated with a positive clinical outcome of cancer. PLoS One. 2013;8(2):e56162.
27. Suenaga Y, Islam SM, Alagu J, Kaneko Y, Kato M, Tanaka Y, et al. NCYM, a Cis-antisense gene of MYCN, encodes a de novo evolved protein that inhibits GSK3beta resulting in the stabilization of MYCN in human neuroblastomas. PLoS Genet. 2014;10(1):e1003996.
28. Begun DJ, Lindfors HA, Thompson ME, Holloway AK. Recently evolved genes identified from Drosophila yakuba and D. erecta accessory gland expressed sequence tags. Genetics. 2006;172(3):1675-81.
29. Begun DJ, Lindfors HA, Kern AD, Jones CD. Evidence for de novo evolution of testis-expressed genes in the Drosophila yakuba/Drosophila erecta clade. Genetics. 2007;176(2):1131-7.
30. Knowles DG, McLysaght A. Recent de novo origin of human protein-coding genes. Genome Res. 2009;19(10):1752-9.
31. Ruiz-Orera J, Hernandez-Rodriguez J, Chiva C, Sabido E, Kondova I, Bontrop R, et al. Origins of De Novo Genes in Human and Chimpanzee. PLoS Genet. 2015;11(12):e1005721.
32. Murphy DN, McLysaght A. De novo origin of protein-coding genes in murine rodents. PLoS One. 2012;7(11):e48650.

936 33. Prabh N, Rodelsperger C. De Novo, Divergence, and Mixed Origin Contribute to the
937 Emergence of Orphan Genes in Pristionchus Nematodes. G3 (Bethesda). 2019;9(7):2277-86.
938 34. Cai J, Zhao R, Jiang H, Wang W. De novo origination of a new protein-coding gene in
939 Saccharomyces cerevisiae. Genetics. 2008;179(1):487-96.
940 35. Carvunis AR, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, et al.
941 Proto-genes and de novo gene birth. Nature. 2012;487(7407):370-4.
942 36. Vakirlis NN, Hebert AS, Opulente DA, Achaz G, Hittinger CT, Fischer G, et al. A
943 molecular portrait of de novo genes in yeasts. Mol Biol Evol. 2017;35(3):631-45.
944 37. Blevins WR, Ruiz-Orera J, Messeguer X, Blasco-Moreno B, Villanueva-Canas JL,
945 Espinar L, et al. Uncovering de novo gene birth in yeast using deep transcriptomics. Nat
946 Commun. 2021;12(1):604.
947 38. Durand E, Gagnon-Arsenault I, Hallin J, Hatin I, Dube AK, Nielly-Thibault L, et al.
948 Turnover of ribosome-associated transcripts from de novo ORFs produces gene-like
949 characteristics available for de novo gene emergence in wild yeast populations. Genome Res.
950 2019;29(6):932-43.
951 39. Poretti M, Praz CR, Sotiropoulos AG, Wicker T. A survey of lineage-specific genes in
952 Triticeae reveals de novo gene evolution from genomic raw 1 material. bioRxiv. 2022.
953 40. Campbell MA, Zhu W, Jiang N, Lin H, Ouyang S, Childs KL, et al. Identification and
954 characterization of lineage-specific genes within the Poaceae. Plant Physiol. 2007;145(4):1311-
955 22.
956 41. Li ZW, Chen X, Wu Q, Hagmann J, Han TS, Zou YP, et al. On the Origin of De Novo
957 Genes in Arabidopsis thaliana Populations. Genome Biol Evol. 2016;8(7):2190-202.
958 42. Lin H, Moghe G, Ouyang S, Iezzoni A, Shiu SH, Gu X, et al. Comparative analyses
959 reveal distinct sets of lineage-specific genes within Arabidopsis thaliana. BMC Evol Biol.
960 2010;10:41.
961 43. Wu DD, Wang X, Li Y, Zeng L, Irwin DM, Zhang YP. "Out of pollen" hypothesis for
962 origin of new genes in flowering plants: study from Arabidopsis thaliana. Genome Biol Evol.
963 2014;6(10):2822-9.
964 44. Xu Y, Wu G, Hao B, Chen L, Deng X, Xu Q. Identification, characterization and
965 expression analysis of lineage-specific genes within sweet orange (Citrus sinensis). BMC
966 Genomics. 2015;16:995.
967 45. Jiang M, Dong X, Lang H, Pang W, Zhan Z, Li X, et al. Mining of Brassica-Specific
968 Genes (BSGs) and Their Induction in Different Developmental Stages and under
969 Plasmodiophora brassicae Stress in Brassica rapa. Int J Mol Sci. 2018;19(7).
970 46. Jiang M, Zhan Z, Li H, Dong X, Cheng F, Piao Z. Brassica rapa orphan genes largely
971 affect soluble sugar metabolism. Hortic Res. 2020;7(1):181.
972 47. Cui X, Lv Y, Chen M, Nikoloski Z, Twell D, Zhang D. Young Genes out of the Male: An
973 Insight from Evolutionary Age Analysis of the Pollen Transcriptome. Molecular plant.
974 2015;8(6):935-45.
975 48. Moyers BA, Zhang J. Evaluating Phylostratigraphic Evidence for Widespread De Novo
976 Gene Birth in Genome Evolution. Mol Biol Evol. 2016;33(5):1245-56.
977 49. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+:
978 architecture and applications. BMC Bioinformatics. 2009;10:421.
979 50. Domazet-Loso T, Brajkovic J, Tautz D. A phylostratigraphy approach to uncover the
980 genomic history of major adaptations in metazoan lineages. Trends in genetics : TIG.
981 2007;23(11):533-9.

982 51. Casola C. From De Novo to "De Nono": The Majority of Novel Protein-Coding Genes
983 Identified with Phylostratigraphy Are Old Genes or Recent Duplicates. Genome Biol Evol.
984 2018;10(11):2906-18.
985 52. Tautz D, Domazet-Loso T. The evolutionary origin of orphan genes. Nat Rev Genet.
986 2011;12(10):692-702.
987 53. Wissler L, Gadau J, Simola DF, Helmkampf M, Bornberg-Bauer E. Mechanisms and
988 dynamics of orphan gene emergence in insect genomes. Genome Biol Evol. 2013;5(2):439-55.
989 54. Weisman CM, Murray AW, Eddy SR. Many, but not all, lineage-specific genes can be
990 explained by homology detection failure. PLoS Biol. 2020;18(11):e3000862.
991 55. Vakirlis N, Carvunis AR, McLysaght A. Synteny-based analyses indicate that sequence
992 divergence is not the main source of orphan genes. Elife. 2020;9.
993 56. Moyers BA, Zhang J. Phylostratigraphic bias creates spurious patterns of genome
994 evolution. Mol Biol Evol. 2015;32(1):258-67.
995 57. Moyers BA, Zhang JZ. Further Simulations and Analyses Demonstrate Open Problems of
996 Phylostratigraphy. Genome Biology and Evolution. 2017;9(6):1519-27.
997 58. Vakirlis N, McLysaght A. Computational Prediction of De Novo Emerged Protein-
998 Coding Genes. Methods Mol Biol. 2019;1851:63-81.
999 59. Washburn JD, Mejia-Guerra MK, Ramstein G, Kremling KA, Valluru R, Buckler ES, et
1000 al. Evolutionarily informed deep learning methods for predicting relative transcript abundance
1001 from DNA sequence. Proc Natl Acad Sci U S A. 2019;116(12):5542-9.
1002 60. Lee D, Zhang J, Liu J, Gerstein M. Epigenome-based splicing prediction using a
1003 recurrent neural network. PLoS Comput Biol. 2020;16(6):e1008006.
1004 61. Wang H, Cimen E, Singh N, Buckler E. Deep learning for plant genomics and crop
1005 improvement. Curr Opin Plant Biol. 2020;54:34-41.
1006 62. Gao Q, Jin X, Xia E, Wu X, Gu L, Yan H, et al. Identification of Orphan Genes in
1007 Unbalanced Datasets Based on Ensemble Learning. Front Genet. 2020;11:820.
1008 63. Zhang X, Xuan J, Yao C, Gao Q, Wang L, Jin X, et al. A deep learning approach for
1009 orphan gene identification in moso bamboo (Phyllostachys edulis) based on the CNN +
1010 Transformer model. BMC Bioinformatics. 2022;23(1):162.
1011 64. Japkowicz N, Stephen S. The class imbalance problem: A systematic study. Intelligent
1012 Data Analysis. 2002;6(5):429-49.
1013 65. Wilson BA, Foy SG, Neme R, Masel J. Young Genes are Highly Disordered as Predicted
1014 by the Preadaptation Hypothesis of De Novo Gene Birth. Nat Ecol Evol. 2017;1(6):0146-146.
1015 66. Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, et al. The genome of the mesopolyploid
1016 crop species Brassica rapa. Nat Genet. 2011;43(10):1035-9.
1017 67. Cheng F, Wu J, Fang L, Wang X. Syntenic gene analysis between Brassica rapa and
1018 other Brassicaceae species. Front Plant Sci. 2012;3:198.
1019 68. Tong C, Wang X, Yu J, Wu J, Li W, Huang J, et al. Comprehensive analysis of RNA-seq
1020 data reveals the complexity of the transcriptome in Brassica rapa. BMC genomics. 2013;14(1):1-
1021 10.
1022 69. Cheng F, Sun R, Hou X, Zheng H, Zhang F, Zhang Y, et al. Subgenome parallel selection
1023 is associated with morphotype diversification and convergent crop domestication in Brassica
1024 rapa and Brassica oleracea. Nat Genet. 2016;48(10):1218-24.
1025 70. Cheng F, Wu J, Cai C, Fu L, Liang J, Borm T, et al. Genome resequencing and
1026 comparative variome analysis in a Brassica rapa and Brassica oleracea collection. Sci Data.
1027 2016;3:160119.

71. Zhang L, Cai X, Wu J, Liu M, Grob S, Cheng F, et al. Improved Brassica rapa reference genome by single-molecule sequencing and chromosome conformation capture technologies. Hortic Res. 2018;5:50.
72. Kang YJ, Yang DC, Kong L, Hou M, Meng YQ, Wei L, et al. CPC2: a fast and accurate coding potential calculator based on sequence intrinsic features. Nucleic Acids Res. 2017;45(W1):W12-W6.
73. Wang HC, Hickey DA. Rapid divergence of codon usage patterns within the rice genome. BMC Evol Biol. 2007;7 Suppl 1:S6.
74. Vakirlis N, Acar O, Hsu B, Castilho Coelho N, Van Oss SB, Wacholder A, et al. De novo emergence of adaptive membrane proteins from thymine-rich genomic sequences. Nat Commun. 2020;11(1):781.
75. Luis Villanueva-Canas J, Ruiz-Orera J, Agea MI, Gallo M, Andreu D, Alba MM. New Genes and Functional Innovation in Mammals. Genome Biol Evol. 2017;9(7):1886-900.
76. Zhao L, Saelao P, Jones CD, Begun DJ. Origin and spread of de novo genes in Drosophila melanogaster populations. Science (New York, NY). 2014;343(6172):769-72.
77. Bitard-Feildel T, Heberlein M, Bornberg-Bauer E, Callebaut I. Detection of orphan domains in Drosophila using "hydrophobic cluster analysis". Biochimie. 2015;119:244-53.
78. Mukherjee S, Panda A, Ghosh TC. Elucidating evolutionary features and functional implications of orphan genes in Leishmania major. Infect Genet Evol. 2015;32:330-7.
79. Schmitz JF, Ullrich KK, Bornberg-Bauer E. Incipient de novo genes can evolve from frozen accidents that escaped rapid transcript turnover. Nat Ecol Evol. 2018;Epub.
80. Basile W, Sachenkova O, Light S, Elofsson A. High GC content causes orphan proteins to be intrinsically disordered. PLoS Comput Biol. 2017;13(3):e1005375.
81. Charte F, Rivera AJ, del Jesus MJ, Herrera F. Addressing imbalance in multilabel classification: Measures and random resampling algorithms. Neurocomputing. 2015;163:3-16.
82. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research. 2002;16:321-57.
83. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, et al. The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. 2012;40(Database issue):D1202-10.
84. Li YR, Liu MJ. Prevalence of alternative AUG and non-AUG translation initiators and their regulatory effects across plants. Genome Res. 2020;30(10):1418-33.
85. Heames B, Schmitz J, Bornberg-Bauer E. A Continuum of Evolving De Novo Genes Drives Protein-Coding Novelty in Drosophila. J Mol Evol. 2020;88(4):382-98.
86. Vakirlis N, Duggan KM, McLysaght A. De novo birth of functional, human-specific microproteins. bioRxiv. 2021.
87. Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210-2.
88. Luo X, Chen S, Zhang Y. PlantRep: a database of plant repetitive elements. Plant Cell Rep. 2022;41(4):1163-6.
89. Laurin-Lemay S, Brinkmann H, Philippe H. Origin of land plants revisited in the light of sequence contamination and missing data. Curr Biol. 2012;22(15):R593-4.
90. Steinegger M, Salzberg SL. Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank. Genome Biol. 2020;21(1):115.

1073 91. Huerta-Cepas J, Szklarczyk D, Heller D, Hernandez-Plaza A, Forslund SK, Cook H, et al.
1074 eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource
1075 based on 5090 organisms and 2502 viruses. Nucleic Acids Res. 2019;47(D1):D309-D14.
1076 92. Stothard P. The sequence manipulation suite: JavaScript programs for analyzing and
1077 formatting protein and DNA sequences. Biotechniques. 2000;28(6):1102, 4.
1078 93. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids
1079 Res. 1999;27(2):573-80.
1080 94. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic
1081 features. Bioinformatics. 2010;26(6):841-2.
1082 95. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open
1083 Software Suite. Trends in genetics : TIG. 2000;16(6):276-7.
1084 96. Kozak M. An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs.
1085 Nucleic Acids Res. 1987;15(20):8125-48.
1086 97. Lutcke HA, Chow KC, Mickel FS, Moss KA, Kern HF, Scheele GA. Selection of AUG
1087 initiation codons differs in plants and animals. Embo J. 1987;6(1):43-8.
1088 98. Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, et al. CDD:
1089 NCBI's conserved domain database. Nucleic Acids Res. 2015;43(Database issue):D222-6.
1090 99. Dosztanyi Z. Prediction of protein disorder based on IUPred. Protein Sci.
1091 2018;27(1):331-40.
1092 100. Uversky VN, Dunker AK. Understanding protein non-folding. Biochim Biophys Acta.
1093 2010;1804(6):1231-64.
1094 101. Moller S, Croning MD, Apweiler R. Evaluation of methods for the prediction of
1095 membrane spanning regions. Bioinformatics. 2001;17(7):646-53.
1096 102. Almagro Armenteros JJ, Salvatore M, Emanuelsson O, Winther O, von Heijne G,
1097 Elofsson A, et al. Detecting sequence signals in targeting peptides using deep learning. Life Sci
1098 Alliance. 2019;2(5).
1099 103. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME Suite. Nucleic Acids Res.
1100 2015;43(W1):W39-49.
1101 104. Bailey TL. STREME: Accurate and versatile sequence motif discovery. Bioinformatics.
1102 2021.
1103 105. Tian F, Yang DC, Meng YQ, Jin J, Gao G. PlantRegMap: charting functional regulatory
1104 maps in plants. Nucleic Acids Res. 2020;48(D1):D1104-D13.
1105 106. Jin J, Tian F, Yang DC, Meng YQ, Kong L, Luo J, et al. PlantTFDB 4.0: toward a central
1106 hub for transcription factors and regulatory interactions in plants. Nucleic Acids Res.
1107 2017;45(D1):D1040-D5.
1108 107. Yilmaz A, Mejia-Guerra MK, Kurz K, Liang X, Welch L, Grotewold E. AGRIS: the
1109 Arabidopsis Gene Regulatory Information Server, an update. Nucleic Acids Res.
1110 2011;39(Database issue):D1118-22.
1111 108. Gan X, Stegle O, Behr J, Steffen JG, Drewe P, Hildebrand KL, et al. Multiple reference
1112 genomes and transcriptomes for Arabidopsis thaliana. Nature. 2011;477(7365):419-23.
1113 109. Bazin J, Baerenfaller K, Gosai SJ, Gregory BD, Crespi M, Bailey-Serres J. Global
1114 analysis of ribosome-associated noncoding RNAs unveils new modes of translational regulation.
1115 Proc Natl Acad Sci U S A. 2017;114(46):E10018-E27.
1116 110. Oughtred R, Rust J, Chang C, Breitkreutz BJ, Stark C, Willems A, et al. The BioGRID
1117 database: A comprehensive biomedical resource of curated protein, genetic, and chemical
1118 interactions. Protein Sci. 2021;30(1):187-200.

111. Lloyd J, Meinke D. A comprehensive dataset of genes with a loss-of-function mutant phenotype in Arabidopsis. Plant Physiol. 2012;158(3):1115-29.
112. Luhua S, Hegie A, Suzuki N, Shulaev E, Luo X, Cenariu D, et al. Linking genes of unknown function with abiotic stress responses by high-throughput phenotype screening. Physiol Plant. 2013;148(3):322-33.
113. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011;12:2825-30.