

## G OPEN ACCESS

**Citation:** Hua Z, Early MJ (2019) Closing target trimming and CTTdocker programs for discovering hidden superfamily loci in genomes. PLoS ONE 14 (7): e0209468. https://doi.org/10.1371/journal. pone.0209468

**Editor:** Christian Schönbach, Nazarbayev University, KAZAKHSTAN

Received: December 5, 2018

Accepted: June 12, 2019

Published: July 2, 2019

**Copyright:** © 2019 Hua, Early. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The full CTT and CTTdocker packages are available in two repositories hosted by GitHub at https://github. com/hua-lab/ctt and https://github.com/hua-lab/ cttdocker, respectively.

**Funding:** This work was supported by a National Science Foundation CAREER award (MCB-1750361 to ZH) and Ohio University Program to Aid Career Exploration (PACE) fund (ZH9299110 to ZH). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. **RESEARCH ARTICLE** 

# Closing target trimming and CTTdocker programs for discovering hidden superfamily loci in genomes

## Zhihua Hua<sup>1,2</sup>\*, Matthew J. Early<sup>1,3</sup>

Department of Environmental and Plant Biology, Ohio University, Athens, Ohio, United States of America,
Interdisciplinary Program in Molecular and Cellular Biology, Ohio University, Athens, Ohio, United States of America,
Department of Electrical Engineering and Computer Science, Ohio University, Athens, Ohio, United States of America

\* hua@ohio.edu

## Abstract

The contemporary capacity of genome sequence analysis significantly lags behind the rapidly evolving sequencing technologies. Retrieving biological meaningful information from an ever-increasing amount of genome data would be significantly beneficial for functional genomic studies. For example, the duplication, organization, evolution, and function of superfamily genes are arguably important in many aspects of life. However, the incompleteness of annotations in many sequenced genomes often results in biased conclusions in comparative genomic studies of superfamilies. Here, we present a Perl software, called Closing Target Trimming (CTT), for automatically identifying most, if not all, members of a gene family in any sequenced genomes on CentOS 7 platform. To benefit a broader application on other operating systems, we also created a Docker application package, CTTdocker. Our test data on the F-box gene superfamily showed 78.2 and 79% gene finding accuracies in two well annotated plant genomes, Arabidopsis thaliana and rice, respectively. To further demonstrate the effectiveness of this program, we ran it through 18 plant genomes and five nonplant genomes to compare the expansion of the F-box and the BTB superfamilies. The program discovered that on average 12.7 and 9.3% of the total F-box and BTB members, respectively, are new loci in plant genomes, while it only found a small number of new members in vertebrate genomes. Therefore, different evolutionary and regulatory mechanisms of Cullin-RING ubiquitin ligases may be present in plants and animals. We also annotated and compared the Pkinase family members across a wide range of organisms, including 10 fungi, 10 metazoa, 10 vertebrates, and 10 additional plants, which were randomly selected from the Ensembl database. Our CTT annotation recovered on average 14% more loci, including pseudogenes, of the Pkinase superfamily in these 40 genomes, demonstrating its robust replicability and scalability in annotating superfamily members in any genomes.

**Competing interests:** The authors have declared that no competing interests exist.

## Introduction

Genome growth and innovation is largely attributed to the expansion of gene superfamilies [1– 3]. Through both whole-genome duplication (WGD) and small-scale duplication (SSD) events, members of a gene family can increase dramatically [4-6]. Newly synthesized members in a family have been commonly recognized to experience either pseudogenization or fixation in a genome through four processes: 1) conservation if gene dosage is beneficial [1], 2) neofunctionalization if a novel function is acquired in one copy [1], 3) subfunctionalization if daughter copies split the function of the ancestral copy [7], or 4) specialization if all daughter copies perform new functions differing from their ancient copy [8]. Previous studies have suggested that dosage-balance constraints guarded duplicated copies against loss by determining the stoichiometry of duplicated products [9, 10]. Interestingly, families involving different cellular functions have various birth over death ratios. Transcription factors tend to have a much higher fixation rate than others, such as G-protein-coupled chemosensory receptors for finding food, odorants and pheromones and the immunoglobulins involving in the primary immune defense in vertebrates, and the nucleotide-binding site-leucine-rich repeat (NBS-LRR) receptors whose functions are largely found in defending pathogens in plants [11-16]. The significant contribution in genome evolution and the striking variance of evolutionary processes of members within and between families place the studies of gene duplication an important role in decoding the function of genomes. To reach this, precisely finding the total members of a gene family in a genome would be the first effort.

Since the era of draft sequencing of human genome, finding genes in massive genomic DNA sequences has drawn significant attention in understanding the structure and function of genomes [17]. Early in this century, many computational methods were developed in order to predict genes in a piece of DNA or in an entire genome precisely. Overall, these methods can be divided into two categories: sequence similarity-based extrinsic methods and machinelearning curated intrinsic approaches. The former employs known query sequences to align a target sequence to discover the structure of an unknown locus, while the latter utilizes dynamic programming to identify genes by locating the gene elements (e.g., specific signals in the 5'- or 3'-ends of genes) [18]. Due to the extreme variability of gene elements, the intrinsic (ab initio) methods often yield more false predictions than the extrinsic evidence-based predictions [18]. For example, prior ab initio methods predicted Caenorhabditis elegans and Homo sapiens genes with an accuracy rate of 50% and 20%, respectively [19, 20]. Certainly, the accuracy of both types of gene annotation methods has been significantly increased since both the scale and the information of known genes are now dramatically expanded and improved. Recently developed next generation mRNA sequencing-based methods have increased the accuracy of de novo genome annotations to an average of 57.6% on C. elegans genes [21, 22]. However, the human ENCODE genome annotation assessment project, EGASP, suggested that the mRNA and protein sequences-based annotation programs still yielded the most accurate prediction of gene models [20].

With the booming of next generation sequencing technologies [23], large-scale genome sequencing has become routine in many individual laboratories, while genome annotation and analysis remains challenging. Considering the importance of gene superfamily studies and the contemporary large accumulation of accomplished genome projects, we developed a Perl package, named Closing Target Trimming (CTT), for superfamily annotation. This package incorporates the well-established sequence similarity-based GeneWise annotation tool [24] to automatically identify and re-annotate most, if not all, members of a gene superfamily in a genome, thus benefiting comparative genomic studies.

The algorithm of CTT was first developed during the course of studying the phylogenetics of the *F-box* gene superfamily in plants [5]. When we retrieved the *F-box* gene members in prior annotations from 18 sequenced plant genomes, including both seedless and seed plants, we noticed that there were many novel *F-box* putative loci left unannotated in genomes. Those loci are either lacking clear gene signals that allow them to be discovered *ab initio* or too close with one another to be uncovered by GeneWise annotation, since the latter only finds the best homologue of a known query protein sequence [5, 24]. Recently, we have also employed this algorithm to study the duplication mechanism of the ubiquitin and ubiquitin-like protein (in whole called ubiquiton) family in 50 sequenced plant genomes [6]. CTT has been demonstrated as an effective tool to retrieve most, if not all, members of a gene superfamily in genomes we have analyzed. To make this program readily available for the biological science community, we organized and improved the scripts used in the previous two studies and compiled it into one integrated package that allows researchers with little programming experience to use.

## Materials and methods

## Package development

The CTT package is written in Perl (Version 5) and has been tested on the CentOS 7 Linux operating system. The CTTdocker is a Docker application package of CTT that has been tested on CentOS 7, Ubuntu 18.04.2 LTS, and macOS Mojave (version 10.14). The results described in this paper are based on the performance of the CTT package running on the CentOS 7 Linux operating system.

## Similarity-based gene annotation

Based on our previous studies on CTT-based re-annotation of *F-box* and *ubiquiton* genes in plant genomes [5, 6], we designed to retrieve a 10,000 nucleotide (nt) genomic sequence sandwiching the start coordinate of a target domain for identifying a novel family locus. Since a non-plant gene may be longer than a plant gene on size, we also examined the effectiveness of CTT prediction by expanding the size of a putative locus to 100,000 nts. A putative novel locus is positioned by tBLANTn search using the query of a collection of seed sequences, which represent the domain signature of a target family, against the genome sequence. Only if a new locus identified by CTT iteration search (see algorithm below) encompasses the tBLASTndefined locus would the genomic sequence be further analyzed. To eliminate a potential false prediction due to the inaccuracy of one reference protein sequence, we applied at least the top two best homologous protein sequences, based on bit scores of BLASTx search [25], to back search the genomic DNA sequence for predicting the transcript model of a putative novel locus using GeneWise [24]. If the locus is consistently predicted as a protein-coding gene by at least two reference sequences, the coding sequence identified with the highest GeneWise score was taken as the best prediction of this putative new locus. The resulting peptide sequence is further examined for the presence of a target family domain using HMMER search (http:// hmmer.org) against the Pfam-A database (Version 32.0, https://pfam.xfam.org, E-value <1).

## CTT iteration search algorithm

The genomic DNA sequence with 10,000 nts is likely to contain more than one locus. This is particularly common in many superfamily genes that were tandemly duplicated. This prevents the identification of a putative new locus by GeneWise because GeneWise only aligns a reference protein sequence to the best homologous region [24]. To expose this putative new locus defined by tBLASTn search, unrelated adjacent DNA sequences are iteratively trimmed and

removed by CTT. First, the 10,000 nt genomic DNA sequence is used to identify a locus that best matches to a homologous protein sequence by GeneWise (GeneWise score  $\geq$  50). The DNA sequence of a predicted locus is trimmed and removed if it does not encompass the tBLASTn-defined region. The remaining DNA sequence is then iterated for a sequential BLASTx-GeneWise-trimming search until the tBLASTn-defined region is identified to be within the GeneWise predicted locus. Otherwise, the tBLASTn-defined region is not considered to be a new family locus.

## Data resources

To test the effectiveness of this CTT package, a genome sequence file, a generic feature format 3 (GFF3) file, and a protein sequence file from in total 63 sequenced genomes (S1 Table) were retrieved from Phytozome (Version 12, https://phytozome.jgi.doe.gov) or Ensembl (Release 94, https://useast.ensembl.org). The seed files for three test superfamilies, the F-box (PF00646), the BTB (PF00651) and the Pkinase (PF00069) protein families, and the Pfam-A database were downloaded from Pfam (Version 32.0, https://pfam.xfam.org).

## Results

## Structure and implementation of CTT

The CTT package is built on the CentOS 7 Linux operating system. The source code is permitted to use and distribute under the GNU General Public License Version 3.0. The package requires the Pfam-A database and six program dependencies, including BioPerl (https:// bioperl.org), BLAST [25], HMMER (http://hmmer.org), GeneWise [24], PfamScan (https:// pfam.xfam.org) and CD-HIT [26]. The users are strongly recommended to follow the steps described in a *README.md* file under the main directory /*ctt* to install the database and the dependencies on CentOS 7. We found that compiling BioPerl could be really problematic for researchers with little programming experience. The instructions provided in "*Install BioPerl*" in the *README.md* file can significantly reduce this hassle. After BioPerl and BioPerl-Run are installed, the "*make all*" function in *makefile* under the directory /*ctt/dependencies* should help to install the Pfam-A database and the remaining program dependencies smoothly.

Before running the program, the users need to collect the information of target genomes and seed sequences for a family of interest as input files into the directories /ctt/species\_databases and /ctt/seeds, respectively. Since we are interested in both known and unknown members of a family, a genome sequence file, a generic feature format 3 (GFF3) file, and a protein sequence file from a previous genome annotation are required to be deposited in the directory of /*ctt/spe*cies\_databases. The family members from prior annotations are used as reference peptide (Ref\_Pep) sequences to predict the gene structure of an unknown locus using GeneWise. Therefore, we recommend adding the same set of three files from well-annotated genomes to increase the prediction accuracy. The package is capable of working on multiple genomes at one time. Once the number of genomes is decided and all the required information is collected, a tab file, named "organismal\_genome\_gff3\_proteome\_files.tab", needs to be prepared with each row containing the names of a genome sequence file, a GFF3 file, and a protein sequence file of one genome, each of which is separated with a tab space. The seed file of a family can be found and downloaded from the family webpage on Pfam (https://pfam.xfam.org). For example, a FASTA-formatted seed file of the F-box family can be found at https://pfam.xfam.org/family/f-BOX#tabview=tab3. A detailed method is also available on protocols.io ([27]; S1 Fig). This file should be placed in the directory /ctt/seeds. All the input sequence files are in a FASTA format, which allows a BioPerl module, Bio::DB::Fasta, to parse the sequences. The species genome and protein sequence files should also be converted to BLASTable databases in the same directory

using the *makeblastdb* function as instructed in the *README.md* file or as described in the manual of BLAST [25]. A step-by-step protocol can be found on protocols.io [27].

The operation code of this program is formulated as one Perl command in the terminal under the directory of the package /*ctt*. The users simply need to enter a code formatted as "*perl ctt.pl -seed family\_seed\_file.txt -f Pfam\_family\_id -superfamily simplified\_family\_id\_you\_named*" to start the program. The *ctt.pl* should automatically read the input files from /*ctt/species\_databases* and /*ctt/seeds* to finish the annotation. There are in total seven steps, each of which is operated by an annotation module saved in the directory /*ctt/annotation\_modules*. In accordance, there are 7 outputs that the program produces and saves stepwise in a final /*ctt/ctt\_output* directory (Table 1).

In Step 1, *prior\_annotation\_search.pm* applies the seed sequences as query to BLASTp [25] a previously-annotated protein database. Pfam domains present in each hit protein sequence are obtained by *PfamScan.pl* search against the Pfam-A database that is automatically downloaded, processed, and saved under the directory */ctt/databases/pfam*. The Pfam domain search in *PfamScan.pl* is essentially done by HMMER3 (http://hmmer.org). The latter applies profile hidden Markov models (profile HMMs) to determine the presence of a sequence homologous to a Pfam domain, whose information is represented by a multiple sequence alignment and a hidden Markov model (HMM) in the Pfam-A database (https://pfam.xfam.org). In the current version of CTT, we apply 1e-5 and 1 as an E-value cutoff for the BLASTp and HMMER searches, respectively. Both values are artificially determined based on our previous studies that may assist to include most known members in both the *F-box* and the *ubiquiton* families [5, 6].

The module of Step 2, named *reference\_pep\_best\_dm\_search.pm*, is designed to organize the full-length protein sequences of family members predicted in Step 1 and retrieve the best family domain sequence in each protein sequence as a new seed sequence. Both the full-length sequences and the combined new and Pfam family seed sequences are subject to CD-HIT [26] to remove redundant ones. The retaining full-length and seed sequences are used as reference and seed sequences for later GeneWise annotation and tBLASTn search, respectively.

Step 3 applies the non-redundant seed sequences as queries to search a genome database using tBLASTn with an e-value cutoff  $\leq$  1e-5. Since the tBLASTn search generates a large number of hits, we considered two hits the same if the middle coordinate of one resides in the region of the other. In addition, we use AWK (comes with centOS) to compare the start and the end coordinates of one tBLASTn hit with those of the previously-annotated genes, whose information is retrieved from the GFF3 file. Since tBLASTn finds a region aligned with a family domain, the hit should reside in an annotated gene if it is part of the gene. Those hits are not considered as new loci and are removed from further analyses. This step is accomplished using a module named *finding\_putative\_new\_loci.pm*.

Step	Module	Output
1	prior_annotation_sesarch	Protein sequences of a target family
2	reference_pep_best_dm_search	Non-redundant reference peptide sequences and combined new and Pfam seed sequences
3	finding_putative_new_loci	Genomic DNA sequences of putative new loci each having 10,000 nts
4	closing_target_trimming	Trimmed genomic DNA sequences only containing the target locus
5	annotate_the_best_model	GeneWise annotation results: the genomic DNA, transcript, and peptide sequences for each protein coding gene are predicted. Pseudogenes are also identified.
6	correct_gdna_coordinates	Fine tune the genomic coordinates of a protein coding gene
7	pfamscan	Finalize the new gene members

Table 1. Annotation modules and outputs of CTT.

Step 4 runs the CTT algorithm which is coded in a module called *closing\_target\_trimming*. *pm*. In order to gain a genomic region long enough to cover a putative new locus, we added a sequence of 5,000 nts on both sides of the start point of a retaining tBLASTn hit in Step 3. Based on our previous study on *F-box* genes, the longest *F-box* locus in plants is less than 10,000 nts [5]. A DNA sequence of 10,000 nts should be sufficient to contain a new locus in plants. For other organisms, the users can modify *finding\_putative\_new\_loci.pm* in Step 3 to retrieve a longer region if needed. This 10,000 nt DNA sequence containing an unknown locus is used to BLASTx the Ref\_Pep database obtained in Step 2. The best hit is then used as a protein template in GeneWise to predict the structure of a new gene. However, the best homology search algorithm may not let GeneWise find the locus containing the target hit if there are multiple homologous sequences in this DNA fragment. To expose the target locus, this module applies up to six iterations of GeneWise annotation and trimming to shorten the genomic sequence until a hit-containing locus is discovered (Fig 1). The iteration stops if the GeneWise score is lower than 50 or no BLASTx hit is obtained.

To reduce the bias caused by a problematic reference sequence, Step 5 calls the module, *annotate\_the\_best\_model.pm*, to utilize the protein sequences of the three best BLASTx hits as GeneWise templates to annotate the gene structure of an unknown target locus uncovered in Step 4. Only if one of the three gene models (peptide-coding gene, pseudogene without Gene-Wise transcript, or pseudogene containing early stop codons) is consistently predicted by at least two reference sequences would we consider it a good annotation. The transcript predicted with the highest GeneWise score is taken as the final prediction for a peptide-coding gene.

Step 6 retrieves the correct genomic coordinates of a putative new member and Step 7 confirms the presence of a family domain in the new annotated protein sequence (Table 1).

## Sensitivity and specificity test

We tested both sensitivity and specificity of CTT on gene finding using the *F-box* gene family (Pfam ID: PF00646) as an example in two well annotated plant genomes, *Arabidopsis thaliana* and rice (*Oryza sativa*). First, we used *prior\_annotation\_search.pm* to identify the full set of *F*-



**Fig 1. A schematic illustration of the CTT algorithm.** The "x" indicates a genomic region that encodes a target family domain. The coordinates of "x" is defined by a tBLASTn search using the query of seed sequences of the target family. "Hit1, 2,...n" indicate the top hits from a BLASTx search using a 10,000 nt genomic DNA sequence containing "x" as query against protein sequences of previously-annotated family members. The process is iterated for up to six times to shorten the genomic sequence until a new locus containing "x" is discovered. The iteration stops if the GeneWise score is lower than 50 or no BLASTx hit is obtained.

Species*	Sens	sitivity		(Sn+Sp)/2		
	Pseudo-genes	Protein Coding	Pident (%)	Integrity (%)	Mean (%)	(%)
Ath	15	59	99.5 ± 1.5	95.1 ± 12.1	97.3 ± 6.2	78.2
Osa	20	65	99.2 ± 2.2	86.9 ± 20	93.0 ± 10	79

#### Table 2. The sensitivity and specificity of CTT in finding plant F-box gene members.

\* The full names of the species along with their abbreviations in Tables 2-6 and 8 are as listed in S1 Table.

https://doi.org/10.1371/journal.pone.0209468.t002

*box* gene members in the previously predicted proteomes of Arabidopsis and rice. To assess the performance, we randomly selected 100 *F-box* members in each species and removed their annotations in the corresponding proteome. We then ran through CTT to examine 1) how many of these 100 *F-box* genes were rediscovered in each genome, and 2) how good their annotations are compared to the previous protein sequences. To increase both the sensitivity and specificity of CTT, we collected the annotation datasets from 18 plant genomes (S1 Table) as we did previously [5].

We defined that a test *F-box* locus was identified if the midpoint coordinate of the CTTannotated locus is within the test locus. We found that 75 and 85 test *F-box* loci were rediscovered by CTT in Arabidopsis and rice genomes, respectively. We also found by CTT that 15 and 20 test *F-box* loci from Arabidopsis and rice, respectively, were predicted to have no transcript available or express a protein with premature stop codons, indicative of pseudogenes (Table 2). This result further confirmed our previous discoveries showing that a significant number of *F-box* loci are pseudogenized [5, 10].

To test the specificity of CTT, we compared both sequence identity and integrity of the predicted protein sequences with previous annotations using BLASTp analysis. Between the two genomes compared, CTT seems to perform slightly better in *Arabidopsis* to predict the full length of a test *F-box* locus (Table 2, *p-value* = 0.005, Student's *t*-test). This is likely due to a close relationship between a reference genome, *A. lyrata*, and *A. thaliana*. However, for all the protein sequences predicted, CTT yielded >99% sequence identities with the original predictions, which are consistent with the outstanding performance of GeneWise annotation if a good reference sequence is available [24]. After averaging both specificity and sensitivity, CTT obtained 78.2 and 79% accuracy in finding a new *F-box* locus in Arabidopsis and rice, respectively (Table 2).

## Size comparison of the F-box and the BTB superfamilies in plants

We next applied CTT to search the members of two gene superfamilies in 18 selected plant genomes. In addition to the *F-box* gene family, we selected the *bric-a-brac/tramtrack/broad complex* (*BTB*) family (Pfam ID: PF00651) for comparison. The criteria of selecting these two families are 1) that both encode a substrate receptor in a Cullin-RING E3 ubiquitin ligase complex important for regulating a vast number of metabolic and signaling pathways in all eukary-otic organisms [28], and 2) that both families are largely expanded in plant genomes [5, 29].

Consistent with our phylogenetic studies on the *F*-box gene superfamily, we obtained a similar number of *F*-box genes in all 18 plant genomes ( $\rho = 0.96$ , *p*-value = 6.7e-06, Spearman's correlation test) (Fig 2A and S1–S4 Files). The slight difference in the number of *F*-box genes found in the two studies may have resulted from 1) the upgrade of genome and F-box seed sequences, and 2) the improvement of CTT in this package. For example, in this package, we incorporated GFF3 files to precisely locate known loci in a prior annotation (Step 3, *finding\_putative\_new\_loci.pm*, Table 1). However, our previous study used BLAT search [30] to locate an annotated gene, which may result in inaccurate genomic coordinates [5].





https://doi.org/10.1371/journal.pone.0209468.g002

After CTT annotation, we discovered that, on average, 12.7 and 9.3% of the total members in the *F*-box and *BTB* families, respectively, are new loci (Tables 3 and 4 and S5–S8 Files). Although a slightly higher proportion of *F*-box genes were discovered than that of *BTB* genes (*p*-value < 0.05, Student's *t*-test), the percentiles of new members in these two families are significantly correlated ( $\rho = 0.69$ , *p*-value = 0.001, Spearman's correlation test) (Fig 2B). Therefore, there are various annotation qualities in different sequenced genomes, further highlighting the effectiveness of this package in helping identify most, if not all, superfamily members in a genome for comparative genomic studies.

# CTT annotation of the *F-box* and the *BTB* superfamilies in non-plant genomes

Since the CTT algorithm was first developed to study the *F-box* gene superfamily in plants [5], we questioned whether this newly designed CTT package was also able to annotate new family members in non-plant genomes. We selected the genomes of human and its two-close relatives, the chimpanzee and mouse, as well as nematode and budding yeast as a test dataset (S1

Species	Prior Annotation	CT	Γ Annotation	Total	New Finding
		Pseudogenes	Protein Coding Genes		(% of total)
Aly	940	63	57	1060	11.3
Ath	771	33	12	816	5.5
Bdi	960	41	33	1034	7.2
Сра	139	53	10	202	31.2
Cre	95	0	0	95	0.0
Csa	210	3	7	220	4.5
Gma	723	39	14	776	6.8
Mes	274	5	4	283	3.2
Mgu	750	49	35	834	10.1
Mtr	1000	81	35	1116	10.4
Osa	831	34	34	899	7.6
Рра	325	6	18	349	6.9
Ptr	534	18	6	558	4.3
Rco	209	12	8	229	8.7
Sbi	772	35	25	832	7.2
Smo	213	80	133	426	50.0
Vvi	140	14	37	191	26.7
Zma	426	36	23	485	12.2

#### Table 3. CTT annotation of *F-box* genes in 18 plant genomes.

https://doi.org/10.1371/journal.pone.0209468.t003

Table). For a strict comparison, we applied CTT to annotate the *F*-box and the *BTB* families in these five genomes.

Unlike the large and unstable expansion of the *F*-box superfamily in plants (Table 3; [5]), we found that its size in three closely related vertebrates is fairly small and stable (Table 5 and

Species	Prior Annotation	СТ	T Annotation	Total	New Finding
		Pseudogenes	Protein Coding Genes		(% of total)
Aly	75	2	3	80	6.3
Ath	86	0	0	86	0.0
Bdi	197	11	1	209	5.7
Сра	43	2	2	47	8.5
Cre	70	0	0	70	0.0
Csa	55	0	0	55	0.0
Gma	194	4	3	201	3.5
Mes	67	3	2	72	6.9
Mgu	60	0	2	62	3.2
Mtr	105	2	2	109	3.7
Osa	173	7	4	184	6.0
Рра	89	1	5	95	6.3
Ptr	163	2	1	166	1.8
Rco	43	2	2	47	8.5
Sbi	171	9	3	183	6.6
Smo	30	2	36	68	55.9
Vvi	45	7	3	55	18.2
Zma	125	20	6	151	17.2

Table 4. CTT annotation of *BTB* genes in 18 plant genomes.

S9–S12 Files), suggesting that plants and vertebrates may have adapted a different mechanism in regulating the evolution and the function of the F-box-mediated protein ubiquitylation system. Consistent with the broad post-translational regulatory function of the ubiquitin-26S proteasome system (UPS), CTT annotation discovered that the three vertebrates have a larger BTB family than many flowering plant species (Table 6 and S13–S16 Files). In addition, the expansion of this family in these genomes is like the *F-box* superfamily showing little size variation. More interestingly, only a small proportion of members from both families were newly discovered by CTT in these five non-plant genomes analyzed (Tables 5 and 6).

We initially thought that the low number of new loci detected by CTT in these five nonplant genomes might be due to a larger genomic size of vertebrate genes than that of plant genes. We revised the *finding\_putative\_new\_loci.pm* module in Step 3 to retrieve a genomic DNA sequence of 100,000 nts instead of 10,000 nts for downstream CTT annotation. This modification did not make any improvement. Indeed, we were able to use the same size of genomic sequences to identify a large proportion of new *Pkinase* members in 40 randomly selected genomes (see below). Therefore, we think a genomic DNA sequence of 10,000 nts, albeit arbitrarily, would be a good size for searching the presence of a superfamily gene. In general, this size can also be rationalized based on the longest locus of known family members in a genome and easily adjusted by changing *\$flank\_head* and *\$flank\_tail* values in the module of Step 3, *finding\_putative\_new\_loci.pm*.

## Replicability and scalability test

Our CTT program was initially designed to study the duplication mechanism of *F*-box genes in plants. In this paper, we demonstrated its capability of finding both *F*-box genes and *BTB* genes in five non-plant genomes as well. To further test its replicability and scalability, we examined its performance in finding a UPS-independent large family, *Pkinase*, in a large set of sequenced genomes, ranging from 10 fungi, 10 metazoa, 10 vertebrates, and 10 additional plants, which were randomly selected from the Ensembl genome collection (https://www. ensembl.org). The Pkinase family proteins involve protein phosphorylation, a process for adding a phosphate group onto a substrate. Within this family, the catalytic domain is called Pkinase, which contains an average of 267 amino acids, 2.5 and 5.5 fold longer than the BTB (108 amino acids) and the F-box domain (48 amino acids), respectively. To benchmark the performance of CTT, we also carefully recorded and compared its execution time in each of the 7 modules among these four groups of genomes (Table 7). As expected, the total execution time varied significantly amongst the four groups, ranging from 2.33 hr to 45.28 hr, which in part depended on the sizes of the proteomes and genomes being annotated. Interestingly, CTT finished the annotation in the fungi and metazoa with a speed similar to that in the plants (0.04 ms/nt) and vertebrates (0.01 ms/nt), respectively, probably reflecting their similar genome organizations.

Species	Prior Annotation	CT	T Annotation	Total	New Finding	
		Pseudogenes Protein Coding Genes			(% of total)	
Cae	388	0	1	389	0.3	
Hom	70	0	0	70	0.0	
Mus	82	0	0	82	0.0	
Pan	69	4	0	73	5.5	
Sac	14	0	0	14	0.0	

Table 5. CTT annotation of F-box genes in 5 non-plant genomes.

Species	Prior Annotation	CT	T Annotation	Total	New Finding
		Pseudogenes	Protein Coding Genes		(% of total)
Cae	126	0	0	126	0.0
Hom	151	0	0	151	0.0
Mus	183	1	0	184	1.1
Pan	155	6	4	165	6.1
Sac	1	0	0	1	0.0

#### Table 6. CTT annotation of BTB genes in 5 non-plant genomes.

https://doi.org/10.1371/journal.pone.0209468.t006

Our annotation results in these four large groups of diverse organisms again showed its merit in finding new superfamily members. We discovered that an average of 4.4%, 15.2%, 27.6%, and 9.9% more *Pkinase* members in the fungi, metazoa, vertebrates, and plants, respectively (Table 8). To validate they were truly new loci, we took *Amborella* as an example to visualize its 34 new peptide loci in JBrowser that is available on Phytozome (https://phytozome.jgi.doe.gov) (S17–S19 Files and Fig 3). Interestingly, no previous loci were duplicated in any one of these new members. Encouragingly, exons in eight of them indeed contain an expression sequence tag (EST), suggesting they are transcribed. Moreover, 16 of them overlap with a BLASTx hit when searched against the proteomes of Arabidopsis, rice and Chlamydomonas, providing additionally evolutionary evidence. Our CTT annotation also discovered that 33.6%, 29.0%, 32.8%, and 64.5% of the new members in fungi, metazoa, vertebrates, and plants, respectively, contain inframe shift and early stop codons, which are characteristics of pseudogenes. Collectively, our CTT annotation of the *Pkinase* family in these 40 genomes further highlighted the importance of comprehensive annotations in understanding the evolutionary process of superfamilies.

## Discussion

The invention of next-generation sequencing technologies revolutionized genomic studies by dramatically reducing the cost and increasing the speed in various genome sequencing

	Fungi	Metazoa	Vertebrates	Plants
Step 1 (hr)*	0.99	1.65	2.92	8.13
Step 2 (hr)	0.00	0.00	0.00	0.00
Step 3 (hr)	0.17	2.13	17.68	6.75
Step 4 (hr)	0.33	2.50	7.92	7.53
Step 5 (hr)	0.82	5.45	15.97	13.67
Step 6 (hr)	0.00	0.05	0.33	0.13
Step 7 (hr)	0.02	0.22	0.47	0.25
Total (hr)	2.33	12.00	45.28	36.47
Size of Proteome (# of proteins)	8.0E+04	2.3E+05	2.9E+05	3.4E+05
Size of Genomes (# of nt)	2.2E+08	5.2E+09	3.2E+10	3.5E+09
Annotating Speed (ms/nt)	0.04	0.01	0.01	0.04

Table 7. Performance benchmark test of CTT in annotating Pkinase genes in 40 genomes.

\*Performance was evaluated based on execution time of each step for annotating *Pkinase* genes in 10 genomes that were randomly selected from each organismal group. Abbreviations: ms, milliseconds; hr, hour; nt, nucleotides. The program was performed in a Dell PowerEdge R710 server with X5670 2 x 12-core CPUs (2.93 GHz) and 64GB ram.

Group	Species*	Prior Annotation	C	<b>FT</b> Annotation	Total	New Finding (% of total)
1			Pseudogenes	Protein Coding Genes	_	
Fungi	Asp	134	3	8	145	7.6
	Bip	116	1	3	120	3.3
	Cer	96	1	2	99	3
	Cry	91	1	2	94	3.2
	Сув	91	3	1	95	4.2
	Ent	40	1	3	45	11.1
	Esc	90	1	2	93	3.2
	Sug	86	0	6	92	6.5
	Tal	186	2	2	190	2.1
	Yar	85	0	0	85	0
Metazoa	Cae	458	27	75	560	18.2
	Cra	239	7	44	290	17.6
	Dro	572	1	0	573	0.2
	Hel	337	1	7	345	2.3
	Lep	178	6	21	205	13.2
	Nem	244	8	31	283	13.8
	Pri	380	22	70	472	19.5
	Sar	167	1	27	195	14.4
	Ste	225	15	193	433	48
	Tri	187	6	4	197	5.1
Vertebrates	Cal	814	64	37	915	11
	Cho	243	216	299	758	67.9
	Сур	680	29	123	832	18.3
	Fic	329	13	98	440	25.2
	Нар	764	13	75	852	10.3
	Ict	558	144	99	801	30.3
	Mes	507	40	90	637	20.4
	Mus	930	40	60	1030	9.7
	Pel	387	10	112	509	24
	Тир	287	165	254	706	59.3
Plants	Amb	460	107	38	607	24.2
	Chl	538	0	0	538	0
	Gal	77	0	0	77	0
	Man	1097	42	39	1178	6.9
	Mus	803	37	20	860	6.6
	Ory	1118	52	8	1178	5.1
	Pru	1105	38	12	1155	4.3
	Sel	784	144	168	1132	30.7
	The	927	10	5	942	1.6
	Vig	693	84	82	859	19.3

### Table 8. CTT annotation results of Pkinase genes in 40 genomes.

 $^{*}\mbox{Genomes}$  were randomly selected from each organismal group available at the Ensembl database.

https://doi.org/10.1371/journal.pone.0209468.t008

projects. The recently-emerging third generation (or long-read) sequencing technologies would further this revolution by sequencing genomes in a much more affordable, accurate, and effective manner [23]. The contemporary genome sequencing technologies have already

#### AmbPK\_N0001

501,000	1.000,000	1.008,000	2.041,84	2,000,000	1.001,000	1.001.000	4,001,000	4,000,000	1.001.000	_
			0000	Arth of 8 waterpoor	- Land vid automot	er Justikes Josities & Gr	4			
2007.600		UNI AND	2.00 (10)		2,001,000	24	al and	2007.000		
Ver Bat Reuts		-	THE BASTANT OF LINKS	annos in the second second	4.88-17)					
			Decision and and and	ACCOUNT NAMES OF TAXABLE	4.50-17					
Tarachi										
Table Investigation (TVT										
BATTA AND Character										

#### AmbPK\_N0004

200 400			1,000	1,00	1.400	1.000	e 2,000	2,000	2.404	1.60	2,896	1.000	1,000	0.4.0	1,600	3,800	4,000	4,00
					D Q	Q Q B	AND VIA and	ar cai - (ar	dest to set				(B)					
4.590			AND			3.754			443			4,00			A.10			
Iter Hot Reults				100.0	RELATION	713 solver!	to because in-range	+3.34-88				_						
				101.0	NUT ANT ANT	CP-1 and other	to be a set of the set of the	- 3.54-86	-	-		_						
					1114031451	and costs	er The ball of the	Aux = 3.4e-5										
Descript																		
PALA Assertiant EXT																		
BATT AutoOhery/Ros					-				-	-	-	-						
AmbP	κı	<b>VOO</b>	07															

3.742.886	2,742,363	2342.600	134,460	3.762.835	2,742,840
iter Bat Analis	PRICES BLACT AND IN A DOMESTIC	Teacher ( )+ Teace + 6.5e-34]			
	ynersep BLAST AN \$1, yr \$1, soulesennys	Teacher (p. Teace + 6.5a-34)			
Terretal					
PhDA have blied EST					
- address and an address					

#### AmbPK\_N0010

Anticervita 113hopods v1	8 - File View Help					
0 200,000 400,00			CO.000 1400.000 2.000.000 2.000.000	LACKING LACKING LACKING C MARKANG MARKANG MILANGKAN (CC)	1000,000 1,200,000 1,400,000	- 1401400 LMILIO 4.001400 4.00
38	641100	642,300	842,300	842.308	642.300	841,400
() User Bast Results		starting BASE And Lot 4 in	selved0004 keeset ja Voien = 1.3e-ME			
		you see BART AND OLD	endfold00044 (entre) (e datas = 1.3e 7/)			
		VALUE OF ALL AND ALL A	auforitioosa teature' (r 2664 - 2 feriol)			
		you see BART ANT (112)	extendiood texture! (# Plake = 1.5e 71)			
O Resortet						
0 PHile Assemblies (107						
a second second second						

#### AmbPK\_N0013

Anborella trkhopoda v1.5	- Fie Vev												
806,800	1,000,000	1,306,800	1.690.201	2,000,000	1061400	3,399,300	4,000,000	4,000,000	1,00,000	5,306,899	4.000,000	6,000,000	1,006,000
			A Contraction of the second	A c	QQ ~	(HILLIGHTHEODINE)	· Arthorizond	WEDERERGOVER, KE	0000 y Co 🖬 🖌				
UNLAW .	4351,605			00,00		4202.875		4,010,000		425	44		4213,250
O that Best Reads				your see BLA	Mach Charles	ADDOT 14 Martinet 24 Fe	A = 3.94140						
				year and \$1.4	steriotioe#	1600014 https://jp.14	64+530143		-				
Terestet													
TROA Assembled ETT													
BATT AND TRANSFORM			-	_	_	_	_	_					

#### AmbPK\_N0016

Anberelia trichepoda v1.0 -	File New 16	*											88 9ws	falsener
9 200,000 400,000	80.20 80.0	an 1,000,000	1200.000 1400.00	0 1.00L000 1.000	000 1,001,000	LINKAR	1,400,000	1,000,000	1.800.000	1.000,000	3,05,000	140,000	3,800,000	1.851,000
			A 9	Q Q Andres			CAROOM AND/S		- 4 0-					
610,000		67.01		812,008		\$13,750			13.800			00,00		
O Ver Red Reals			VINCTION 10, AUX AV	VALUE ADDRESS AND ADDRESS AND	unt le line - 1 de d		•	-						
			PROTON BLADT AN	as shinefor \$1.0	141 p-10.e = 13+3	4								
O Taxorde														
O AREA REPORTED	-					-								
O BATT A 44 Charyflur	-													

#### AmbPK\_N0019

1.000,000	4,000,000 0	100.000	1,000,000	1000.000	1.000.000	4,000,000	1,00,000	12,093,000	1.00
		Q Q Q Q Q Q	And year and see	66 - anit ut & andre		66 🌛 🕞			
AMUST	0.451,000	140,313		8.400.301	8.401.0	13	5.481,730		640,575
) that Bed Reads		PER MICH.A	T ANT I TO MANAGED TO	Instant (Printer - Carl	10				
		Plan and BLA	er kalt (st. e. juskassoo	10000010-0000-000	4				
Datarist									
111 Init Associated III									
BATTANETCharyRise				_					

#### AmbPK\_N0022

2008 40,00 50,00	ARLAND LARLAND		LANCON CONCERNMENT	THE LANSE LEADER	LANDON LANDON LONGING LANDON		
2.86.251 The Had Rey's	2201500	LOM/DO	L 10.00	A.101.050	A START		
		Planning BLADT AND UT D, ADMONTO	ETT Neurol ( Hide + Q	_			
Twartst		in Prosterior Contenter	Tank				
TALA Association (ST		-					
		_					

#### AmbPK\_N0025

Ancorsta s'shop		110 100															
a 200.000	400,000	603,505	A00.001	1,005,000	1,000,000	1,406,000	1.000.000	1,800,000	2,005,098	1,251,200	146.00	2.805.000	1.003,000	1.000	3,095,000	1401800	3,800,000
				Ð	(-) 0	0.0.	a	1.0.00000000	ALC: NO.	a.xahattees	(2110).011		0.0+			_	
CATE ALL		2,015,400			2.818,500			13/8,500			13340		-	201,010		2,810	1/90
O that Best Results				70	CHIQUEARTANT	propaga	SOLA PRIMA		a .								
· hearist																	
© PhSA hearthfiel (07																	
	-	_	_			_										_	

#### AmbPK\_N0028

Anberella 552 epoda v1.0 - File	View Help					oothen Maxw
e xxx,000 1,000,000 1,00	0.000 1.000,000 1.000 000 000 000 000 000 000 000 00	C. C. C. C. MARCHE	4303,000 5,00,000	LIND.202 6.000,400 6.000,000	100,000 1,000,000	A200203 A200200
0.000.000	4,812,005	execution playing Relation to a surface	6.807.000	6.00.00	6.062.005	
O Tenologi		ever 37 maint deals yet 6 mailtainth	an Jin an Art	alai. Aw <sup>15</sup> , v <sup>1</sup> 2, acadhadodor (2.39) v <sup>1</sup> .8, acadhadolor (2.39)		
O Pala Asserbiat 007						
C ButTX AndetOrlany/Rea				-		

#### AmbPK\_N0031

1 100,000	1.000,000	1,000,000	2,000,000	2,800,008	1.09,00	2,805,800	6,000,000	4,000,000	4,010,000	1,00,000
			DQQQ	Q AND IN E AND	NUMBER + AND VIE AND	Part 1010000000000000000000000000000000000	NORA (2) (10) 🔤 📣 🔤	-		
1.0%	766	2214,000		2,314,400	1.	DN.800	110		3,275,008	
) for Eatherals	2	onterview of the	urbiologia ketral (a)	vision = 1.30-502		1.1.1.1				
O Terecipt										
O Bild Assessibled ET										
BATTANE/DanyRis										
Amelo										
AMDPK	_N00	34								
AMDPK	_N00	34								State Falance
	_N003	5 <b>4</b>	138.80 1483	100 100	100.001 2,00.00	1.00.000 1.4	0.00 1.00.00	1.000	64 10 1.01.30	Mars Palacian Lettion p
	_NO0	34 ⊛⊜		0 100.00 0 (milior and	LANCOM LONGING	LINUX LA	na Lanan Korfai 🎜 De	100.00 I 100.	00 1.01.00	Stars Fallenas
	_NO03	34 ⊛⊸		00 100000 Q (mitro 1 par	1.000 000 2.000.000 1000000 = 1.000.000 2.000.000	LINUM 14	8.00 L800.00 norfas <b>ai G</b> e 87.00	110.00	100 LINL 00	New Palaces 140030 J
AMDPK 1903 - 40,695 Jan Jan Ver Bart haufs	_NO03	34 • • •		en unione part	1.000-000 2.000.000 0000009[ + [AAY5 x1-3 aug 2.000-00 000079 (based (s-2000 + 2		0.00 1.000.00 00 / 01 ∳ (D 02.00	180.00 Dom.	00 1101.300	Dan Palacoa 3.405300 3 1203.0
AMDPK recents to the point of all 100 M 40.000 100 M 40.0000 100 M 40.0000	_N003	34 ©@			1.000.001 1.000.000 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.0000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.00 1.000.000	LINUM LA	1000 10000 100-3 (a) 2 (b) 2 Anto 7 Anto y 12 puedrados 7 Anto y 12 puedrados	2,000,000 0,000 2,000,000 1,000,000 (2,000,0,0,00 20,000,000 (2,000,0,0,0,0) 20,000,000 (2,000,0,0,0,0)	64 19 LIN.30	Dat Nilota Lettor J
AMDP'K	_NOU	34 			1.000000 1.000.000 1.000.000 1.000.00 1.000.00 1.000.00 1.000.00	Landor V. Jacob and V. All	ID 200 LAND 201 ID 20 1 Cm P Cm ID 201 1 Cm ID 201 Cm	1.000.000 1.0000 1.000.001 (or Vision = 1.10. 10 Master of (or Vision = 1.10. 10	-	Start Polacina 1.405300 3 LPILX
AMDERA Internals scorepole str.d 100000 40000 Unit Part Annals Tencipe	_NOU	34 ©@			1.00000 1.00000 100000 + (AAN 0.0 pp 1.00000 100000 backst (rither of 100000 backst (rither of 100000 rither of	Landon La Barray Rati Promy Rati	10.00 LINO.00 100 - Car J Car 100 - Car	2.000.000 3.000. 	-	Statt Palarsas 140300 J
AMDDPK	NOO	54 ©@		en union Olimitativa par Independent and out and Transformed and out a control	1.00000 1.000000 000000 + [.acti u-0.000 1.00000 000000 100000 1.00000 + 0 000000 100	Lance La and the second second second from the lance second second second from the lance second s	1000 LANDON 100 A AL A CA AT AN 7 An To A AL ANALASIO 7 An To A AL ANALASIO 7 An To A ANA	2000000 0 5000 120000 21 Malant (s Mar + 1 (s 21 Malant (s Mar + 1 (s 21 Mar + 1 (s)	ed RO LINJO	Start Palarsee

#### AmbPK\_N0002

		æ		0.0.00	A 447 -10	entretori e l	and and a surface	COLUMN TRANSPORT	Track of the loss of	1 CP	1000.000	
	1.88.88		Innan	quan	3.364	76		7.88.85		1.001.000		1.807.628
Var Biet Rends				PAR INCIDENTIAL	01100608	The Ballant Ja- Villa	127448					
				PRO INCIDENT AN	or to shake	In Maket (Phila	+17+81	-				
NOA Assembled ET												
BATTA AND TRANSPORT												

#### AmbPK\_N0005

1,890,000	100,00	0 4,000,000	6,001,000 6,000	200 7,006,000	0,000,000	8,800,008	10,000,000 11	,006,880
			AND ALL AND ADDRESS +	Avb. +1.0. watercook 25# (KN.)	Caman di Ca 🥔 🖉 -			
1.00	0.300	2,360,794	2.366.000	2.366,210	2.341	200	3.348,750	
Our Bait Results		what had BLAST and		1.5e-15Q				
		year and Boated Arrite		1.5e-150				
Savariat								
Non Assessment OF								
But The Andre Charge Rose								

#### AmbPK\_N0008

MEAN LOOK	0 1.00.00	1.00.00 LAN.000	1.00.000 1.00.000	4.00.00 4.00.000	1.00.00 1.00.00	4.000,000	4,500,000	1,000,000	7.606.008	
			QQQQ me		WONDOORS READERS, ROOMER # DO	🥔 CP				_
630,70	4.325.008	4.30	70	4.301.500	4,338,750		6.20.205		4,304	24
C Deer Had Results										
		THE INC. N. A. LAND. M.D.	automotical leadership in the second	(e-1)						
		COLUMN STATE	automatical leaders in the state of	teal?						
D Transfel										
O MM Assembled ETT									-	
O BATTAND/Dhayftor										

#### AmbPK\_N0011

Andorella trichopeda v1.0 ·																
200.00 400.000	806,000	80.00	1,006,000	1,296,281	1,400,000	1.803.000	1,806,000	2,000,000	2,206,048	2,400,000	2400,000	2,000,000	1,000,000	1296,000	140.00	1.000.0
			(	(->) (0)	0.00	Anti-uti	and the second	- janh yas			- C 100	()+				
	205-004		10	4.00		267.6	*		200,750			240300			241,212	
iter Hat famile					Pharmal R.	CT AND ITS	A CONSIGNAL	estaret je klalas estaret je klalas	-14-52	:						
Turneript						849.35			4	•	- ·					
Phid Associated EXT													-			
BATTA MICHAEL THE					-	_	-									

#### AmbPK\_N0014

200.008 400,000 600.000	006,000 1,000,000 1,006,000	1.60.300 1.806.000 1.803.000 40	X-040 2,348,000 3,400,000 2,800,	00 180306 330600 33083	00 3,400,208 3,806,000	180,00 4,00,00	
	(E) (=		ANTHRONY + ANT ILL MATHROOM	FRANKET, FRANKER, M. Co. M. C. LEW			
LINKATS	1.000.000	1,848.00	1.949,210	Lastura	Lastance		1.848
) User Rist Results		PRESS NO. ANT ANT OVER	x1404000002 Autors1 (0-1004 x 5.54-104)				
		WAR HAD BEET AND YOR	and setting second in loss of the line	-			
Name at							
O Risk learning (SP							
D B. ACT Available and the							

#### AmbPK\_N0017



### AmbPK\_N0020



Anborella trici	opodav1.9 · r	ie Vew Neip								
	105.000	1,040,000	1000.000	2,800,000	2,805,890	1,09,000	1.00.000	4,00,00	4,000,000	1,001,000
				QQQQ -		11, 11 8, AND 18 10 19 10 10		CO-		
	41,200	AL3	10	#1.750		AD.300	86,216		41.88	
O Dar Bat Area				A SHART AND STORE	NATIONAL REALING IN VALUE	-2.1e8()	-			
				A REAL PROPERTY AND AND	NAMES OF TAXABLE PARTY.	-2.1em	-			
D Tereorge							-			
O PAGA Assembles	67									
O BATTANAIS	inter Tex						-			

#### AmbPK\_N0026

Anborelia trichopoda v1	A - PM	New Help										a Dana Adarmen
500,000	1.000,000	1,500,000	1000,000	2306.000	100,000 3,000,000	4,000,000	4,501,000	1,00,000	5.501.000	4,000,000	4,501,50	2,000,000
			A A	O O B	And of a material	01 - ANT A	August and a second	ermer i se	4 0-			
KATLED.		6.611,200		641.00		641.734		AND DOL		8.812.200		A412.500
O they Had Reads			your sec	RATING COM	uddell maaret je daar + 3				•			
O Teneript					en.HneeAnit.rt.	100T100017300						
O TABLA annubied EET			-					-	_			
O BATTAND/Dansfilm			_						-			

#### AmbPK\_N0029

Reserves createred at														
201,000	40,00	100,000	800,000	1.000.000	1,205 018	LANCING	1,401,000	1,800,000	2,005,000	2,206,000	1,416,000	2,830,200	2,600,000	3.305,499
			(-)	$(\Rightarrow)$ $(0)$	QQQ	ARCHINE		10,83846000	4 1010404 AD1300	6 IN - 4 15				
1212,045		1,212	200		1,213,380		LEAN			1213.345		1,213,360		1,21
O Uner Mark Results			you may like th	ANTON COM		1004123487	-							
O Teneral O														
O REALASSINGLE														

#### AmbPK\_N0032

## **CTT Annotation** PASA Assemble EST **Prior Annotation** BLASTx Ath/Osa/Cre

## AmbPK\_N0003

Line Red Reads	Liques	Ulaise	Linde	Usava	Usking	1363,560
		parang Built Arthough	production to become 1 (in Value or 1.5en 17)			
Neorigt						
Pick Resident CV						

#### AmbPK\_N0006

	500,000	0.00,000	Campion	2,806,200	1.560,000	100,000 1,000	100.00	4,50,000	Lasting	A ACK JOINT	4,000,000	1.301.000	7,000,000	1348,400	1.00
				۲		aaq -	C. yl A staffeddiri's	· Anitoria and	Autor-34617-59	6800-091 Get	📣 (Br			_	
5.857,590			10.00		6.802.808		8.105.000		8.301.58			549,900		6.887	
O loar Ba	e baselo			Tenen netr	and she wanted	territorial language and	indi								
				-1											
				HAR BE BLAST	NT. 11. 10740	OCCI Maturel (a-Value +1	(9-4)	0.02			1111				
							;	the said fl. and Area	OT LOOK HOURS	betural (s-Webe	1234-45				
O heart	*	even 37 model Apr?	Autor	60-0.309											
													_		
U PAGA AN	ankied D7			•									-		
O BARTS.															

#### AmbPK\_N0009

		096.08 107040110 + [Av	1.490.008	1.800.000	1,801,000	2,006,000	2,295,000	
	QQ MAINE	oner a law	h +1.0 methodoriss 790	P1.76202-7-6 Cr	1 1 1 m			
		261,820		281,768		261.00		41,200
DATING BLACE AND CO.	Louise in the second second	inter-Ide-60		-	-			
included and the local	, sudodiiriis tedarri ja i	ave - Mertij		_	•			
				-				
			-	-	_			-
-		_	_		-			
	ise any BATATY.	pro an 3017 APT - 5 - 5 - 5 - 5 - 5 - 5 - 5 - 5 - 5 -						

#### AmbPK\_N0012

		View Netp											
995,000	1,000,000	1,005,000	5,000.000	2.000.000 2.00	10.000 L.M.R.200	4.001,000	1,000.000	5,005,000	5.500-000	4.000.000	6.000,000	1.000	
				0000	anti yi 2 aufusto	10 + Anhoticus	ATTACA CO. A STATE		📣 (Gr)				
N,125	4.1	204,258	4.55	14,815	4.004.000		4,034,825			50A,796		4304,87	
the fact leads				and and BUAS	And of Example Control of Control	antin kan si a di antin kan si a di antin ta sakadat antin ta sakadat antin ta sakadat antin ta sakadat	n F Theorem (c) in Value – Theorem (c) in Value – CHI National (c) in Value ADDCHI National (c) in Value	8.14-16 8.14-16 8.14-26 4.14-26					
Sworigt													
PIGA Assertive ESP													
BACK AutoOtany Stor				-			_	-					

#### AmbPK\_N0015

9 MK,00 1,001,000 1,001,000 2,001,000			0 6.000,000 6.000,000	6.000,000 7.000,000 7.000,000 6.000	NO 1,00,000 1,00,0	00 L.005
LEST, AN	1,247,628	1,287,760	1,007,075	1,206,000	1,204,108	
⊖ ther Field Budts	phonomy BLAST AND phonomy BLAST AND phonomy BLAST AND phonomy BLAST AND	V.S. posteletitic for factors (p. Name - de 7 V.S. posteletitic) for the figure of the V.S. posteletitic for factors (p. Name - d. fig- V.S. posteletitic) for the figure - de 7 V.S. posteletitic for factors (p. Name - de 7 V.S. posteletitic) for factors (p. Name - d 16)	6 44 49 49 49 50 mm (8) 50 mm (8)	17 Mart, yr 6 ywdraddodd hawwr ( i'r 1944 y 1 de ef 17 Mart, yr 6 ywdraddodd hawwr ( i'r 1944 y 1 de ef		
() Teneript						
() Mid-Assembled BT						
AmbPK_N0018	3					

2 855 850	2 100 100	2.000.000	2.045.000	2.000.000	2 680 100
Rec But Keals			Net ( ( 1996 - 17.8-27)) Instit ( 1996 - 1.9-1) Net ( 1996 - 1.9-1)		
Percept					
This bounded DY					

#### AmbPK\_N0021

NO.00 1,000,000	1,00,000	2,20,000	2,506,000	1000,000	1,501,000	A005.000 A005.00	1.000.000	1,100,000	4,000,000	8,500,000	2,000,000	2,000.000	1,00,000	8,500,000	1,00,00
				→ C	0000	and yith automation	10 - ANTLO	1.00000000		P ( 10	100				
6.754	1,062,000			1,966,010		1.961.500		1.0	10,00		1,864	208		1,064,000	
O for Hat Reals			POLICE BUT	ITANK POINT	AND NO.	nt je take - 1									
			you may be a	ITAN DOLD	Avenue was	ni je tiske e lij									
O Descript						am.,21	uarter, stat	NUMBER OF STREET							
D PALA Assembled EDT															
O BLACK And Chargeline			_		-	_			-	_	-				

#### AmbPK\_N0024

200,000	40,00	#00,000	801,000	1,000,000 1,30	1,00000	1,400,000	1.600.000	1.00.00	2,306,809	1.49
			Q Q Q	Q Anti-una aurosa	THE R AND ALL ADDRESS		A .c.			
734.403	TOR ADD		100,000	756.8	a	796,065		134.328	I	04,360
Iner Bast Results		your and \$2,527	Self-Self-metallette	i haiwel (n kisar + 6.7a 76)	_					
Securit										
No. A contract Day			-		_		-	_		-
ALTERAND COMPANY					_	_	-			

#### AmbPK\_N0027

100,000	206,890	200,000	406,000	100.005	806,800	766,000	805,000	800,000	1,000,000	1,100,000	1,000,000	1,000,000	1,400,000	1,000,000
	_			→ Q	000	AND 110,108	HEE'CO + AN	n via automor	AD INVESTIGATION OF	14.00 - 4.0	1. A			
44	518.1	N0		119,300		379,292			10.00		326,760		3413	
Tanacht														
NGA Assembled 217														
and the second second second														

#### AmbPK\_N0030

98,300 1,30,000 1,80,300	1.000.000 LINE.000 3.000	ON LOLDS ADDING ADD	LIO LINCOL LINLIN 430	100,001 00,004 00,0	7.505.000 4.00.200 4.505.000 4.002.000
	(*)	QQQQ mene	Auguste - (with yild aufbeloop from	99.7794194 (c) 🤞 😂	
7,70,30	1.783.300	1,755,623	7,796,798	1,104,815	1.738,000
Dar Bad Reach		star and Build And on Line	dissioned because 1 (in Value = 3.1 ar 101)		
Paracript					
Phil Associat UT					
BAS7LAndolDiarryRice					

#### AmbPK\_N0033

		Q. B. C. Anix or C. stationer of	* ANTI VID ASPREETIS 200001 2090	ce / 64 🖌 🥔 🖽	
1311.53	100.00	140.300	1.04.80	2,252,362	2.226,068
Core Blast Results		and the set of the set of the production of the set of	je Value - bile dilj		
	÷.	e sag BLAST Arth, vi à professions haves	l ja Value o B. Ja-		
Terestal	and 17 made hits with	and address of			
INGA Assembled (07					
Butorit and Otherspillor					

Fig 3. JBrowse images showing 34 truly new *Pkinase* loci identified by CTT in the *Amborella* genome. The transcript of each CTT-annotated *Pkinase* locus was used as query to BLASTn the *Amborella* genome. The best hit was visualized in JBrowse available on phytozome. Species abbreviation: *Ath, Arabidopsis thaliana; Osa, Oryza sativa; Cre, Chlamydomonas reinhartii.* 

https://doi.org/10.1371/journal.pone.0209468.g003

made it possible to sequence the genomes of every organism on the earth, such as the Earth BioGenome Project (EBP) initiated this year [31]. However, a sequencing machine only produces a lengthy genome book composed of 4 or 5 letters, i.e., A, T, C, G, or N. To understand the organization, evolution, functions, and interactions of numerous genomes, post genome sequencing data analysis is much more challenging than ever before. For example. many biological laboratories are not yet able to take full advantage of the power of whole-genome sequencing because of the lack of bioinformatic tools and/or skills.

While many state-of-the-art bioinformatic tools have been developed to annotate genomes, our studies have shown that superfamily members were often missing in various accomplished genome projects [5, 6] (Tables 3 and 4). Studying gene duplications of a family across genomes can shed light on the evolutionary and functional mechanisms of intracellular regulatory pathways [4, 6, 10]. Yet, a biased conclusion could be made if a comprehensive understanding of the members of a gene family in genomes is not achieved. For example, it was concluded that herbaceous annual plants have a larger *F-box* gene family than woody perennial plants from a study of few plant genomes [32]. However, our later CTT-based comprehensive re-annotation of the *F-box* genes in a broad range of plant genomes did not fully support this conclusion [5]. Our previous data and the data in this work clearly show that the size of the *F-box* superfamily is not significantly correlated with the growth behavior of a plant species, indicative of a genomic drift evolutionary mechanism ([5]; Table 3). The *F-box* family in herbaceous plants, papaya (*Carica papaya*), cucumber (*Cucumis sativus*), and *Castor bean* (*Ricinus communis*), is smaller than that in the woody plants *Populus trichocarpa* and *Medicago truncatula* (Table 3).

The effectiveness of CTT in annotating the *F-box* and the *BTB* superfamilies in both plants and non-plant organisms suggests a broad application of this software. Our further analysis of its performance in annotating the *Pkinase* family in 40 diverse genomes demonstrated its robust replicability and scalability (Tables 7 and 8 and Fig 3). The availability of this CTT package will benefit both evolutionary and functional genomic studies of superfamily genes in eukaryotic organisms.

## Package availability

The full CTT and CTTdocker packages are available in two repositories hosted by GitHub at https://github.com/hua-lab/ctt and https://github.com/hua-lab/cttdocker, respectively.

## Supporting information

**S1** Table. List of 18 plant genomes and 5 non-plant genomes used in this work. (XLSX)

S1 Fig. Stepwise description for retrieving a FASTA-formatted seed sequence file of a family.

(TIF)

S1 File. Protein sequences of the *F-box* genes identified by CTT in prior annotations of 18 plant species.

(FA)

**S2** File. Protein sequences of novel *F-box* genes reannotated by CTT in 18 plant species. (FA)

**S3** File. Transcript sequences of novel *F-box* genes reannotated by CTT in 18 plant species. (FA)

S4 File. Genomic DNA sequences of novel *F-box* genes reannotated by CTT in 18 plant species.

(FA)

S5 File. Protein sequences of the *BTB* genes identified by CTT in prior annotations of 18 plant species.

(FA)

S6 File. Protein sequences of novel *BTB* genes reannotated by CTT in 18 plant species. (FA)

S7 File. Transcript sequences of novel *BTB* genes reannotated by CTT in 18 plant species. (FA)

S8 File. Genomic DNA sequences of novel *BTB* genes reannotated by CTT in 18 plant species.

(FA)

S9 File. Protein sequences of the *F*-box genes identified by CTT in prior annotations of five non-plant species.

(FA)

S10 File. Protein sequences of novel *F-box* genes reannotated by CTT in five non-plant species.

(FA)

S11 File. Transcript sequences of novel *F-box* genes reannotated by CTT in five non-plant species.

(FA)

S12 File. Genomic DNA sequences of novel *F-box* genes reannotated by CTT in five non-plant species.

(FA)

S13 File. Protein sequences of the *BTB* genes identified by CTT in prior annotations of five non-plant species.

(FA)

S14 File. Protein sequences of novel *BTB* genes reannotated by CTT in five non-plant species.

(FA)

S15 File. Transcript sequences of novel *BTB* genes reannotated by CTT in five non-plant species.

(FA)

S16 File. Genomic DNA sequences of novel *BTB* genes reannotated by CTT in five nonplant species. (FA) **S17** File. Protein sequences of novel *Pkinase* genes reannotated by CTT in *Amborella*. (FA)

**S18** File. Transcript sequences of novel *Pkinase* genes reannotated by CTT in *Amborella*. (FA)

S19 File. Genomic DNA sequences of novel *Pkinase* genes reannotated by CTT in *Amborella*.

(FA)

## Acknowledgments

We thank each genome-sequencing project for providing the annotation and genome sequences and the Pfam database for its data resources. We also thank numerous researchers for providing the free open source codes in the five dependencies adopted in this package.

## **Author Contributions**

Conceptualization: Zhihua Hua.

Data curation: Zhihua Hua.

Formal analysis: Zhihua Hua, Matthew J. Early.

Funding acquisition: Zhihua Hua.

**Investigation:** Zhihua Hua.

Methodology: Zhihua Hua, Matthew J. Early.

Software: Zhihua Hua.

Supervision: Zhihua Hua.

Validation: Zhihua Hua.

Visualization: Zhihua Hua.

Writing - original draft: Zhihua Hua.

Writing - review & editing: Zhihua Hua.

## References

- 1. Ohno S. Evolution by Gene Duplication. Springer-Verlag, New York. 1970.
- 2. Li W-H. Evolution of duplicate genes and pseudogenes. In: Evolution of Genes and Proteins (Nei M and Koehn RK eds). 1983: Sinauer Associates, Sunderland, MA. pp 14–37.
- Panchy N, Lehti-Shiu M, Shiu SH. Evolution of Gene Duplication in Plants. Plant Physiol. 2016; 171 (4):2294–316. Epub 2016/06/12. https://doi.org/10.1104/pp.16.00523 PMID: 27288366; PubMed Central PMCID: PMC4972278.
- Li Z, Defoort J, Tasdighian S, Maere S, Van de Peer Y, De Smet R. Gene Duplicability of Core Genes Is Highly Consistent across All Angiosperms. Plant Cell. 2016; 28(2):326–44. <u>https://doi.org/10.1105/tpc. 15.00877</u> PMID: 26744215; PubMed Central PMCID: PMC4790876.
- Hua Z, Zou C, Shiu SH, Vierstra RD. Phylogenetic comparison of F-Box (FBX) gene superfamily within the plant kingdom reveals divergent evolutionary histories indicative of genomic drift. PLoS One. 2011; 6(1):e16219. Epub 2011/02/08. https://doi.org/10.1371/journal.pone.0016219 PMID: 21297981; PubMed Central PMCID: PMC3030570.
- Hua Z, Doroodian P, Vu W. Contrasting duplication patterns reflect functional diversities of ubiquitin and ubiquitin-like protein modifiers in plants. Plant J. 2018; 95(2):296–311. Epub 2018/05/09. https://doi. org/10.1111/tpj.13951 PMID: 29738099.

- Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. Preservation of duplicate genes by complementary, degenerative mutations. Genetics. 1999; 151(4):1531–45. PMID: 10101175; PubMed Central PMCID: PMC1460548.
- He X, Zhang J. Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics. 2005; 169(2):1157–64. https://doi.org/10.1534/genetics.104. 037051 PMID: 15654095; PubMed Central PMCID: PMC1449125.
- Conant GC, Birchler JA, Pires JC. Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. Current Opinion in Plant Biology. 2014; 19:91–8. https://doi.org/10.1016/j.pbi.2014.05.008 PMID: 24907529.
- Hua Z, Pool JE, Schmitz RJ, Schultz MD, Shiu SH, Ecker JR, et al. Epigenomic programming contributes to the genomic drift evolution of the F-Box protein superfamily in Arabidopsis. Proc Natl Acad Sci U S A. 2013; 110(42):16927–32. Epub 2013/10/02. https://doi.org/10.1073/pnas.1316009110 PMID: 24082131; PubMed Central PMCID: PMC3801079.
- McHale L, Tan X, Koehl P, Michelmore RW. Plant NBS-LRR proteins: adaptable guards. Genome Biol. 2006; 7(4):212. Epub 2006/05/09. doi: gb-2006-7-4-212 [pii] https://doi.org/10.1186/gb-2006-7-4-212 PMID: 16677430; PubMed Central PMCID: PMC1557992.
- Nei M, Gu X, Sitnikova T. Evolution by the birth-and-death process in multigene families of the vertebrate immune system. Proc Natl Acad Sci USA. 1997; 94:7799–806. https://doi.org/10.1073/pnas.94. 15.7799 PMID: 9223266
- Nei M, Niimura Y, Nozawa M. The evolution of animal chemosensory receptor gene repertoires: roles of chance and necessity. Nat Rev Genet. 2008; 9(12):951–63. Epub 2008/11/13. doi: nrg2480 [pii] https:// doi.org/10.1038/nrg2480 PMID: 19002141.
- Shiu S-H, Shih M-C, Li W-H. Transcription factor families have much higher expansion rates in plants than in animals. Plant Physiol. 2005; 139(1):18–26. <u>https://doi.org/10.1104/pp.105.065110</u> PMID: 16166257.
- Zou C, Lehti-Shiu MD, Thibaud-Nissen F, Prakash T, Buell CR, Shiu S-H. Evolutionary and expression signatures of pseudogenes in *Arabidopsis* and rice. Plant Physiol. 2009; 151(1):3–15. <u>https://doi.org/ 10.1104/pp.109.140632 PMID: 19641029.</u>
- Marone D, Russo MA, Laido G, De Leonardis AM, Mastrangelo AM. Plant nucleotide binding site-leucine-rich repeat (NBS-LRR) genes: active guardians in host defense responses. Int J Mol Sci. 2013; 14 (4):7302–26. https://doi.org/10.3390/ijms14047302 PMID: 23549266; PubMed Central PMCID: PMC3645687.
- Burge CB, Karlin S. Finding the genes in genomic DNA. Curr Opin Struct Biol. 1998; 8(3):346–54. Epub 1998/07/17. PMID: <u>9666331</u>.
- Mathe C, Sagot MF, Schiex T, Rouze P. Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Res. 2002; 30(19):4103–17. Epub 2002/10/05. https://doi.org/10.1093/nar/ gkf543 PMID: 12364589; PubMed Central PMCID: PMC140543.
- Coghlan A, Fiedler TJ, McKay SJ, Flicek P, Harris TW, Blasiar D, et al. nGASP—the nematode genome annotation assessment project. BMC Bioinformatics. 2008; 9:549. Epub 2008/12/23. https://doi.org/10. 1186/1471-2105-9-549 PMID: 19099578; PubMed Central PMCID: PMC2651883.
- Guigo R, Flicek P, Abril JF, Reymond A, Lagarde J, Denoeud F, et al. EGASP: the human ENCODE Genome Annotation Assessment Project. Genome Biol. 2006; 7 Suppl 1:S2 1–31. Epub 2006/08/24. https://doi.org/10.1186/gb-2006-7-s1-s2 PMID: 16925836; PubMed Central PMCID: PMC1810551.
- Schweikert G, Zien A, Zeller G, Behr J, Dieterich C, Ong CS, et al. mGene: accurate SVM-based gene finding with an application to nematode genomes. Genome Res. 2009; 19(11):2133–43. Epub 2009/07/ 01. https://doi.org/10.1101/gr.090597.108 PMID: 19564452; PubMed Central PMCID: PMC2775605.
- Behr J, Bohnert R, zZeller G, Schweikert G, Hartmann L, Ratsch G. Next generation genome annotation with mGene.ngs. BMC Bioinformatics. 2010; 11(Suppl 10):08.
- van Dijk EL, Jaszczyszyn Y, Naquin D, Thermes C. The Third Revolution in Sequencing Technology. Trends Genet. 2018; 34(9):666–81. Epub 2018/06/27. https://doi.org/10.1016/j.tig.2018.05.008 PMID: 29941292.
- Birney E, Clamp M, Durbin R. GeneWise and Genomewise. Genome Res. 2004; 14(5):988–95. Epub 2004/05/05. https://doi.org/10.1101/gr.1865504 PMID: 15123596; PubMed Central PMCID: PMC479130.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of Molecular Biology. 1990; 215(3):403–10. https://doi.org/10.1016/S0022-2836(05)80360-2 PMID: 2231712.
- Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006; 22(13):1658–9. Epub 2006/05/30. doi: btl158 [pii] https://doi.org/10. 1093/bioinformatics/btl158 PMID: 16731699.

- Hua Z. Using CTT for comprehensive superfamily gene annotations. Protocolsio. 2019. <u>https://doi.org/10.17504/protocols.io.zf4f3qw</u>
- Hua Z, Vierstra RD. The cullin-RING ubiquitin-protein ligases. Annu Rev Plant Biol. 2011; 62:299–334. Epub 2011/03/05. https://doi.org/10.1146/annurev-arplant-042809-112256 PMID: 21370976.
- Gingerich DJ, Hanada K, Shiu SH, Vierstra RD. Large-scale, lineage-specific expansion of a bric-abrac/tramtrack/broad complex ubiquitin-ligase gene family in rice. Plant Cell. 2007; 19(8):2329–48. Epub 2007/08/28. <u>https://doi.org/10.1105/tpc.107.051300</u> PMID: <u>17720868</u>; PubMed Central PMCID: PMC2002615.
- Kent WJ. BLAT—the BLAST-like alignment tool. Genome Res. 2002; 12(4):656–64. Epub 2002/04/05. https://doi.org/10.1101/gr.229202 Article published online before March 2002. PMID: 11932250; PubMed Central PMCID: PMC187518.
- Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, et al. Earth BioGenome Project: Sequencing life for the future of life. Proc Natl Acad Sci U S A. 2018; 115(17):4325–33. Epub 2018/04/25. https://doi.org/10.1073/pnas.1720115115 PMID: 29686065; PubMed Central PMCID: PMC5924910.
- Yang X, Kalluri UC, Jawdy S, Gunter LE, Yin T, Tschaplinski TJ, et al. The F-box gene family is expanded in herbaceous annual plants relative to woody perennial plants. Plant Physiol. 2008; 148 (3):1189–200. Epub 2008/09/09. https://doi.org/10.1104/pp.108.121921 PMID: 18775973; PubMed Central PMCID: PMC2577272.