MAKER-P: a tool-kit for the rapid creation, management, and quality control of plant genome annotations

Michael S. Campbell¹, MeiYee Law¹², Carson Holt³, Joshua C. Stein⁹, Gaurav D. Moghe⁴, David E. Hufnagel⁵, Jikai Lei⁶, Rujira Achawanantakun⁶, Dian Jiao⁷, Carolyn J. Lawrence⁸, Doreen Ware¹⁰¹¹, Shin-Han Shiu⁴⁵, Kevin L. Childs⁵, Yanni Sun⁶, Ning Jiang⁴⁹, Mark Yandell¹

1. Eccles Institute of Human Genetics, University of Utah, Salt Lake City, UT 84112
2. Department of Biomedical Informatics, University of Utah, Salt Lake City, UT 84112
3. Ontario Institute for Cancer Research, Toronto, ON M5G 1L7
4. Genetics Program, Michigan State University, East Lansing, MI 48824
5. Department of Plant Biology, Michigan State University, East Lansing, MI 48824
6. Department of Computer Science and Engineering, Michigan State University, East Lansing, MI 48824
7. University of Texas at Austin, Texas Advanced Computing Center, Austin TX 78758
8. USDA-ARS CICGRU and Department of Genetics, Development and Cell Biology, Iowa State University, Ames, IA 50011
9. Department of Horticulture, Michigan State University, 1066 Bogue Street, East Lansing, MI 48824
10. The iPlant Collaborative, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA
11. USDA-ARS NAA, Robert W. Holley Center for Agriculture and Health, Tower Road, Ithaca, NY, 14853, USA
Abstract

We have optimized and extended the widely used annotation-engine MAKER to in order to better support plant genome annotation efforts. New features include better parallelization for large repeat-rich plant genomes, ncRNA annotation capabilities, and support for pseudogene identification. We have benchmarked the resulting software toolkit, MAKER-P, using the *A. thaliana* and *Z. mays* genomes. Here we demonstrate the ability of the MAKER-P toolkit to automatically update, extend, and revise the *A. thaliana* annotations in light of newly available data; and to annotate pseudogenes and ncRNAs absent from the TAIR10 build. Our results demonstrate that MAKER-P can be used to manage and improve the annotations of even *A. thaliana*, perhaps the best-annotated plant genome. We have also installed and benchmarked MAKER-P on the Texas Advanced Computing Center (TACC). We show that this public resource can de novo annotate the entire *Arabidopsis thaliana* and *Zea mays* genomes in less than three hours, and produce annotations of comparable quality to those of the current TAIR10 and *Z. mays* V2 annotation builds.

Introduction

Because high-throughput genome sequencing technology has become widely available, many genome projects are now carried out by small groups with little prior experience in genome annotation. A major challenge for these researchers is the generation and dissemination of high quality gene structure annotations for downstream applications. This is especially true for plant genomics researchers given that plant genomes can be difficult targets for annotation: they are unusually rich in transposable elements (Feschotte et al., 2002; Schnable et al., 2009; Kejnovsky et al., 2012), have high rates of pseudogenization (Thibaud-Nissen et al., 2009; Zou et al., 2009; Hua et al., 2011) and contain many novel protein-coding and non-coding RNA (ncRNA) genes as revealed through RNA-Seq and proteomics studies (Campbell et al., 2007; Hanada et al., 2007; Jiang et al., 2009; Yang et al., 2009; Li et al., 2010; Lin et al., 2010; Donoghue et al., 2011; Garg et al., 2011; Boerner and McGinnis, 2012; Moghe et al., 2013). Plant genomes are also relatively large compared to other eukaryotes, representing some of the largest genomes in existence (Pellicer et al., 2010; Birol et al., 2013; Nystedt et al., 2013), meaning that the time required to annotate a large plant genome can be measured in months rather than hours. Moreover, different plant genomes—and in some cases, even the same plant genome—have been annotated using very different procedures, and to very different levels of accuracy. The plant genomics community is thus in need of an annotation engine that will scale to extremely large datasets; can produce accurate annotations in a repeat- and ncRNA-rich genomic landscape; integrate computational predictions and transcriptome data; and compare, evaluate, merge, and update legacy annotations. Most importantly, this software must be easy-to-use, as many of today’s plant genome sequencing groups have only limited bioinformatics expertise and computational resources.
To achieve these goals we have optimized and extended an established genome annotation-engine, MAKER (Holt and Yandell, 2011), for the plant genome research community. Not only is MAKER portable and easy to use, it is already in wide use by the animal, and fungal research communities (Kumar et al., 2012; Amemiya et al., 2013; Eckalbar et al., 2013; Scharldl et al., 2013; Smith et al., 2013). MAKER, unlike existing pipelines, can produce accurate annotations even in the absence of training data (Holt and Yandell, 2011). Importantly, MAKER generates a set of quality control measures to compare, evaluate, merge, and update legacy annotations (Cantarel et al., 2008; Eilbeck et al., 2009; Holt and Yandell, 2011).

We have extended MAKER for better performance on plant genomes, developing means for annotation of pseudogenes and ncRNAs, and optimized its parallelization for maximal performance on large, repeat-rich plant genomes. The resulting software is available for download, and a MAKER-P module is installed the Texas Advanced Computing Center using the iPlant Cyberinfrastructure (Goff et al., 2011).

Here we benchmark MAKER-P’s accuracy and speed using two previously annotated plant genomes: *A. thaliana* and *Z. mays*. Our *A. thaliana* results demonstrate that MAKER-P can be used to manage and improve the annotations of what is arguably the best-annotated plant genome. Using, a massively parallel version of MAKER-P on the Texas Advanced Computing Center (TACC), we also show that MAKER-P can *de novo* annotate the *A. thaliana* and *Zea mays* genomes in less than three hours, and that the resulting annotations are of comparable quality to the current TAIR10 and *Z. mays* V2 annotation builds. Collectively these results demonstrate that MAKER-P provides the plant genomics community with a very rapid and effective means for both *de novo* annotation of new plant genomes, and for management of existing plant genome annotations.

**Results & Discussion**

**Choice of target species.** We chose to benchmark MAKER-P using *A. thaliana* because it has a well-assembled reference genome and its genome annotations have been subject to extensive computational and manual curation (Lamesch et al., 2012). In addition, there is a large pool of experimental evidence available to aid the annotation of the *A. thaliana* genome including traditional ESTs, full-length cDNAs, and vast amounts of RNA-Seq data (Rounsley et al., 1996; Paz-Ares, 2002; Seki et al., 2002; Yamada et al., 2003). Moreover, The *A. thaliana* Informatics Resource (TAIR;(Lamesch et al., 2012)) has put great effort into assigning evidence-based quality values to each annotation via their five-star rating system (The Arabidopsis Information Resource, 2009) in the current release of the *A. thaliana* annotation set (TAIR10) (Lamesch et al., 2012). Thus,
the A. thaliana genome provides a perfect opportunity to benchmark the performance of MAKER-P.

**Gene-level accuracies.** We first used the TAIR10 annotations as a gold standard with which to determine gene-level accuracies of the \textit{ab initio} gene finders SNAP (Korf, 2004) and Augustus (Stanke and Waack, 2003; Stanke et al., 2008). To do so, we ran SNAP and Augustus trained for A. thaliana both with and without MAKER-P. When run within, MAKER-P can pass SNAP and Augustus additional information regarding protein, EST, and RNA-seq evidence allowing these programs to modify their predictions based on the evidence (Holt and Yandell, 2011). Results of this analysis are reported in Table 1. As can be seen, all three approaches achieve similar gene-level accuracies. These results demonstrate an established fact of gene finding: given sufficient training data, good gene-level accuracies are relatively easy to obtain (Guigó et al., 2006; Yandell and Ence, 2012). However, often no training data is available for novel genomes. In such cases \textit{ab initio} gene finders perform poorly, requiring an evidence-driven means of genome annotation (Yandell and Ence, 2012). This phenomenon is illustrated by the penultimate column in Table 1 wherein we have run SNAP using the \textit{Zea mays} HMM as a surrogate for a poorly trained gene finder. In this case the gene level accuracy is much poorer—70% compared to 82% using the A. thaliana HMM. This demonstrates that attempts to leverage training data from other plants, \textit{Zea mays} in this example, are fraught with difficulty, a fact that is well established (Korf, 2004; Holt and Yandell, 2011; Yandell and Ence, 2012). The last column of Table 1 reports the impact of running the same version of SNAP trained for \textit{Zea mays} within the MAKER software harness along with the RNA-Seq, EST/cDNA, and protein evidence datasets as described in methods section. This column of Table 1 demonstrates that MAKER-P's evidence-driven functions allow it to achieve high gene-level accuracies even using poorly trained \textit{ab initio} gene-finders, an observation consistent with previous work using animal genomes (Holt and Yandell, 2011), and one that demonstrates the utility of MAKER-P as a means to annotate novel plant genomes.

**Using AED to measure Exon-level accuracy.** Gene-level accuracy is only the first step toward producing a well-annotated genome. Gene annotations must do more than simply overlap genes, as downstream applications require that their intron-exon structures and predicted protein sequences to also be correct. The accuracy of intron-exon structures is usually assessed by means of exon-level or nucleotide-level accuracy calculations using gold-standard annotations (for review see Yandell and Ence, 2012). One question that naturally arises in such analyses is how to assess the accuracy of the gold standard annotations themselves. MAKER-P, like its parent application MAKER (Holt and Yandell, 2011), provides an automated means for addressing both these questions. MAKER-P uses Annotation Edit Distance (AED) (Cantarel et al., 2008; Eilbeck et al., 2009; Holt and Yandell, 2011) to measure the goodness of fit of an annotation to the evidence supporting it. AED is a number between 0 and 1, with
an AED of zero denoting perfect concordance with the available evidence and a value of one indicating a complete absence of support for the annotated gene model (Eilbeck et al., 2009). AED can be calculated relative to any specific sort of evidence: EST and protein alignments, ab initio gene predictions, or RNA-seq data. In each case, the AED score provides a measure of each annotation’s congruency with a particular type or types of evidence. By plotting the cumulative distribution function (CDF) of AED across all annotations (Holt and Yandell, 2011), a genome-wide perspective of how well the annotations and/or ab initio gene predictions reflect the EST, protein, and RNA-Seq evidence can be obtained. Importantly, this can be done even in the absence of a gold-standard set of reference annotations for that genome (see Supplemental Figure 1 for an example comparing gene models produced by the ab initio gene finder Augustus run with and without MAKER supervision). Similarly, the same procedure can be used to evaluate the goodness of fit between a gold-standard annotation dataset and the evidence used to produce it. See Eilbeck et al., 2009; Holt and Yandell, 2011; Yandell and Ence, 2012 for additional information on AED.

**Cross-genome validation.** AED also makes possible cross-genome assessments of annotation datasets in the context of each genome’s own supporting evidence (Eilbeck et al., 2009; Holt and Yandell, 2011). An example is shown in Figure 1, which provides a genome-wide overview of the goodness of fit of the TAIR10 annotations to the evidence datasets used for our benchmarking analyses (see methods for evidence dataset details). As can be seen, A. thaliana is a very well annotated genome; overall the congruency of the TAIR10 annotations with this evidence is roughly equivalent to that of the human RefSeq annotations, in that greater than 85% of annotations have an AED score less than 0.5 when compared to a previously published analysis of Human RefSeq annotations (Lander et al., 2001; Venter et al., 2001) (see methods for details of dataset). Figure 1 also demonstrates that our evidence set provides support for 90% of the annotated genes in the TAIR10 dataset.

**Comparison of AED and TAIR’s 5-star system.** One advantage of using the TAIR10 annotations to benchmark MAKER-P is that each TAIR10 annotation has already been assigned a quality score via TAIR’s five-star ranking system (The Arabidopsis Information Resource, 2009), whereby the best supported genes are afforded five stars or four stars, with less well supported annotations assigned three, two, and one-star status. Annotations with no external support are classified as ‘no-star’. Table 2 provides a breakdown of TAIR10 annotations by their star-rating in the context of their supporting evidence using the evidence datasets used for our benchmarking analyses. Also shown in Table 2 is the cumulative support for the TAIR10 annotations in toto and for the MAKER standard annotation build produced using the same evidence (see methods for details). Importantly, these results demonstrate that (2) MAKER-P can automatically produce a de novo genome-annotation dataset of very similar quality to the highly curated TAIR10 annotations, and (2) that there is good concordance between TAIR10 star rating and the degree of evidence support.
Next we sought to determine the ability of MAKER-P to revise and improve upon the preexisting TAIR10 annotations when fed new evidence. We first used MAKER-P’s update functionality (Holt and Yandell, 2011) to automatically update each of the TAIR10 annotations, bringing each gene model into better agreement with the available evidence, by means of extending and modifying the exon coordinates of each existing TAIR10 gene annotation in light of RNA-Seq-based transcript assembly data, EST, cDNA, and protein evidence (see methods for details). Next we ran MAKER-P as we would to annotate a novel genome using the same evidence dataset, allowing MAKER-P to create a new or de novo set of gene annotations based upon the same evidence that we used to update the TAIR10 annotations.

Figure 2 displays the cumulative AED distributions for the MAKER de novo, the MAKER updated TAIR10 annotations, and the original TAIR10 A. thaliana annotations as a reference. As can be seen, both the updated and the de novo MAKER-P datasets are in better agreement with supporting evidence than the original TAIR10 annotations. Much of the improvement, especially in the case of the MAKER-P de novo annotations, is due to the absence of poorly supported TAIR10 genes in the MAKER-P de novo gene-build. The MAKER-P de novo gene-build, for example, contains 1,250 fewer genes than the TAIR10 dataset. In total there are 2,368 genes present in TAIR 10 that are absent from the MAKER de novo gene-build. 60% of the absent models are single-exon genes; 53% are 1- or no-star gene-models, but 96% of all TAIR 5-, 4-, 3- and 2-star transcripts are present. We also evaluated MAKER-P’s performance using a subset of genes with a one-to-one relationship between the TAIR10 and MAKER-P de novo annotations shown in Figure 2 and allowed MAKER-P to update the TAIR10 annotations. These results are shown in Supplemental Figure 2 and demonstrate that MAKER-P’s improvements to the TAIR10 gene models are not solely due to having culled the unsupported TAIR10 gene models; rather the improvements are made across the entire TAIR10 dataset. Figure 3 demonstrates this fact quite clearly. There is excellent agreement between the TAIR10 manually curated evidence classifications and MAKER’s automatic AED-based quality control schema, cross validating both MAKER-P’s AED and TAIR10’s star rating approaches to assigning confidence levels to individual annotation. For 5-star TAIR10 genes, 94% have AED scores of less than 0.5, whereas only 33% of 1-star genes have an AED less than 0.5. Note that the 4- and 5-star genes’ AED curves are very similar. This is because under the TAIR system, genes supported entirely by a single piece of evidence (usually a single full-length cDNA) are afforded 5-star status, whereas an annotation completely supported by tiled evidence is afforded 4-star status. MAKER-P’s AED calculation makes no such distinction; hence the two curves are quite similar.

Figure 3 also demonstrates another important point: the greatest improvements are made to the highest confidence TAIR10 gene models. The dotted lines denote the AED curves for the MAKER updated TAIR10 annotations. Note that
the greatest MAKER-P-mediated improvements to the TAIR10 gene models are seen for 2-star through 5-star genes. While this may seem a paradoxical result, it is wholly expected. Single-star and no-star genes by definition have little supporting evidence; hence, there is little raw material available to MAKER-P with which to affect revisions. In contrast, the better-supported genes (2-star through 5-star annotations, for example) have correspondingly more evidence—some supporting, some contradicting the TAIR10 models. It is thus to the best-supported gene-models under the TAIR10 classification system that MAKER-P is able to make the most positive changes. This is an important point, and it demonstrates a key strength of MAKER-P. Highly supported, highly expressed genes often have some data that strongly supports a given transcript model. A single full-length cDNA, for example, may confirm the entire exon-intron structure of the annotated transcript, affording that model 5-star status. Contradictory evidence is not considered under the TAIR schema; it is however considered by MAKER-P. This means the resulting MAKER-P transcript structure is not necessarily a perfect match to any given piece of evidence but rather reflects the best-possible gestalt of all of the evidence for that gene. Consequently, no matter how well supported a gene model, it will have an AED > 0 if other evidence contradicts that model. The ability of AED to take into account both confirming and contradictory evidence is a key strength of the MAKER-P approach. The fact that MAKER-P is able to effect positive revisions to what would appear to be the best annotated genes in the TAIR10 datasets (5 and 4 star genes) demonstrates the strength of the AED approach to quality control. Further insight into the nature of these revisions is provided in Table 3, which focuses on gene-models with alternatively spliced transcripts.

**Alternative splicing.** MAKER-P annotates only the most certain of alternately spliced transcripts—those with clear support for differential internal exon (cassette splicing); hence, the number of alternatively spliced transcripts is very limited compared to TAIR10. MAKER-P’s update functionality, on the other hand, provides a means to update individual alternatively spliced transcripts. MAKER-P deleted or merged 184 alt-spliced transcripts, and added an average of 19 5’-UTR nucleotides, and 32 3’-UTR nucleotides per transcript genome-wide. The cumulative effects of the revisions are shown in the last column of Table 3; prior to revision, 79% of TAIR10 transcripts had an AED less than 0.2. After revision, the proportion of gene models with AED less than 0.2 has climbed to 82%. MAKER-P thus provides a rapid and automated means to improve even intensively manually-curated alternatively spliced gene-models.

**Repeats.** Plant genomes can be difficult targets for annotation because they can be unusually rich in transposable elements (Bennetzen, 2005; Schnable et al., 2009), have high rates of pseudogenization (Zou et al., 2009; Hua et al., 2011) and contain many novel non-coding RNA genes as revealed through RNA-Seq (Fahlgren et al., 2007; Sunkar et al., 2008). We have attempted to address these points with the MAKER-P project. Although MAKER-P employs RepeatMasker (Smit et al.) as well as its own internal repeat finding method (Cantarel et al.,
novel genomes, especially plant genomes, often contain new classes of repeats absent from both RepBase (Jurka et al., 2005) and from MAKER’s internal repeat library (Cantarel et al., 2008). Failure to identify, annotate, and mask repeats during the gene finding stages of annotation can result in spurious gene calls and can lead to the creation of gene models containing portions of transposons and retrotransposons in the form of exons derived from transposon sequences fused to legitimate protein-coding genes. Although there exist several packages to identify repeats and to construct repeat libraries for new genomes [see (Lerat, 2010) for discussion], many MAKER users report that these tools are difficult to use. Moreover, the resulting output of existing packages often contains non-transposon gene or gene fragments, which may lead to the masking of bona fide genes. To address this point, the MAKER-P toolkit now contains two guided tutorials, walking users through a series of steps necessary to create their own custom repeat library. The Basic tutorial describes the process of generating a species specific repeat library suitable for repeat masking prior to protein coding gene annotation with MAKER or MAKER-P. The advanced tutorial explains how to classify repeats identified using the basic tutorial into families. See Table-5 for the web-addresses for both tutorials. We used the approach outlined in the basic tutorial to construct a novel A. thaliana repeat library and then assayed the impact of using it for de novo annotation of A. thaliana, using AED to evaluate the results. These data are shown in Supplemental Figure 3. In this case, we found little difference in MAKER-P’s performance. However, A. thaliana is not an ideal genome to demonstrate the effect of repeats on gene annotation because the A. thaliana genome contains the fewest repeats among all the sequenced plant genomes with the exception of the carnivorous bladderwort plant Utricularia gibba (The Arabidospsis Genome Initiative, 2000; Slotkin et al., 2012; Ibarra-Laclette et al., 2013).

Pseudogenes. With MAKER-P we have also extended MAKER to include means for the annotation of pseudogenes and ncRNAs. These tools are included in the MAKER-P tool kit (see methods). We benchmarked them on the A. thaliana genome. The MAKER-P pseudogene tools define pseudogenes as unannotated genomic regions with significant resemblance to annotated protein sequences from the genome in question, e.g. A. thaliana (see Methods). In total, we identified 4,204 pseudogenes. Among these presumed pseudogenes, 2,277 have at least one premature stop and/or frameshift (referred to as disabling substitutions). Although the rest are without disabling substitutions, the median pseudogene length is 175 bp (see Supplemental Figure 4), significantly shorter than those of TAIR10 genes and annotated pseudogenes. Thus they are severely truncated genes that likely have no function. Because our method relied on the use of annotated protein coding genes, all pseudogene annotations have significant similarities to known A. thaliana proteins. Nonetheless, 18% have RNA-Seq coverage. If the analysis pipeline is applied to the whole genome, 2.5% and 0.6% of currently annotated protein-coding genes are identified as pseudogenes due to the presence of misidentified stops and frameshifts, respectively. Indicating that the false positive rate of our pipeline is 3.1%
Assuming the pseudogene and its most closely related functional gene are paralogous, we found that the most commonly occurring domains in progenitors that gave rise to pseudogenes are F-box and related domains, RNase H, and protein kinase. Although the size of a domain family with annotated genes generally correlates with the number of pseudogenes, families differ significantly in their pseudogene to gene ratio. For example, the pseudogene to gene ratios differ significantly between F-box (152:567) and protein kinases (54:1021) (p<2.2xe-16), demonstrating that these families differ greatly in their loss rates.

**ncRNAs.** Using nine small RNA-Seq data sets of *A. thaliana* (**Supplemental Table 3**), the MAKER-P ncRNA tools identified 807 ncRNAs in total. The intersections of our predictions and TAIR10 annotations are summarized in **Table 4** for tRNA, rRNA, snoRNA, miRNA, and other types of ncRNA genes. It is worth noting that the number of identified ncRNAs, especially miRNAs, heavily depends on the RNA-Seq data. Some previously annotated ncRNAs are not transcribed or have extremely low transcription level (e.g., 1 mapped read) in the RNA-Seq data we used for our analyses.

**Community availability.** Web addresses, download sites, and passwords (where applicable) for all tools, datasets, and online documentation described in this report are listed in **table 5**. MAKER-P like its parent package MAKER, is a multi-threaded, fully MPI-compliant annotation engine (Holt and Yandell, 2011). MAKER-P was specifically optimized for improved functionality on the iPlant infrastructure relative to MAKER, and is packaged with the necessary launch scripts to ensure optimal performance. MAKER-P also includes integrated means for tRNA and snoRNAs. MAKER-P is available to iPlant users as a supported module on the TACC’s Lonestar cluster (see table 5 for usage instructions; specifically ‘iPlant MAKER-P documentation’). The MAKER-P toolkit is freely available for academic use; see **table 5** for download information.

**Speed benchmarks.** We first used the *A. thaliana* genome to benchmark MAKER-P’s performance on the Texas Advanced Computing Center (TACC), which hosts the iPlant compute infrastructure. Using 600 CPUs we are able to complete the entire *de novo* annotation of the *A. thaliana* assembly (~120 Megabases) in 2 hrs and 44 minutes. Even faster compute times can be achieved using additional CPUs and/or by launching multiple instances of MAKER-P (chromosome-by-chromosome, for example). By doing so we were able to perform the same annotation in 1 hour 27 minutes on 1,500 CPUs. An additional benchmarking analysis using the *Zea mays* assembly (~2 gigabases) and 2,172 CPUs finished in 2 hours 53 minutes. See **figure 4**. Run times are both a function of the evidence dataset presented for alignment as well as the gene density of a genome, but the observed throughput of greater than 500 megabases per hour demonstrates that even the largest of plant genomes could be annotated in a reasonable time frame by leveraging MAKER-P’s scalability. **Supplemental figure 6** compares the resulting MAKER-P Z. mays annotations to those of the current Chromosome 10 V2 annotations available at MaizeGDB.
As can be seen the MAKER-P results compare favorably to the V2 annotations, with MAKER–P generating 3,059 gene annotations on this chromosome—an additional 365 gene annotations compared to the current V2 build. All of the 365 additional MAKER–P annotations are supported by either RNA-seq, EST, protein, or Pfam domain evidence; and have overall better AED scores (see Supplemental figure 6). Moreover, MAKER-P’s annotation of alternatively spliced transcripts (see Supplemental table 7) mirrors its performance on the A. thaliana genome (table 3), further demonstrating that MAKER-P can produce highly accurate Z. mays annotations, and that it can systematically improve upon the quality of the existing V2 annotation build. Collectively, these results demonstrate that using MAKER-P, a single investigator can carry out the de novo annotation of a grass genome and/or update its existing genome annotations with new RNA-Seq data in a few hours.

Redistribution of annotations. Dissemination of genome annotations, especially those of novel genomes to the wider biological community, is often a bottleneck for genome annotation projects. To remedy this problem we have worked with the Web-Apollo project (Helt et al.) to provide MAKER and MAKER–P users with easy means to distribute their annotation datasets to the wider community. MAKER–P’s outputs are fully WebApollo ready; thus a WebApollo database can be constructed and placed online within hours of finishing an annotation run using either the downloadable version of MAKER–P run locally on a user’s machine, or using the community iPlant version installed on the TACC. As proof of principle, we constructed a WebApollo database containing the TAIR10, MAKER–P de novo, MAKER–P updated annotations, the pseudogene and ncRNA annotations and their associated protein and RNA-seq evidence described in this report. This database is available online at http://weatherby.genetics.utah.edu:8080/WebApollo_A_thaliana (username: MAKER-P, password: marksentme). For example, click the edit button on the first page; then drag and drop any dataset shown on the left-hand panel into the JBrowse central frame. See table 5 for additional details and dataset download locations. WebApollo has many features that will benefit to the plant genomes community. WebApollo, for example, provides functionality for remote editing of the annotations; and supports concurrent users. Meaning that it can be easily deployed in the classroom for purposes of hands-on instruction and rapidly deployed in support of distributed genome jamborees that aim to rapidly curate all or a specific subset of the gene annotations. Figure 5 shows a screen shot for the TAIR10 AT5G03540 gene from the database. Note that this TAIR10 gene has three annotated transcripts—two 4-star and one 2-star; as expected the MAKER–P default model summarizes these with a single consensus transcript (minus the fourth exon of AT5G03540.3 for which there is no RNA-Seq, EST, or cDNA evidence); whereas the MAKER–P update of the TAIR-10 gene model maintained all three transcripts—each containing additional 5- and 3-prime UTR sequence as suggested by the RNA-Seq data, improving the overall AED of this gene model to 0.04 compared to the AED of 0.06 of the original TAIR10 gene model.
Conclusions

Today the evidence for genome annotations evolves more rapidly than the annotations. In many cases annotations fall out of sync with the available evidence almost as soon as they are created. MAKER-P provides a solution to this problem, providing a means to rapidly update a genome’s annotations, bringing them into sync with the latest datasets. As we have demonstrated, the greatest revisions are accomplished at those genes with the most evidence. In such cases, the quantity and complexity of RNA-seq data supporting and contradicting even the most established gene models can confound attempts by human annotators to produce consistent, coherent gene models. MAKER-P, in contrast, guarantees a constant, complete analysis of these data, resulting in demonstrable improvements to the annotations of even the well-annotated A. thaliana genome. Moreover, our time trials using the maize genome demonstrate that even large, complex plant genomes can be annotated in only a few hours using the version of MAKER-P installed on iPlant resources at TACC. The availability of MAKER-P within the iPlant Cyberinfrastructure will grant independent plant genome researchers the ability to rapidly annotate new plant genomes, to revise and manage existing ones, and to create online databases for distribution of their results. MAKER-P thus provides the plant genome research community with a basic resource that democratizes genome annotation.

Methods

Evidence sources and assembly. Sequence evidence used for annotation by MAKER-P consisted of SwissProt protein data, EST and cDNA sequences from A. thaliana, and transcript assemblies derived from publicly available RNA-Seq data sets. A SwissProt data file containing only protein sequences from plants was obtained from UniProt (release-2011_12). All A. thaliana proteins were removed from this file, and only the non-A. thaliana plant proteins were used when running MAKER-P. A file of A. thaliana EST sequences (ATH_EST_sequences_20101108.fas) was obtained from TAIR (Lamesch et al., 2012). Full-length A. thaliana cDNA sequences were downloaded from the National Center for Biotechnology Information (NCBI) Nucleotide database (Benson et al., 2013). Forty-seven RNA-Seq datasets derived from different A. thaliana tissues and/or grown under different conditions were collected from NCBI’s Short Read Archive (SRA; Supplemental Table 1) (Wheeler et al., 2008). The reads from each file were cleaned using programs from the FASTX-Toolkit (version 0.0.13; http://hannonlab.cshl.edu/fastx_toolkit/). Fastx_clipper removed Illumina adapter sequences, and fastx_artifacts_filter removed any aberrant reads. Finally, fastx_quality_trimmer removed nucleotides with Phred scores less than 30 and discarded reads less than 20 bases long. The Trinity transcript assembly package (r2011-11-26) was used to generate transcript assemblies with lengths of 150 nucleotides or longer (Grabherr et al., 2011). The 47 RNA-
RNA-Seq data sets were from 17 SRA Studies and were thus assembled into 17 different transcript assemblies (Supplemental Table 1). All RNA-Seq data were treated as single-end reads in order to avoid aligning transcripts with stretches of N's. The same procedures were used for the *Z. mays* datasets detailed in Supplemental Table 6.

*H. sapiens* annotations for release 37.2 were downloaded from NCBI. AED metrics were computed using all mouse proteins from release 37.1, all Uniprot/SwissProt proteins minus human proteins, and all human ESTs in dbEST.

**Repeat library.** In this study, we established two protocols to satisfy the demand of different users. For the basic protocol (see table 5 for web-address of the tutorial), RepeatModeler ([http://www.repeatmasker.org/RepeatModeler.html](http://www.repeatmasker.org/RepeatModeler.html)) was used to process the genomic sequences with all *A. thaliana* repeats excluded from the RepeatMasker repeat library so that the *A. thaliana* genome would act as a “novel” genome. Among the repetitive sequences generated by RepeatModeler, some are classified and they are considered as TEs. Sequences with unknown identity from RepeatModeler were searched against a transposase database (without *A. thaliana* transposase) and sequences matching transposases were considered as transposons belonging to the relevant superfamily. Many transposable elements carry genes or gene fragments. To exclude gene fragments, all repeats were searched against a plant protein database with transposon-proteins excluded. Sequences matching plant proteins as well as 50 bp of flanking sequence were excluded. After the exclusion, if the remaining portion of the sequence was shorter than 50 bp, the entire sequence was excluded.

For the advanced protocol (see table 5 for web-address of the advanced tutorial), we used a combination of structure-based and homology-based approaches to maximize the opportunity for repeat collection. Briefly, sequences of miniature inverted repeat transposable elements (MITEs) were collected using MITE-Hunter ([Han and Wessler, 2010](http://www.repeatmasker.org/RepeatModeler.html)) with all default parameters. LTR retrotransposons were collected using LTR-harvest and LTR-digest ([Ellinghaus et al., 2008](http://www.repeatmasker.org/RepeatModeler.html); [Steinbiss et al., 2009](http://www.repeatmasker.org/RepeatModeler.html)), followed by a filtering to exclude false positives. To reduce redundancy, representative sequences (exemplars) were chosen as previously described ([Schnable et al., 2009](http://www.repeatmasker.org/RepeatModeler.html)). To collect other repetitive sequences, the genomic sequence was then masked using the LTR and MITE sequences. The unmasked sequence was extracted and processed by RepeatModeler. The gene fragments contained in all repetitive sequences were excluded as described above. More details can be found in Advanced repeat library construction tutorial; its web-location is given in Table 5. The libraries made through different protocols masked different percentages of the genome (Supplemental Table 2), however, the use of the basic protocol versus the advanced protocol did not significantly affect the overall AED distribution or gene-level accuracy. The resulting annotation with the Basic TE library is a possible
exception generating a slightly lower accuracy and slightly higher overall AED scores (Supplemental Figure 2).

**MAKER-P de novo annotation of *A. thaliana***. MAKER-P 2.27 r1020 was run on *A. thaliana* (TAIR10 assembly) using the assembled *A. thaliana* mRNA-seq data, a set of traditional ESTs and full-length cDNAs, and a set of plant proteins from UniProt/SwissProt as evidence. Repetitive regions were masked using a custom repeat library. The details surrounding evidence and repeat library generation was described earlier in this section. Additional areas of low complexity were soft masked (Korf et al., 2003) using RepeatMasker to prevent seeding of evidence alignments in those regions but still allowing extension of evidence alignments through them (Korf et al., 2003; Cantarel et al., 2008). Genes were predicted using SNAP (Korf, 2004) and Augustus (Stanke and Waack, 2003; Stanke et al., 2008) trained for *A. thaliana* or *Z. mays* using MAKER-P in an iterative fashion as described for MAKER in Cantarel et al. (Cantarel et al., 2008).

**Generating MAKER-P default, standard, and max builds.** When using MAKER-P to generate *de novo* annotations for a genome, users can choose from three different options to produce their final annotation dataset: default, standard, and max. The MAKER-P *default build* consists only of those gene-models that are supported by the evidence (i.e., AED <1.0). The default build is thus very conservative. The MAKER-P *standard build* (which was used in Figure 2 and Tables 1 and 2, for example), includes every gene-model in the default build, plus every *ab initio* gene prediction that (1) encodes a Pfam domain as detected by InterProScan (Quevillon et al., 2005), and (2) does not overlap an annotation in the MAKER default set. The MAKER-P *max build* includes every gene-model in the default build plus every *ab initio* gene prediction that does not overlap an annotation in the MAKER default set, regardless of whether or not it encodes a Pfam domain. When using TAIR10 as a gold standard, the MAKER-P default build had the highest specificity, the MAKER-P max build had the highest sensitivity, and the MAKER-P standard build balances sensitivity and specificity to give the highest overall accuracy, which is why we used it for the comparisons in this paper (see Sup. Figure 4). MAKER-P annotation of alt transcripts was not evoked unless specified in the text.

**Generating AED scores for TAIR10 and gene finders only.** AED scores for the TAIR10 annotation set were generated using MAKER-P 2.27 r1020. The TAIR10 annotations were passed to MAKER-P as gene models in a GFF file and evaluated against the same evidence and repeat library used for the MAKER-P *de novo* annotation. This allowed MAKER-P to calculate AED scores for each of the TAIR10 annotations without allowing MAKER-P to modify the annotation in any way. This same procedure was used to generate AED scores for the *ab initio* gene predictions generated without MAKER-P supervision.
**MAKER-P update of TAIR10.** The TAIR10 gene models were passed to MAKER-P as gene predictions with the same evidence and repeat library used for the MAKER-P de novo annotation. This allows MAKER-P to update the TAIR10 annotations to better match the evidence.

**Pseudogene identification.** We adapted a previously published pseudogene pipeline for use with MAKER-P (Zou et al., 2009). To identify genomic regions likely to be pseudogenes, we first searched the *A. thaliana* genome using all *A. thaliana* annotated protein sequences as queries. The output was filtered based on the following thresholds: E-Value < 1e-5, % identity > 40%, match length >30 aa and coverage > 5% of the query sequence. The filtered matches provide pseudo-exon definitions. These pseudo-exons that are <457 bp (95th percentile of the intron length distribution) from each other and having matches to the same protein are concatenated together to form putative pseudogenes. Pseudogenes overlapping with annotated protein coding regions were removed from the dataset. Finally, pseudogenes with significant similarity to known Viridiplantae repeats (Cutoff=300, Divergence=30; RepeatMasker 3.3.0) were discarded. This MAKER-P pseudogene identification pipeline is available for download at the location given in table-5.

**tRNA and snoRNA annotation.** MAKER-P features integrated means for annotation of tRNAs and snoRNAs. tRNAs are identified using tRNAScan-SE (Lowe and Eddy, 1997), and snoRNAs with snoscan (Lowe, 1999). Both tools are now supported and integrated into within the MAKER-P software harness and their outputs are included in MAKER-P’s GFF outputs, where they are described using the Sequence Ontology terms tRNA, snoRNA, respectively.

**miRNA annotation.** Our ncRNA annotation pipeline uses multiple ncRNA homology search tools (described below) and small RNA RNA-Seq data to identify transcribed ncRNAs. There are three major components in the pipeline. First, we employ Infernal (Nawrocki et al., 2009), a stochastic context-free grammar (SCFG) based general ncRNA search tool to identify ncRNA homologs to annotated ncRNA families in Rfam (Gardner et al., 2009). The output of this step provides candidate ncRNA genes. However, it is known that genome-scale SCFG searches can incur high false positive rates. In order to discard false predictions, we evaluate the expression levels of the candidate ncRNAs in the second step. As the expression of many types of ncRNAs is condition and tissue specific, we quantified the expression levels of these putative ncRNAs in multiple small RNA-seq datasets (see Supplemental Table 1), which were sequenced from different tissues and conditions. All ncRNAs that were expressed in at least one RNA-Seq dataset were validated using family-specific properties. tRNAScan-SE (Lowe and Eddy, 1997) and snoscan (Lowe, 1999) were applied to candidate tRNAs and snoRNAs, respectively. For miRNAs, we used our own miRNA identification tool miR-PREFeR. miR-PREFeR and its documentation are available for download at [https://github.com/hangelwen/miR-PREFeR](https://github.com/hangelwen/miR-PREFeR). When running this tool on *A. thaliana* we used the properties that are associated with
the biogenesis of miRNA maturation as features and trained an AD-tree based classification model to distinguish true from false stem loops. The features we examined include the expression pattern of the mature miRNA and miRNA*, 3’ overhang, secondary structure, minimum free energy, existence of the regulation target (miRNA target finding), number of samples in which the miRNA are expressed, and expression level change across multiple RNA-seq samples. All ncRNAs that pass the three-step pipeline are reported in TABLE 4. The total runtime for miR-PREFeR on A. thaliana was 12 hours 21 minutes using 4 processing cores and nine RNA seq samples.

Acknowledgements. This work was supported by NSF IOS-1126998 to SS, YS, NG, KC and MY. We gratefully acknowledge the TACC support personnel. We also acknowledge and thank Chris Towne of JCVI for helpful discussion and feedback as well as Suzie Lewis at UC Berkley and the rest of the WebApollo team for their efforts to ensure WebApollo compatibility with MAKER-P outputs. JCS and DW received support from the iPlant Collaborative, which is funded by a grant from the National Science Foundation Plant Cyberinfrastructure Program (#DBI-0735191).

References


Donoghue MT, Keshavaiah C, Swamidatta SH, Spillane C (2011) Evolutionary origins of Brassicaceae specific genes in Arabidopsis thaliana. BMC evolutionary biology 11: 47


Hanada K, Zhang X, Borevitz JO, Li W, Shiu S (2007) A large number of novel coding small open reading frames in the intergenic regions of the Arabidopsis thaliana genome are transcribed and/or under purifying selection. Genome research 17: 632–40


Lerat E (2010) Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. Heredity 104: 520–33


Tables

Table 1: The effect of MAKER-P’s Supervision of Gene Finders on Genome level Sensitivity and Specificity.

<table>
<thead>
<tr>
<th></th>
<th>MAKER default*</th>
<th>MAKER standard*</th>
<th>MAKER max*</th>
<th>Augustus trained for A. thaliana run outside of MAKER</th>
<th>SNAP Trained for A. thaliana run outside of MAKER</th>
<th>SNAP Trained for Z. mays run outside of MAKER</th>
<th>SNAP Trained for Z. mays run inside of MAKER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sensitivity</td>
<td>0.88</td>
<td>0.91</td>
<td>0.93</td>
<td>0.91</td>
<td>0.84</td>
<td>0.47</td>
<td>0.67</td>
</tr>
<tr>
<td>Specificity</td>
<td>0.93</td>
<td>0.91</td>
<td>0.81</td>
<td>0.93</td>
<td>0.81</td>
<td>0.92</td>
<td>0.94</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.90</td>
<td>0.91</td>
<td>0.87</td>
<td>0.92</td>
<td>0.82</td>
<td>0.70</td>
<td>0.80</td>
</tr>
</tbody>
</table>

*MAKER default, standard and max, refer to different MAKER gene-build options. See methods and supplemental figure 4.

Table 2. Breakdown of evidence types supporting TAIR10 and MAKER-P annotations

<table>
<thead>
<tr>
<th>Star Rating</th>
<th>Fraction of annotation s with Pfam domains</th>
<th>Fraction of annotation s with eukaryotic RefSeq protein homology</th>
<th>Fraction of annotation s with spliced RNA-seq support</th>
<th>Fraction of annotation s with any RNA support (mRNA-seq, EST, cDNA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>*****</td>
<td>0.76</td>
<td>0.81</td>
<td>0.42</td>
<td>0.79</td>
</tr>
</tbody>
</table>
Table 3: Features of alternatively spliced genes in the MAKER-P De Novo Annotation of *A. thaliana*, TAIR10, and a MAKER-P update of TAIR10

<table>
<thead>
<tr>
<th></th>
<th>MAKER-P</th>
<th>TAIR10</th>
<th>MAKER-P update TAIR10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of alt. spliced genes</td>
<td>3,024</td>
<td>5,804</td>
<td>5,726</td>
</tr>
<tr>
<td>Number of alt. spliced transcripts</td>
<td>7,190</td>
<td>13,774</td>
<td>13,590</td>
</tr>
<tr>
<td>Ave. exons/transcript</td>
<td>10.18</td>
<td>7.79</td>
<td>7.82</td>
</tr>
<tr>
<td>Total transcripts with 5’ UTR</td>
<td>5,708</td>
<td>12,714</td>
<td>12,352</td>
</tr>
<tr>
<td>Total transcripts with 3’ UTR</td>
<td>6,195</td>
<td>13,148</td>
<td>13,198</td>
</tr>
<tr>
<td>Average nucleotides/transcripts</td>
<td>2,029.87</td>
<td>1,737.20</td>
<td>1,788.76</td>
</tr>
<tr>
<td>Average nucleotides/CDS</td>
<td>1,617.68</td>
<td>1,333.73</td>
<td>1,333.26</td>
</tr>
<tr>
<td>Ave. 5’ UTR length</td>
<td>169.18</td>
<td>160.13</td>
<td>179.41</td>
</tr>
<tr>
<td>Ave. 3’ UTR length</td>
<td>243</td>
<td>243.34</td>
<td>275.22</td>
</tr>
<tr>
<td>Fraction of transcripts with AED &lt; 0.2</td>
<td>0.81</td>
<td>0.79</td>
<td>0.82</td>
</tr>
</tbody>
</table>

Comparison of structural features between alternatively spliced genes generated by MAKER run *de novo*, TAIR10, and MAKER updating TAIR10.

**TABLE 4. ncRNA annotations.**

<table>
<thead>
<tr>
<th></th>
<th>TAIR10</th>
<th>MAKER-P</th>
<th>Annotated by TAIR10 and MAKER-P</th>
</tr>
</thead>
<tbody>
<tr>
<td>tRNA</td>
<td>631</td>
<td>633</td>
<td>628</td>
</tr>
<tr>
<td>rRNA</td>
<td>4</td>
<td>18</td>
<td>4</td>
</tr>
</tbody>
</table>

Percentage of MAKER standard and TAIR 10 annotations broken down by star rating with Pfam domains, homology to eukaryotes in RefSeq, or various combinations of RNA seq/EST/cDNA evidence.
<table>
<thead>
<tr>
<th>Type</th>
<th>TAIR10</th>
<th>MAKER-P</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>snoRNA</td>
<td>71</td>
<td>70</td>
<td>64</td>
</tr>
<tr>
<td>miRNA</td>
<td>180</td>
<td>348</td>
<td>131</td>
</tr>
<tr>
<td>Others</td>
<td>480</td>
<td>38</td>
<td>19</td>
</tr>
</tbody>
</table>

Number of non-coding RNA annotations broken down by type in the TAIR10 and MAKER-P annotation sets. Last column shows the number of each type of non-coding RNAs annotated in both sets.

**Table 5. Location of all software and datasets.** Column 1: software, user tutorials or datasets Column 2: download location and password if applicable.

<table>
<thead>
<tr>
<th>Software</th>
<th>Download Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAKER-P (v2.29) download</td>
<td><a href="http://www.yandell-lab.org/software/maker-p.html">http://www.yandell-lab.org/software/maker-p.html</a></td>
</tr>
<tr>
<td>WebApollo download</td>
<td><a href="https://code.google.com/p/apollo-web/downloads/list">https://code.google.com/p/apollo-web/downloads/list</a></td>
</tr>
<tr>
<td>TAIR10, MAIZE, and MAKER-P annotation gff3 files</td>
<td><a href="http://weatherby.genetics.utah.edu/A_thaliana/username:MAKER-P">http://weatherby.genetics.utah.edu/A_thaliana/username:MAKER-P</a> password:marksentme</td>
</tr>
<tr>
<td>iPlant MAKER-P documentation</td>
<td><a href="https://pods.iplantcollaborative.org/wiki/display/sciplant/MAKER-P+Documentation">https://pods.iplantcollaborative.org/wiki/display/sciplant/MAKER-P+Documentation</a></td>
</tr>
<tr>
<td>Pseudogene pipeline download and tutorial</td>
<td><a href="http://shiulab.plantbiology.msu.edu/wiki/index.php/Protocol:Pseudogene">http://shiulab.plantbiology.msu.edu/wiki/index.php/Protocol:Pseudogene</a></td>
</tr>
<tr>
<td>miR-PREFeR</td>
<td><a href="https://github.com/hangelwen/miR-PREFeR">https://github.com/hangelwen/miR-PREFeR</a></td>
</tr>
<tr>
<td>tRNAscan-SE</td>
<td><a href="http://selab.janelia.org/software.html">http://selab.janelia.org/software.html</a></td>
</tr>
<tr>
<td>Snoscan</td>
<td><a href="http://lowelab.ucsc.edu/snoscan/">http://lowelab.ucsc.edu/snoscan/</a></td>
</tr>
</tbody>
</table>
Figure legends

Figure 1. AED CDF for TAIR10 annotations compared to human RefSeq annotations. Annotation Edit Distance (AED) can be used to assess how well an annotation set agrees with its associated evidence. When plotted as a cumulative AED distribution multiple annotation sets can be visualized on the same plot. Here we have included the AED CDF for the TAIR10 (orange line) annotation of *A. thaliana* and the human RefSeq (purple line) annotations of human for purposes of comparison.

Figure 2. MAKER-P *de novo* annotation and update of TAIR10 annotations. AED CDF curves for MAKER-P run as a *de novo* plant annotation engine (green curve), and when used to update the existing TAIR10 gene annotation dataset (blue curve), bringing it into better agreement with the evidence. Both MAKER-P datasets improve upon the existing TAIR10 annotations (orange curve).

Figure 3. MAKER-P improvements in AED are distributed across the entire TAIR10 dataset. This figure shows the cumulative AED distributions for the TAIR10 representative transcripts broken down by TAIR star rating system. Note the excellent agreement between the TAIR10 manually curated evidence classifications and MAKER's automatic AED-based quality control schema. The dotted lines denote the AED curves for the MAKER-P updated TAIR10 annotations.

Figure 4. MAKER-P runtimes on the entire *Zea mays* V2 genome assembly vs. the number of processors used. Increasing the number of processors given to MAKER-P decreases the runtime. Runtime is less than 4 hours using fewer than 500 CPUs, decreasing to less than 3 hours with 1092 CPUs.

Figure 5. MAKER-P annotations can be easily visualized using WebApollo. This view from WebApollo shows the original TAIR10 AT5G03540 gene transcripts (orange), the MAKER-P *de novo* gene annotation at that locus (blue), and the MAKER-P updated AT5G03540 gene transcripts (green). A subset of the mRNA-seq and EST/cDNA data is shown in beige.
Cumulative Fraction of Annotations

- TAIR10 rep transcripts (27,206)
- Human_NM (29,123)
Cumulative Fraction of Annotations

- TAIR10 rep transcripts (27,206)
- MAKER de novo (25,956)
- MAKER update of TAIR10 (26,885)