Maximizing prediction of orphan genes in assembled genomes

Arun Seetharam[1], Urminder Singh[2], Jing Li[2], Priyanka Bhandary[2], Zeb Arendsee[2], and Eve Syrkin Wurtele[2,*]

[1]Genome Informatics Facility, Iowa State University, Ames, IA 50014, USA
[2]Department of Genetics, Development and Cell Biology, Center for Metabolic Biology, Iowa State University, Ames, IA 50014
[*]mash@iastate.edu

ABSTRACT

The evolutionary rapid emergence of new genes gives rise to “orphan genes” that share no sequence homology to genes in closely related genomes. These genes provide organisms with a reservoir of genetic elements to quickly respond to changing selection pressures. Gene annotation pipelines that combine *ab initio* machine-learning with sequence homology-based searches are efficient in identifying basal genes with a long evolutionary history. However, their ability to identify orphan genes and other young genes has not been systematically evaluated. Here, we classify the phylostrata of curated *Arabidopsis thaliana* genes and use these to assess the ability of two of the most prevalent annotation pipelines, MAKER and BRAKER, to predict orphans and other young genes. MAKER predictions are highly dependent on the RNA-Seq evidence, predicting between 11% and 60% of the orphan-genes and 95% to 98% of basal-genes in the annotated genome of Arabidopsis. In contrast, BRAKER consistently predicts 33% of orphan-genes and 98% of basal-genes. A less used method to identify genes is by directly aligning RNA-Seq data to the genome sequence. We present a Findable, Accessible, Interoperable and Reusable (FAIR) approach, called BIND, that mitigates the under-prediction of orphan genes. BIND combines BRAKER predictions with direct evidence-based inference of transcripts based on RNA-Seq alignments to the genome. BIND increases the number and accuracy of orphan gene predictions, identifying 68% of Araport11-annotated orphan genes and 99% of the conserved genes.

Introduction

Eukaryotic and prokaryotic genomes contain genes ("orphan genes") whose proteins are recognizable only in a single species; some of these have emerged *de novo* from the genome, while others have diverged so quickly as to be unrecognizable by homology[1–9]. Of the billions of extant orphan genes in eukaryotic species[10], the precise function of only a few is understood[2,10–12,12–19,19–25]. However, orphans appear to play a crucial role in adaptation to new biological niches. Forward and reverse genetics studies from vertebrates, annelids, insects, fungi, and plants show that extant proteins encoded by orphans interact with prey or predators, or integrate into established internal metabolic and developmental pathways[2,6,11,12,15–19,19–23,26–28]. A subset of orphans are retained in the genome and continue to evolve, such that each genome contains a mixture of genes of different ages (phylostrata)[1,3]. Phylostratigraphic analyses[3] indicates orphans played a role in development of new reproductive and neural structures across evolutionary time[29–31]. Thus, the advent of new genes may provide a critical enabler of speciation. Without the ability to accurately predict orphan genes and other young lineage-specific genes we are missing a big piece in understanding how life evolves.

Gene annotation is a fundamental step in genome sequencing projects. However, no standard best practice has been established, and protocols are diverse. Prevailing methods combine *homology-based analysis*, which compares a new genome to previously-identified genes from other species, with *ab initio prediction* of genes from the genome sequence[32–40]. Each approach may have an inherent bias against young genes[41]. Homology-based methods (content sensors) assume that genes have identifiable orthologs in other species. Because orphan genes have no homology to those of other species, homology-based methods are clearly not useful in predicting them. *Ab initio*-based predictors (signal sensors) assume a pre-defined gene structure for all the genes of an organism[42–45]. Typically, the predictor is trained using extrinsic RNA-Seq evidence, and possibly additional evidence specific to the species being annotated, to provide a gene model[33–35,37–39] . The nucleic acid signatures by which genes are predicted can include untranslated regions (UTR), translation start sites, termination sites, and intron-exon boundaries with standard splice donor and acceptors. Various *ab initio* algorithms show weaker predictions for short genes and exons[39,46], genes with alternative splicing[46], genes with long introns[47], highly conserved introns and non-coding sequences[48], overlapping/nested genes[49] and/or non-canonical splice sites[50].

A more straightforward approach to identify genes is to directly align RNA-Seq data to genomes[51]. Although this approach is important for predicting non-coding RNAs[52,53] and young genes[1,51,54–56], with some exceptions it has not been widely adopted for genome annotation[33,57]. This may be in part because direct inference predicts many genes that have not been annotated[55].

It has been suggested that canonical signatures are less well-defined in young genes, and therefore presumed that *ab initio* methods would be less likely to detect them[1,55,58–64]. However, the ability of *ab initio* approaches to detect genes of recent origin had not been directly evaluated. If young genes are indeed missed, diversifying the RNA-Seq and protein sequence that is extrinsically-supplied as training data might improve prediction of young genes. Here again, to our knowledge, the influence of extrinsically-supplied training data on the prediction outcome for young genes, or indeed for genes overall, had not been assessed.

Based on the postulates that prediction of younger genes is generally challenging, and that diverse RNA-Seq and protein sequence data used as evidence would increase detection of younger genes, we leverage the well-annotated *A. thaliana* genome, and RNA-seq and protein datasets of different sizes and depths, to evaluate the efficacy of five gene prediction pipelines (**Fig. 1**). From the multitude of variations of *ab initio* gene prediction pipelines, we focused on two of the most extensively-used software: MAKER[65], and BRAKER[66]. We also developed in-house pipelines that incorporate evidence-based gene inference directly from RNA-Seq data.

We show that, regardless of method or extrinsic evidence, younger genes were vastly under-detected. Furthermore, we quantify the extent to which diverse, RNA-Seq evidence can improve the annotation of genes in a genome. We provide a pipeline for improved gene annotation, BIND (**B**RAKER-**In**ferred **D**irectly); BIND combines *ab initio* prediction with predictions based on directly aligning RNA-Seq data to the genome.

Results

Phylostratograpy of Araport11-annotated *A. thaliana* genes

Phylostratigraphic designations of gene age are contingent on the ability, borne out by extensive evolutionary simulations[67], of the protein-BLAST algorithm to infer time of gene origin, spanning even temporally-distant homologs. The phylostratigraphic origins of the annotated protein-coding genes of the *A. thaliana* genome were inferred using the R-based pipeline, *phylostratr*[68], and the most recent community-curated, evidence-based annotation for *A. thaliana*, Araport11[69,70] (See **Supplementary File 1**). We similarly used *phylostratr* to infer the age of the potential proteins that would be translated from the longest ORF in any sequence predicted as a gene by a prediction pipeline, but not annotated in Araport11.

Phylostratographic analysis predicted 1,038 orphan genes, as compared to the 1,084 orphans inferred in a determination using the older TAIR10 *A. thaliana* gene annotations and less recent genome and protein sequence data[2]. The resultant phylostratal assignments were used to bench-mark gene prediction pipelines for their efficacy in predicting genes according to their inferred phylostrata.

RNA-Seq and protein used as extrinsic evidence for gene prediction scenario

Ab initio protocols are highly relied on for genome annotation. RNA-Seq provides training data to guide the performance of *ab initio* gene prediction programs[35,65,66,71]. There is little guidance in the literature on how to formulate the training data, with some genome annotation projects using raw reads from only 1-3 RNA-Seq samples as training evidence.Furthermore, in many new genome annotation projects only limited amounts of RNA-Seq data are available, and often these are developed as part of the annotation project.

To assess the effect of extrinsic evidence on prediction methods, we assembled six datasets of varied sizes and compositions (**Supplementary Table 1**). These are: a "Small", "Medium" and "Large" dataset, based on size; a “Pooled” dataset consisting of all of these three; an “orphan-rich" dataset comprised of the 38 samples (selected from a total of 1074 samples in SRA) that are most highly represented in Araport11-annotated orphan genes; and, a "ground truth" dataset composed of models of all Arabidopsis genes/proteins as annotated in Araport11 (“Ara11” dataset).

The "Small", "Medium", "Large", and Pooled datasets are real-world extrinsic evidence of similar sizes as might be used in genome annotation. The "orphan-rich" dataset is designed to maximize orphan gene representation. We reasoned that selecting samples which contain a breadth of orphan gene transcripts could be important because many, though by no means all[10], orphans are highly expressed under only a very limited set of conditions[10,14,54,72–78]. The "orphan-rich" RNA-Seq samples were chosen from over a thousand samples in SRA; over 60% of all Araport11-annotated orphans are transcribed in each sample.

Because Arabidopsis is a model species with high quality annotation, we were able to select a diverse RNA-Seq dataset comprised of samples rich in expression of orphan genes by determining the expression of all Araport11-annotated orphan genes across many samples. However, many species do not have a highly-vetted set of annotated orphan genes. For a species

that has an available body of RNA-Seq samples, an alternative strategy to identify diverse RNA-Seq samples rich in orphan gene expression would be: 1) run an initial gene annotation; 2) infer the phylostratum of each predicted gene; 3) realign raw reads from all available RNA-Seq samples to the predicted genes; 4) select those RNA-Seq samples with maximal representation of predicted orphan genes.

Raw reads of each dataset were aligned to the genome before being provided to BRAKER (which uses genome-aligned RNA-Seq data as training data) or being used as direct-inference evidence (which predicts genes by aligning RNA-Seq data directly to the genome). Raw reads of each dataset were aligned to the genome and then assembled into transcripts before being provided to MAKER (which uses assembled transcripts of RNA-Seq datasets along with predicted protein sequence, as training data). Protein sequence data used for MAKER was generated either by prediction from the RNA-Seq training data, or downloaded from Phytozyme. See **Supplementary Table 2** for the full set of prediction scenarios.

Gene predictions by MAKER

MAKER is a standard gene prediction pipeline, and the original impetus of this research was to test Maker's ability to predict orphan genes. An initial prediction annotated only 11 percent of the orphan genes present in Araport11. This led us to consider whether providing more diverse datasets as evidence for training MAKER's machine learning algorithms might improve the predictions.

Seven combinations of transcript and protein evidence were provided to MAKER (**Table 1** and **Supplementary Table 2**). Predictions were decreased when no protein data was provided **Supplementary Table 3**. Depending on the RNA-Seq plus protein training data, MAKER predicted between 80% to 94% of Araport11-annotated genes. For example, 22,065 of the Araport11-annotated genes were predicted with training data of the Small RNA-Seq dataset plus its predicted proteins, whereas 25,649 of the Araport11-annotated genes were predicted with training data from the Pooled dataset plus its predicted proteins (**Figure 2** and **Supplementary Table 3**).

Regardless of the evidence supplied, MAKER's ability to predict genes (**Fig. 2** and **Supplementary Table 3**) was greatest for the genes of the oldest phylostratum (Cellular Organisms, PS1) and progressively decreased for younger phylostrata.

The numbers of genes in younger phylostrata that were predicted varied up to five-fold, depending on the evidence supplied. 11% of the Araport11-annotated orphan genes were predicted if MAKER was provided the Small RNA-Seq dataset, versus 41% predicted from the Pooled dataset, and 60% for the orphan-rich dataset (**Fig. 2** and **Supplementary Table 3**).

The best predictions were obtained by providing MAKER with the complete set of Araport11-annotated genes and proteins as training data (i.e., 27,635 genes, 48,338 transcripts, with corresponding protein sequences); MAKER then predicted nearly 96% of all the genes. However, even with this gold-standard gene set, which contains all the Araport11-annotated orphans, only 70% of the Araport11-annotated orphans were predicted by MAKER.

The greater the diversity of the RNA-Seq evidence provided to MAKER, the more genes MAKER predicted that did not match any Araport11 gene annotation, for example, 2,048 for the Small dataset (1474 of which are orphan ORFs) and 14,147 for the Orphan-rich dataset (11,470 of which are orphan ORFs)(**Supplementary Table 4**). Precision and overall accuracy were lowest with the Orphan-rich dataset (**Supplementary Table 7**), because: 1) many predictions did not match any Araport11 gene, and 2) fewer ancient genes were predicted.

Gene predictions by BRAKER

Six RNA-Seq datasets were provided to BRAKER (**Table 1**and **Supplementary Table 2**). BRAKER predicted 93%-94% of all genes annotated by Araport11 (**Fig. 2** and **Supplementary Tables 3 and 7**). Regardless of training data, BRAKER yielded predictions that differed in quantity and accuracy by less that two percent. BRAKER uses RNA-Seq alignments to the genome for unsupervised training of GeneMark-ES/ET (**Supplementary Fig. 1**); however, BRAKER then selects a *subset* of the predicted protein coding genes to train AUGUSTUS, which makes the final gene predictions. This filtering of the extrinsic evidence prior to making the final predictions might explain why BRAKER provides similar predictions even with large variations in the makeup and sizes of the RNA-Seq dataset provided as extrinsic evidence. BRAKER's filtering of the RNA-Seq evidence enables good predictions from even the Small RNA-Seq dataset, but also means it can't take full advantage of very large, diverse RNA-Seq datasets, when used on its own.

For each prediction scenario, BRAKER's ability to predict genes was greatest for the genes of the oldest phylostratum ("Cellular organisms") and progressively decreased for younger phylostrata (**Fig. 2** and **Supplementary Table 3**). As a specific example, when provided the Pooled dataset as training data, BRAKER predicted 98% of genes that traced back to Cellular Organisms (PS1), but only 34% of the orphan genes (**Fig. 2**). 8,204 of the genes predicted by BRAKER were not annotated by Araport11 (1,498 of these were orphan-ORFs) (**Supplementary Tables 4 and 6**).

Gene predictions inferred directly from transcript evidence

Prediction by direct alignment of transcriptomic evidence to the genome is only occasionally used for gene annotation[40]. However, use of cDNAs and ESTS was standard for early annotations. We reasoned that non-canonical gene sequences could

be better identified by inferring them directly from RNA-Seq alignment. To use the transcript evidence directly, rather than as training data, RNA-Seq raw read data was aligned to the reference Arabidopsis genome and assembled. After concatenating the transcripts and removing the redundant ones, the ORF(s) in each inferred transcript were determined. The resultant transcripts were processed by Mikado[79], which removes redundant transcripts, consolidates information in a relational database, and selects representative transcripts for each locus. Because this approach directly relies on RNA-Seq alignments to the genome, if an RNA is not expressed under the conditions sampled for RNA-Seq evidence, it will not be detected. Thus, providing a wide variety of RNA-Seq data collected from various developmental stages, tissues and under various conditions is particularly important. For example, the Orphan-rich dataset, by far the most diverse RNA-Seq dataset used in this study, predicted nearly 96% of all Araport11-annotated genes, and 69% of Araport11-annotated orphans, while the Pooled dataset predicted about 70% of all Araport11-annotated genes, and 41% of the orphans (**Fig. 2** and **Supplementary Table 3**).

Gene predictions combining *ab initio* predictions with transcripts inferred directly (BIND or MIND)

Because *ab initio* methods can identify very high proportions of the conserved genes, while direct inference can identify those genes that are expressed in the RNA-Seq data evidence without regard to canonical structure or homology to genes in other organisms, we tested whether combining both approaches would maximize gene predictions across phylostrata. **(Fig. 1)**.

When BRAKER predictions were combined with transcripts INferred Directly (BIND), genes matching Araport11 annotations increased over either method alone. Similar results were observed when MAKER predictions were combined with transcripts INferred Directly (MIND) (**Fig. 2** and **Supplementary Table 3**). BIND and MIND, using the Orphan-rich dataset as input, predicted more Araport11-annotated orphan genes than did either *ab initio* predictor alone. Basal F1 scores for overall prediction accuracy were comparable for BIND and MIND (75-77%). MIND predicted 11,642 genes that did not match any Araport11 genes, whereas BIND predicted 6,686. Irrespective of RNA-Seq data input, the most accurate representation of all genes, and young genes in particular, was BIND, which combined direct inference with *ab initio* detection by BRAKER. (**Supplementary Table 6**).

Discussion

Structural annotation of genes in a genome is critical to making genomics data useful to the research community. In experimental studies, standard practice is to align RNA-Seq data only to the annotated genes; thus, any gene that is not annotated will be completely missed. However, genes are annotated by a wide variety of protocols, which are at times sparsely or unclearly documented, with a strong emphasis on *ab initio* prediction from the genome. Although RNA-Seq is almost universally supplied as training data, a wide variety of sizes and diversities of RNA-Seq data have been used. Few annotation pipelines are benchmarked against "ground-truth" genes and genomes. Many genome annotation projects routinely filter from the list of predicted genes sequences that are "too short" (eliminating some orphan genes), have only one exon (eliminating many orphan genes), and/or that have no known homologs (eliminating all orphan genes). Although these strategies minimize false positives, they also exclude many bonafide genes. Here, we deploy the highly curated, community-based gene annotations from the model species, *A. thaliana*, as gold standard[70,80] to explicitly illustrate the challenge of annotating young genes. The Arabidopsis RNA-Seq datasets compiled for this study, along with the phylostratal designations of genes, are available for benchmarking other gene prediction pipelines.

As we demonstrate by their low visibility to *ab initio* annotation, young genes have a less recognizable sequence signature than their more ancient counterparts, which have undergone hundreds of millions of years of selection. The efficacy of every gene prediction scenario we tested, regardless of the pipeline, or the amount or quality of the evidence supplied, was strongly dependent on the phylostratal origin of the gene. This confirms earlier speculation that young genes have minimal canonical structure and thus may be more difficult to identify by *ab initio* approaches[1–3,55,81,82]. The clear trend is that the more ancient a gene, the more likely it is to be predicted.

Our study reveals that the selection of RNA-Seq training data evidence greatly affects *ab initio* MAKER genome annotation, particularly for young genes. Using RNA-Seq data with high representation of orphan gene expression dramatically improves the efficacy of gene prediction by MAKER, by the Direct Inference pipeline, and by the BIND and MIND combination pipelines. Thus, the accuracy of an initial genome annotation will be maximized by selecting RNA-Seq data from diverse samples, including samples that typically express high levels of young genes, e.g., reproductive tissues and stressed tissues[10,14,54,72–78,83]. Additionally, reannotation[70,84,85] outcomes can be optimized by utilizing evidence from diverse samples in the expanded body of public RNA-Seq datasets.

Each scenario that we tested predicted genes that are not annotated in Araport11. Compelling RNA-seq evidence from wide range of phyla, ranging from bacteria, to yeast, to humans, shows that many sequences that are not annotated as genes are highly transcribed and translated[1,55,56,62,86–90]. For example, in a study of RNA-Seq data from over 3000 diverse yeast

samples, about half of the transcripts are not annotated as genes by Saccharomyces Genome Database (SGD)[55]. A large portion of these unannotated transcripts might be “transcriptional noise"[51] or, as suggested by[62], fodder for *de novo* gene evolution. Other sequences may be on a continuum from protogene to gene to pseudogene and/or to oblivion[1,91]. No clear criteria have been established to assign these products of this "dark transcriptome" to any particular category.

Essential protein-mediated functions have been experimentally demonstrated for some transcripts that had not been annotated as genes[92,93]. The potential that translated transcripts that encode orphan proteins could be useful, although they have only "recently" been subjected to selection pressure, has been reinforced by synthetic biology research revealing that randomly-generated or evolutionarily-selected peptides with no clear homology to other proteins are often able to bind small molecules *in vivo*[94], and can have beneficial consequences when expressed *in vivo*, inducing developmental, stress-resistance, and longevity phenotypes[94,95]. If information on the expression of the dark transcriptome was more easily accessible, potential roles of these transcripts could be better considered.

We argue that best practice would be a universal, evidence-based standard for gene annotation, along with a more inclusive annotation of predicted genes (See Supplementary Table 12 for an example). This could enable researchers to prioritize genes (or candidate genes) for experimental study. A case for routinely incorporating translation evidence has been logically described[61,96]. *In silico* evidence could include: (*ab initio* prediction; homology; direct inference; inferred phylostrata; syntenic analysis; functional annotation). Experimental evidence could include: transcription (RNA-Seq, cDNA); translation (Ribo-Seq, proteomics); genetic screens; targeted experimental studies. This more inclusive approach would provide both annotated genes and expressed transcripts (candidate genes), together with their evidence. Thus, the signal of candidate genes would be retained in the processed data. Ultimately, providing broad, straightforward access to information on predicted genes would facilitate understanding of genome evolution and function.

In summary, we demonstrate that orphan genes and other young genes often elude annotation in new genomes. Our findings support annotating genes by: 1) combining an *ab initio*/ homology-based pipeline with a direct inference approach, e.g., BIND; 2) testing out new gene prediction pipelines on well-sequenced genomes with “gold standard” gene annotations for their ability to predict “annotated genes; 3) including diverse transcriptomic evidence; and 4) iteratively reannotating the genome, leveraging new RNA-Seq and other evidence.

Methods

Complete methods, all scripts used in this study, and all results files are documented on GitHub (https://github.com/eswlab/orphan-prediction).

Data download

RNA-Seq raw read data were downloaded from NCBI SRA[97] as pre-filtered reads using the SRA-toolkit (v2.8.0)[97]. RNA-Seq used as training data input to *ab initio* predictors and for direct alignments is specified in **Table 1** and **Supplementary Table 1**. To create the orphan-rich dataset, RNA-Seq samples were queried from the NCBI SRA web interface with the expression: ("Arabidopsis thaliana"[Organism] AND "filetype fastq"[Filter] AND "paired"[Layout] AND "illumina"[Platform] AND "transcriptomic"[Source]) This query returned 1074 individual RNA-Seq samples, which we downloaded and converted to fastq format using SRA-toolkit. The reads from each sample were mapped and number of reads mapping to each Araport11-annotated CDS sequence were computed using Kallisto[98]. The 1074 RNA-Seq samples were ranked by number of expressed Araport11-annotated orphan genes (TPM > 0 was considered expressed), and the 38 samples with the greatest number of expressed orphans (more than 60% of the annotated orphans were expressed in each sample) were chosen to create a “orphan-rich” dataset. All other RNA-Seq datasets were generated by downloading SRRs directly from NCBI SRA using BioProject IDs.

The *A. thaliana* reference genome version (Araport11), GFF3 files, transcripts, and protein sequences for *A. thaliana* (Araport11 version) were downloaded from TAIR[70]. Protein sequence used as evidence in MAKER[65] were: 1) generated by assembling RNA-Seq reads using Trinity (v2.6.6)[99], followed by open reading frame (ORF) prediction and translation using TransDecoder (v3.0.1)[99] or, 2) downloaded from Phytozome[100] as predicted protein sequences for nine species: *Arabidopsis thaliana*, (*Glycine max*, *Populus trichocarpa*, *Arabidopsis lyrata*, *Conradina grandiflora*, *Setaria italica*, *Oryza sativa*, *Physcomitrella patens*, *Chlamydomonas reinhardtii*, and *Brassica rapa*).

BRAKER

RNA-Seq raw reads were mapped to the indexed *Arabidopsis* genome using HiSat2 aligner (v2.1.0)[101] (default settings). The resultant SAM files were sorted and converted to BAM format. BAM files from each set of RNAseq samples were combined using SAMTools (v1.9)[102] and provided as training for the BRAKER (v2.1.2) pipeline[66], along with the unmasked *A. thaliana* genome (Araport11). BRAKER is an automated pipeline to predict genes using GeneMark-ET (v4.33)[38] and

AUGUSTUS (v3.3.1)[103]. Briefly, GeneMark-ET is used for iterative training of AUGUSTUS, by generating initial gene predictions. GeneMark-ET-predicted genes are filtered and provided for AUGUSTUS training, followed by AUGUSTUS prediction, integrating the RNA-Seq information, to generate the final gene predictions (Fig. 1). In additional analyses, we compared results using the translated protein sequence from the transcripts as evidence. BRAKER results with and without protein evidence were virtually identical; the BRAKER User Guide notes that protein evidence is useful only if there is minimal RNA-Seq evidence.

MAKER

To implement the MAKER (v2.31.10)[65] pipeline, RNA-Seq data was assembled into a transcriptome using Trinity (v2.6.6)[99]; this CDS evidence, was supplied along with the unmasked *A. thaliana* genome (Araport11). Depending on the case study, either CDS-only; CDS and translated proteins; or CDS and Phytozome proteins were supplied (Supplementary **Tables 1 and 2**).

MAKER was run in two successive rounds with default settings (**Fig. 1**). In round one, transcriptome and protein data were aligned to the reference genome to generate crude gene predictions. These crude predictions were then used for training SNAP (release 2006-07-28)[37] and AUGUSTUS (v3.2.1) *ab initio* gene predictors with default options. In round two, the Hidden Markov Models (HMM) for *ab initio* gene predictors, along with self-trained HMM of GeneMark-ES (v4.32) were used within MAKER to predict genes. MAKER finalizes the comprehensive sets of genes from all three predictors were ranked using Annotation Edit Distance (AED)[104]; the highest-ranking genes were retained for the final set of predictions. MAKER's default output includes key metadata about gene predictions (evidence scores supporting each prediction, name of the component(s) within MAKER that generated the prediction).

Computer allocations and ease of use for MAKER and BRAKER

BRAKER had much simpler prerequisites than MAKER, which had a large overhead, starting from installation through finalizing the predictions. BRAKER only required a user to execute a single command-line operation (once the raw reads were mapped to the genome), whereas MAKER required longer preparation time (assembling transcripts, collecting evidence, training gene predictors) running multiple iterations and has moderate to heavy scripting requirements. Run-time for BRAKER (359 CPU hours) is several orders of magnitude shorter than for MAKER (1536 CPU hours, not including the Trinity/TransDecoder run-time), if each is provided with similar training data. BRAKER is more efficient in terms of disk usage, disk I/O (Input/Output), and unlike MAKER, it does not create millions of intermediary files. BRAKER is easier to install and use, and has fewer computational prerequisites than MAKER. New approaches are being developed to facilitate MAKER use[105].

Direct inference-based

Raw RNA-seq reads were assembled using four genome-guided transcriptome assemblers: *viz.* Class2, Trinity, StringTie and CuffLinks[99,106–108]. The BAM file generated by mapping reads to the Araport11-annotated indexed genome using HiSat2 (v2.1.0)[101] was provided as training for the assemblers. The resultant assembled transcripts were used to predict ORFs using Transdecoder[99], and for BLAST[109] alignments against the Araport11-annotated genome (XML format). We selected those ORFs over 150 nt. (Other user requirements might include: transcript length, number of exons, exon length, intron length, start/end codon, expression value, or presence of UTRs.) These data files, along with splicing junctions identified from the alignments using Portcullis[110] were provided as input to Mikado[79]. Except for the number of threads (set to 28), Mikado pick was run with default parameters.

Combined gene predictions by MIND and BIND

The total predictions from Direct Inference were integrated with BRAKER predictions (BIND) or with MAKER predictions (MIND), using Evidence Modeler (EVM)[111] to merge predictions. Briefly, the weights file was prepared by setting the direct evidence to 2 and *ab initio* predictions to 1. EVM was run as per the guidelines, with default options, using GFF3 files, genome file and weights file as input. The merged predictions were finalized in GFF3 format.

Accuracy calculations

Accuracy calculations were based on sensitivity (Sn), precision (Pr), and the combined accuracy score, F1[112]. Sn is a measure of the percent of genes identified by a given set of predictions relative to the genes in the gold-standard Araport11 annotations (true positives), and calculated as [predictions/all genes in Araport11* 100]. Precision (Pr) is a measure of how specific the predictions are, and was calculated as [Pr= true positives/(all predictions) * 100]. The F1 score combines the sensitivity and precision as a measure of accuracy [2(sensitivity*precision)/(sensitivity+precision)].

Phylostratography

Phylostratographic analysis was implemented using *phylostratr* software (v0.2.0)[68]. Phylostratal origin of genes based on the homology of predicted proteins to proteins in clades of increasing depth (age) was inferred using the default settings of

phylostratr[68]. The focal species was set as "3702" (NCBI taxa id for *A. thaliana*). A database of protein sequences was created by *phylostratr* from 138 species, including custom species selected based on evolutionary clade and genome/proteome quality. A customized, more comprehensive, set of proteins was provided for several species, and protein sequences of the remaining species were downloaded from Uniprot

Comparing prediction scenarios

The results of each gene prediction pipeline scenario were compared to the Araport11 annotations using Mikado Compare[79]; gene structure annotation predictions were provided as GFF3 files. Similarity statistics are reported for each gene locus individually. Compiled reports of similarity statistics, phylostratography of predictions and reference, were generated using custom bash scripts and final plots were generated using custom R-Script (see https://github.com/eswlab/orphan-prediction for code).

References

1. Carvunis, A. R. *et al.* Proto-genes and de novo gene birth. *Nature* **487**, 370–374 (2012).
2. Arendsee, Z. W., Li, L. & Wurtele, E. S. Coming of age: orphan genes in plants. *Trends plant science* **19**, 698–708 (2014).
3. Tautz, D. & Domazet-Loso, T. The evolutionary origin of orphan genes. *Nat. Rev. Genet.* **12**, 692–702 (2011).
4. Gibson, A. K., Smith, Z., Fuqua, C., Clay, K. & Colbourne, J. K. Why so many unknown genes? partitioning orphans from a representative transcriptome of the lone star tick amblyomma americanum. *BMC genomics* **14**, 135 (2013).
5. Fischer, D. & Eisenberg, D. Finding families for genomic orfans. *Bioinforma. (Oxford, England)* **15**, 759–762 (1999).
6. Khalturin, K., Hemmrich, G., Fraune, S., Augustin, R. & Bosch, T. C. More than just orphans: are taxonomically-restricted genes important in evolution? *Trends Genet.* **25**, 404–413 (2009).
7. Gollery, M. *et al.* What makes species unique? the contribution of proteins with obscure features. *Genome biology* **7**, R57 (2006).
8. Wilson, G. *et al.* Orphans as taxonomically restricted and ecologically important genes. *Microbiology* **151**, 2499–2501 (2005).
9. Arendsee, Z. *et al.* fagin: synteny-based phylostratigraphy and finer classification of young genes. *BMC bioinformatics* **20**, 1–14 (2019).
10. Bhandary, P., Seetharam, A. S., Arendsee, Z. W., Hur, M. & Wurtele, E. S. Raising orphans from a metadata morass: A researcher's guide to re-use of public'omics data. *Plant science* **267**, 32–47 (2018).
11. Li, L. *et al.* Identification of the novel protein qqs as a component of the starch metabolic network in arabidopsis leaves. *The Plant J.* **58**, 485–498 (2009).
12. Hahnel, S., Parker-Manuel, R., Dissous, C., Cailliau, K. & Grevelding, C. G. First characterization of smopg1, a novel protein involved in gonad-associated processes in schistosoma mansoni. *Mol. Biochem. Parasitol.* **213**, 22–25 (2017).
13. Ni, F. *et al.* Wheat ms2 encodes for an orphan protein that confers male sterility in grass species. *Nat. Commun.* **8** (2017).
14. Perochon, A. *et al.* Tafrog encodes a pooideae orphan protein that interacts with snrk1 and enhances resistance to the mycotoxigenic fungus fusarium graminearum. *Plant physiology* **169**, 2895–2906 (2015).
15. Li, D. *et al.* A de novo originated gene depresses budding yeast mating pathway and is repressed by the protein encoded by its antisense strand. *Cell Res.* **20**, 408–420 (2010).
16. Xiao, W. *et al.* A rice gene of de novo origin negatively regulates pathogen-induced defense response. *PloS ONE* **4**, e4603 (2009).
17. Luhua, S., Ciftci-Yilmaz, S., Harper, J., Cushman, J. & Mittler, R. Enhanced tolerance to oxidative stress in transgenic *Arabidopsis* plants expressing proteins of unknown function. *Plant Physiol.* **148**, 280–292 (2008).
18. Jones, D. C. *et al.* A Clade-Specific *Arabidopsis* Gene Connects Primary Metabolism and Senescence. *Front Plant Sci* **7**, 983 (2016).
19. Mayer, M. G. & Sommer, R. J. Nematode orphan genes are adopted by conserved regulatory networks and find a home in ecology. *Worm* **4**, e1082029 (2015).
20. DeVries, A. L. Glycoproteins as biological antifreeze agents in antarctic fishes. *Science* **172**, 1152–1155 (1971).

21. Chen, L., DeVries, A. L. & Cheng, C. H. Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish. *Proc. Natl. Acad. Sci. U.S.A.* **94**, 3811–3816 (1997).

22. Zhuang, X. & Cheng, C. H. ND6 gene "lost" and found: evolution of mitochondrial gene rearrangement in Antarctic notothenioids. *Mol. Biol. Evol.* **27**, 1391–1403 (2010).

23. Brinkman, D. L. *et al.* Venom proteome of the box jellyfish Chironex fleckeri. *PLoS ONE* **7**, e47866 (2012).

24. Wilkins, K. A. *et al.* Self-incompatibility-induced programmed cell death in field poppy pollen involves dramatic acidification of the incompatible pollen tube cytosol. *Plant physiology* **167**, 766–779 (2015).

25. Qi, M. *et al.* Qqs orphan gene and its interactor nf-yc 4 reduce susceptibility to pathogens and pests. *Plant biotechnology journal* **17**, 252–263 (2019).

26. Luhua, S. *et al.* Linking genes of unknown function with abiotic stress responses by high-throughput phenotype screening. *Physiol Plant* **148**, 322–333 (2013).

27. Li, L. & Wurtele, E. S. The qqs orphan gene of arabidopsis modulates carbon and nitrogen allocation in soybean. *Plant biotechnology journal* **13**, 177–187 (2015).

28. Calvete, J. J. Venomics: integrative venom proteomics and beyond. *Biochem. J.* **474**, 611–634, DOI: 10.1042/BCJ20160577 (2017). http://www.biochemj.org/content/474/5/611.full.pdf.

29. Šestak, M. S., Božičević, V., Bakarić, R., Dunjko, V. & Domazet-Lošo, T. Phylostratigraphic profiles reveal a deep evolutionary history of the vertebrate head sensory systems. *Front. zoology* **10**, 18 (2013).

30. Neme, R. & Tautz, D. Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC genomics* **14**, 117 (2013).

31. Lei, L., Steffen, J. G., Osborne, E. J. & Toomajian, C. Plant organ evolution revealed by phylotranscriptomics in arabidopsis thaliana. *Sci. reports* **7**, 7567 (2017).

32. Yandell, M. & Ence, D. A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.* **13**, 329–342 (2012).

33. Cook, D. E. *et al.* Long-read annotation: automated eukaryotic genome annotation based on long-read cdna sequencing. *Plant physiology* **179**, 38–54 (2019).

34. Reid, I. *et al.* Snowyowl: accurate prediction of fungal genes by using rna-seq and homology information to select among ab initio models. *BMC bioinformatics* **15**, 229 (2014).

35. Mudge, J. M. & Harrow, J. The state of play in higher eukaryote gene annotation. *Nat. Rev. Genet.* **17**, 758–772 (2016).

36. Borodovsky, M., Rudd, K. E. & Koonin, E. V. Intrinsic and extrinsic approaches for detecting genes in a bacterial genome. *Nucleic acids research* **22**, 4756–4767 (1994).

37. Korf, I. Gene finding in novel genomes. *BMC bioinformatics* **5**, 59 (2004).

38. Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y. O. & Borodovsky, M. Gene identification in novel eukaryotic genomes by self-training algorithm. *Nucleic acids research* **33**, 6494–6506 (2005).

39. Steijger, T. *et al.* Assessment of transcript reconstruction methods for rna-seq. *Nat. methods* **10**, 1177 (2013).

40. Pilkington, S. M. *et al.* A manually annotated actinidia chinensis var. chinensis (kiwifruit) genome highlights the challenges associated with draft genomes and gene prediction in plants. *BMC genomics* **19**, 257 (2018).

41. Klasberg, S., Bitard-Feildel, T. & Mallet, L. Computational identification of novel genes: current and future perspectives. *Bioinforma. Biol. insights* **10**, BBI–S39950 (2016).

42. Veeckman, E., Ruttink, T. & Vandepoele, K. Are we there yet? reliably estimating the completeness of plant genome sequences. *The Plant Cell* **28**, 1759–1768 (2016).

43. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic dna. *J. Mol. Biol.* **268**, 78–94 (1997).

44. Keilwagen, J., Hartung, F., Paulini, M., Twardziok, S. O. & Grau, J. Combining rna-seq data and homology-based gene prediction for plants, animals and fungi. *BMC bioinformatics* **19**, 189 (2018).

45. Keilwagen, J., Hartung, F., Paulini, M., Twardziok, S. O. & Grau, J. Combining rna-seq data and homology-based gene prediction for plants, animals and fungi. *BMC bioinformatics* **19**, 189 (2018).

46. Foissac, S. & Schiex, T. Integrating alternative splicing detection into gene prediction. *BMC bioinformatics* **6**, 25 (2005).

47. Wang, J. *et al.* Vertebrate gene predictions and the problem of large genes. *Nat. Rev. Genet.* **4**, 741–749 (2003).

48. Duret, L., Dorkeld, F. & Gautier, C. Strong conservation of non-coding sequences during vertebrates evolution: potential involvement in post-transcriptional regulation of gene expression. *Nucleic acids research* **21**, 2315–2322 (1993).

49. Quesada, V., Ponce, M. R. & Micol, J. L. Otc and aul1, two convergent and overlapping genes in the nuclear genome of arabidopsis thaliana. *FEBS letters* **461**, 101–106 (1999).

50. Burset, M., Seledtsov, I. & Solovyev, V. Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic acids research* **28**, 4364–4375 (2000).

51. Pertea, M. *et al.* Thousands of large-scale rna sequencing experiments yield a comprehensive new human gene list and reveal extensive transcriptional noise. *bioRxiv* 332825 (2018).

52. Ulitsky, I. Evolution to the rescue: using comparative genomics to understand long non-coding rnas. *Nat. Rev. Genet.* **17**, 601 (2016).

53. Rahman, R.-U. *et al.* Oasis 2: improved online analysis of small rna-seq data. *BMC bioinformatics* **19**, 54 (2018).

54. Zhao, L., Saelao, P., Jones, C. D. & Begun, D. J. Origin and spread of de novo genes in drosophila melanogaster populations. *Science* **343**, 769–772 (2014).

55. Li, J., Arendsee, Z., Singh, U. & Wurtele, E. S. Recycling rna-seq data to identify candidate orphan genes for experimental analysis. *bioRxiv* 671263 (2019).

56. Wang, X. *et al.* Full-length transcriptome reconstruction reveals a large diversity of RNA and protein isoforms in rat hippocampus. *Nat. Commun.* **10**, 1–15, DOI: 10.1038/s41467-019-13037-0 (2019).

57. Tretina, K. *et al.* Re-annotation of the theileria parva genome refines 53% of the proteome and uncovers essential components of n-glycosylation, a conserved pathway in many organisms. *bioRxiv* DOI: 10.1101/749366 (2019). https://www.biorxiv.org/content/early/2019/08/31/749366.full.pdf.

58. Hoff, K. J. & Stanke, M. Predicting genes in single genomes with augustus. *Curr. protocols bioinformatics* **65**, e57 (2019).

59. Wu, B. & Knudson, A. Tracing the De Novo Origin of Protein-Coding Genes in Yeast. *mBio* **9**, e01024–18, DOI: 10.1128/mBio.01024-18 (2018).

60. Gonzalez-Munoz, E. *et al.* The maize (Zea mays ssp. mays var. B73) genome encodes 33 members of the purple acid phosphatase family. *Front Plant Sci* **6**, 341 (2015).

61. Ruiz-Orera, J. & Albà, M. M. Conserved regions in long non-coding rnas contain abundant translation and protein–rna interaction signatures. *NAR Genomics Bioinforma.* **1**, e2–e2 (2019).

62. Ruiz-Orera, J. & Albà, M. Frequent translation of small open reading frames in evolutionary conserved lncrna regions. *bioRxiv* DOI: 10.1101/348326 (2018). https://www.biorxiv.org/content/early/2018/06/16/348326.full.pdf.

63. Blevins, W. R. *et al.* Frequent birth of de novo genes in the compact yeast genome. *bioRxiv* 575837 (2019).

64. Lu, T.-C., Leu, J.-Y. & Lin, W.-C. A Comprehensive Analysis of Transcript-Supported De Novo Genes in Saccharomyces sensu stricto Yeasts. *Mol. Biol. Evol.* **34**, 2823–2838, DOI: 10.1093/molbev/msx210 (2017).

65. Holt, C. & Yandell, M. Maker2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinforma.* **12**, 491, DOI: 10.1186/1471-2105-12-491 (2011).

66. Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M. & Stanke, M. Braker1: Unsupervised rna-seq-based genome annotation with genemark-et and augustus. *Bioinformatics* **32**, 767, DOI: 10.1093/bioinformatics/btv661 (2016).

67. Domazet-Lošo, T. *et al.* No evidence for phylostratigraphic bias impacting inferences on patterns of gene emergence and evolution. *Mol. biology evolution* **34**, 843–856 (2017).

68. Arendsee, Z. *et al.* phylostratr: a framework for phylostratigraphy. *Bioinformatics* DOI: 10.1093/bioinformatics/btz171 (2019). http://oup.prod.sis.lan/bioinformatics/advance-article-pdf/doi/10.1093/bioinformatics/btz171/28311890/btz171.pdf.

69. Cheng, C.-Y. *et al.* Araport11: a complete reannotation of the arabidopsis thaliana reference genome. *The Plant J.* **89**, 789–804 (2017).

70. Berardini, T. Z. *et al.* The arabidopsis information resource: making and mining the "gold standard" annotated reference plant genome. *genesis* **53**, 474–485 (2015).

71. Wang, Z., Gerstein, M. & Snyder, M. Rna-seq: a revolutionary tool for transcriptomics. *Nat. reviews genetics* **10**, 57–63 (2009).

72. Kaur, N. *et al.* Transcriptome analysis of the whitefly, bemisia tabaci meam1 during feeding on tomato infected with the crinivirus, tomato chlorosis virus, identifies a temporal shift in gene expression and differential regulation of novel orphan genes. *BMC genomics* **18**, 370 (2017).

73. Li, L. & Wurtele, E. S. The QQS orphan gene of *Arabidopsis* modulates carbon and nitrogen allocation in Soybean. *Plant Biotechnol. J.* **13**, 177–187 (2015).

74. Donoghue, M. T., Keshavaiah, C., Swamidatta, S. H. & Spillane, C. Evolutionary origins of brassicaceae specific genes in arabidopsis thaliana. *BMC evolutionary biology* **11**, 47 (2011).

75. Khraiwesh, B. *et al.* Genome-wide expression analysis offers new insights into the origin and evolution of physcomitrella patens stress response. *Sci. reports* **5** (2015).

76. Colbourne, J. K. *et al.* The ecoresponsive genome of daphnia pulex. *science* **331**, 555–561 (2011).

77. Doughty, T. W. *et al.* Young genes are more responsive to environmental stress than ancient genes in budding yeasts. *BioRxiv* 660274 (2019).

78. Li, G. *et al.* Orphan genes are involved in drought adaptations and ecoclimatic-oriented selections in domesticated cowpea. *J. Exp. Bot.* (2019).

79. Venturini, L., Caim, S., Kaithakottil, G., Mapleson, D. L. & Swarbreck, D. Leveraging multiple transcriptome assembly methods for improved gene structure annotation. *bioRxiv* 216994 (2017).

80. Consortium, I. A. I. *et al.* Arabidopsis bioinformatics resources: The current state, challenges, and priorities for the future. *Plant Direct* **3**, e00109 (2019).

81. Werner, M. S. *et al.* Young genes have distinct gene structure, epigenetic profiles, and transcriptional regulation. *Genome research* **28**, 1675–1687 (2018).

82. Stein, J. C. *et al.* Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus oryza. *Nat. genetics* **50**, 285 (2018).

83. Dion-Cote, A.-M. A hotspot for new genes. *eLife* **8** (2019).

84. Yang, H. *et al.* Re-annotation of eight drosophila genomes. *Life science alliance* **1**, e201800156 (2018).

85. Mitchell, N. M. *et al.* Proteogenomic re-annotation of coccidioides posadasii strain silveira. *Proteomics* **18**, 1700173 (2018).

86. Lin, D., Yin, X., Wang, X., Zhou, P. & Guo, F.-B. Re-annotation of protein-coding genes in the genome of saccharomyces cerevisiae based on support vector machines. *PLOS ONE* **8**, 1–6, DOI: 10.1371/journal.pone.0064477 (2013).

87. Ji, Z., Song, R., Regev, A. & Struhl, K. Many lncrnas, 5'utrs, and pseudogenes are translated and some are likely to express functional proteins. *eLife* **4**, e08890, DOI: 10.7554/eLife.08890 (2015).

88. Gao, H. *et al.* Long non-coding rnas associated with metabolic traits in human white adipose tissue. *EBioMedicine* **30**, 248–260 (2018).

89. Ruiz-Orera, J. *et al.* Origins of de novo genes in human and chimpanzee. *PLoS genetics* **11**, e1005721 (2015).

90. Delcourt, V., Staskevicius, A., Salzet, M., Fournier, I. & Roucou, X. Small proteins encoded by unannotated orfs are rising stars of the proteome, confirming shortcomings in genome annotations and current vision of an mrna. *Proteomics* **18**, 1700058 (2018).

91. Van Oss, S. B. & Carvunis, A.-R. De novo gene birth. *PLoS genetics* **15**, e1008160 (2019).

92. Ji, Z., Song, R., Regev, A. & Struhl, K. Many lncRNAs, 5'UTRs, and pseudogenes are translated and some are likely to express functional proteins. *eLife* **4**, e08890 (2015).

93. Andrews, S. J. & Rothnagel, J. A. Emerging evidence for functional peptides encoded by short open reading frames. *Nat. Rev. Genet.* **15**, 193 (2014).

94. Bao, Z., Clancy, M. A., Carvalho, R. F., Elliott, K. & Folta, K. M. Identification of novel growth regulators in plant populations expressing random peptides. *Plant physiology* **175**, 619–627 (2017).

95. Neme, R., Amador, C., Yildirim, B., McConnell, E. & Tautz, D. Random sequences are an abundant source of bioactive RNAs or peptides. *Nat. Ecol. & Evol.* **1**, 0127+ (2017).

96. Ren, Z. *et al.* Improvements to the rice genome annotation through large-scale analysis of rna-seq and proteomics data sets. *Mol. & Cell. Proteomics* **18**, 86–98 (2019).

97. Leinonen, R., Sugawara, H. & Shumway, M. The sequence read archive. *Nucleic acids research* gkq1019 (2010).

98. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic rna-seq quantification. *Nat. biotechnology* **34**, 525 (2016).

99. Haas, B. J. *et al.* De novo transcript sequence reconstruction from rna-seq using the trinity platform for reference generation and analysis. *Nat. protocols* **8**, 1494–1512 (2013).

100. Goodstein, D. M. *et al.* Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.* **40**, D1178, DOI: 10.1093/nar/gkr944 (2012).

101. Kim, D., Langmead, B. & Salzberg, S. L. Hisat: a fast spliced aligner with low memory requirements. *Nat. methods* **12**, 357–360 (2015).

102. Li, H. *et al.* The sequence alignment/map format and samtools. *Bioinformatics* **25**, 2078–2079 (2009).

103. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cdna alignments to improve de novo gene finding. *Bioinformatics* **24**, 637–644 (2008).

104. Eilbeck, K., Moore, B., Holt, C. & Yandell, M. Quantitative measures for the management and comparison of annotated genomes. *BMC bioinformatics* **10**, 67 (2009).

105. Wang, L., Lu, Z., Van Buren, P. & Ware, D. Sciapps: a cloud-based platform for reproducible bioinformatics workflows. *Bioinformatics* **34**, 3917–3920 (2018).

106. Song, L., Sabunciyan, S. & Florea, L. Class2: accurate and efficient splice variant annotation from rna-seq reads. *Nucleic acids research* **44**, e98–e98 (2016).

107. Pertea, M. *et al.* Stringtie enables improved reconstruction of a transcriptome from rna-seq reads. *Nat. biotechnology* **33**, 290 (2015).

108. Trapnell, C. *et al.* Differential gene and transcript expression analysis of rna-seq experiments with tophat and cufflinks. *Nat. protocols* **7**, 562 (2012).

109. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. molecular biology* **215**, 403–410 (1990).

110. Mapleson, D., Venturini, L., Kaithakottil, G. & Swarbreck, D. Efficient and accurate detection of splice junctions from rna-seq with portcullis. *GigaScience* **7**, giy131 (2018).

111. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using evidencemodeler and the program to assemble spliced alignments. *Genome biology* **9**, R7 (2008).

112. Powers, D. M. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *J. Mach. Learn. Technol.* **2**, 37–63 (2011).

Acknowledgements

We thank Andrew J. Severin, Genome Informatics Facility, Iowa State University, for assistance with data analysis and helpful discussions. We are grateful to the Center for Metabolic Biology and the High Performance Computing Facility, and Research IT at Iowa State University for providing support and cyberinfrastructure facilities that made this study possible. Many thanks to Urminder Singh, Jing Li and Priyanka Bhandary who provided valuable insights and participated in discussions. This study is based upon work supported by the National Science Foundation under Grant No. IOS 1546858.

Author contributions statement

AS designed and performed the experiments, analyzed the data, and wrote the paper.
ZA performed parts of the phylostratagraphic analysis and wrote related parts of the paper. ESW conceived the study, contributed annotations and analysis tools, and wrote the paper.

Additional information

To include, in this order: **Accession codes** (where applicable); **Competing interests** (mandatory statement).

The corresponding author is responsible for submitting a competing interests statement on behalf of all authors of the paper. This statement must be included in the submitted article file.

Figures and tables can be referenced in LaTeX using the ref command, e.g. Figure 2 and Table 1

ShortName	Analysis Pipeline	Description	SRR samples used	Datasize (Gbp)
Ara11	Araport11	Araport11 version of *A. thaliana* annotations	NA	NA
Make-small+trans	MAKER	Maker predictions using Small dataset (assembled transcripts and translated proteins)	12	12.8
Make-pool+trans		Maker predictions using Pooled dataset (assembled transcripts and translated proteins)	77	241.4
Brake-small	BRAKER	Braker predictions using Small RNAseq dataset (raw RNA-Seq)	12	12.8
Brake-pool		Braker predictions using Pooled RNAseq dataset (raw RNA-Seq)	77	241.4
DirInf-pool	Direct Inference	Direct Inference using Pooled RNAseq dataset (assembled transcripts and translated proteins)	38	595.1
DirInf-orph		Direct Inference using Orphan-rich RNAseq dataset (assembled transcripts and translated proteins)	38	595.1
BIND-orph	Combined (BIND or MIND)	Direct-Inference using Orphan-rich dataset plus BRAKER orphan enriched predictions	38	595.1
MIND-orph		Direct-inference using Orphan-rich RNAseq dataset plus Maker orphan enriched predictions	38	595.1

Table 1. Gene prediction scenario and the dataset properties used for benchmarking.

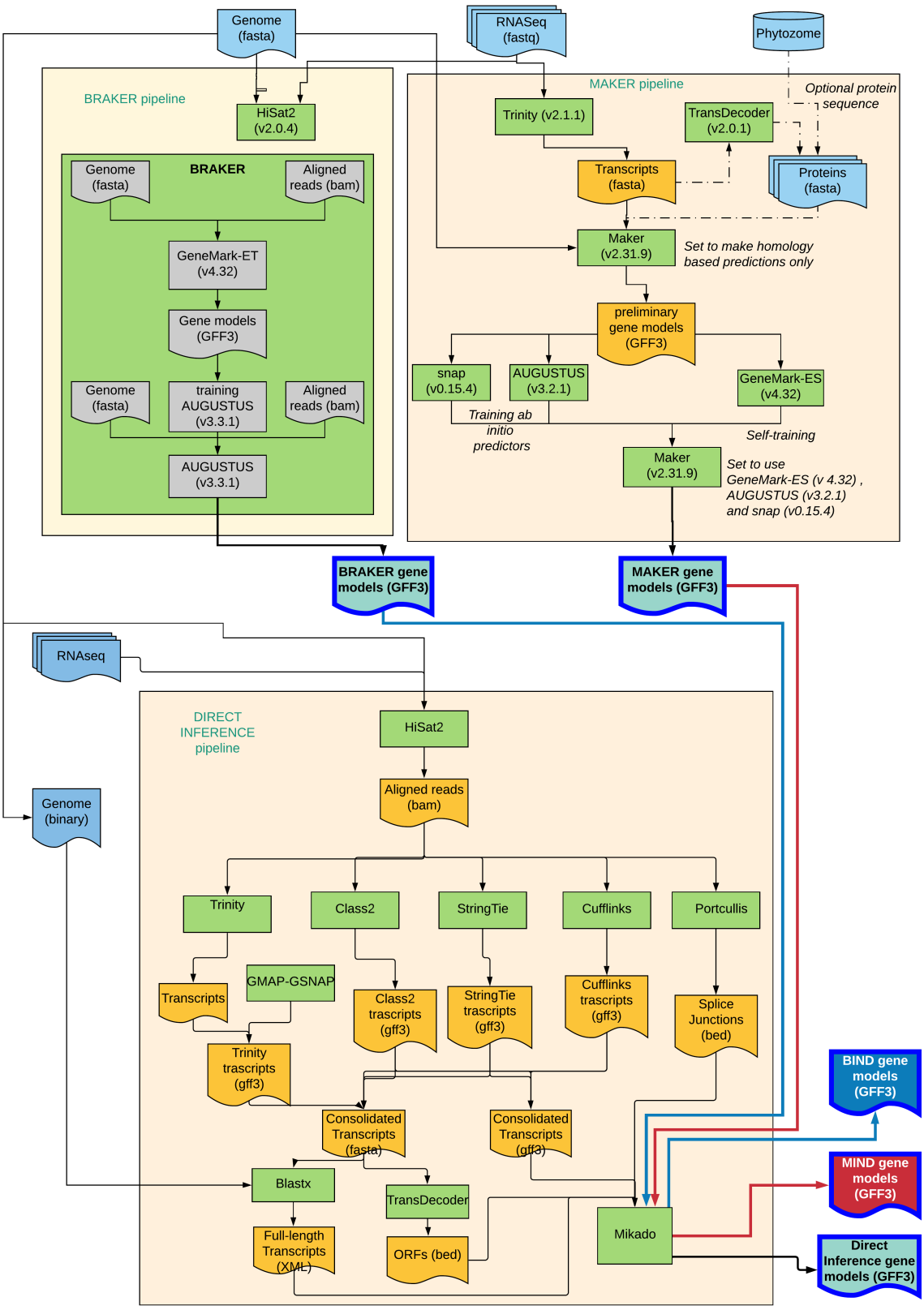

Figure 1. Schematic diagram of gene prediction pipelines. BRAKER (top left) and MAKER (top right) were installed and ran according to developers specifications. The Direct Inference pipeline (lower diagram) was run using custom scripts (see GitHub repository). The BIND, and MIND models combines predictions from the Direct Inference pipeline with predictions of MAKER (red line and box) or BRAKER (blue line and box).

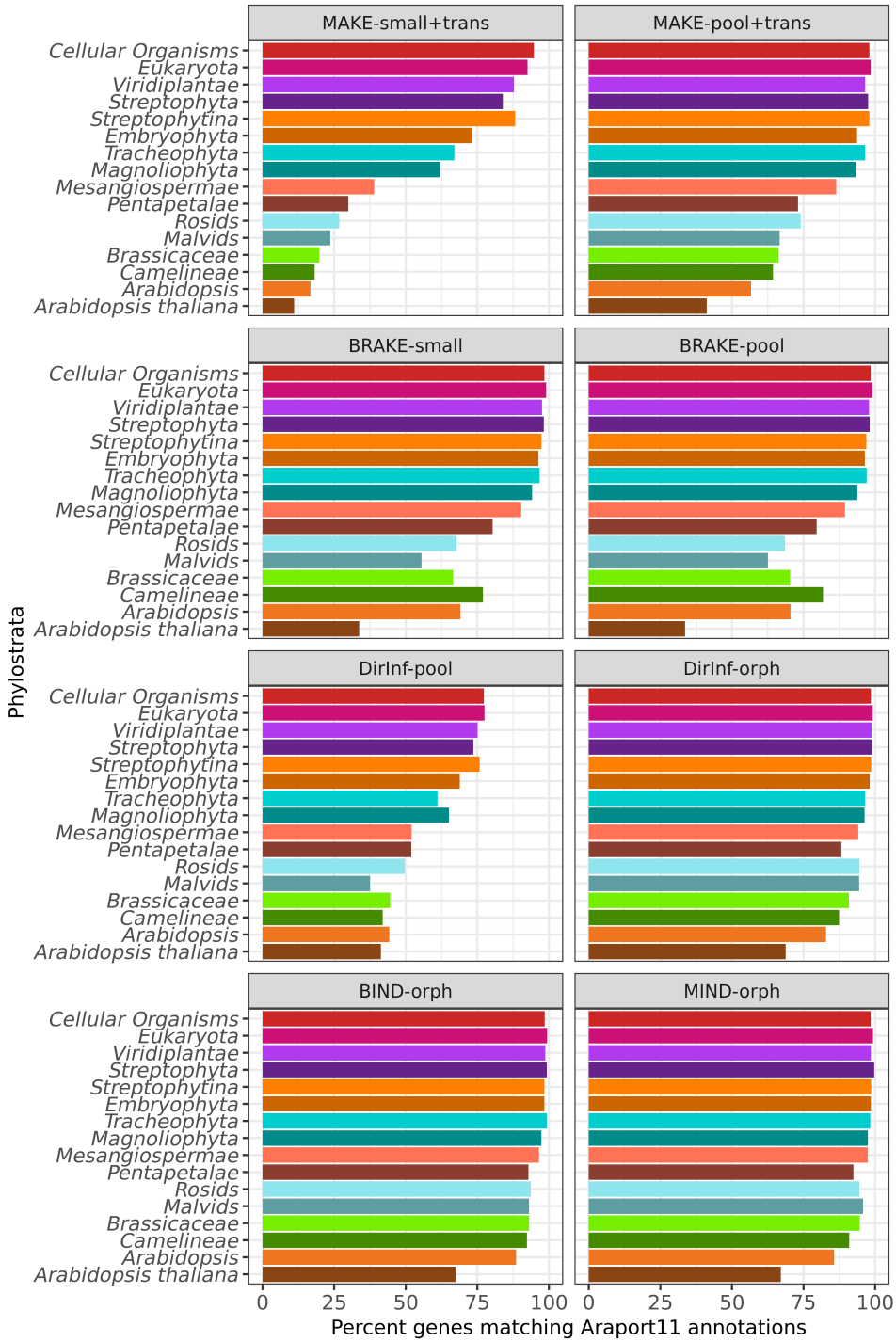

Figure 2. Percent Araport11-annotated genes predicted across different scenario by phylostratal designation. (**See Supplementary Table 3 for predictions of all tested scenario**)