SOFTWARE TOOL ARTICLE

REVISED A2TEA: Identifying trait-specific evolutionary adaptations [version 2; peer review: 2 approved]

Tyll Stöcker, Carolin Uebermuth-Feldhaus, Florian Boecker, Heiko Schoof

Crop Bioinformatics, University of Bonn, Bonn, NRW, 53115, Germany

v2 **First published:** 05 Oct 2022, **11**:1137
https://doi.org/10.12688/f1000research.126463.1

Latest published: 12 Apr 2023, **11**:1137
https://doi.org/10.12688/f1000research.126463.2

Abstract

Background: Plants differ in their ability to cope with external stresses (e.g., drought tolerance). Genome duplications are an important mechanism to enable plant adaptation. This leads to characteristic footprints in the genome, such as protein family expansion. We explore genetic diversity and uncover evolutionary adaptation to stresses by exploiting genome comparisons between stress tolerant and sensitive species and RNA-Seq data sets from stress experiments. Expanded gene families that are stress-responsive based on differential expression analysis could hint at species or clade-specific adaptation, making these gene families exciting candidates for follow-up tolerance studies and crop improvement.
Software: Integration of such cross-species omics data is a challenging task, requiring various steps of transformation and filtering. Ultimately, visualization is crucial for quality control and interpretation. To address this, we developed A2TEA: Automated Assessment of Trait-specific Evolutionary Adaptations, a Snakemake workflow for detecting adaptation footprints in silico. It functions as a one-stop processing pipeline, integrating protein family, phylogeny, expression, and protein function analyses. The pipeline is accompanied by an R Shiny web application that allows exploring, highlighting, and exporting the results interactively. This allows the user to formulate hypotheses regarding the genomic adaptations of one or a subset of the investigated species to a given stress.
Conclusions: While our research focus is on crops, the pipeline is entirely independent of the underlying species and can be used with any set of species. We demonstrate pipeline efficiency on real-world datasets and discuss the implementation and limits of our analysis workflow as well as planned extensions to its current state. The A2TEA workflow and web application are publicly available at: https://github.com/tgstoecker/A2TEA.Workflow and https://github.com/tgstoecker/A2TEA.WebApp, respectively.

Keywords

plants, crops, adaptation, evolution, stress, workflow, software

Open Peer Review

Approval Status ✓ ✓

	1	2
version 2 (revision) 12 Apr 2023	✓ view	✓ view
version 1 05 Oct 2022	? view	? view

1. **Manuel Aranda**, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

   Octavio Salazar Moya, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

2. **Rahul Siddharthan**, The Institute of Mathematical Sciences, Chennai, India

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the Genomics and Genetics gateway.

This article is included in the Plant Science gateway.

This article is included in the Plant Computational and Quantitative Genomics collection.

Corresponding authors: Tyll Stöcker (tyll.stoecker@gmail.com), Heiko Schoof (schoof@uni-bonn.de)

Author roles: Stöcker T: Conceptualization, Data Curation, Formal Analysis, Investigation, Methodology, Project Administration, Software, Validation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Uebermuth-Feldhaus C**: Conceptualization, Formal Analysis, Investigation, Methodology, Writing – Review & Editing; **Boecker F**: Software, Writing – Review & Editing; **Schoof H**: Conceptualization, Funding Acquisition, Resources, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The author(s) declared that no grants were involved in supporting this work.

How to cite this article: Stöcker T, Uebermuth-Feldhaus C, Boecker F and Schoof H. **A2TEA: Identifying trait-specific evolutionary adaptations [version 2; peer review: 2 approved]** F1000Research 2023, **11**:1137 https://doi.org/10.12688/f1000research.126463.2

First published: 05 Oct 2022, **11**:1137 https://doi.org/10.12688/f1000research.126463.1

REVISED **Amendments from Version 1**

Dear readers, based on the reviewer's comments, we have made improvements to the manuscript, and we have also added new features to our software.

Fixed editing errors: garbled text, missing citations, improved visualizations.

Paragraph 1 of the introduction has been rewritten to more clearly separate prior approaches and the need to combine them as we do with A2TEA.

New features of the software:

The pipeline has become more flexible since it is no longer required to provide RNA-seq data for all species – this allows species to be included in the phylogenetic analyses for which no RNA-seq data for the stress/treatment under investigation is available.

Tables in the WebApp can now be downloaded in their un-/filtered entirety.

Updates to plots such as better labels, positioning of cutoff lines behind plotted data, and removal of small 0 bars in the log2FC layer of the tea plots.

Any further responses from the reviewers can be found at the end of the article

Introduction

While genomic resources for crop species continuously increase, with more and more high-quality reference genome sequences and transcriptome datasets becoming available, the lack of integrated trait and functional information limits the ability to interpret genomic-scale datasets and discover genotype-phenotype associations. Methods such as differential expression analysis have led to the discovery of many candidate genes e.g., involved in tolerance to stresses or, more generally, central to the physiological reaction pattern towards a specific experimental treatment.[1,2] However, the interpretation of large lists of candidate genes is challenging and does not provide insight into evolutionary adaptation to the treatment under investigation. In contrast, comparative genomics approaches allow the identification of genomic footprints of adaptation,[3] such as protein family expansions.[4] Gene duplication is a major driver of molecular evolution,[5,6] and in plants, whole-genome duplication events are frequent (reviewed in Ref. 7), but tandem and transposon-mediated duplications also play a role (reviewed in Ref. 8). Most gene duplicates are lost or silenced,[9] but retained duplicates may hint at some evolutionary advantage and may be targets of adaptation.[10] However, associating evolutionary retention with functions relating to complex traits of a species is not possible without considering further information, such as insight into condition-specific gene usage, e.g., in the form of expression data. These approaches thus have clear limitations when used in isolation.

While many efforts have focused on individual (model) species and the outlined singular methodological approaches, the increasing availability of omics data for many more or less related genomes opens opportunities to explore genetic diversity through multi-genome comparisons. To overcome the aforementioned limitations, we propose a pipeline that combines differential expression analysis and comparative genomics to prioritize genes that were targets of evolutionary adaptation, thereby facilitating their application in crop improvement. Especially for the adaptation of regulatory networks, duplication allows for neo- or subfunctionalization,[11] which form an evolutionary scenario that can be observed based on our integration of phylogeny and differential expression under treatment/stress data. However, we also consider differential expression under a given treatment in any species as a functional link to the given treatment, even if there is not sufficient data to confirm neo- or subfunctionalization. This allows us to filter for gene family expansions functionally linked to the given treatment, as not all adaptations and thus retained duplicates in a genome need to relate to e.g., tolerance to the given stress, other traits not under analysis will also show adaptation and thus protein family expansions. Our approach allows for a more comprehensive understanding of trait adaptation (in both plants and other organisms) and can guide the development of strategies for improving crops (Figure 1).

The challenge for this multi-genome approach is the cross-species integration of multiple types of omics data, which requires several software tools and various custom steps of transformation and filtering. To promote the exploration of genomics and transcriptomics data and the association of genotype with phenotype data in order to address adaptation, we developed **A2TEA** (**A**utomated **A**ssessment of **T**rait-specific **E**volutionary **A**daptations). Our software aims at identifying candidate genes for stress adaptation in plant species and enables GUI-based exploration of the results but is suitable for gene family expansion analysis integrated with differential expression data in any set of genomes. It is composed of a Snakemake workflow and an R (Shiny) package working in tandem to automate and ease all bioinformatics and analysis tasks involved.

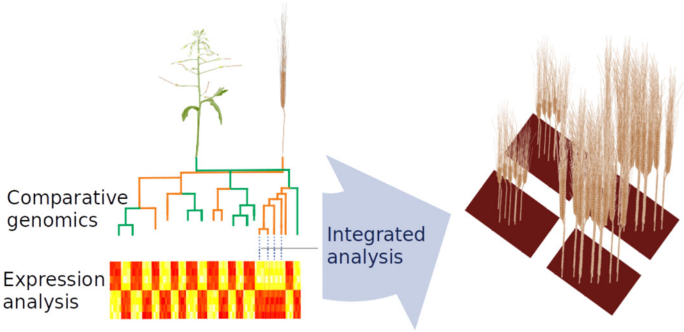


Figure 1. Identification of interesting gene families for crop improvement by integration of differential gene expression with gene family expansion.

The A2TEA.Workflow functions as a one-stop processing pipeline, integrating the prediction of gene families in the form of orthologous groups (OGs) with the analysis of their phylogeny, protein function, and expression, using RNA-seq data from all species. It allows the user to formulate adaptation hypotheses as specific scenarios of gene family expansion in one or several of the genomes, for example, based on a classification of species as stress-tolerant or sensitive, or to identify clade-specific adaptations. As input, the workflow requires for each species a protein FASTA file for orthologous group prediction and RNA-seq reads suitable for differential expression analysis (control vs. treatment), together with either a genomic FASTA file with appropriate gene annotation or a transcriptomic/cDNA FASTA file. Functional information for each species can be provided by the user or can be optionally inferred by our tool AHRD (https://github.com/groupschoof/AHRD) during runtime. The single compressed output is ready for analysis with the R programming language[12] as we took care to create well-structured objects and easy-to-parse outputs. In addition, in order to facilitate immediate and easy exploration and visualization of the results, we created the A2TEA.WebApp written in R Shiny,[13] which allows exploring, highlighting, and exporting the results interactively.

The A2TEA.Workflow combines state-of-the-art bioinformatics software with custom integration steps to combine inferred gene family expansion events with expression results and functional associations.

The workflow is designed as a complete solution starting with raw data and performing upstream quality controls and data transformations automatically. We also took care to allow for a high degree of customizability - e.g., RNA-seq analysis can be performed either alignment-based or using pseudoalignment, and tool-specific parameters can be tweaked in one central config file. Importantly, the workflow is designed to answer biological questions and as such requires the definition of hypotheses in form of combinations of the species of interest. For each hypothesis, the user needs to adjust parameters related to the definition and cutoffs of expansion events. This allows computation of results for several combinations in parallel and facilitates the investigation of many hypotheses downstream e.g., expansion in all tolerant species, in only a specific species, or in all species of a clade.

The A2TEA.WebApp provides an interactive web interface to explore, filter, and visualize the previously generated results via a straightforward tab-structured dashboard design. We took care to create a user-friendly mouse-controlled experience in order to extend the usability from bioinformaticians to experimentalists. The user first uploads the output file of the workflow and chooses the specific “hypothesis” to investigate. This generates a general information tab providing an overview of phylogeny, expression, and set sizes of orthologous groups (OGs) passing the thresholds. The user is then able to switch to dedicated analysis tabs relating to 1) filtering and analyzing OGs with associated data, 2) set size comparisons and tests, and 3) gene ontology (GO) term enrichment analyses. Reactively rendered tables and visualizations are dynamically populated with links to databases such as Ensembl[14] and AmiGO[15] to allow for an immediate follow-up exploration of interesting genes. Tables and graphs can be exported in a variety of formats. The web application also provides a bookmarking system that facilitates the collection and export of the most interesting genes and OGs.

To extend the usability of the workflow by allowing for further species-specific exploration of gene and geneset functional enrichments we integrated the creation of GeneTonic input data files into the A2TEA.Workflow. GeneTonic is a web application that serves as a comprehensive toolkit for streamlining the interpretation of functional enrichment analyses from RNA-seq data.[16] As our workflow is built on Snakemake,[17] the addition of further analyses or outputs allows for modular expansion of its current state. We also intend to add further analyses and features to the A2TEA.WebApp web application.

A2TEA combines best practices in both choice of tools as well as reproducibility and offers a one-stop solution for the integration of genome comparisons with expression and functional data to unravel candidate genes for natural adaptation, e.g. in stress-tolerant plant species. The web application empowers users to explore stress-specific gene family expansions combined with transcriptomic data from their own or published stress experiments by providing interactive visualizations, statistical tests, and dynamically generated database queries.

Both the A2TEA.Workflow and A2TEA.WebApp are freely available at https://github.com/tgstoecker/A2TEA.Workflow and https://github.com/tgstoecker/A2TEA.WebApp, respectively, and archived in Zenodo.[59,60] For demonstration purposes, we also made a public instance of the web application available at https://tgstoecker.shinyapps.io/A2TEA-WebApp.

Methods

Implementation

The A2TEA.Workflow is written in Python and makes use of the Snakemake workflow framework. It leverages the bioconda project channel[18] of the conda package manager to handle software installation and dependency management. Another tool from our lab, AHRD, is integrated as a Git submodule and can be optionally used to infer protein function annotation for any of the species under investigation.

The typical use case for running the workflow consists of cloning the GitHub repository, configuring it to specific needs, and then starting the analyses with either installation of software and dependencies during runtime or usage of a Docker/Singularity container (Figure 2). Modification of the workflow is performed by changing dedicated configuration files controlling samples, species, hypotheses, and tool-specific options. With "hypotheses" we refer to the definition

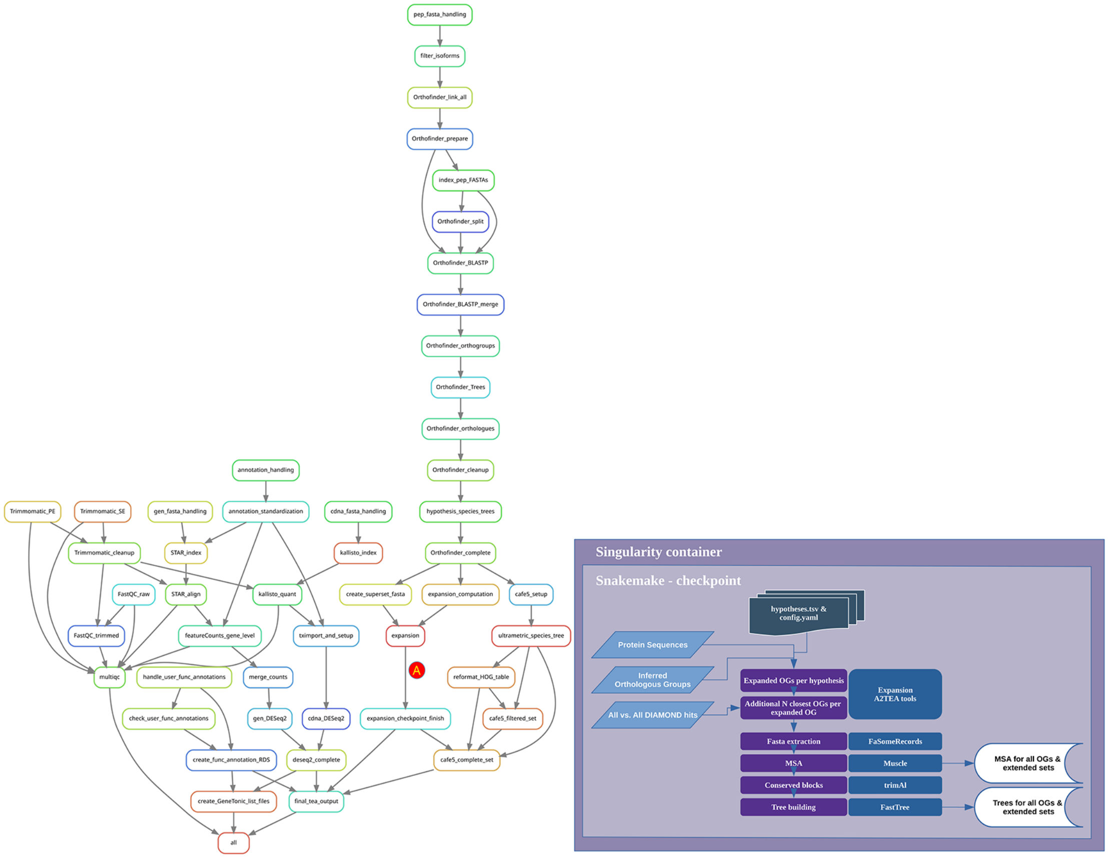


Figure 2. Overview of the A2TEA.Workflow. Workflow diagram of the A2TEA.Workflow displayed as Snakemake rulegraph. After computation of expanded orthologous groups (OGs) (rule expansion - marked with A) the directed acyclic graph (DAG) is re-evaluated since the results are not known beforehand. This Snakemake checkpoint then uses the reciprocal best hits computed by Orthofinder to find the N most similar additional OGs per OG, where N is a variable set by the user. For each OG and additional set of 1 to N additional OGs, multiple sequence alignments and phylogenetic trees are built and used in the downstream steps.

of "gene family expansion" in the set of species under investigation. Several hypotheses can be run in parallel. This multi-hypothesis structure permits the investigation of several defined biological questions, for instance, gene family expansion in stress-tolerant compared to stress-sensitive species. For each hypothesis, we always require the definition of a set of one or more species that should be checked for expansion compared to a second set of one or more species that should not show expansion. For each hypothesis, the user is able to set several options, such as the ratio or the minimal number of genes in a species, to qualify as an expanded OG. The hypotheses.tsv file is structured column-wise with both an index number and a "name" variable used to identify the choices throughout the workflow. Generally, the connection between files and workflow rules is achieved by the species names (e.g., "Arabidopsis_thaliana"). Many hypotheses can be computed in a single workflow with a single final output object that contains all results. This facilitates easy comparisons in the downstream web application, which is especially useful to check the parameter choices for the definition of gene family expansion or when it is necessary to work with unclear trait classification of some species.

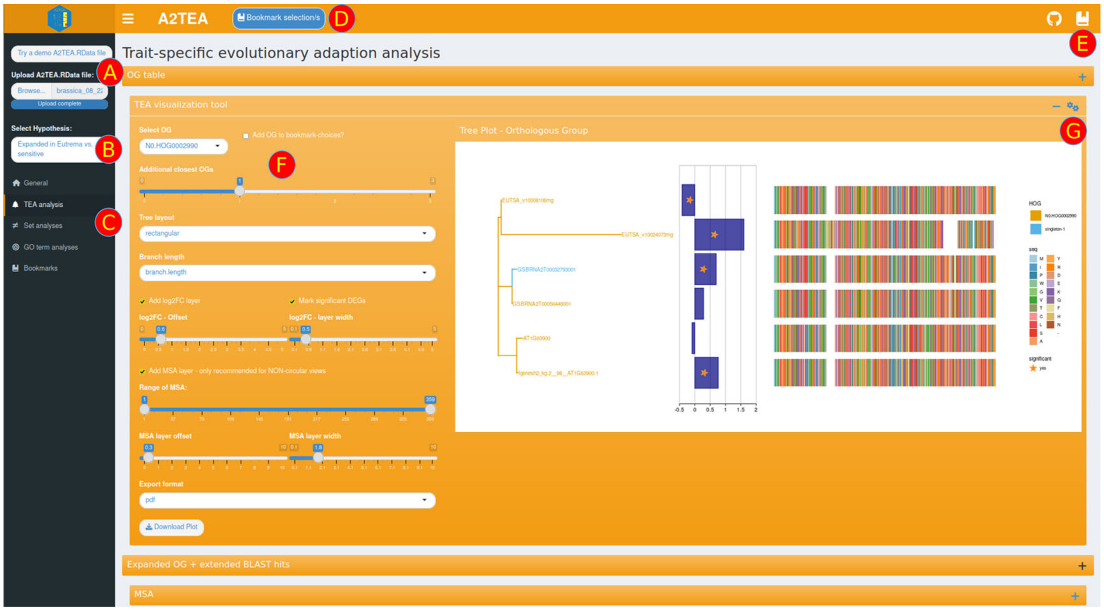
The final output generated by the workflow is a single .RData file that can be loaded into an active R environment with the load() command. This provides several separate objects containing all results in a compact form factor:

- HYPOTHESES.a2tea - List object with one S4 object per hypothesis. Each S4 object contains several layers of nested information. E.g., HYPOTHESES.a2tea$hypothesis_2@expanded_OGs$N0.HOG0001225 refers to a specific expanded OG and S4 data object that contains:
  - blast_table (complete BLAST/DIAMOND[19,20] results for OG genes & extended hits)
  - add_OG_analysis (includes multiple sequence alignment (MSA), phylogenetic tree, and gene info for expanded OG and additional OGs based on best BLAST/DIAMOND hits)
- HOG_level_list - List object with one tibble per hypothesis. Information includes OG, number of genes per species, boolean expansion info, number of significant differentially expressed genes (DEGs), and more. The last N list element is a non-redundant superset of all species analyzed over all formulated hypotheses. This makes it easy to create a comparison set e.g., conserved OGs of all species to which the hypothesis subset can then be compared.
- HOG_DE.a2tea - Tibble of DESeq2[21] results for all genes + additional columns.
- A2TEA.fa.seqs - Non-redundant list object containing corresponding amino acid FASTA sequences of all genes/transcripts in the final analysis (this includes those of expanded OGs + those in additional BLAST hits & additional OGs based on user-chosen parameters).
- SFA/SFA_OG_level - Gene/transcript level tables that contain functional predictions (human readable descriptions & GO terms inferred by AHRD).
- hypotheses - A copy of the user-defined hypotheses definitions for the underlying workflow run.
- all_speciesTree - Phylogenetic tree of all species in the workflow run (a non-redundant superset of hypotheses) as inferred by Orthofinder/STAG/Stride.

The .RData output can be investigated inside an R session or via the A2TEA.WebApp, which was specifically designed to allow for interactive inspection, visualization, filtering, and export of the results and subsets. We feature a tutorial for its usage and details on how to work with the results of an A2TEA.Workflow analysis run in the Use Case section and in the project's pkgdown site.

The A2TEA.WebApp is written in the R programming language[12] and uses the Shiny[13] framework to facilitate interactivity with the data. It expects the user to upload a .RData file created by the A2TEA.Workflow. The web application comes with a test dataset that can be loaded with a single click so that interested users can try out its functionality before having to finish an A2TEA.Workflow run.

We developed the web application following community standards and have set up a continuous integration system with GitHub actions that performs build checks of both the package itself and the associated pkgdown site[22] hosted via GitHub pages.



Trait-specific evolutionary adaption analysis

Upload A2TEA.RData file: brassica_08_22 — Upload complete

Select Hypothesis: Expanded in Eutrema vs. sensitive

Select OG: N0.HOG0002990

□ Add OG to bookmark-choices?

Additional closest OGs: 1

Tree layout: rectangular

Branch length: branch.length

☑ Add log2FC layer

☑ Mark significant DEGs

log2FC - Offset: 0.6

log2FC - layer width: 0.5

☑ Add MSA layer - only recommended for NON-circular views

Range of MSA: 1 – 359

MSA layer offset: 0.3

MSA layer width: 1.8

Export format: pdf

Figure 3. Screenshot of the trait evolutionary analysis tab in the A2TEA.WebApp. The user can either decide to load the included test dataset or upload a .RData result object (A). Other options in the sidebar menu are a selector of the hypothesis (meaning: species comparison) to display or change to another analysis tab (C). Genes, transcripts, or orthologous groups (OGs) can be marked in tables or boxes ticked and then bookmarked with a dedicated button (D). Bookmarks have their dedicated tab but can also be displayed as a sidebar window anywhere for quick reference purposes (E). Analysis- or plot-specific parameters (F) are displayed to the left of the visualization, and a box-specific sidebar window for aesthetic parameters can be opened by clicking the gears icon (G). The underlying dataset investigates drought tolerance among four Brassicaceae species. Displayed here are the maximum likelihood phylogenies of a gene family showing potential subfunctionalization of *Eutrema salsugineum* homologs (top two genes in the tree). Blue bars show log2 (fold change) of gene expression between drought and control conditions, stars mark adj. p ≤ 0.1 significance cutoff, and the multiple sequence alignment of amino acid sequences is displayed to the right. We provide this particular OG as a bookmarked subset .RData file (see Underlying Data).

The interface is structured in tabs with shinydashboard[23] and shinydashboardPlus,[24] providing the layout infrastructure (shown in Figure 3). The main functionality includes a selector to choose which hypothesis to display (Figure 3B), a sidebar menu that enables the user to switch between different analysis types (Figure 3C), and tool-specific options for parameters, visuals, and export (Figure 3F).

We designed the interface to allow the focus to be put on an individual analysis or plot to gain insight from the data. Plots and tables are contained in collapsible boxes, leaving it up to the user to decide how much information should be displayed at once. Additionally, we tried to separate important parameters from purely aesthetic choices in the plot options, with main options always visible at the side and aesthetic choices reactively displayed after the user switches a box toggle.

Since the exploration of data can be a lengthy process with many iteration cycles, we looked for ways of aiding the user in storing the observations made. Following the example set by the GeneTonic web application,[16] we integrated a bookmarking system that temporarily stores interesting genes/transcripts and OGs. For this, the user needs to mark the respective ID in one of the tables and then click the dedicated bookmarking button displayed at the top of the interface (Figure 3D). All bookmarks are rendered in two reference tables, both in a dedicated tab as well as a pop-up window that can be displayed on every analysis tab. This quick reference is convenient when performing filtering operations in the tables or choosing an interesting OG to display. While these tables can of course, be downloaded, we also implemented a subsetting feature on the bookmarks tab that creates a smaller .RData file with information only pertaining to the bookmarks and associated data. These smaller subsets are fully functioning complete inputs and can be loaded into the application again at a later time - for re-plotting purposes for instance. With this feature, it is straightforward to extract, store and share all information about interesting genes, transcripts, or OGs.

Operation

The A2TEA.Workflow has been primarily designed for use within Linux and requires a standard bash environment with working installations of Snakemake[17] and conda/mamba (Ref. 18, https://github.com/mamba-org/mamba). The former facilitates compatibility with common cluster setups such as SLURM or LSF. Instructions for a minimal setup are described in the project's README.

For each species, the A2TEA.Workflow should be provided with input RNA-Seq reads (both paired-/single-end possible) suitable for differential expression analysis (control vs. treatment), either a genomic or transcriptomic FASTA file, annotations, and a peptide FASTA file. Since the latest release (v1.1), RNA-Seq data is not required for a workflow run because downstream inferences can still be made by the user in cases where, for one or several species, no expression data for the investigated conditions are available. The user can provide functional information per species, or it can be optionally inferred by our tool AHRD during the workflow. Control of the workflow is handled by several configuration files, which the user needs to adapt to their specific inputs and scientific questions.

The samples.tsv table needs to list all RNA-seq FASTQ files with the columns providing additional information based on which the workflow can infer associations such as species, replicate, and the correct steps to perform. For instance, by leaving out the column for the second paired sample, it is automatically inferred that single-end options have to be used (trimming, mapping, etc.). Operations such as recognizing that files are gzipped and need to be handled appropriately are performed automatically as well. The species.tsv table functions similarly and needs to provide per species information on the FASTA and annotation files, the ploidy of the species, and the location of a file providing the functional information, in the form of GO terms, per protein. If no functional information can be provided the user can choose to add "AHRD" instead of a file path which will trigger a sub-workflow during computation that will create an appropriate file via our functional annotation tool AHRD. Based on whether the user provides a genomic or cDNA FASTA file for a species, the workflow will perform either traditional alignment with STAR[25] or pseudoalignment with kallisto.[26]

The config.yaml controls parameters such as thread usage for individual steps, tool-specific parameters, and parameters relating specifically to the A2TEA.Workflow. Two other very important choices that have to be considered are whether or not automatic filtering for the longest representative isoforms of the peptide FASTA files should be performed and whether gene or transcript level quantification is wanted. Choosing automatic isoform filtering will create a subset peptide FASTA file with only the longest isoform per gene; the header will be shortened to just the gene name identifier. This option must be used in conjunction with gene level quantification since otherwise matching both types of data is not possible.

The notion behind the hypotheses.tsv table is outlined in the Implementation section due to its central importance to the expansion calculation. Here, we briefly want to present some of the available choices the user can consider. Besides defining sets of species that should be analyzed for expansion compared to other sets of species, the user is able to specify the required numerical differences between the two and which OGs to disregard immediately. For instance, "Nmin_expanded_in" takes as input an integer value that defines the minimum number of the investigated species that need to fulfill the expansion criteria in order for the gene family to be called "expanded". These criteria are for instance "min_expansion_factor" and "min_expansion_difference" with which either the minimum factor of expansion or the minimum number of additional genes compared to the non-expanded set of species can be defined. To complement these broad cutoffs, the workflow also integrates a hypothesis-specific CAFE analysis,[27] with which changes in gene family size are analyzed in a way that accounts for phylogenetic history and provides a statistical foundation for evolutionary inferences.

After all choices have been made, the workflow can be started with a single Snakemake command. A2TEA.Workflow will then perform all previously listed steps and merge results into the final output file described in the Implementation section (Figure 2). The user is then able to investigate the integrated and condensed results.

We offer several ways of starting an A2TEA.WebApp instance for downstream investigation of the data: 1) installation with R devtools from our GitHub repository, 2) a docker container with the latest release installed, and lastly, 3) a demo instance hosted on shinyapps.io (https://tgstoecker.shinyapps.io/A2TEA-WebApp/). As the A2TEA.WebApp is an interactive tool with an explorative focus and no strict work order, we illustrate its core operative features in the dedicated Use cases section of this manuscript.

Use cases

In this section, we will illustrate the functionality of the A2TEA.WebApp, using the A2TEA.Workflow results of a three-species analysis of *Hordeum vulgare* (barley), *Zea mays* (maize), and *Oryza sativa japonica* (rice) that investigates adaptive processes in barley to drought stress. Details on the files used as well as their respective publication and SRA accession numbers are listed in detail in both GitHub repositories and the Source data section.[56,57,58]

We integrated this dataset into the workflow and the web application to illustrate the software's setup and to allow for a quick exploration of the tools' functionalities. After cloning the A2TEA.Workflow repository, an additional script can be run (get_test_data.sh) that quickly sets up the experiment by downloading the required input files. Peptide FASTA files are reduced to 2000 proteins; the transcriptomic data is subsampled to 2M reads to allow for a quicker runtime. The functional annotations are precomputed by AHRD. The differential expression analysis is set to be performed on the gene level and two comparisons are performed as defined in the hypotheses.tsv table. These are "Expanded in barley compared to rice and maize" and "Expanded in barley compared to maize". For both, expansion is defined as "number of genes species A ≥ 2 × number of genes of species B".

The final output produced by the workflow is also integrated into the current release of the A2TEA.WebApp and can be loaded via clicking the "Try a demo A2TEA.RData file" at the top of the interface.

Initial inspection of integrated data

The general analysis tab is the default view inside the A2TEA.WebApp. Once input is loaded, reactive information boxes display the number of species, the number of expanded OGs, and the number of DEGs for the currently selected hypothesis. Changing the hypothesis (Figure 3B) e.g., to the second hypothesis in our test set ("Expanded in barley compared to maize"), changes the statistics and all other sets/plots to reflect only the species considered in the hypothesis. Two tables display gene-level differential expression results and functional annotation information (human readable descriptions and GO terms), which allow, for example, the exploration of genes related to a particular function. Also displayed are an inferred phylogenetic tree of the species in the hypothesis subset and an intersection plot (Venn/UpSet) which displays the number of conserved (OG with ≥1 gene from every species), overlapping, or species unique OGs and singleton genes. Importantly, a table describing the details of the currently displayed hypothesis is also displayed. All of this facilitates a broad overview of the data and allows the user to spot errors such as faulty hypothesis definitions or cutoffs that are too strict.

Exploring expansion events with annotated phylogenetic trees

The main feature of the TEA (trait-specific evolutionary adaptation) tab is a comprehensive toolkit for the visualization of maximum likelihood phylogenies of expanded OGs and associated information such as the log2(fold change) of the displayed genes and an MSA of the respective protein sequences (Figure 3). The MSA can be added as a geometric layer to the tree plots[28] or displayed separately with additional options such as a conservation bar (Figure 3A). To make an informed decision of which OGs are most worthwhile to investigate closer, a table showcasing the total and significantly differentially expressed genes per OG is also provided. With this, the user is enabled to apply several filters, for example, to select all expanded OGs that possess at least 1 DEG and more than 4 genes from *Hordeum vulgare*. The last table on the tab provides insight into the reciprocal BLAST/DIAMOND hits for the currently chosen OG and the additional most similar OGs. Notably, this table also provides the identifiers given to the proteins by Orthofinder,[29] making it easy to relate insight gained in the web application back to other outputs created by Orthofinder in the A2TEA.Workflow, such as the list of putative xenologs.

Comparing sets of orthologous groups

To describe adaptive processes at a larger scale, we also integrated functionality to visualize distributions of user-defined OG sets and test for their over-representation; e.g., "What is the frequency of OGs that show expansion and at least 1 DEG in *Hordeum vulgare* in all conserved OGs?" and "Is this set over-represented within the background distribution of conserved OGs with at least 1 DEG from any species?". We took care to make answering such questions very accessible by providing the user with text-based choices of which sets to plot or compare. Currently integrated are an enrichment analysis suite allowing for Fisher-Tests and a corresponding circular set plot (Figure 4B) that visualizes the chosen sets. Also provided is a tool for comparing the size distributions of the OGs (Figure 4C) with which group size effects can be checked; e.g., "Do we see differences in the number of DEGs in OGs of a certain size range between the set of interest and the background set?".

Performing functional enrichment tests

The last analysis tab provides options for performing GO term over-representation analysis based on the topGO R package.[30] Functions that occur more often than expected can be identified by setting several parameters that specify the set of OGs the user wants to analyze. With our test data, the user could, for instance, be interested in enriched molecular functions of OGs that are expanded in *Hordeum vulgare* and also possess at least 2 DEGs of *Hordeum vulgare*. Once computed, a table is displayed that shows the top significantly enriched GO terms and also contains dynamically created links for these to AmiGO2.[15] A second table contains information on the corresponding OGs and genes so that the user can follow up on a particular enriched GO term and inspect the underlying data. We also provide two visualizations that summarize the results. The first is a GO enrichment dotplot (Figure 4D) straightforwardly showcasing the overall results,

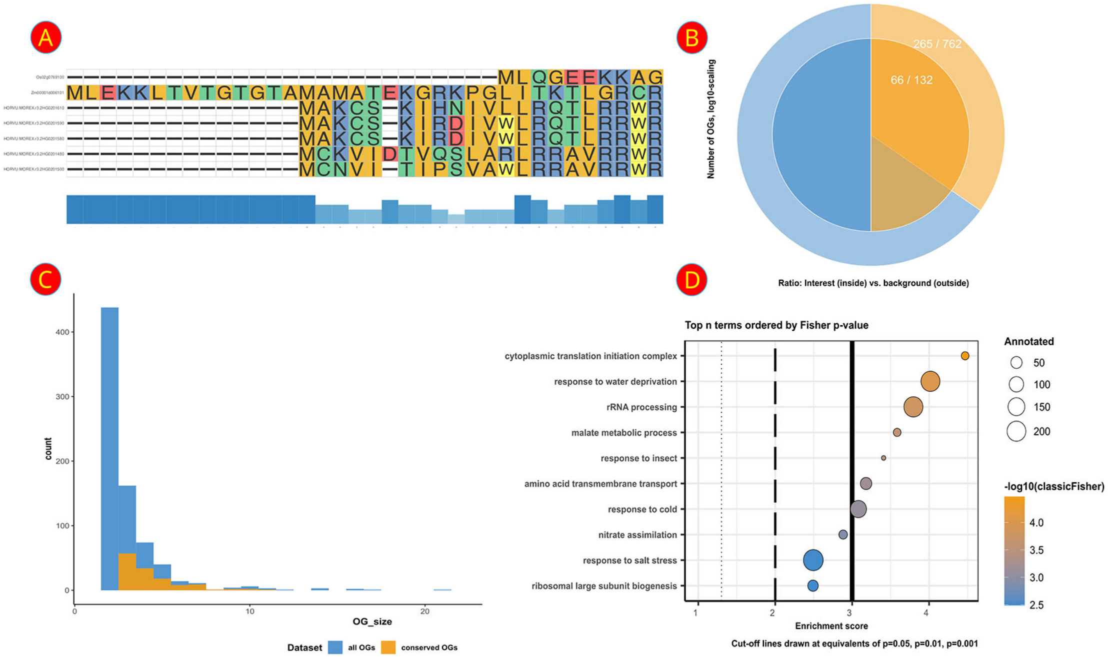


Figure 4. Overview of several analysis plots featured in the A2TEA.WebApp. (A) Multiple sequence alignment of an expanded orthologous group (OG) + additional most similar singletons or OGs. Bars at the bottom represent the degree of conservation. (B) Visual representation of a hypergeometric test for over-representation - colors/layers represent sets in the urn model. The outer ring shows the background set (light blue; "complete background") and the subset in the background set (light orange; "success in background"). The inner ring displays the set of interest (blue; "sample size") and the subset in the set of interest (orange; "success in sample"). (C) Barplot of the total number of OGs per group size (number of genes) between all OGs (blue) and only those OGs that are conserved among the analyzed species (orange). (D) Dotplot of the top over-represented biological processes in the subset of OGs of interest compared to the background; dot size indicates the number of OGs with the respective GO term; color displays negative log [base 10] of the p-values from the enrichment test.

and the second is a GO subgraph of selected top N enriched GO terms. With the latter, we provide the user with an insightful way of investigating how the significant GO terms are distributed over the GO graph.

Export options, bookmarking & ending a session

Tables can be downloaded as .tsv files, and plots are exportable into various formats, such as.pdf, .png, or.svg, allowing the user to easily save and share the observations and results. However, even a relatively small set of species, like the three Poaceae species in our test data, lead to several OGs that are worthwhile to investigate, substantiating the need for the bookmarking system outlined in the Implementation section. It quickly becomes very valuable to bookmark, e.g., all OGs annotated with the top 5 enriched BP GO terms in the OGs expanded in *Hordeum vulgare* if the intention is to return to the analysis later or to generate a list to use with another tool quickly. Relating this to the previous sub-sections, we want to emphasize that bookmarking is integral to using the A2TEA.WebApp and is fully featured on all analysis tabs except the "Set analyses" since here individual genes or OGs are not the focus. To further aid users in the bookmarking process, we also added informative pop-up messages to indicate for instance, that all selected genes/OGs have already been saved. Since the bookmarks can also be used to export a completely functional .RData subset file, only the most relevant information is kept while the processing speed is increased, and all relevant results of the integrative effort are kept. If, for instance, during the analysis, it turned out that hypothesis 2 in our example data ("Expanded in barley compared to maize") is, in fact, not of interest anymore, subsetting the .RData file to interesting OGs of hypothesis 1 completely removes the unneeded "bloat" of hypothesis 2. Similarly, the user could create 2 .RData files (one for each hypothesis) and run a custom script on each separately, efficiently producing hypothesis-specific results.

Discussion

Classic transcriptomic studies produce large lists of gene regulatory information for which, traditionally, pathway or GO term analyses are used to discover the overall molecular trends caused by the experimental treatments.[31] We propose that we can identify novel genes relevant for stress adaptation by comparing same-stress experiments of several plant species with different levels of stress adaptation in combination with evolutionary footprints in the form of protein family expansion. As illustrated in the Methods & Use Cases sections, our novel software tool A2TEA facilitates the identification of genes associated with the evolution of a trait in a species or a group of related species. Based on the rediscovery of known genes related to the trait, we believe that also novel genes discovered through A2TEA are related to the trait, but experimental verification is in progress, see below. As an example, Figure 3 presents a possible subfunctionalization of gene duplicates in *Eutrema salsugineum*, discovered from data of drought tolerance among four Brassicaceae species (details see Underlying Data). The *A. thaliana* homolog is involved in drought stress response.[32]

Several approaches have been employed to identify potential candidate genes that could provide a genetic basis for more resilient crops. This includes forward genetics approaches such as identifying causative genes for advantageous mutant phenotypes,[33] finding common regulators for several stresses via traditional transcriptomics,[34] usage of Quantitative Trait Locus (QTL) mapping and Genome-wide association studies (GWAS) incl. potential integration with expression data,[35] the combination of expression data with functional information and clustering methods[36,37] and also machine learning based approaches that employ transcriptomic or phenomic data as the basis of their candidate gene predictions.[38,39]

The underlying methods are manifold and include approaches such as Bulked-Segregant analysis,[40] k-means clustering,[41] WGCNA,[42] co-expression networks[43] and set analyses of DEGs often in combination with pathway or GO term enrichment analyses.[31] While most studies share the approach of reducing a list of regulated genes via secondary criteria, to our knowledge, A2TEA is the first openly available tool that specifically combines stress-specific expression data from several species with gene family expansion to unravel candidate genes for stress adaptation in stress-tolerant species.

With A2TEA, we present software that simplifies the complex bioinformatics workflow for the user and provides an interactive web interface for analysis of the results. By using Snakemake as a bioinformatic workflow manager, we remove the need for step-by-step handling of raw data (including software setup and dependencies necessary for computations) and ensure FAIR (findable, accessible, interoperable, and reusable) computational analysis standards.[44] The downstream analysis and visualization framework makes the navigation of the resulting large sets of tabular data faster, more intuitive, and more practicable for scientists without programming skills.[16,31] It offers a variety of summary statistics on the levels of gene family expansion, differential expression, and functional enrichment to ensure quality control. The Shiny framework provides interactivity regarding the visualization and the analysis of the results, and this interactivity highly facilitates the exploration of scientific questions.

Based on user experiences with the web application, we have included analyses and visualizations to allow detecting problems in the bioinformatic predictions, e.g., of orthologous groups (OGs). In order to spot potential misassignments of Orthofinder, close homologs to members of an OG are detected by similarity search and displayed with phylogenetic trees and multiple sequence alignments. A typical case is the non-inclusion of a singleton gene of a species due to a significant portion of protein sequence missing in the annotation, caused e.g., by gaps or sequencing errors in the genome sequence or errors in gene prediction. Similarly, false expansions based on a putative paralog that has only very limited alignment overlap with other members of the OG can be detected. These could be actual duplicates but degenerated through pseudogenization or partial duplication, e.g., the action of transposable elements.

We designed A2TEA with extendability in mind. Both the Snakemake-based A2TEA.Workflow and the A2TEA Shiny App can be easily expanded in a modular fashion to integrate novel features. We are currently testing several additional visualization and testing options. This includes the option for positive selection tests concerning a particular OG, e.g., by calculating the ratio of non-synonymous amino-acid substitutions over synonymous amino-acid substitutions (dN/dS)), distribution comparisons between random and actual DEG-containing OGs, and visualizations for the analysis of general gene/transcript regulation trends. The GO term enrichment functionality is aimed at discovering general trends in the adaptation to the particular stress under investigation. At the moment, the implemented enrichment tests provide options for single over-representation analysis as implemented in the R topGO package.[30] It will be interesting to evaluate and potentially implement further options for functional enrichment analysis in A2TEA, such as modern ensembl approaches[45] or simplification strategies that aid in summarization.[46] Lastly, we intend to implement the option to download a comprehensive RMarkdown/Quarto report summarizing plots and statistics for all bookmarked genes and OGs. This has been demonstrated to be a significant step forward in guaranteeing the portability of results once an interactive session is concluded.[16]

While our research focus is on crops, from a software perspective, A2TEA is entirely independent of the underlying species and can be used with any set of species. This opens the question of how feasible applying the A2TEA methodology to species from other kingdoms might be. Our motivation for developing A2TEA is primarily rooted in the notion that genome duplication played a major role in the evolutionary past of plants.[8] Plant comparative genomics research has shown that gene families are mostly conserved across great evolutionary timescales, comprising even the diversification of all angiosperms and nonflowering plants.[47,48] Fascinatingly, this conservation of gene families is combined with lineage-specific fluctuations in gene family size, which are frequent among taxa.[8,47–50] This suggests that since comparatively few novel gene families arose, much of the great diversity and phenotypic variation seen in land plants may have arisen primarily due to duplication and adaptive specialization of already existing genes.[48]

While whole genome duplication events are expected and reported less frequently in the animal kingdom and thus gene duplication as a driver of protein family expansion does not play as prominent a role in animals as in plants,[51,52] protein family expansion is still an important driver of adaptation.[53] We expect that A2TEA will be useful in non-plant species, even if protein family expansion only represents a small portion of adaptive changes, with other sources of variation, like alternative splicing playing a potentially more important role.[54]

Currently, we are investigating several publicly available genomic and transcriptomic datasets from various groups of plant species with A2TEA. While we expect to detect candidate genes relevant to adaptation to the stress being investigated, this assumption is based on the rediscovery of known genes. One important follow-up step is thus to experimentally verify the impact of selected candidate genes in vivo. To this end, we perform stress experiments in plants bearing knockout mutations in candidate genes predicted by A2TEA, using sequence-indexed mutant collections such as BonnMu.[55] This will allow us to assess the phenotypic impact of these mutations and, thus, the role of these genes in the tolerance of the stress. While testing all candidates will not be feasible, the rate of genes relevant to the trait under investigation among tested candidates will represent an estimate of the prediction performance.

Conclusions

With the availability of multiple genome sequence and RNA-seq data sets, it is now possible to combine comparative evolutionary analyses, in our case protein family expansion, with differential expression to predict genes involved in adaptive traits. However, running the required bioinformatics analyses and data integration tasks as well as summarizing and visualizing the results, remains challenging. A2TEA only requires standard data files as input, follows best practice software standards for both reproducibility and portability, and provides a user-friendly web application for interactive exploration and selection of the most promising candidate genes. We show that genes known to be involved in stress tolerance can be detected in datasets of stress-tolerant and stress-sensitive plants, but we expect A2TEA to be useful in a broader scope when analyzing protein families and their expression in multiple genomes as the parameters for selecting interesting families are very flexible. A2TEA follows a positive trend in modern research software development that provides easy installation and execution through the use of container and workflow technologies as well as interactive visualization and exploration tools for the generated results. Combined, this facilitates better reproducibility, communication, and shareability of comprehensive analyses.

Data availability

Source data

Poaceae test data:

Transcriptomics:

Hordeum vulgare: SRR6782243, SRR6782247, SRR6782257, SRR6782249, SRR6782250, SRR6782254

Zea mays: SRR2043219, SRR2043217, SRR2043190, SRR2043220, SRR2043226, SRR2043227

Oryza sativa japonica: SRR5134063, SRR5134064, SRR5134065, SRR5134066

These correspond to the following studies relating on drought stress:

Hordeum vulgare: https://doi.org/10.1186/s12864-019-5634-0

Zea mays: https://doi.org/10.1104/pp.16.01045

Oryza sativa japonica: https://doi.org/10.3389/fpls.2017.00580

Assemblies & annotations hosted on EnsemblPlants:

Hordeum vulgare: cDNA FASTA, GTF, Peptide FASTA

Zea mays: Genome FASTA, GTF, Peptide FASTA

Oryza sativa japonica: cDNA FASTA, GTF, Peptide FASTA

An archived version of the complete grasses test data (reduced as used in the examples) is deposited here:

https://zenodo.org/record/7089022[56]

Data used in the Brassicaceae example:

Transcriptomics:

Eutrema salsugineum: SRR7624684, SRR7624685, SRR7624692, SRR7624687, SRR7624721, SRR7624722

Arabidopsis lyrata: SRR7624680, SRR7624702, SRR7624703, SRR7624732, SRR7624733, SRR7624742

Arabidopsis thaliana: SRR7624694, SRR7624696, SRR7624697, SRR7624710, SRR7624714, SRR7624723

Brassica napus: SRR12429701, SRR12429702, SRR12429703, SRR12429698, SRR12429699, SRR12429700

These correspond to the following studies on drought stress response: *Eutrema salsugineum, Arabidopsis lyrata, Arabidopsis thaliana*: https://doi.org/10.1111/nph.15841

Brassica napus: PRJNA656507

Assemblies & annotations hosted on EnsemblPlants:

Eutrema salsugineum: Genome FASTA, GTF, Peptide FASTA

Arabidopsis lyrata; Genome FASTA, GTF, Peptide FASTA

Arabidopsis thaliana: Genome FASTA, GTF, Peptide FASTA

Brassica napus: Genome FASTA, GTF, Peptide FASTA

Underlying data

The results generated by the A2TEA.Workflow which are also used for demonstrating the A2TEA.WebApp's functionality presented in this work are available at https://zenodo.org/record/7089608[57] and https://zenodo.org/record/7089606.[58]

Data are available under the terms of the Creative Commons Attribution 4.0 International license (CC-BY 4.0).

Software availability

Both the A2TEA.Workflow and the A2TEA.WebApp are available as MIT licensed open source softwares.

- Software available from: https://tgstoecker.github.io/A2TEA.WebApp
- Source code available from: https://github.com/tgstoecker/A2TEA.Workflow, https://github.com/tgstoecker/A2TEA.WebApp
- Archived source code at time of publication: https://zenodo.org/record/7725859,[59] https://zenodo.org/record/7750290[60]
- License: MIT

References

1. Muktar MS, Lübeck J, Strahwald J, *et al.*: **Selection and validation of potato candidate genes for maturity corrected resistance to Phytophthora infestans based on differential expression combined with SNP association and linkage mapping.** *Front. Genet.* 2015; **6**: 294.
PubMed Abstract | Publisher Full Text | Free Full Text

2. Lovell JT, Mullen JL, Lowry DB, *et al.*: **Exploiting differential gene expression and epistasis to discover candidate genes for drought-associated QTLs in Arabidopsis thaliana.** *Plant Cell.* 2015; **27**(4): 969–983.
PubMed Abstract | Publisher Full Text | Free Full Text

3. Lin D, Lan L, Zheng T, *et al.*: **Comparative genomics reveals recent adaptive evolution in Himalayan giant honeybee Apis laboriosa.** *Genome Biol. Evol.* 2021; **13**(10): evab227.
Publisher Full Text

4. Wang J, Hu H, Liang X, *et al.*: **High-quality genome assembly and comparative genomic profiling of yellowhorn (*Xanthoceras sorbifolia*) revealed environmental adaptation footprints and seed oil contents variations.** *Front. Plant Sci.* 2023; **14**: 976.
PubMed Abstract | Publisher Full Text | Free Full Text

5. Ohno S: *Evolution by gene duplication.* Springer Science & Business Media; 2013.

6. Taylor JS, Raes J: **Duplication and divergence: The evolution.** *Annu. Rev. Genet.* 2004; **38**: 615–643.
PubMed Abstract | Publisher Full Text

7. Qiao X, Zhang S, Paterson AH: **Pervasive genome duplications across the plant tree of life and their links to major evolutionary innovations and transitions.** *Comput. Struct. Biotechnol. J.* 2022; **20**: 3248–3256.
PubMed Abstract | Publisher Full Text

8. Panchy N, Lehti-Shiu M, Shiu S-H: **Evolution of gene duplication in plants.** *Plant Physiol.* 2016; **171**(4): 2294–2316.
PubMed Abstract | Publisher Full Text

9. Adams KL, Wendel JF: **Polyploidy and genome evolution in plants.** *Curr. Opin. Plant Biol.* 2005; **8**(2): 135–141.
Publisher Full Text

10. Maere S, De Bodt S, Raes J, *et al.*: **Modeling gene and genome duplications in eukaryotes.** *Proc. Natl. Acad. Sci.* 2005; **102**(15): 5454–5459.
PubMed Abstract | Publisher Full Text

11. Voordeckers K, Pougach K, Verstrepen KJ: **How do regulatory networks evolve and expand throughout evolution?** *Curr. Opin. Biotechnol.* 2015; **34**: 180–188.
PubMed Abstract | Publisher Full Text

12. R Core Team: *R: A Language and Environment for Statistical Computing.* 2022.
Reference Source

13. Chang W, Cheng J, Allaire JJ, *et al.*: *shiny: Web Application Framework for R.* 2022. R package version 1.7.2.
Reference Source

14. Cunningham F, Allen JE, Allen J, *et al.*: **Ensembl 2022.** *Nucleic Acids Res.* 2022; **50**(D1): D988–D995.

15. Carbon S, Ireland A, Mungall CJ, *et al.*: **Amigo: online access to ontology and annotation data.** *Bioinformatics.* 2009; **25**(2): 288–289.
PubMed Abstract | Publisher Full Text

16. Marini F, Ludt A, Linke J, *et al.*: **Genetonic: an r/bioconductor package for streamlining the interpretation of rna-seq data.** *BMC Bioinform.* 2021; **22**(1): 1–19.

17. Mölder F, Jablonski KP, Letcher B, *et al.*: **Sustainable data analysis with snakemake.** *F1000Res.* 2021; **10**: 33.
PubMed Abstract | Publisher Full Text

18. Grüning B, Dale R, Sjödin A, *et al.*: **Bioconda: sustainable and comprehensive software distribution for the life sciences.** *Nat. Methods.* 2018; **15**(7): 475–476.
PubMed Abstract | Publisher Full Text

19. Altschul SF, Gish W, Miller W, *et al.*: **Basic local alignment search tool.** *J. Mol. Biol.* 1990; **215**(3): 403–410.
PubMed Abstract | Publisher Full Text

20. Buchfink B, Xie C, Huson DH: **Fast and sensitive protein alignment using DIAMOND.** *Nat. Methods.* 2015; **12**(1): 59–60.
PubMed Abstract | Publisher Full Text

21. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.** *Genome Biol.* 2014; **15**(12): 1–21.
PubMed Abstract | Publisher Full Text | Free Full Text

22. Wickham H, Hesselberth J, Salmon M: *pkgdown: Make Static HTML Documentation for a Package.* 2022. R package version 2.0.6.
Reference Source

23. Chang W, Ribeiro BB: *shinydashboard: Create Dashboards with 'Shiny'.* 2021. R package version 0.7.2.
Reference Source

24. Granjon D: *shinydashboardPlus: Add More 'AdminLTE2' Components to 'shinydashboard'.* 2021. R package version 2.0.3.
Reference Source

25. Dobin A, Davis CA, Schlesinger F, *et al.*: **Star: ultrafast universal rna-seq aligner.** *Bioinformatics.* 2013; **29**(1): 15–21.
PubMed Abstract | Publisher Full Text

26. Bray NL, Pimentel H, Melsted P, *et al.*: **Near-optimal probabilistic rna-seq quantification.** *Nat. Biotechnol.* 2016; **34**(5): 525–527.
Publisher Full Text

27. Mendes FK, Vanderpool D, Fulton B, *et al.*: **Cafe 5 models variation in evolutionary rates among gene families.** *Bioinformatics.* 2021; **36**(22-23): 5516–5518.

28. Yu G, Smith DK, Zhu H, *et al.*: **ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data.** *Methods Ecol. Evol.* 2017; **8**(1): 28–36.
Publisher Full Text

29. Emms DM, Kelly S: **Orthofinder: phylogenetic orthology inference for comparative genomics.** *Genome Biol.* 2019; **20**(1): 1–14.

30. Alexa A, Rahnenführer J: **Gene set enrichment analysis with topgo.** *Bioconductor Improv.* 2009; **27**: 1–26.

31. Supek F, Škunca N: **Visualizing go annotations.** *The Gene Ontology Handbook.* New York, NY: Humana Press; 2017; pp. 207–220.

32. Ling Q, Jarvis P: *Current Biology, and undefined 2015. Regulation of chloroplast protein import by the ubiquitin e3 ligase sp1 is important for stress tolerance in plants.* Elsevier; 2015.
Reference Source

33. Kirschner GK, Rosignoli S, Guo L, *et al.*: **Enhanced gravitropism 2 encodes a sterile alpha motif–containing protein that controls root growth angle in barley and wheat.** *Proc. Natl. Acad. Sci.* 2021; **118**(35): e2101526118.
PubMed Abstract | Publisher Full Text

34. Sham A, Moustafa K, Al-Ameri S, *et al.*: **Identification of arabidopsis candidate genes in response to biotic and abiotic stresses using comparative microarrays.** *PLoS One.* 2015; **10**(5): e0125666.
PubMed Abstract | Publisher Full Text

35. Guo T, Yang J, Li D, *et al.*: **Integrating gwas, qtl, mapping and rna-seq to identify candidate genes for seed vigor in rice (oryza sativa l.).** *Mol. Breed.* 2019; **39**(6): 1–16.

36. Sewelam N, Brilhaus D, Bräutigam A, *et al.*: **Molecular plant responses to combined abiotic stresses put a spotlight on unknown and abundant genes.** *J. Exp. Bot.* 2020; **71**(16): 5098–5112.
PubMed Abstract | Publisher Full Text

37. Kar S, Mai H-J, Khalouf H, *et al.*: **Comparative transcriptomics of lowland rice varieties uncovers novel candidate genes for adaptive iron excess tolerance.** *Plant Cell Physiol.* 2021; **62**(4): 624–640.
PubMed Abstract | Publisher Full Text

38. Shaik R, Ramakrishna W: **Machine learning approaches distinguish multiple stress conditions using stress-responsive genes and identify candidate genes for broad resistance in rice.** *Plant Physiol.* 2014; **164**(1): 481–495.
PubMed Abstract | Publisher Full Text

39. Braun IR, Yanarella CF, Lawrence-Dill CJ: **Computing on phenotypic descriptions for candidate gene discovery and crop improvement.** 2020.
Publisher Full Text | Reference Source | Reference Source

40. Michelmore RW, Paran I, Kesseli RV: **Identification of markers linked to disease-resistance genes by bulked segregant analysis: A rapid method to detect markers in specific genomic regions by using segregating populations.** *Proc. Natl. Acad. Sci. U. S. A.* 1991; **88**: 9828–9832. 00278424.
Publisher Full Text | Reference Source

41. Likas A, Vlassis N, Verbeek JJ: **The global k-means clustering algorithm.** *Pattern Recogn.* 2003; **36**(2): 451–461.
Publisher Full Text

42. Langfelder P, Horvath S: **Wgcna: an r package for weighted correlation network analysis.** *BMC Bioinform.* 2008; **9**(1): 1–13.

43. Aoki K, Ogata Y, Shibata D: **Approaches for extracting practical information from gene co-expression networks in plant biology.** *Plant Cell Physiol.* 2007; **48**(3): 381–390.
PubMed Abstract | Publisher Full Text

44. Wratten L, Wilm A, Göke J: **Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers.** *Nat. Methods.* 2021; **18**(10): 1161–1168.
PubMed Abstract | Publisher Full Text

45. Alhamdoosh M, Ng M, Wilson NJ, *et al.*: **Combining multiple tools outperforms individual methods in gene set enrichment analyses.** *Bioinformatics.* 2017; **33**(3): 414–424.
PubMed Abstract | Publisher Full Text

46. Zuguang G, Hübschmann D: **Simplify enrichment: A bioconductor package for clustering and visualizing functional enrichment results.** *Genom. Proteom. Bioinf.* 2022.
Publisher Full Text

47. Rensing SA, Lang D, Zimmer AD, *et al.*: **The physcomitrella genome reveals evolutionary insights into the conquest of land by plants.** *Science.* 2008; **319**(5859): 64–69.
PubMed Abstract | Publisher Full Text

48. Flagel LE, Wendel JF: **Gene duplication and evolutionary novelty in plants.** *New Phytol.* 2009; **183**(3): 557–564.
Publisher Full Text

49. Velasco R, Zharkikh A, Troggio M, *et al.*: **A high quality draft consensus sequence of the genome of a heterozygous grapevine variety.** *PLoS One.* 2007; **2**(12): e1326.
PubMed Abstract | Publisher Full Text

50. Ming R, Hou S, Feng Y, *et al.*: **The draft genome of the transgenic tropical fruit tree papaya (carica papaya linnaeus).** *Nature.* 2008; **452**(7190): 991–996.
PubMed Abstract | Publisher Full Text

51. Mable BK: **'why polyploidy is rarer in animals than in plants': myths and mechanisms.** *Biol. J. Linn. Soc.* 2004; **82**(4): 453–466.
Publisher Full Text

52. Murat F, Van de Peer Y, Salse J: **Decoding plant and animal genome plasticity from differential paleo-evolutionary patterns and processes.** *Genome Biol. Evol.* 2012; **4**(9): 917–928.
PubMed Abstract | Publisher Full Text

53. Demuth JP, De Bie T, Stajich JE, *et al.*: **The evolution of mammalian gene families.** *PLoS One.* 2006; **1**(1): e85.
PubMed Abstract | Publisher Full Text

54. Kim E, Magen A, Ast G: **Different levels of alternative splicing among eukaryotes.** *Nucleic Acids Res.* 2007; **35**(1): 125–131.
PubMed Abstract | Publisher Full Text

55. Marcon C, Altrogge L, Win YN, *et al.*: **Bonnmu: a sequence-indexed resource of transposon-induced maize mutations for functional genomics studies.** *Plant Physiol.* 2020; **184**(2): 620–631.
PubMed Abstract | Publisher Full Text

56. Stöcker T: **A2TEA.Workflow test data (v1.0.0) [Data set]. Zenodo.** 2022.
Publisher Full Text

57. Stöcker T: **A2TEA.Workflow Poaceae reduced example data (v1.0.0) [Data set]. Zenodo.** 2022.
Publisher Full Text

58. Stöcker T: **A2TEA Brassicaceae example data (v.1.0.0) [Data set]. Zenodo.** 2022.
Publisher Full Text

59. Stöcker T: **tgstoecker/A2TEA.Workflow: No transcriptomes required (v1.1.0). Zenodo.** 2023.
Publisher Full Text

60. Stöcker T: **tgstoecker/A2TEA.WebApp: v1.1.5 (v1.1.5). Zenodo.** 2023.
Publisher Full Text

Open Peer Review

Current Peer Review Status: ✓ ✓

Version 2

Reviewer Report 15 May 2023

https://doi.org/10.5256/f1000research.146089.r169362

✓

Rahul Siddharthan
Computational Biology, The Institute of Mathematical Sciences, Chennai, Tamil Nadu, India

I am satisfied by the changes made by the authors and have no further comment.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Computational biology: regulatory genomics, chromatin, algorithms

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 03 May 2023

https://doi.org/10.5256/f1000research.146089.r169361

✓

Octavio Salazar Moya
Biological and Environmental Sciences and Engineering Division (BESE), Red Sea Research Center (RSRC), King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

Manuel Aranda
Biological and Environmental Sciences and Engineering Division (BESE), Red Sea Research Center (RSRC), King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

The authors addressed most of our comments adequately and we have only a few comments left.

Specific comments:

1 . Introduction: “These approaches thus have clear limitations when used in isolation.” We suggest changing it to: These approaches, thus, have clear limitations when used in isolation.

2. Related to our previous comment:

“GO term analyses tab: At GO term analyses, changing the algorithm does not change the column name at GO term set choices. It always says classicFisher. Same in enrichment plots. Also, in the enrichment plots, it might be better to plot the cut-off lines after plotting the terms, as sometimes the terms are small due to a low number of annotated genes belonging to them and might be covered by the line. Changing the x-axis label for the bar plot to the number of significant orthogroups might be better.”

The authors answered:

The choice of the topGO algorithm concerns how the GO graph is analyzed (independence vs. consideration parent terms, etc.) and is a separate option from the downstream test statistic (https://bioconductor.org/packages/release/bioc/vignettes/topGO/inst/doc/topGO.pdf). In the current release, we only allow classicFisher as the test statistic. As such, the column does not need to change in our opinion. However, we agree that information of the used algorithm should be kept. We have now implemented that the name of the algorithm is part of the output tables that the user is able to download.

New comment:

The authors mention that only classicFisher is allowed as the test statistic. However, under GO term set choices, there is a drop down menu allowing the user to specify the Algorithm as classic, elim, weight, weight01, and parentchild. Changing this option leads to different results at the Go term set choices window (Also Go should be changed to GO), suggesting that the current release of A2TEA does support the use of other algorithms other than classicFisher. Moreover, the p-values do change. Which would then come back to our previous comment of having the column name change according to the algorithm used at the Go term set choices output.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genomics

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 01 February 2023

https://doi.org/10.5256/f1000research.138877.r156575

?

Rahul Siddharthan

[1] Computational Biology, The Institute of Mathematical Sciences, Chennai, Tamil Nadu, India
[2] Computational Biology, The Institute of Mathematical Sciences, Chennai, Tamil Nadu, India

The software tool A2TEA is well described and motivated. The idea of identifying genes involve in adaptive traits, such as stress-tolerance, using inter-species comparisons, is sound. A2TEA integrates the bioinformatics (RNA-seq, orthologous group computation, functional information inference, etc.

I do not work in this field so have not tested out the tool. The webapp looks well laid out and user-friendly but I did not try it on actual data.

Several boxes in the flowchart in figure 2 have garbled text (text replaced with boxes). I also suggest highlighting a couple of possible paths through this figure in an actual workflow (eg, a single workflow would not use both kallisto and STAR?).

Minor comment: one block of text, starting with "The A2TEA.WebApp is written in the R programming language..." (two paragraphs) is repeated (as a single paragraph immediately after).

Is the rationale for developing the new software tool clearly explained?

Yes

Is the description of the software tool technically sound?

Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Computational biology: regulatory genomics, chromatin, algorithms

I confirm that I have read this submission and believe that I have an appropriate level of

expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 24 Mar 2023

Tyll Stöcker

We thank all reviewers for providing valuable input to the article that helped us to improve it further. We have tried to address their comments and made efforts to incorporate their insightful criticisms and suggestions. Similar suggestions and criticisms have been grouped together in order to address them concisely.

- **Several boxes in the flowchart in figure 2 have garbled text (text replaced with boxes). I also suggest highlighting a couple of possible paths through this figure in an actual workflow (eg, a single workflow would not use both kallisto and STAR?).**

  We have made sure that the figure is now without the noted garbled text. Regarding the paths through the diagram: The reviewer is correct that in a normal differential expression analysis one would not choose to process some of the samples with one tool and the rest with another. However, in our field of plant/crop research, we have encountered many cases of early assembly and annotation versions, with many cases of only a transcriptome or a genome assembly being available. Due to this, the workflow does support running one species through a pseudoalignment and another through classic alignment-based quantification to provide more flexibility when combining data from different sources. In our Workflow README, we point out that for runtime and resource purposes, we recommend kallisto/pseudoalignment, and it should always be preferred if possible.

- **Minor comment: one block of text, starting with "The A2TEA.WebApp is written in the R programming language..." (two paragraphs) is repeated (as a single paragraph immediately after).**

  We thank the reviewer very much for pointing out this mistake that occurred during our editing process.

Competing Interests: No competing interests were disclosed.

Reviewer Report 25 November 2022

https://doi.org/10.5256/f1000research.138877.r154237

?

Manuel Aranda
[1] Biological and Environmental Sciences and Engineering Division (BESE), Red Sea Research Center (RSRC), King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia
[2] Biological and Environmental Sciences and Engineering Division (BESE), Red Sea Research Center (RSRC), King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

Octavio Salazar Moya
[1] Biological and Environmental Sciences and Engineering Division (BESE), Red Sea Research Center (RSRC), King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia
[2] Biological and Environmental Sciences and Engineering Division (BESE), Red Sea Research Center (RSRC), King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

General comments:

The manuscript submitted by Stöcker et *et al*. describes an integrated pipeline for the identification of genes potentially involved in adaptation by analyzing gene duplications and gene family expansions. The pipeline combines RNA-Seq expression analyses via DESeq2, identification and phylogenetic analysis of orthologous groups with OrthoFinder, statistical analyses of gene family changes taking into account phylogeny inference with Cafe5, and gene ontology enrichment analyses with TopGO. The results of the pipeline can be visualized using the R shiny package, which allows for easy exploration of the results even for experimentalists with little bioinformatic background. A2TEA is a simple yet useful tool for the visualization and integration of multiple analyses and datasets. A few things need to be revised in the manuscript, such as a corrupted Figure 2, repeated paragraphs, and the addition of missing references. The implementation of a “Download All Results” option in the general tab of the shiny app would be helpful. Allowing the pipeline to run without the addition of RNA-Seq could be quite useful to allow analyses of species without available RNA-Seq data from the same condition. Additionally, it would be appropriate to put an emphasis on the broader use of the app instead of focusing on plant stress, as it would make the app more appealing to a broader audience.

Specific comments

- Abstract: “It functions as a one-stop processing pipeline, integrating protein family, phylogeny, expression, and protein function analysis”. Change analysis to analyses.

- Abstract: “The pipeline is accompanied by an R Shiny web application that”. Jump in line, fix editing.

- Introduction: “Most of the duplicates are lost or silenced....”. Maybe change it to something like: Most of the gene duplicates are lost or silenced, but retained duplicates may hint at some evolutionary advantage, which may be targets of adaptation.

- Introduction: The first paragraph of the introduction needs some work. It is unclear if the authors are explaining the different approaches that can be used to identify genes related to adaptation or if they are referring to their pipeline, as they use “we” and “us”. It would be appropriate to first describe the different approaches and their benefits and then the need to combine them. The second and third paragraphs already properly address what the pipeline is doing.

- Introduction: Change & for and at: “working in tandem to automate and ease all bioinformatics & analysis tasks involved.”

- Introduction: Change “gene families in form of” to gene families in the form of...

- Figure 1 (and the whole article). While the tool was designed for the identification of genes for crop improvement, it is not limited to that application. As the name of the application indicates, it is designed for trait-specific evolutionary adaptations. It might be a good idea to generalize the possible applications of this tool rather than focus only on plant stress-specific responses. The authors can specify that an example of its use is in crop improvement while making sure to maintain a general tone for the use of A2TEA for a broader audience.

- Cite DIAMOND, DESeq2, BLAST.

- Cite Cafe in the main text (only found in the reference list).

- This paragraph is repeated: “The A2TEA.WebApp is written in the R programming language and uses the Shiny framework to facilitate interactivity with the data. It expects the user to upload a .RData file created by the A2TEA.Workflow. The web application comes with a test dataset that can be loaded with a single click so interested users can try out its functionality before having to finish an A2TEA.Workflow run.”

- Figure 2 is full of typos like orthoander instead of orthofinder and many of the boxes don’t have text but empty squares, probably an error while formatting.

- Figure 3. Italicize Eutrema salsugineum.

- Figure 4. Increase the resolution or the font size, as words in panels B, C, and D are blurry.

- Comments on the A2TEA Shiny App.
  - General tab: Differential expression and Functional annotation require a “Download Full Results” button and not only a download current page.

  - GO term analyses tab: At GO term analyses, changing the algorithm does not change the column name at GO term set choices. It always says classicFisher. Same in enrichment plots. Also, in the enrichment plots, it might be better to plot the cut-off lines after plotting the terms, as sometimes the terms are small due to a low number of annotated genes belonging to them and might be covered by the line. Changing the x-axis label for the bar plot to the number of significant orthogroups might be better.

- Discussion: “We propose that we can identify novel genes relevant for stress adaptation by comparing same-stress experiments of several plant species with different levels of stress adaptation in combination with evolutionary footprints in the form of protein family expansion”. Again, it would be better to keep the usability of this tool as general as possible, not only focusing on plat stress, as it could be used for any species and conditions. Also, using the word condition or treatment instead of stress would be better.

- Additional comments. It could be beneficial if the App could also be run without RNA-Seq data for the development of a general hypothesis when comparing species for which RNA-Seq data under the same conditions is not available.

Is the rationale for developing the new software tool clearly explained?
Yes

Is the description of the software tool technically sound?
Yes

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?
Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?
Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Genomics

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Author Response 24 Mar 2023

Tyll Stöcker

We thank all reviewers for providing valuable input to the article that helped us to improve it further. We have tried to address their comments and made efforts to incorporate their insightful criticisms and suggestions. Similar suggestions and criticisms have been grouped together in order to address them concisely.

- **Abstract: "It functions as a one-stop processing pipeline, integrating protein family, phylogeny, expression, and protein function analysis". Change analysis to analyses.**

- **Abstract: "The pipeline is accompanied by an R Shiny web application that". Jump in line, fix editing.**

- **Introduction: “Most of the duplicates are lost or silenced....”. Maybe change it to something like: Most of the gene duplicates are lost or silenced, but retained duplicates may hint at some evolutionary advantage, which may be targets of adaptation.**

- **Introduction: Change & for and at: “working in tandem to automate and ease all bioinformatics & analysis tasks involved.”**

- **Introduction: Change “gene families in form of” to gene families in the form of...**

  We thank the reviewer very much for taking the time and point out several minor formatting and phrasing errors, all of which we have corrected.

- **Introduction: The first paragraph of the introduction needs some work. It is unclear if the authors are explaining the different approaches that can be used to identify genes related to adaptation or if they are referring to their pipeline, as they use “we” and “us”. It would be appropriate to first describe the different approaches and their benefits and then the need to combine them. The second and third paragraphs already properly address what the pipeline is doing.**

  On re-reading the first paragraph of the introduction with the reviewer's comments in mind, we strongly agreed that a more logical separation was necessary. We have restructured and rewritten parts of it to more clearly separate the prior approaches and then – in a second step – the need to combine them as we do with our software.

- **Cite DIAMOND, DESeq2, BLAST.**

- **Cite Cafe in the main text (only found in the reference list).**

  We thank the reviewer for pointing out important missing citations in our text and have added these as well as additional ones. We have repositioned our citation of CAFE5 immediately next to the software’s name in the text (prior it was at the end of the sentence).

- **This paragraph is repeated: “The A2TEA.WebApp is written in the R programming language and uses the Shiny framework to facilitate interactivity with the data. It expects the user to upload a .RData file created by the A2TEA.Workflow. The web application comes with a test dataset that can be loaded with a single click so interested users can try out its functionality before having to finish an A2TEA.Workflow run.”**

  We thank the reviewer very much for pointing out this mistake that occurred during our editing process.

- **Figure 2 is full of typos like orthoander instead of orthofinder and many of the boxes don’t have text but empty squares, probably an error while formatting.**

Unsure of the cause, we have uploaded the figure in a different format which seems to have solved the garbled text.

- **Figure 3. Italicize Eutrema salsugineum.**

  We italicized this species in the figure text and took care to check the manuscript for similar formatting flaws.

- **Figure 4. Increase the resolution or the font size, as words in panels B, C, and D are blurry.**

  We re-exported the image in higher resolution to increase the overall legibility.

- **Comments on the A2TEA Shiny App.**

  General tab: Differential expression and Functional annotation require a "Download Full Results" button and not only a download current page.

  The reviewer was correct in pointing out the need for this feature. For all tables, we have now implemented the possibility of downloading all results and not just the displayed page via the suggested buttons. If the user did not perform any filtering, the complete table is downloaded; if the user performs filter operations (e.g. | log2FC | > 1) the complete, filtered table is downloaded.

  GO term analyses tab: At GO term analyses, changing the algorithm does not change the column name at GO term set choices. It always says classicFisher. Same in enrichment plots. Also, in the enrichment plots, it might be better to plot the cut-off lines after plotting the terms, as sometimes the terms are small due to a low number of annotated genes belonging to them and might be covered by the line. Changing the x-axis label for the bar plot to the number of significant orthogroups might be better.

  The choice of the topGO algorithm concerns how the GO graph is analyzed (independence vs. consideration parent terms, etc.) and is a separate option from the downstream test statistic (https://bioconductor.org/packages/release/bioc/vignettes/topGO/inst/doc/topGO.pdf). In the current release, we only allow classicFisher as the test statistic. As such, the column does not need to change in our opinion. However, we agree that information of the used algorithm should be kept. We have now implemented that the name of the algorithm is part of the output tables that the user is able to download.

  We have changed the order of the plotting in the enrichment plot – now, the lines plot behind the terms, which ensures clearer visualizations.

  We have also adapted the reviewer's suggestion for a better x-axis label in the enrichment bar plot.

- **Figure 1 (and the whole article). While the tool was designed for the identification of genes for crop improvement, it is not limited to that application. As the name of the application indicates, it is designed for trait-specific evolutionary adaptations. It might be a good idea to generalize the possible applications of this tool rather than focus only on plant stress-specific responses. The authors can specify that an example of its use is in crop improvement while making sure to maintain a general tone for the use of A2TEA for a broader audience.**

- **Discussion: "We propose that we can identify novel genes relevant for stress adaptation by comparing same-stress experiments of several plant species with different levels of stress adaptation in combination with evolutionary footprints in the form of protein family expansion". Again, it would be better to keep the usability of this tool as general as possible, not only focusing on plant stress, as it could be used for any species and conditions. Also, using the word condition or treatment instead of stress would be better.**

  We agree with the reviewer that the tone and focus of our manuscript can be somewhat loosened to emphasize the general technical feasibility of applying our approach to research outside of the plant kingdom. We have 1. altered parts of the manuscript to incorporate the words treatment/condition instead of the word stress (to underline broader context applicability) and 2. added an additional sentence to the end of the first paragraph emphasizing the general technical applicability in other organisms than plants.

  However, as we point out in two paragraphs near the end of the discussion, our particular approach stems from the support of previous research that genome duplication was a decisive factor in the evolutionary history of plants and much of the phenotypic variation in land plants may have arisen primarily due to duplication and adaptive specialization of already existing genes. Since our method specifically focuses on the analysis of protein family expansion events, it might be much less applicable in other groups of species where other evolutionary processes, such as alternative splicing or pathway/network rewiring, could play a proportionally much more important role. These considerations made us specifically target the "Plant Computational and Quantitative Genomics" collection because the feasibility of our software for research in other kingdoms is unclear.

- **It could be beneficial if the App could also be run without RNA-Seq data for the development of a general hypothesis when comparing species for which RNA-Seq data under the same conditions is not available.**

  We agree with the reviewer's notion of not requiring RNA-seq data for all species in a workflow run. The latest release of the A2TEA.Workflow and WebApp now allow for this. The README of the Workflow explains that if the user does not possess transcriptomic data for a species, they can leave the genomic/cDNA fasta and annotation positions in the species.tsv file empty. We updated the WebApp to handle such cases of missing data. Therefore our whole pipeline has become considerably more flexible.

Competing Interests: No competing interests were disclosed.