TOOLS AND TRAINING FOR GENOME ANNOTATION AND ANALYSIS

by

Michael Stephen Campbell

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Human Genetics

The University of Utah

December 2015

Copyright © Michael Stephen Campbell 2015

All Rights Reserved

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Michael Stephen Campbell**

has been approved by the following supervisory committee members:

Mark Yandell	, Chair	**7/7/2015** Date Approved
Carl Thummel	, Member	**7/7/2015** Date Approved
Anthea Letsou	, Member	**7/7/2015** Date Approved
Ellen Pritham	, Member	**7/7/2015** Date Approved
Lisa Joss-Moore	, Member	**7/7/2015** Date Approved

and by **Lynn Jorde** , Chair/Dean of

the Department/College/School of **Human Genetics**

and by David B. Kieda, Dean of The Graduate School.

ABSTRACT

The MAKER genome annotation and curation software tool was developed in response to increased demand for genome annotation services, secondary to decreased genome sequencing costs. MAKER currently has over 1000 registered users throughout the world. This wide adoption of MAKER has uncovered the need for additional functionalities. Here I addressed moving MAKER into the domain of plant annotation, expanding MAKER to include new methods of gene and noncoding RNA annotation, and improving usability of MAKER through documentation and community outreach.

To move MAKER into the plant annotation domain, I benchmarked MAKER on the well-annotated *Arabidopsis thaliana* genome. MAKER performs well on the Arabidopsis genome in de novo genome annotation and was able to improve the current TAIR10 gene models by incorporating mRNA-seq data not available during the original annotation efforts. In addition to this benchmarking, I annotated the genome of the sacred lotus *Nelumbo Nucifera*.

I enabled noncoding RNA annotation in MAKER by adding the ability for MAKER to run and process the outputs of tRNAscan-SE and snoscan. These functionalities were tested on the Arabidopsis genome and used MAKER to annotate tRNAs and snoRNAs in *Zea mays*. The resulting version of MAKER was named MAKER-P. I added the functionality of a combiner by adding EVidence Modeler to the MAKER code base.

As the number of MAKER users has grown, so have the help requests sent to the MAKER developers list. Motivated by the belief that improving the MAKER documentation would obviate the need for many of these requests, I created a media wiki that was linked to the MAKER download page, and the MAKER developers list was made searchable. Additionally I have written a unit on genome annotation using MAKER for *Current Protocols in Bioinformatics*. In response to these efforts I have seen a corresponding decrease in help requests, even though the number of registered MAKER users continues to increase.

Taken together these products and activities have moved MAKER into the domain of plant annotation, expanded MAKER to include new methods of gene and noncoding RNA annotation, and improved the usability of MAKER through documentation and community outreach.

TABLE OF CONTENTS

LIST OF FIGURES

Figures

LIST OF TABLES

Tables

CHAPTER 1

INTRODUCTION

Abstract

Genome projects have evolved from large international undertakings to tractable endeavors for a single lab. Accurate structural genome annotation is critical for successful genomic and molecular biology experiments. These annotations can be generated using a number of approaches and available software tools. This unit describes methods for structural genome annotation and a number of software tools commonly used in structural gene annotation. Submitted to Current Protocols in Bioinformatics.

Introduction

Genome projects have evolved rapidly over the past quarter of a century. Projects that once required large international consortiums, multimillion-to-billion-dollar budgets, and a decade plus of effort[1–4] can now be completed by a small lab using startup funds in a matter of months[5–9]. These changes are a direct result of the decreased sequencing costs spurred on by introduction of second generation sequencing technologies and computational algorithms that can make sense of the data these technologies generate.

Genome projects can be broadly broken into three successive parts: 1. Generate a reference genome assembly; 2. Generate a set of gene annotations reporting the intron/exon structure of the genes in the assembly; 3. Design and carry out experiments.

Every experiment in part 3 is doomed to failure if the results of parts 1 or 2 are inaccurate. This chapter focuses on part 2, structural gene annotation. Gene annotation is not gene prediction. A gene prediction is a prediction of the intron/exon structure of a gene based on a mathematical model, while an annotation is the synthesis of multiple lines of evidence; including gene predictions, expression data (often in the form of mRNA seq data), protein homology, and repetitive elements into the intron/exon structure of a gene while maintaining a trail of supporting evidence. Many gene predictors blur the lines between gene prediction and annotation by using hints from aligned evidence to inform and update their mathematical models, resulting in more accurate predictions[10,11]. However they lack the evidence trail associated with full annotation pipelines. Structural genome annotation is a two-step process, including evidence generation and synthesis. There are a number of ways to generate evidence and synthesize it into final gene annotations. The simplest approaches either run a single gene predictor across the genome, or use transcript or protein alignments to generate the gene models. These methods are fast but suffer from low accuracy[12]. Full annotation pipelines improve accuracy by incorporating multiple tools and approaches to generate gene annotations but require more computational resources[12]. The varying levels of complexity in gene prediction/annotation, as well as common workflows, are illustrated in Figure 1.1.

Evidence generation

Defined in loose terms, evidence is any information that can be used to identitfy/inform the exon/intron structure of a gene. Repeat masking, transcript and protein alignments, gene predictions, and whole genome alignment of closely related species in some combination are commonly used as evidence in genome annotation.

Repeat masking

Transposable elements and low-complexity repeats can combine in obnoxious ways that wreak havoc on gene annotation. Transposable elements often contain open reading frames that can be mistaken for exons by gene predictors and added to nearby gene models. Additionally, many real transcripts and protein products contain stretches of low-complexity sequence such as short tandem repeats and homopolymer runs. These low-complexity regions sprinkled throughout the genome can result in spurious evidence alignments[13] giving false support for gene annotations. Soft masking is the preferred method for handling low-complexity repeats. Soft masking is accomplished by changing the case in the FASTA sequence from upper case to lower case while hard masking changes the sequence to Ns[13]. Lower case letters serve as a signal to the aligner to not seed alignments in the region while preserving the sequence identity, allowing alignments to be extended through these regions, preventing off-target alignments. Seg and dust filtering are native to BLAST and are examples of algorithms that identify and soft mask low-complexity sequences[13].

Transposable elements are more challenging to overcome. RepeatMasker will mask transposable elements in a given genome given a library of known transposable elements[14]. RepBase is a collection of repetitive/transposable elements from a wide range of species, and it is commonly used to mask newly assembled genomes[15]. However, if RepBase does not contain transposable elements from one's species of interest, unmasked transposable elements are likely to remain. These unmasked transposable elements can be either species-specific or highly diverged from those in RepBase. Generating a species-specific repeat library can mitigate this problem. Common tools for identifying and

classifying repetitive elements include RepeatModeler[16], RECON[17], RepeatScout[18], TRF[19], LTRharvest[20], and MITE-hunter[21]. When generating a species-specific repeat library, it is important to screen the resulting sequences for "real" genes. Many of these tools use homology-based methods to identify repetitive elements, so the final library often contains more than just transposable elements, including highly conserved genes, such as histones. This filtering can be accomplished by aligning the repeat library to a set of known proteins—such as SWISS-PROT—and using the ProtExcluder package to remove sequences that are highly similar to nontransposable element proteins[22]. Hard masking transposable elements is typically the first step in genome annotation.

Transcript and protein sequence alignment and polishing

mRNA-seq data from the organism of interest and proteins from closely related species can be used to identify putative exons by indicating if they are expressed or evolutionarily conserved. mRNA-seq data in the form of short reads can be aligned to the genome directly using a gapped short read aligner (such as Tophat2)[23], NovoAline[24], and GSNAP[25]). They can be further processed into putative transcripts using tools such as Cufflinks[26], StringTie[27], and Trinity[28]; or assembled *de novo* into transcripts using tools such as Trinity[28], and Trans-ABySS[29], and aligned to the genome using BLASTN. A complete view of an organism's transcriptome would require mRNA-seq data from every tissue at every developmental stage under all possible conditions. This level of sequencing is cost prohibitive for most organisms, and in the case of endangered/protected species, impossible to obtain. To complement or, in some cases, replace expression evidence, it is helpful to use the whole proteomes of several well-annotated closely related species, as well as a set of curated proteins, such as those found

in SWISS-PROT[30]. Raw transcript and protein alignments are not very useful on their own. These alignments are unreliable around exon boundaries often extending into introns, and at times, missing whole exons, or aligned to the wrong strand[13]. Exonerate[31] and GeneWise[32] are splice-aware tools used to polish BLAST alignments. These polished alignments can then be used to inform the annotation of the coding sequence, splice sites, three prime and five prime untranslated regions (UTR) of genes and, in some of the simplest pipelines, serve as the final gene models. mRNA-seq data from a closely related species can be used as evidence when transcript data is not available for the species being annotated. Aligning these data in nucleotide space often results in low scoring alignments, so using TBLASTX or an equivalent to align these sequences in protein space produces much more sensitive alignments. However, translating and aligning the query and subject into all six possible reading-frames is computationally expensive. Moreover, any support for five prime and three prime untranslated regions is lost. Thus, it is more efficient to use the proteome of a closely related species in place of the transcriptome when possible. See Table 1.1 for a list of selected assemblers and aligners.

Gene prediction (*ab initio* and evidence driven)

Ab initio gene predictors provide a fast and easy way to identify genes in newly assembled genomes. These gene predictors rely on mathematical models of intron/exon structure to identify genes. Generating the mathematical models these gene predictors rely on (also known as training) requires a large number of high quality gene models. Once trained, these tools can perform quite well, approaching 100% accuracy in well-studied genomes for which ample training data is available[33,34]. This level of training is hard to achieve in newly assembled genomes. Using a gene predictor trained on a closely

related species is an option, but many newly sequenced genomes were chosen for sequencing because they are not closely related to currently annotated genomes. Moreover, even between closely related species there can be big differences in intron size distribution and GC content, resulting in low accuracy. These differences can also be seen in the same genome; an example of this is the honeybee genome, which was recently re-annotated with variable GC content in mind[35]. Some gene predictors have evolved over time to take advantage of mRNA-seq data in the absence of pre-existing gene models, and offer online services for gene-finder training[36,37]. The MAKER annotation pipeline also provides means for simplifying gene-finder training using the aligned transcript and protein evidence generated by the pipeline[38]. See Support Protocol 1 in unit 4.11 for instructions on how to use MAKER to train the gene finder SNAP. Many of the command lines in this protocol are specific to SNAP, but the process is applicable to a number of gene finders. See Table 1.2 for a list of selected gene prediction tools.

Closely related species whole genome/exome alignments

Many of the newly sequenced genomes are of great evolutionary interest because they are not closely related to species with sequenced well-annotated genomes[39–41]. However, a number of closely related species are also being sequenced. Many of these are microbial pathogens with relatively small genomes that are only expected to differ slightly. For these genomes, whole genome alignment approaches are highly specific, cost effective, and time saving. Mugsy-annotator[42] and CONTRAST[43] are examples of tools that annotate orthologs through whole-genome multiple alignments. For closely related large genomes, exome alignment decreases some of the overhead associated with large whole-genome alignments. Projector[44] selects known genes from one species and

predicts the corresponding genes in another species. GASS[45] is the first example of a tool that uses whole-exome alignment and a shortest-path model to generate gene predictions. Both Projector and GASS require collinearity between the known and predicted genes. These approaches are limited in that species-specific genes will be missed, and errors in the annotation of the closely related species may be propagated forward into the new annotation. See Table 1.3 for a list of selected multiple alignment/projection annotation tools.

Synthesis

Once the evidence has been generated, the daunting task of the annotator is to synthesize this information into gene annotations. Many of the early genome projects including Human, Arabidopsis, and Drosophila dedicated years to manual annotation and curation where evidence and gene predictions were clustered visually into gene regions from which the annotators identified/created gene models that best represented the evidence[1–4]. The small genome communities of today do not have the funding or time to support these kinds of manual annotation efforts, opting instead for automated annotation pipelines. There are a number of approaches to automated annotation. Two common themes among all automated annotation pipelines are: 1. identifying gene regions based on evidence clustering, and 2. using the aligned RNA and protein evidence to improve the accuracy of the gene predictors. At this point, some pipelines will use a combiner/chooser algorithm to choose the combination of exons that best represents the evidence. GLEAN[46], the EVidence Modeler (EVM)/Program to Assemble Spliced Alignments (PASA)[47] pipeline, and recent versions of the MAKER pipeline that uses EVM internally weight different types of evidence based on known error profiles and

user input, then choose the combination of exons that minimizes the error. In the nGASP competition, designed to identify the most accurate gene prediction/annotation tools, the combiner approaches outperformed machine-learning gene prediction approaches[34]. To improve gene predictions further, the MAKER and Ensembl pipelines supply hints from the protein and RNA-seq alignments to gene predictors at runtime, allowing them to update the mathematical models they use to generate gene predictions. This approach is particularly attractive for annotating assemblies with variable GC content. MAKER, EVM/PASA, and Ensembl will also add untranslated regions to gene annotations based on RNA-seq data to further increase accuracy. See Tables 1.4 and 1.5 for a list of selected choosers/combiners and full annotation pipelines respectively.

So which annotation approach should one use? The simplest approaches are fast but suffer from low accuracy. Full annotation pipelines improve accuracy by incorporating multiple tools and approaches to generate gene annotations, but require more computational resources and often more time. Some things to consider when choosing an annotation method are related to the genome of interest, others are related to the computational resources available for use, and both of these should be viewed through the lens of balancing effort and accuracy. If one is annotating a genome with a closely related annotated species, a projection tool may perform well. If the organism of interest has no closely related annotated species, then a pipeline that can utilize RNA-seq and protein evidence will generate more accurate annotations.

The most advanced annotation pipelines are designed to run in a highly parallel manner using large multicore servers or computing clusters. This decreases the wall time required to run them dramatically. For example, the MAKER pipeline annotated the

loblolly pine genome in less than 14.6 hours using 8,640 CPUs at the Texas Advanced Computing Center (TACC)[48]. That same annotation would have taken months on a 48-core server and would still be running today if started on a single CPU. Many universities have one or more computing clusters with shared resources available to researchers. Additionally, many large compute clusters such as those at TACC will give allocations upon request. iPlant also provides access to large computational resources and provides multiple tools for genome annotation including the MAKER pipeline[49]. Cloud computing resources are a good alternative to local compute clusters. It is easy to set up cloud-based computational resources through services like Amazon EC2[50], if one has a basic understanding of unix, that scale cost effectively to large genome annotation projects. Additionally, once the cloud-based resource is no longer needed, it can be shut down; this translates to huge cost savings, making these analyses possible for small genome communities with limited resources.

Quality control

A multitude of failed experiments can be traced directly back to incorrect gene annotations. Even with an exon accuracy of 90% (rarely achieved by a gene predictor alone) the majority of genes in a given genome will have at least one incorrectly annotated exon. These incorrect gene models can affect not only experiments designed for the organism of interest, but can be passed on to other genomes projects that use them as evidence in their annotation efforts. It is important to have some measure of quality associated with individual annotations as well as the annotation set as a whole, including the evidence used to generate them, to prevent these time wasting and expensive mistakes.

Quality metrics

Since high-quality reference gene models are not available for most newly assembled genomes, it is impossible to use the standard metrics of sensitivity, specificity, and accuracy to assess annotation quality. Different organism communities and annotation pipelines have approached quality assignment in different ways. These approaches are typically based on the agreement of an annotation to the aligned RNA/protein evidence or homology and synteny to closely related species. The Arabidopsis Informatics Resource (TAIR)[51] developed a good example of an aligned evidence-based system for the *Arabidopsis thaliana* genome. In this system, each annotated transcript was assigned a number of stars ranging from zero stars, where there was no support from aligned evidence, to five stars, where every exon was supported and splice site confirmed by a single full length cDNA[52]. The maize genome community uses a comparative genomics approach to quality control, in which annotations that share sequence similarity with annotations in other grass species are considered higher quality if they are syntenic in relation to the five prime and three prime flanking genes. Annotations with sequence similarity but not synteny are considered lower quality, those that have no homology are the lowest[53]. This method works well in grasses because of the well-conserved synteny in the group, but would not work well for species that are phylogenetically isolated from other annotated species, or where synteny is not well conserved, such as with gibbons in the primate lineage[54].

The MAKER pipeline uses a metric called Annotation Edit Distance (AED), developed by The Sequence Ontology (SO)[55]. AED is a value between zero and one. An AED of zero means that there is no distance between the aligned evidence and the

annotation: every nucleotide is supported by an alignment and there are no alignments to nucleotides outside of the annotation (perfect accuracy). An AED of one indicates that there is no aligned evidence support or the aligned evidence is completely at odds with the annotation. AED has been shown to agree well with the TAIR star system and the maize synteny approach[53].

The MAKER annotation pipeline provides AED for each annotated transcript, along with other quality metrics, including the number of splice sites confirmed by RNA-seq evidence, exons confirmed by protein or RNA-seq data, as well as the length of the five and three prime UTRs[38]. MAKER can also be used to add these quality metrics to annotations from other sources. See Basic Protocol 2 in unit 4.11 for instructions.

Protein family domains can also be a good indicator of annotation quality. An annotation containing an identifiable protein domain is more likely to code for a functional protein than one that does not. Therefore, protein domains can be used to rescue gene models that would have been given a low quality score from lack of aligned evidence support. Unfortunately this method is binary labeling each annotation as good or bad, when one would like to know how good or how bad when assessing the quality of a set of annotations as a whole. Fortunately the fraction of annotations containing a protein family domain in a given genome is quite similar across genomes at about 0.69, allowing for a genome-wide measure of quality[12]. MAKER and Ensembl commonly report this fraction using accessory scripts outside of their annotation pipelines.

Community curation

As sophisticated as computational quality control can be, there is no real substitute for biologists looking at the annotations of their favorite genes in a genome

browser. Because of this, many genome communities will hold annotation jamborees, where members of the community come together and visually inspect every annotation. If it is not possible to visually inspect every annotation, it is important to visualize and correct (if necessary), annotations before designing experiments based on them. Tools for visualizing annotations include Gbrowse[56], Jbrowse[57], and IGV[58]; tools for visualizing and editing include WebApollo[59], GenomeView[60], and Artemis[61]. See Table 1.6 for a select list of genome visualization and annotation editing tools.

Annotation updating/management

As time passes, sequencing technology and tools advance and additional data becomes available. These advances lead to improved assemblies and more evidence to inform annotations. The majority of published genome projects are drafts and are expected to improve over time. Advancing long-read technologies are facilitating scaffolding and gap filling while RNA-seq experiments are producing large amounts of transcript evidence. Taken alone or together, these data can dramatically improve gene annotations. The challenge then becomes incorporating this new data into existing annotations. Additionally, it is common for the same genome to be annotated by multiple groups using different pipelines of varying degrees of sophistication; this results in annotations of variable accuracy. So now the task is to combine and update annotations and document the process.

As updates in assemblies and the generation of new data are ongoing, it is essential that the process of updating annotations be built into the maintenance of these community resources. Unfortunately, funding for resource maintenance is hard to obtain, therefore it is important that these updates are as automatable as possible. Several

existing tools can be used to report differences between annotation sets, including GLEAN[46], PASA[47], and BEDtools[62]. Ensembl[63] and PASA[47] can update annotations in the light of additional RNA-seq data. Ensembl can also merge annotation sets to create a consensus set. MAKER[12] performs all of these functionalities and has the added ability to map annotations forward to a new assembly and maintain a documented evidence trail.

Future considerations

Inexpensive sequencing has relaxed the selective pressure on the standard genome project allowing it to evolve from a single reference based endeavor to population-level interrogation of single species or multiple species[64,65]. These data have brought to light the shortcomings tied to a linear reference genome. Once ploidy exceeds one a linear reference fails to truly represent even a single individual, let alone a population. These shortcomings are propagated forward to the gene annotations, where a single based difference between homologous chromosomes or individuals can cause big changes in the structure of a gene. Add insertions, deletions, and structural rearrangements to the mix, and even more dramatic effects on an individual's complement of genes become apparent. This is particularly challenging in plants, where two accessions from the same species can have megabase level deletions large enough to contain multiple protein-coding genes[66]. Addressing this problem will require new representations for population level genomic sequences as well as new tools that can operate on these data.

References

1. Venter, J. C. *et al.* The Sequence of the Human Genome. *Science* **291,** (2001).

2. Adams, M. D. *et al.* The genome sequence of Drosophila melanogaster. *Science* **287,** 2185–95 (2000).

3. The Arabidospsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408,** 796–815 (2000).

4. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).

5. Vonk, F. J. *et al.* The king cobra genome reveals dynamic gene evolution and adaptation in the snake venom system. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 20651–6 (2013).

6. Castoe, T. *et al.* The Burmese python genome reveals the molecular basis for extreme adaptation in snakes. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 20645–50 (2013).

7. Smith, C. D. *et al.* Draft genome of the globally widespread and invasive Argentine ant (Linepithema humile). *Proc. Natl. Acad. Sci. U. S. A.* **108,** 5673–5678 (2011).

8. Suen, G. *et al.* The genome sequence of the leaf-cutter ant Atta cephalotes reveals insights into its obligate symbiotic lifestyle. *PLoS Genet.* **7,** (2011).

9. Smith, C. R. *et al.* Draft genome of the red harvester ant Pogonomyrmex barbatus. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 5667–5672 (2011).

10. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24,** 637–44 (2008).

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

12. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12,** 491 (2011).

13. Korf, I., Yandell, M. & Bedell, J. *Blast*. (O'Reilly, 2003).

14. Smit, A. F. ., Hubley, R. & Green, P. RepeatMasker. at <http://repeatmasker.org>

15. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110,** 462–7 (2005).

16. Smit, A. & Hubley, R. RepeatModeler. at <http://www.repeatmasker.org/RepeatModeler.html>

17. Bao, Z. & Eddy, S. R. Automated de novo identification of repeat sequence families in sequenced genomes. *Genome Res.* **13,** 1269–1276 (2003).

18. Price, A. L., Jones, N. C. & Pevzner, P. a. De novo identification of repeat families in large genomes. *Bioinformatics* **21,** 351–358 (2005).

19. Benson, G. Tandem repeats: a program to analyze DNA sequences. *Nucleic Acids Res.* **27,** 573–580 (1999).

20. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9,** 18 (2008).

21. Han, Y. & Wessler, S. R. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res.* **38,** e199 (2010).

22. Jiang, N. ProtExcluder. at <http://weatherby.genetics.utah.edu/MAKER/wiki/index.php/Repeat_Library_Construction-Basic>

23. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14,** R36 (2013).

24. NovoAlign. at <http://www.novocraft.com/products/novoalign/>

25. Wu, T. D. & Nacu, S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26,** 873–881 (2010).

26. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7,** 562–78 (2012).

27. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33,** (2015).

28. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29,** 644–52 (2011).

29. Robertson, G. *et al.* De novo assembly and analysis of RNA-seq data. *Nat. Methods* **7,** 909–912 (2010).

30. Bairoch, a & Apweiler, R. The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.* **28,** 45–48 (2000).

31. Slater, G. S. C. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6,** 31 (2005).

32. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res.* **14,** 988–995 (2004).

33. Guigó, R. *et al.* EGASP: the human ENCODE Genome Annotation Assessment Project. *Genome Biol.* **7 Suppl 1,** S2.1–31 (2006).

34. Coghlan, A. *et al.* nGASP--the nematode genome annotation assessment project. *BMC Bioinformatics* **9,** 549 (2008).

35. Elsik, C. G. *et al.* Finding the missing honey bee genes: lessons learned from a genome upgrade. *BMC Genomics* **15,** 86 (2014).

36. Hoff, K. J. & Stanke, M. WebAUGUSTUS--a web service for training AUGUSTUS and predicting genes in eukaryotes. *Nucleic Acids Res.* **41,** 123–128 (2013).

37. Solovyev, V., Kosarev, P., Seledsov, I. & Vorobyev, D. Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biol.* **7 Suppl 1,** S10.1–12 (2006).

38. Cantarel, B. L. *et al.* MAKER : An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res*. 188–196 (2008). doi:10.1101/gr.6743907.1

39. Smith, J. J. *et al.* Sequencing of the sea lamprey (Petromyzon marinus) genome provides insights into vertebrate evolution. *Nat. Genet.* **45,** 415–21, 421e1–2 (2013).

40. Amemiya, C. T. *et al.* The African coelacanth genome provides insights into tetrapod evolution. *Nature* **496,** 311–6 (2013).

41. Ming, R. *et al.* Genome of the long-living sacred lotus (Nelumbo nucifera Gaertn.). *Genome Biol.* **14,** R41 (2013).

42. Angiuoli, S. V, Hotopp, J. C. D., Salzberg, S. L. & Tettelin, H. Improving pan-genome annotation using whole genome multiple alignment. *BMC Bioinformatics* **12,** 272 (2011).

43. Gross, S. S., Do, C. B., Sirota, M. & Batzoglou, S. CONTRAST: a discriminative, phylogeny-free approach to multiple informant de novo gene prediction. *Genome Biol.* **8,** R269 (2007).

44. Meyer, I. M. & Durbin, R. Gene structure conservation aids similarity based gene prediction. *Nucleic Acids Res.* **32,** 776–783 (2004).

45. Wang, Y., Chen, L., Song, N. & Lei, X. GASS: genome structural annotation for Eukaryotes based on species similarity. *BMC Genomics* **16,** 1–14 (2015).

46. GLEAN. at <http://glean-gene.sourceforge.net/>

47. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9,** R7 (2008).

48. Wegrzyn, J. L. *et al.* Unique features of the loblolly pine (Pinus taeda L.) megagenome revealed through sequence annotation. *Genetics* **196,** 891–909 (2014).

49. Goff, S. a *et al.* The iPlant Collaborative: Cyberinfrastructure for Plant Biology. *Front. Plant Sci.* **2,** 34 (2011).

50. Amazon EC2. at <http://aws.amazon.com/ec2/>

51. Lamesch, P. *et al.* The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.* **40,** D1202–10 (2012).

52. The Arabidopsis Information Resource. Documentation for the TAIR gene model and exon confidence ranking system. **2009,** 2–6 (2009).

53. Campbell, M. *et al.* MAKER-P: a tool-kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiol.* **164,** 513–524 (2013).

54. Carbone, L. *et al.* Gibbon genome and the fast karyotype evolution of small apes. *Nature* **513,** 195–201 (2014).

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

56. Stein, L. D. *et al.* The generic genome browser: A building block for a model organism system database. *Genome Res.* **12,** 1599–1610 (2002).

57. Skinner, M. E., Uzilov, A. V, Stein, L. D., Mungall, C. J. & Holmes, I. H. JBrowse : A next-generation genome browser. *Genome Res.* **19,** 1630–1638 (2009).

58. Robinson, J. T. *et al.* Integrative genomics viewer. *Nature biotechnology* **29,** 24–26 (2011).

59. Lee, E. *et al.* Web Apollo: a web-based genomic annotation editing platform. *Genome Biol* **14,** R93 (2013).

60. Abeel, T., Van Parys, T., Saeys, Y., Galagan, J. & Van De Peer, Y. GenomeView: A next-generation genome browser. *Nucleic Acids Res.* **40,** 1–10 (2012).

61. Carver, T., Harris, S. R., Berriman, M., Parkhill, J. & McQuillan, J. a. Artemis: An integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* **28,** 464–469 (2012).

62. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–842 (2010).

63. Curwen, V. *et al.* The Ensembl automatic gene annotation system. *Genome Res.* **14,** 942–950 (2004).

64. Shapiro, M. D. *et al.* Genomic diversity and evolution of the head crest in the rock pigeon. *Science* **339,** 1063–7 (2013).

65. Prado-Martinez, J. *et al.* Great ape genetic diversity and population history. *Nature* **499,** 471–5 (2013).

66. Schatz, M. C. *et al.* Whole genome de novo assemblies of three divergent strains of rice, Oryza sativa, document novel gene space of aus and indica. *Genome Biol.* **15,** 506 (2014).

67. Kent, W. J. BLAT — The BLAST -Like Alignment Tool. *Genome Res.* **12,** 656–664 (2002). doi:10.1101/gr.229202.

68. Kapustin, Y., Souvorov, A., Tatusova, T. & Lipman, D. Splign: algorithms for computing spliced alignments with identification of paralogs. *Biol. Direct* **3,** 20 (2008).

69. Wang, K. *et al.* MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Res.* **38,** 1–14 (2010).

70. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29,** 15–21 (2013).

71. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19,** ii215–ii225 (2003).

72. Lomsadze, A., Burns, P. D. & Borodovsky, M. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Res.* **42,** 1–8 (2014).

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

74. Ter-hovhannisyan, V., Lomsadze, A., Chernoff, Y. O. & Borodovsky, M. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised

training Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. **18,** 1979–1990 (2008). doi:10.1101/gr.081612.108

75. Souvorov, A *et al.* Gnomon–NCBI eukaryotic gene prediction tool. *Natl. Cent. Biotechnol. Inf.* 1–24 (2010). at < http://www.ncbi.nlm.nih.gov/core/assets/genome/files/Gnomon-description.pdf>

76. Schweikert, G. *et al.* mGene: Accurate SVM-based gene finding with an application to nematode genomes. *Genome Res.* **19,** 2133–2143 (2009).

77. Liu, Q., Mackey, A. J., Roos, D. S. & Pereira, F. C. N. Evigan: A hidden variable model for integrating gene evidence for eukaryotic gene prediction. *Bioinformatics* **24,** 597–605 (2008).

78. Allen, J. E. & Salzberg, S. L. JIGSAW: Integration of multiple sources of evidence for gene prediction. *Bioinformatics* **21,** 3596–3603 (2005).

79. Foissac, S. *et al.* Genome Annotation in Plants and Fungi: EuGene as a Model Platform. *Curr. Bioinform.* **3,** 87–97 (2008).

80. Campbell, M. S. *et al.* MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiol.* **164,** 513–524 (2014).

81. Thibaud-Nissen, F., Souvorov, A., Murphy, T., DiCuccio, M. & Kitts, P. Eukaryotic Genome Annotation Pipeline. 169439 (2013). at <http://www.ncbi.nlm.nih.gov/books/NBK169439/>

82. Argo. at <http://www.broadinstitute.org/annotation/argo>

83. Thorvaldsdóttir, H., Robinson, J. T., Turner, D. & Mesirov, J. P. A genomic data viewer for iPad. *Genome Biol.* **16,** 1–6 (2015).

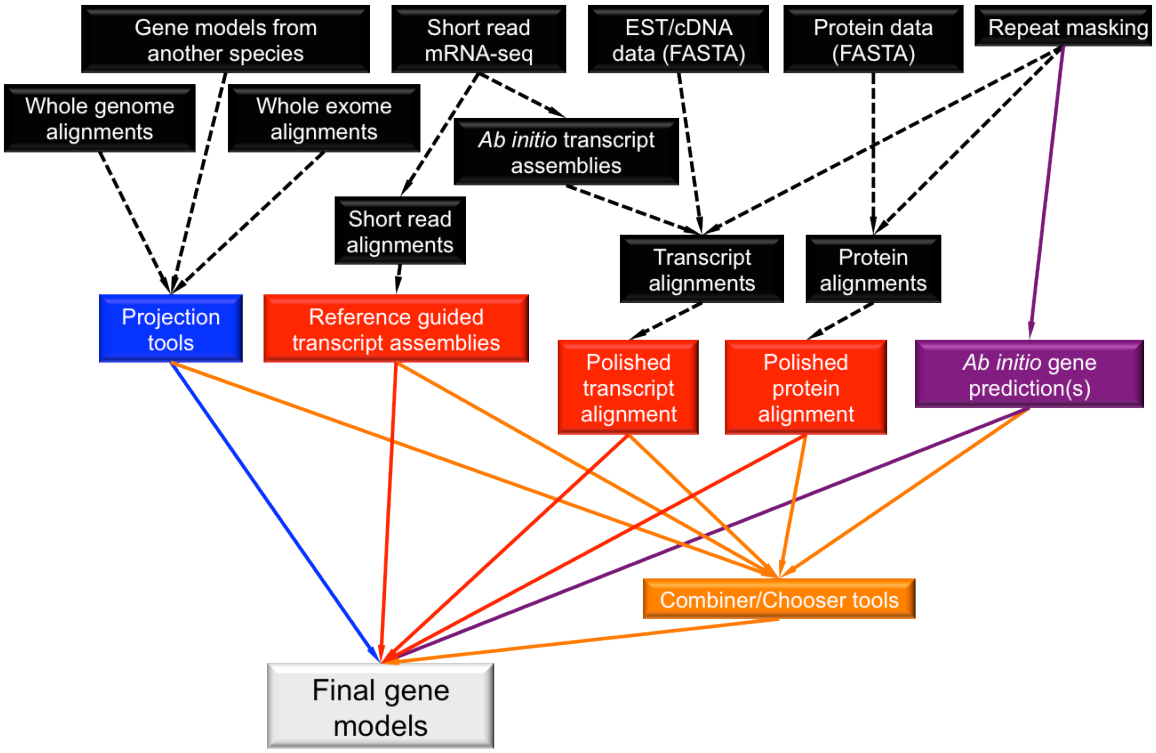

Figure 1.1. Multiple methods can be combined to generate structural gene annotations. Each black node in this flow chart represents an evidence type that can be used to inform gene annotation. The colored nodes represent classes of tools that can generate gene models. Arrows indicate inputs and outputs for each class of tools. Final gene models can be generated by projector tools (blue), from evidence alignments (red), *Ab initio* gene predictions (purple), and combiner tools (orange). Full annotation pipelines automate these processes while providing quality metrics and an evidence trail.

Table 1.1.

Evidence aligners and assemblers

Software package	Features	Reference
BLAST	A suite of tools that can align any combination of protein and nucleotide sequences. Uses Karlin-Altschule statistics.	13
BLAT	Faster than BLAST but not as configurable.	67
Tophat2	Memory efficent splice junction mapper for RNA-seq reads.	23
StringTie	Assembles transcripts from Tophat aligned RNA-seq reads and estimates transcript abundance. Designed as the successor to Cufflinks.	27
Trinity	Assembles transcripts *de novo* or with reference guidance.	28
NovoAlign	Aligns RNA and DNA short read sequences. Can use ambiguous nucleotide codes in the reference sequence. Requires purchased license.	24
GSNAP	Single nucleotide variant tolerant aligner for splice site detection. Available as part of the GMAP package.	25
Splign	Combines global and local alignment algorithms in a splice aware manner to align transcript sequences to a reference.	68
MapSplice	Splice junction mapper for RNA-seq reads.	69
STAR	Very fast and accurate RNA-seq aligner uses sequential mappable seed search in uncompressed suffix arrays.	70
Exonerate	Aligns proteins and assembled transcripts to a reference in a splice aware manner.	31

Table 1.2.

Gene predictors

Software Package	Features	Reference
Augustus	Can incorporate mRNA-seq data. Predicts alternatively spliced transcripts.	10,36,71
Genemark	Self-training. Performs well on fungal genomes. Versions available for prokaryotic and eukaryotic gene prediction.	72–74
Fgenesh	Run locally or through a webservice. Fee for use. Trained by softberry (no local training option).	37
SNAP	Easily trained. Incorporates hints from mRNA-seq and protein alignments.	11
Gnomon	Uses a combination of *ab initio* modeling and homology searching. Accepts mRNA-seq and protein data.	75
mGene	Utilizes multiple machine learning techniques including generalized hidden Markov models and Support Vector Machines.	76

Table 1.3.

Projection tools

Software package	Features	Reference
GASS	Uses a shortest path algorithm to predict protein-coding genes based on exon alignments from a closely related species.	45
Projector	Uses nucleotide and structural conservation between closely related species to predict genes.	44
CONTRAST	Predicts protein-coding genes from multiple genomic alignments. Source code is available but the developers don't recommend running it locally	43
Mugsy-Annotator	Identifies orthologs and assesses annotation quality though whole genome multiple alignments. Developed for prokaryotes.	42

Table 1.4.

Choosers and combiners

Software package	Features	Reference
EVidenceModeler	Combines aligned protein and transcript evidence with gene predictions in GFF3 format into weighted consensus gene models.	47
GLEAN	Combines multiple evidence types into a consensus gene model using a latent class statistical model.	46
Evigan	Uses a dynamic Bayesian network to generate a consensus gene model from multiple lines of evidence.	77
JIGSAW	Combines multiple evidence types into a consensus gene model. Can use non-linear models (training required) or a weighted linear combiner (no training required) to choose the best consensus model.	78

Table 1.5.

Genome annotation pipelines

Software package	Features	Reference
EuGene	Annotation pipeline that integrates multiple evidence types using a C++ based plugin system.	79
MAKER	Annotation pipeline that aligns and polishes protein and transcriptome data with BLAST and Exonerate, provides evidence-based hints to gene predictors, and provides an evidence trail and quality metrics for each annotation. Highly parallelizable.	12,38,80
Ensembl	Annotation pipeline that builds gene models from aligned and polished protein and transcript data. Identical transcripts are merged and a non-redundant set of transcripts is reported for each gene. Approximately a 4 month process.	63
NCBI	Annotation pipeline that aligns and polishes protein and transcript data. Generates Gnomon gene predictions. Weights gene models generated from manually-curated evidence higher than computationally derived models.	81
PASA	Annotation pipeline that aligns transcripts to the genome using BLAT, GMAP, or sim4. Can generate annotatons based on transcript data alone or preexisting gene models/gene predictions.	47

Table 1.6.

Genome browsers for community curation

Software package	Features	Reference
WebApollo	Web based plug in for Jbrowse with an editable user created annotation track. Edits are visible in real time to all currators.	59
Argo	Stand alone Java application for viewing and editing gene annotations.	82
IGV	Genome viewer that supports a variety of data times including bam and array based data. Also available for iPad.	58,83
GenomeView	Stand alone genome viewer and editor. Supports visualization of syntenic mapping and multiple-alignment data.	60
Artemis	Browser and annotation tool than can read EMBL and GENBANK database entries; sequence in FASTA format (indexed or raw); and other features in EMBL, GENBANK, or GFF format.	61
Jbrowse	Fast embeddable genome browser. Supports multiple data formats including VCF visualization.	57
Gbrowse	Feature-rich, highly customizable, web-based genome browser. Predecessor of Jbrowse.	56

CHAPTER 2

MAKER-P: A TOOL KIT FOR THE RAPID CREATION, MANAGEMENT, AND QUALITY CONTROL OF PLANT GENOME ANNOTATIONS

The following is a reprint of an article coauthored by myself, MeiYee Law, Carson Holt, Joshua C. Stein, Gaurav D. Moghe, David E. Hufnagel, Jikai Lei, Rujira Achawanantakun, Dian Jiao, Carolyn J. Lawrence, Doreen Ware, Shin-Han Shiu, Kevin L. Childs, Yanni Sun, Ning Jiang, and Mark Yandell. This article is originally published in *Plant Physiology* 2014, 164: 513-524 and is used with permission.

Personal contribution

I developed the benchmarking strategy and performed the experiments for the protein-coding gene benchmarks. I incorporated the ncRNA and pseudogene data into the final annotation sets. I wrote an early draft of the manuscript and contributed significantly to the text of the final publication. I edited the manuscript based on reviewer comments and edited and submitted the final version.

Breakthrough Technologies

MAKER-P: A Tool Kit for the Rapid Creation, Management, and Quality Control of Plant Genome Annotations[1][W][OPEN]

Michael S. Campbell, MeiYee Law, Carson Holt, Joshua C. Stein, Gaurav D. Moghe, David E. Hufnagel, Jikai Lei, Rujira Achawanantakun, Dian Jiao, Carolyn J. Lawrence, Doreen Ware, Shin-Han Shiu, Kevin L. Childs, Yanni Sun, Ning Jiang, and Mark Yandell*

Eccles Institute of Human Genetics (M.S.C., M.L., M.Y.) and Department of Biomedical Informatics (M.L.), University of Utah, Salt Lake City, Utah 84112; Ontario Institute for Cancer Research, Toronto, Ontario, Canada M5G 1L7 (C.H.); Genetics Program (G.D.M., S.-H.S., N.J.), Department of Plant Biology (D.E.H., S.-H.S., K.L.C.), Department of Computer Science and Engineering (R.A., Y.S.), and Department of Horticulture (J.C.S., N.J.), Michigan State University, East Lansing, Michigan 48824; University of Texas, Texas Advanced Computing Center, Austin, Texas 78758 (D.J.); United States Department of Agriculture-Agricultural Research Service Corn Insects and Crop Genetics Research Unit and Department of Genetics, Development, and Cell Biology, Iowa State University, Ames, Iowa 50011 (C.J.L.); iPlant Collaborative, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724 (D.W.); and United States Department of Agriculture-Agricultural Research Service North Atlantic Area, Robert W. Holley Center for Agriculture and Health, Ithaca, New York 14853 (D.W.)

We have optimized and extended the widely used annotation engine MAKER in order to better support plant genome annotation efforts. New features include better parallelization for large repeat-rich plant genomes, noncoding RNA annotation capabilities, and support for pseudogene identification. We have benchmarked the resulting software tool kit, MAKER-P, using the Arabidopsis (*Arabidopsis thaliana*) and maize (*Zea mays*) genomes. Here, we demonstrate the ability of the MAKER-P tool kit to automatically update, extend, and revise the Arabidopsis annotations in light of newly available data and to annotate pseudogenes and noncoding RNAs absent from The Arabidopsis Informatics Resource 10 build. Our results demonstrate that MAKER-P can be used to manage and improve the annotations of even Arabidopsis, perhaps the best-annotated plant genome. We have also installed and benchmarked MAKER-P on the Texas Advanced Computing Center. We show that this public resource can de novo annotate the entire Arabidopsis and maize genomes in less than 3 h and produce annotations of comparable quality to those of the current The Arabidopsis Information Resource 10 and maize V2 annotation builds.

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 noncoding 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 with 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 data sets; 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

[1] This work was supported by the National Science Foundation (grant no. IOS–1126998 to S.-H.S., K.L.C., Y.S., N.J., and M.Y. and grant no. DBI–0735191 to the iPlant Collaborative).

* Address correspondence to myandell@genetics.utah.edu.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Mark Yandell (myandell@genetics.utah.edu).

[W] The online version of this article contains Web-only data.

[OPEN] Articles can be viewed online without a subscription.

www.plantphysiol.org/cgi/doi/10.1104/pp.113.230144

Downloaded from www.plantphysiol.org on December 17, 2014 - Published by www.plant.org
Copyright © 2014 American Society of Plant Biologists. All rights reserved.

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; Schardl 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 the 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 at the Texas Advanced Computing Center (TACC) using the iPlant Cyberinfrastructure (Goff et al., 2011).

Here, we benchmark MAKER-P's accuracy and speed using two previously annotated plant genomes: Arabidopsis (*Arabidopsis thaliana*) and maize (*Zea mays*). Our Arabidopsis 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 TACC, we also show that MAKER-P can de novo annotate the Arabidopsis and maize genomes in less than 3 h and that the resulting annotations are of comparable quality to the current The Arabidopsis Information Resource 10 (TAIR10) and maize 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 the management of existing plant genome annotations.

RESULTS AND DISCUSSION

Choice of Target Species

We chose to benchmark MAKER-P using Arabidopsis 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 Arabidopsis genome, including traditional ESTs, full-length complementary DNAs (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 Arabidopsis Information Resource (TAIR; Lamesch et al., 2012) has put great effort into assigning evidence-based quality values to each annotation via its five-star rating system (The Arabidopsis Information Resource, 2009) in the current release of the Arabidopsis annotation set (TAIR10; Lamesch et al., 2012). Thus, the Arabidopsis 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 ab initio gene finders Semi Hidden Markov model [HMM]-Based Nucleic Acid Parser (SNAP; Korf, 2004) and Augustus (Stanke and Waack, 2003; Stanke et al., 2008). To do so, we ran SNAP and Augustus trained for Arabidopsis 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). The results of this analysis are reported in Table I. 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 are available for novel genomes. In such cases, 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 I, wherein we have run SNAP using the maize HMM as a surrogate for a poorly trained gene finder. In this case, the gene-level accuracy is much poorer: 70% compared with 82% using the Arabidopsis HMM. This demonstrates that attempts to leverage training data from other plants, maize 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 I reports the impact of running the same version of SNAP trained for maize within the MAKER software harness along with the RNA-Seq, EST/cDNA, and protein evidence data sets, as described in "Materials and Methods." This column of Table I demonstrates that MAKER-P's evidence-driven functions allow it to achieve high gene-level accuracies even using poorly trained 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 Annotation Edit Distance 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

Downloaded from www.plantphysiol.org on December 17, 2014 - Published by www.plant.org
Copyright © 2014 American Society of Plant Biologists. All rights reserved.

Table I. *Effects of MAKER-P's supervision of gene finders on genome-level sensitivity and specificity*

MAKER default, standard, and max refer to different MAKER gene-build options (see "Materials and Methods" and Supplemental Fig. S5).

Parameter	MAKER Default	MAKER Standard	MAKER Max	Augustus Trained for Arabidopsis Run outside of MAKER	SNAP Trained for Arabidopsis Run outside of MAKER	SNAP Trained for Maize Run outside of MAKER	SNAP Trained for Maize Run inside of MAKER
Sensitivity	0.88	0.91	0.93	0.91	0.84	0.47	0.67
Specificity	0.93	0.91	0.81	0.93	0.81	0.92	0.94
Accuracy	0.90	0.91	0.87	0.92	0.82	0.70	0.80

structures and predicted protein sequences 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 0 denoting perfect concordance with the available evidence and a value of 1 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 (for an example comparing gene models produced by the ab initio gene finder Augustus run with and without MAKER supervision, see Supplemental Fig. S1). Similarly, the same procedure can be used to evaluate the goodness of fit between a gold standard annotation data set and the evidence used to produce it. For additional information on AED, see Eilbeck et al. (2009), Holt and Yandell (2011), and Yandell and Ence (2012).

Cross-Genome Validation

AED also makes possible cross-genome assessments of annotation data sets 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 data sets used for our benchmarking analyses (for evidence data set details, see "Materials and Methods"). As can be seen, Arabidopsis 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 with a previously published analysis of human RefSeq annotations

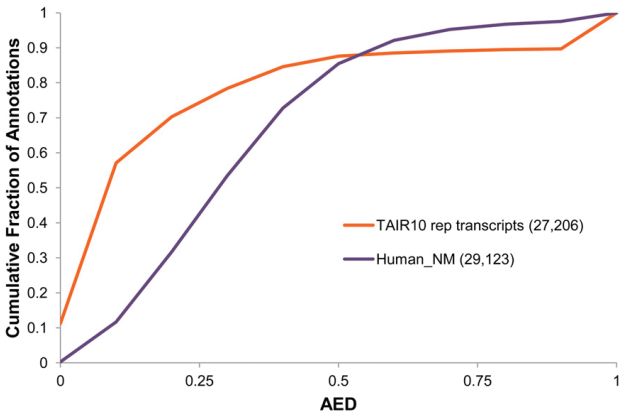

Figure 1. AED CDF for TAIR10 annotations compared with human RefSeq annotations. AED can be used to assess how well an annotation set agrees with its associated evidence. When plotted as a cumulative AED distribution, multiple annotation sets can be visualized on the same plot. Here, we have included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

Downloaded from www.plantphysiol.org on December 17, 2014 - Published by www.plant.org
Copyright © 2014 American Society of Plant Biologists. All rights reserved.

(Lander et al., 2001; Venter et al., 2001; for details of the data set, see "Materials and Methods"). Figure 1 also demonstrates that our evidence set provides support for 90% of the annotated genes in the TAIR10 data set.

Comparison of AED and TAIR's Five-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 II provides a breakdown of TAIR10 annotations by their star rating in the context of their supporting evidence using the evidence data sets used for our benchmarking analyses. Also shown in Table II is the cumulative support for the TAIR10 annotations in total and for the MAKER standard annotation build produced using the same evidence (for details, see "Materials and Methods"). Importantly, these results demonstrate that (1) MAKER-P can automatically produce a de novo genome annotation data set of very similar quality to the highly curated TAIR10 annotations and (2) there is good concordance between the 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 (for details, see "Materials and Methods"). Then we ran MAKER-P as we would to annotate a novel genome using the same evidence data set, 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 Arabidopsis annotations as a reference. As can be seen, both the updated and the de novo MAKER-P data sets 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 data set. In total, there are 2,368 genes present in TAIR10 that are absent from the MAKER de novo gene build. Sixty percent of the absent models are single-exon genes; 53% are one- or no-star gene-models; but 96% of all TAIR five-, four-, three-, and two-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 S2 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 data set. 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 scheme, cross validating both MAKER-P's AED and TAIR10's star rating approaches to assigning confidence levels to individual annotations. For five-star TAIR10 genes, 94% have AED scores of less than 0.5, whereas only 33% of one-star genes have an AED less than 0.5. Note that the four- and five-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 five-star status, whereas an annotation completely supported by tiled evidence is afforded four-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

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

The percentage of MAKER standard and TAIR10 annotations are broken down by star rating with Pfam domains, homology to eukaryotes in RefSeq, or various combinations of RNA-Seq/EST/cDNA evidence.

Star Rating	Fraction of Annotations with Pfam Domains	Fraction of Annotations with Eukaryotic RefSeq Protein Homology	Fraction of Annotations with Spliced RNA-Seq Support	Fraction of Annotations with RNA-Seq Support	Fraction of Annotations with Any RNA Support (mRNA-Seq, EST, cDNA)
Five stars (n = 7,880)	0.76	0.81	0.42	0.79	0.96
Four stars (n = 12,654)	0.87	0.84	0.94	0.95	0.99
Three stars (n = 2,087)	0.80	0.80	0.53	0.71	0.85
Two stars (n = 2,188)	0.81	0.79	0.64	0.69	0.80
One star (n = 1,788)	0.54	0.61	0.06	0.22	0.40
No star (n = 604)	0.14	0.22	0.02	0.04	0.07
TAIR10 representative transcripts (n = 27,206)	0.79	0.79	0.65	0.80	0.90
MAKER standard (n = 25,956)	0.79	0.72	0.66	0.82	0.93

Downloaded from www.plantphysiol.org on December 17, 2014 - Published by www.plant.org
Copyright © 2014 American Society of Plant Biologists. All rights reserved.

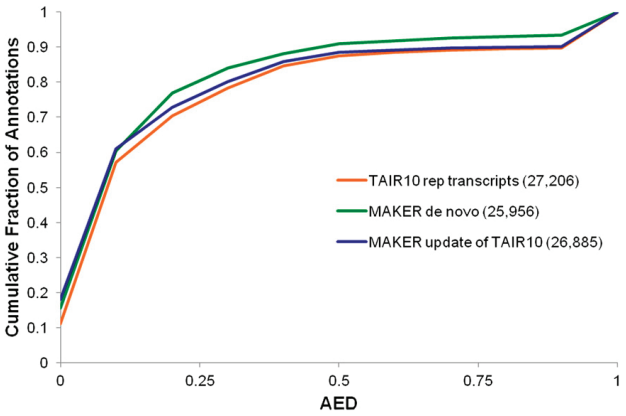

Figure 2. MAKER-P de novo annotation and update of TAIR10 annotations. AED CDF curves are shown for MAKER-P run as a de novo plant annotation engine (green curve) and when used to update the existing TAIR10 gene annotation data set (blue curve), bringing it into better agreement with the evidence. Both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

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 two-star through five-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 effect revisions. In contrast, the better supported genes (two-star through five-star annotations) 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 support 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 five-star status. Contradictory evidence is not considered under the TAIR scheme; however, it is considered by MAKER-P. This means that 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 greater than 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

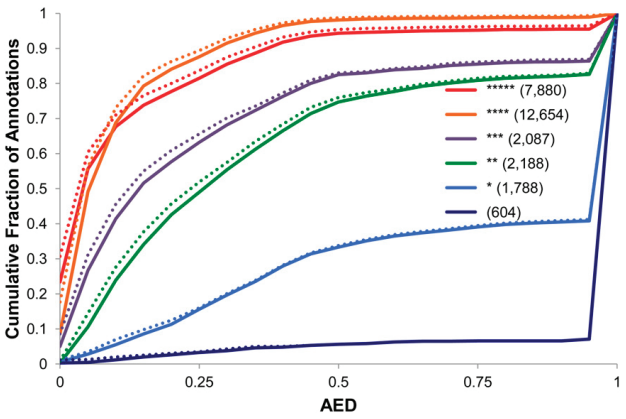

Figure 3. MAKER-P improvements in AED are distributed across the entire TAIR10 data set. The cumulative AED distributions for the TAIR10 representative transcripts are broken down by the TAIR star rating system. Note the excellent agreement between the TAIR10 manually curated evidence classifications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

Downloaded from www.plantphysiol.org on December 17, 2014 - Published by www.plant.org
Copyright © 2014 American Society of Plant Biologists. All rights reserved.

TAIR10 data sets (five- and four-star genes) demonstrates the strength of the AED approach to quality control. Further insight into the nature of these revisions is provided in Table III, which focuses on gene models with alternatively spliced transcripts

Alternative Splicing

MAKER-P annotates only the most certain of alternatively spliced transcripts, those with clear support for differential internal exon (cassette splicing); hence, the number of alternatively spliced transcripts is very limited compared with 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 alternatively spliced transcripts and added an average of 19 5′ untranslated region (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 III; 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 ncRNA 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 (A.F. Smit, R. Hubley, and P. Green, unpublished data) as well as its own internal repeat-finding method (Cantarel et al., 2008), 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 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 (for discussion, see Lerat, 2010), many MAKER users report that these tools are difficult to use. Moreover, the resulting output of existing packages often contains nontransposon genes or gene fragments, which may lead to the masking of bona fide genes. To address this point, the MAKER-P tool kit 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. For the Web addresses for both tutorials, see Table IV. We used the approach outlined in the basic tutorial to construct a novel Arabidopsis repeat library and then assayed the impact of using it for de novo annotation of Arabidopsis, using AED to evaluate the results. These data are shown in Supplemental Figure S3. In this case, we found little difference in MAKER-P's performance. However, Arabidopsis is not an ideal genome to demonstrate the effect of repeats on gene annotation, because the Arabidopsis genome contains the fewest repeats among all the sequenced plant genomes with the exception of the carnivorous bladderwort plant *Utricularia gibba* (Arabidopsis 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 "Materials and Methods"). We

Table III. *Features of alternatively spliced genes in the MAKER-P de novo annotation of Arabidopsis, TAIR10, and a MAKER-P update of TAIR10*

Comparison is shown for structural features between alternatively spliced genes generated by MAKER run de novo, TAIR10, and MAKER updating TAIR10.

Feature	MAKER-P	TAIR10	MAKER-P Update of TAIR10
No. of alternatively spliced genes	3,024	5,804	5,726
No. of alternatively spliced transcripts	7,190	13,774	13,590
Average exons per transcript	10.18	7.79	7.82
Total transcripts with 5′ UTR	5,708	12,714	12,352
Total transcripts with 3′ UTR	6,195	13,148	13,198
Average nucleotides per transcript	2,029.87	1,737.20	1,788.76
Average nucleotides per coding sequence	1,617.68	1,333.73	1,333.26
Average 5′ UTR length	169.18	160.13	179.41
Average 3′ UTR length	243	243.34	275.22
Fraction of transcripts with AED less than 0.2	0.81	0.79	0.82

Downloaded from www.plantphysiol.org on December 17, 2014 - Published by www.plant.org
Copyright © 2014 American Society of Plant Biologists. All rights reserved.

benchmarked them on the Arabidopsis 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. Arabidopsis; see "Materials and Methods"). In total, we identified 4,204 pseudogenes. Among these presumed pseudogenes, 2,277 have at least one premature stop and/or frame shift (referred to as disabling substitutions). Although the rest are without disabling substitutions, the median pseudogene length is 175 bp (Supplemental Fig. S4), 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 Arabidopsis 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 frame shifts, respectively, indicating that the false-positive rate of our pipeline is 3.1%. Assuming that 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:gene ratio. For example, the pseudogene:gene ratios differ significantly between F-box (152:567) and protein kinases (54:1,021; $P < 2.2 \times$ e-16), demonstrating that these families differ greatly in their loss rates.

ncRNAs

Using nine small RNA-Seq data sets of Arabidopsis (Supplemental Tables S1 and S2), the MAKER-P ncRNA tools identified 807 ncRNAs in total. The intersections of our predictions and TAIR10 annotations are summarized in Table V for tRNA, ribosomal RNA, small nucleolar RNA (snoRNA), microRNA (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 levels (e.g. one 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, data sets, and online documentation described in this report are listed in Table IV. MAKER-P, like its parent package MAKER, is a multithreaded, fully message passing interface-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 Lonestar cluster (for usage instructions [specifically "iPlant MAKER-P documentation"], see Table IV). The MAKER-P tool kit is freely available for academic use; for download information, see Table IV.

Speed Benchmarks

We first used the Arabidopsis genome to benchmark MAKER-P's performance on the TACC, which hosts the iPlant compute infrastructure. Using 600 central processing units (CPUs), we were able to complete the entire de novo annotation of the Arabidopsis assembly (approximately 120 Mb) in 2 h and 44 min. Even faster compute times can be achieved using additional CPUs and/or by launching multiple instances of MAKER-P (e.g. chromosome by chromosome). By doing so, we were able to perform the same annotation in 1 h and

Table IV. *Locations of all software and data sets*

Software, User Tutorials, or Data Sets	Download Location and Password if Applicable
MAKER-P (version 2.29) download	http://www.yandell-lab.org/software/maker-p.html
WebApollo download	https://code.google.com/p/apollo-web/downloads/list
TAIR10, maize, and MAKER-P annotation GFF3 files	http://weatherby.genetics.utah.edu/A_thaliana/ (username, MAKER-P; password, marksentme)
iPlant MAKER-P documentation	https://pods.iplantcollaborative.org/wiki/display/sciplant/MAKER-P+Documentation
Basic MAKER tutorial	http://weatherby.genetics.utah.edu/MAKER/wiki/index.php/MAKER_Tutorial
Pseudogene pipeline download and tutorial	http://shiulab.plantbiology.msu.edu/wiki/index.php/Protocol:Pseudogene
miR-PREFeR	https://github.com/hangelwen/miR-PREFeR
tRNAscan-SE	http://selab.janelia.org/software.html
snoscan	http://lowelab.ucsc.edu/snoscan/
Basic repeat library construction tutorial	http://weatherby.genetics.utah.edu/MAKER/wiki/index.php/Repeat_Library_Construction-Basic
Advanced repeat library construction tutorial	http://weatherby.genetics.utah.edu/MAKER/wiki/index.php/Repeat_Library_Construction-Advanced

Downloaded from www.plantphysiol.org on December 17, 2014 - Published by www.plant.org
Copyright © 2014 American Society of Plant Biologists. All rights reserved.

Table V. *ncRNA annotations*

The numbers of ncRNA annotations broken down by type in the TAIR10 and MAKER-P annotation sets are shown. The last column shows the number of each type of ncRNA annotated in both sets.

RNA	TAIR10	MAKER-P	Annotated by TAIR10 and MAKER-P
tRNA	631	633	628
Ribosomal RNA	4	18	4
snoRNA	71	70	64
miRNA	180	348	131
Others	480	38	19

27 min on 1,500 CPUs. An additional benchmarking analysis using the maize assembly (approximately 2 Gb) and 2,172 CPUs finished in 2 h and 53 min (Fig. 4). Run times are both a function of the evidence data set presented for alignment as well as the gene density of a genome, but the observed throughput of greater than 500 Mb h^{-1} demonstrates that even the largest of plant genomes could be annotated in a reasonable time frame by leveraging MAKER-P's scalability. Supplemental Figure S5 compares the resulting MAKER-P maize annotations with those of the current chromosome 10 V2 annotations available at MaizeGDB. As can be seen, the MAKER-P results compare favorably with the V2 annotations, with MAKER-P generating 3,059 gene annotations on this chromosome, an additional 365 gene annotations compared with the current V2 build. All of the 365 additional MAKER-P annotations are supported by RNA-Seq, EST, protein, or Pfam domain evidence and have overall better AED scores (Supplemental Fig. S6). Moreover, MAKER-P's annotation of alternatively spliced transcripts (Supplemental Table S3) mirrors its performance on the Arabidopsis genome (Table III), further demonstrating that MAKER-P can produce highly accurate maize 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 WebApollo project (Lee et al., 2013) to provide MAKER and MAKER-P users with easy means to distribute their annotation data sets 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 ether 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, and 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 data set shown on the left-hand panel into the JBrowse central frame. For additional details and data set download locations, see Table IV. WebApollo has many features that will benefit the plant genomes community. For example, WebApollo 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

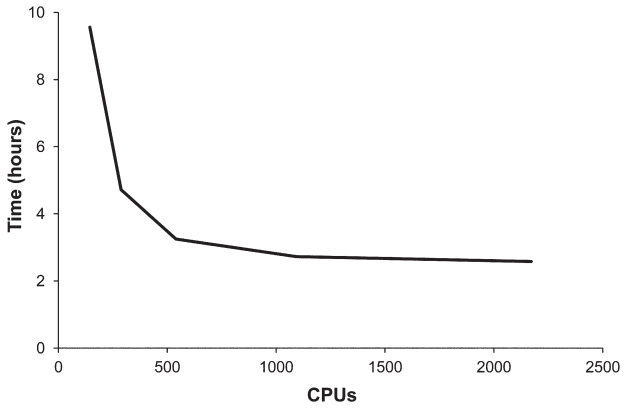

Figure 4. MAKER-P run times on the entire maize V2 genome assembly versus the number of processors used. Increasing the number of processors given to MAKER-P decreases the run time. Run time is less than 4 h using fewer than 500 CPUs, decreasing to less than 3 h with 1,092 CPUs.

Downloaded from www.plantphysiol.org on December 17, 2014 - Published by www.plant.org
Copyright © 2014 American Society of Plant Biologists. All rights reserved.

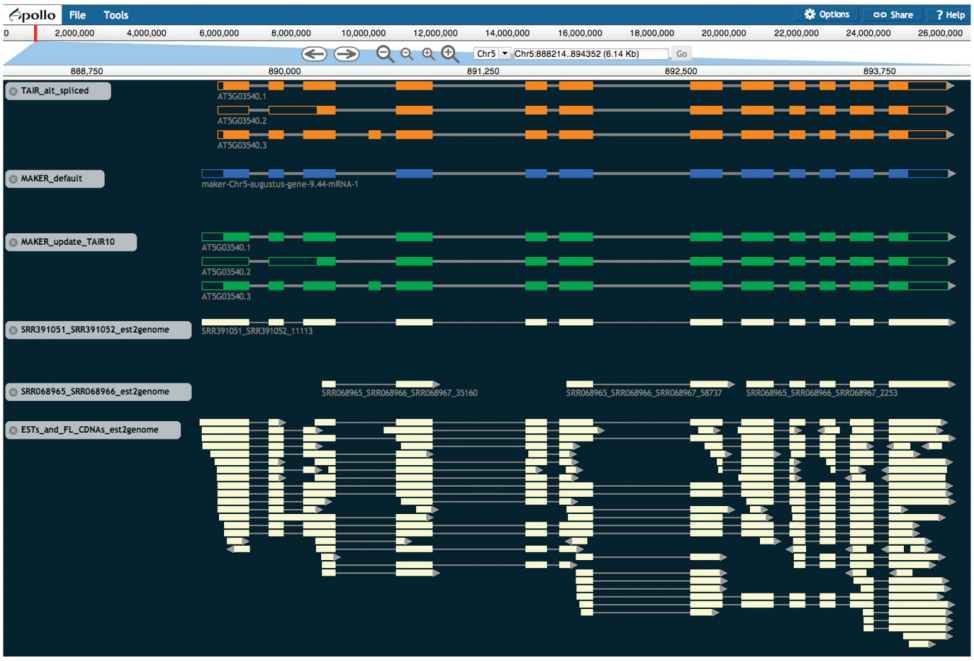

Figure 5. MAKER-P annotations can be easily visualized using WebApollo. This view from WebApollo shows the original TAIR10 *AT5G03540* gene transcripts (orange), the MAKER-P de novo gene annotation at that locus (blue), and the MAKER-P-updated *AT5G03540* gene transcripts (green). A subset of the mRNA-Seq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

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 four-star and one two-star transcripts; 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). The MAKER-P update of the TAIR10 gene model maintained all three transcripts, each containing additional 5′ and 3′ UTR sequences, as suggested by the RNA-Seq data, improving the overall AED of this gene model to 0.04 compared with the AED of 0.06 of the original TAIR10 gene model.

CONCLUSION

Today, the evidence for genome annotations evolves more rapidly than the annotations. In many cases, annotations fall out of synchronization 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 synchronization with the latest data sets. As we have demonstrated, the greatest revisions are accomplished for 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 Arabidopsis 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 the 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 the distribution of their results. MAKER-P thus provides the plant genome research community with a basic resource that democratizes genome annotation.

Downloaded from www.plantphysiol.org on December 17, 2014 - Published by www.plant.org
Copyright © 2014 American Society of Plant Biologists. All rights reserved.

MATERIALS AND METHODS

Evidence Sources and Assembly

Sequence evidence used for annotation by MAKER-P consisted of SwissProt protein data, EST and cDNA sequences from Arabidopsis (*Arabidopsis 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 Arabidopsis proteins were removed from this file, and only the non-Arabidopsis plant proteins were used when running MAKER-P. A file of Arabidopsis EST sequences (ATH_EST_sequences_20101108.fas) was obtained from TAIR (Lamesch et al., 2012). Full-length Arabidopsis cDNA sequences were downloaded from the National Center for Biotechnology Information (NCBI) Nucleotide database (Benson et al., 2013). Forty-seven RNA-Seq data sets derived from different Arabidopsis tissues and/or grown under different conditions were collected from the NCBI Short Read Archive (Supplemental Table S4; Wheeler et al., 2008). The reads from each file were cleaned using programs from the FASTX tool kit (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-Seq data sets were from 17 Short Read Archive studies and were thus assembled into 17 different transcript assemblies (Supplemental Table S4). All RNA-Seq data were treated as single-end reads in order to avoid aligning transcripts with stretches of Ns. The same procedures were used for the maize (*Zea mays*) data sets detailed in Supplemental Table S5.

Human annotations for release 37.2 were downloaded from the 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 demands of different users. For the basic protocol (for the Web address of the tutorial, see Table IV), RepeatModeler (http://www.repeatmasker.org/RepeatModeler.html) was used to process the genomic sequences with all Arabidopsis repeats excluded from the RepeatMasker repeat library so that the Arabidopsis genome would act as a "novel" genome. Among the repetitive sequences generated by RepeatModeler, some are classified, and they are considered as transposable elements. Sequences with unknown identity from RepeatModeler were searched against a transposase database (without Arabidopsis 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 (for the Web address of the advanced tutorial, see Table IV), 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 were collected using MITE-Hunter (Han and Wessler, 2010) with all default parameters. Long terminal repeat retrotransposons were collected using LTR-harvest and LTR-digest (Ellinghaus et al., 2008; Steinbiss et al., 2009), followed by a filtering to exclude false positives. To reduce redundancy, representative sequences (exemplars) were chosen as described previously (Schnable et al., 2009). To collect other repetitive sequences, the genomic sequence was then masked using the long terminal repeat and miniature inverted repeat transposable element 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 the advanced repeat library construction tutorial; its Web location is given in Table IV. The libraries made through different protocols masked different percentages of the genome (Supplemental Table S2); 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 transposable element library is a possible exception, generating a slightly lower accuracy and slightly higher overall AED scores (Supplemental Fig. S3).

MAKER-P de Novo Annotation of Arabidopsis

MAKER-P 2.27 r1020 was run on Arabidopsis (TAIR10 assembly) using the assembled Arabidopsis 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 were described above. 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 the 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 Arabidopsis or maize using MAKER-P in an iterative fashion as described for MAKER by 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 data set: default, standard, and max. The MAKER-P default build consists only of those gene models that are supported by the evidence (i.e. AED less than 1.0). The default build is thus very conservative. The MAKER-P standard build (which was used in Fig. 2 and Tables I and II) 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 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 (Supplemental Fig. S5). MAKER-P annotation of alternative 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 Arabidopsis genome using all Arabidopsis annotated protein sequences as queries. The output was filtered based on the following thresholds: E value < 1e-5, identity greater than 40%, match length greater than 30 amino acids, and coverage greater than 5% of the query sequence. The filtered matches provide pseudoexon definitions. These pseudoexons that are less than 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 data set. Finally, pseudogenes with significant similarity to known Viridiplantae repeats (cutoff = 300, divergence = 30; RepeatMasker 3.3.0) were discarded.

Downloaded from www.plantphysiol.org on December 17, 2014 - Published by www.plant.org
Copyright © 2014 American Society of Plant Biologists. All rights reserved.

This MAKER-P pseudogene identification pipeline is available for download at the location given in Table IV.

tRNA and snoRNA Annotation

MAKER-P features integrated means for the annotation of tRNAs and snoRNAs. tRNAs are identified using tRNAScan-SE (Lowe and Eddy, 1997) and snoRNAs with snoscan (Lowe and Eddy, 1999). Both tools are now supported and integrated 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 and 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-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 stochastic context-free grammar 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 data sets (Supplemental Tables S1, S2, and S7), which were sequenced from different tissues and conditions. All ncRNAs that were expressed in at least one RNA-Seq data set were validated using family-specific properties. tRNAScan-SE (Lowe and Eddy, 1997) and snoscan (Lowe and Eddy, 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. When running this tool on Arabidopsis, we used the properties that are associated with the biogenesis of miRNA maturation as features and trained an Alternating-Decision-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* (for the RNA strand that does not go on to become the active miRNA), 3′ overhang, secondary structure, minimum free energy, existence of the regulation target (miRNA target finding), number of samples in which the miRNA is expressed, and expression-level change across multiple RNA-Seq samples. All ncRNAs that pass the three-step pipeline are reported in Table V. The total run time for miR-PREFeR on Arabidopsis was 12 h and 21 min using four processing cores and nine RNA-Seq samples.

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure S1. Performance of an ab initio gene finder improves when supervised by MAKER.

Supplemental Figure S2. MAKER-P's improvements to the TAIR-10 gene models are not limited to culling of poorly supported gene models or merging gene models.

Supplemental Figure S3. Basic versus advanced repeat library generation has little effect on overall AED in Arabidopsis.

Supplemental Figure S4. Length distributions of genic and pseudogene features.

Supplemental Figure S5. Benchmarks of MAKER-P using the TAIR10 annotation dataset.

Supplemental Figure S6. Maize chromosome 10 analysis of V2 gene models.

Supplemental Table S1. RNA-Seq data sources used for miRNA identification from the NCBI's Sequence Read Archive.

Supplemental Table S2. RNA-Seq data sources used for miRNA identification from Massively Parallel Signature Sequencing Database.

Supplemental Table S3. Features of alternatively spliced genes in the MAKER-P de novo annotation of maize chromosome 10.

Supplemental Table S4. RNA-Seq data sources used for Arabidopsis benchmarks.

Supplemental Table S5. RNA-Seq data sources used for maize benchmarks.

Supplemental Table S6. Percentage of genomic sequences masked by different repeat libraries.

Supplemental Table S7. RNA-Seq data sources used for small RNA identification from the NCBI's Gene Expression Omnibus.

ACKNOWLEDGMENTS

We gratefully acknowledge the TACC support personnel. We also acknowledge and thank Chris Towne of J. Craig Venter Institute for helpful discussion and feedback as well as Suzie Lewis at the University of California, Berkeley, and the rest of the WebApollo team for their efforts to ensure WebApollo compatibility with MAKER-P outputs.

Received October 8, 2013; accepted November 26, 2013; published December 4, 2013.

LITERATURE CITED

Amemiya CT, Alföldi J, Lee AP, Fan S, Philippe H, Maccallum I, Braasch I, Manousaki T, Schneider I, Rohner N, et al (2013) The African coelacanth genome provides insights into tetrapod evolution. Nature **496:** 311–316

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature **408:** 796–815

Bennetzen JL (2005) Transposable elements, gene creation and genome rearrangement in flowering plants. Curr Opin Genet Dev **15:** 621–627

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2013) GenBank. Nucleic Acids Res **41:** D36–D42

Birol I, Raymond A, Jackman SD, Pleasance S, Coope R, Taylor GA, Yuen MM, Keeling CI, Brand D, Vandervalk BP, et al (2013) Assembling the 20 Gb white spruce (Picea glauca) genome from whole-genome shotgun sequencing data. Bioinformatics **29:** 1492–1497

Boerner S, McGinnis KM (2012) Computational identification and functional predictions of long noncoding RNA in Zea mays. PLoS ONE **7:** e43047

Campbell MA, Zhu W, Jiang N, Lin H, Ouyang S, Childs KL, Haas BJ, Hamilton JP, Buell CR (2007) Identification and characterization of lineage-specific genes within the Poaceae. Plant Physiol **145:** 1311–1322

Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, Holt C, Alvarado AS, Yandell M (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res **18:** 188–196

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

Eckalbar WL, Hutchins ED, Markov GJ, Allen AN, Corneveaux JJ, Lindblad-Toh K, Di Palma F, Alföldi J, Huentelman MJ, Kusumi K (2013) Genome reannotation of the lizard Anolis carolinensis based on 14 adult and embryonic deep transcriptomes. BMC Genomics **14:** 49

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

Ellinghaus D, Kurtz S, Willhoeft U (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics **9:** 18

Fahlgren N, Howell MD, Kasschau KD, Chapman EJ, Sullivan CM, Cumbie JS, Givan SA, Law TF, Grant SR, Dangl JL, et al (2007) High-throughput sequencing of Arabidopsis microRNAs: evidence for frequent birth and death of MIRNA genes. PLoS ONE **2:** e219

Feschotte C, Jiang N, Wessler SR (2002) Plant transposable elements: where genetics meets genomics. Nat Rev Genet **3:** 329–341

Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, Wilkinson AC, Finn RD, Griffiths-Jones S, Eddy SR, et al (2009) Rfam: updates to the RNA families database. Nucleic Acids Res **37:** D136–D140

Garg R, Patel RK, Jhanwar S, Priya P, Bhattacharjee A, Yadav G, Bhatia S, Chattopadhyay D, Tyagi AK, Jain M (2011) Gene discovery and tissue-

Downloaded from www.plantphysiol.org on December 17, 2014 - Published by www.plant.org
Copyright © 2014 American Society of Plant Biologists. All rights reserved.

specific transcriptome analysis in chickpea with massively parallel pyrosequencing and Web resource development. Plant Physiol **156:** 1661–1678

Goff SA, Vaughn M, McKay S, Lyons E, Stapleton AE, Gessler D, Matasci N, Wang L, Hanlon M, Lenards A, et al (2011) The iPlant Collaborative: cyberinfrastructure for plant biology. Front Plant Sci **2:** 34

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol **29:** 644–652

Guigó R, Flicek P, Abril JF, Reymond A, Lagarde J, Denoeud F, Antonarakis S, Ashburner M, Bajic VB, Birney E, et al (2006) EGASP: the human ENCODE Genome Annotation Assessment Project. Genome Biol (Suppl 1) **7:** S2

Han Y, Wessler SR (2010) MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Res **38:** e199

Hanada K, Zhang X, Borevitz JO, Li WH, Shiu SH (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 Res **17:** 632–640

Holt C, Yandell M (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics **12:** 491

Hua Z, Zou C, Shiu SH, Vierstra RD (2011) Phylogenetic comparison of F-Box (FBX) gene superfamily within the plant kingdom reveals divergent evolutionary histories indicative of genomic drift. PLoS ONE **6:** e16219

Ibarra-Laclette E, Lyons E, Hernández-Guzmán G, Pérez-Torres CA, Carretero-Paulet L, Chang TH, Lan T, Welch AJ, Juárez MJ, Simpson J, et al (2013) Architecture and evolution of a minute plant genome. Nature **498:** 94–98

Jiang SY, Christoffels A, Ramamoorthy R, Ramachandran S (2009) Expansion mechanisms and functional annotations of hypothetical genes in the rice genome. Plant Physiol **150:** 1997–2008

Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J (2005) Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res **110:** 462–467

Kejnovsky E, Hawkins J, Feschotte C (2012) Plant Transposable Elements: Biology and Evolution. *In* J Wendel, J Greilhuber, J Dolezel, IJ Leitch, eds, Plant Genome Diversity, Vol 1: Plant Genomes, Their Residents, and Their Evolutionary Dynamics. Springer, New York, pp 17–34

Korf I (2004) Gene finding in novel genomes. BMC Bioinformatics **5:** 59

Korf I, Yandell M, Bedell J (2003) BLAST. O'Reilly, Sebastopol, CA

Kumar S, Kushwaha H, Bachhawat AK, Raghava GPS, Ganesan K (2012) Genome sequence of the oleaginous red yeast Rhodosporidium toruloides MTCC 457. Eukaryot Cell **11:** 1083–1084

Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, et al (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res **40:** D1202–D1210

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al (2001) Initial sequencing and analysis of the human genome. Nature **409:** 860–921

Lee E, Helt GA, Reese JT, Munoz-Torres MC, Childers CP, Buels RM, Stein L, Holmes IH, Elsik CG, Lewis SE (2013) Web Apollo: a web-based genomic annotation editing platform. Genome Biol **14:** R93

Lerat E (2010) Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. Heredity (Edinb) **104:** 520–533

Li X, Wu HX, Southerton SG (2010) Comparative genomics reveals conservative evolution of the xylem transcriptome in vascular plants. BMC Evol Biol **10:** 190

Lin H, Moghe G, Ouyang S, Iezzoni A, Shiu SH, Gu X, Buell CR (2010) Comparative analyses reveal distinct sets of lineage-specific genes within Arabidopsis thaliana. BMC Evol Biol **10:** 41

Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res **25:** 955–964

Lowe TM, Eddy SR (1999) A computational screen for methylation guide snoRNAs in yeast. Science **283:** 1168–1171

Moghe GD, Lehti-Shiu MD, Seddon AE, Yin S, Chen Y, Juntawong P, Brandizzi F, Bailey-Serres J, Shiu SH (2013) Characteristics and significance of intergenic polyadenylated RNA transcription in Arabidopsis. Plant Physiol **161:** 210–224

Nawrocki EP, Kolbe DL, Eddy SR (2009) Infernal 1.0: inference of RNA alignments. Bioinformatics **25:** 1335–1337

Nystedt B, Street NR, Wetterbom A, Zuccolo A, Lin YC, Scofield DG, Vezzi F, Delhomme N, Giacomello S, Alexeyenko A, et al (2013) The Norway spruce genome sequence and conifer genome evolution. Nature **497:** 579–584

Paz-Ares J, Valencia A, Costantino P, Vittorioso P, Davies B, Gilmartin P, Giraudat J, Parcy F, Reindl A, Sablowski R, et al (2002) REGIA, an EU project on functional genomics of transcription factors from Arabidopsis thaliana. Comp Funct Genomics **3:** 102–108

Pellicer J, Fay M, Leitch I (2010) The largest eukaryotic genome of them all? Bot J Linn Soc **165:** 10–15

Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R (2005) InterProScan: protein domains identifier. Nucleic Acids Res **33:** W116–W120

Rounsley SD, Glodek A, Sutton G, Adams MD, Somerville CR, Venter JC, Kerlavage AR (1996) The construction of Arabidopsis expressed sequence tag assemblies: a new resource to facilitate gene identification. Plant Physiol **112:** 1177–1183

Schardl CL, Young CA, Hesse U, Amyotte SG, Andreeva K, Calie PJ, Fleetwood DJ, Haws DC, Moore N, Oeser B, et al (2013) Plant-symbiotic fungi as chemical engineers: multi-genome analysis of the Clavicipitaceae reveals dynamics of alkaloid loci. PLoS Genet **9:** e1003323

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves T, et al (2009) The B73 maize genome: complexity, diversity, and dynamics. Science **326:** 1112–1115

Seki M, Narusaka M, Kamiya A, Ishida J, Satou M, Sakurai T, Nakajima M, Enju A, Akiyama K, Oono Y, et al (2002) Functional annotation of a full-length Arabidopsis cDNA collection. Science **296:** 141–145

Slotkin R, Nuthikatuu S, Jaing N (2012) Plant Genome Diversity, Vol 1: Plant Genomes, Their Residents, and Their Evolutionary Dynamics. Springer, New York, pp 35–58

Smith JJ, Kuraku S, Holt C, Sauka-Spengler T, Jiang N, Campbell MS, Yandell MD, Manousaki T, Meyer A, Bloom OE, et al (2013) Sequencing of the sea lamprey (Petromyzon marinus) genome provides insights into vertebrate evolution. Nat Genet **45:** 415–421

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

Stanke M, Waack S (2003) Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics (Suppl 2) **19:** ii215–ii225

Steinbiss S, Willhoeft U, Gremme G, Kurtz S (2009) Fine-grained annotation and classification of de novo predicted LTR retrotransposons. Nucleic Acids Res **37:** 7002–7013

Sunkar R, Zhou X, Zheng Y, Zhang W, Zhu JK (2008) Identification of novel and candidate miRNAs in rice by high throughput sequencing. BMC Plant Biol **8:** 25

The Arabidopsis Information Resource (2009) Documentation for the TAIR gene model and exon confidence ranking system. ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR_gene_confidence_ranking/DOCUMENTATION_TAIR_Gene_Confidence.pdf (June 22, 2012)

Thibaud-Nissen F, Ouyang S, Buell CR (2009) Identification and characterization of pseudogenes in the rice gene complement. BMC Genomics **10:** 317

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al (2001) The sequence of the human genome. Science **291:** 1304–1351

Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, Dicuccio M, Edgar R, Federhen S, et al (2008) Database resources of the National Center for Biotechnology Information. Nucleic Acids Res **36:** D13–D21

Yamada K, Lim J, Dale JM, Chen H, Shinn P, Palm CJ, Southwick AM, Wu HC, Kim C, Nguyen M, et al (2003) Empirical analysis of transcriptional activity in the Arabidopsis genome. Science **302:** 842–846

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

Yang X, Jawdy S, Tschaplinski TJ, Tuskan GA (2009) Genome-wide identification of lineage-specific genes in Arabidopsis, Oryza and Populus. Genomics **93:** 473–480

Zou C, Lehti-Shiu MD, Thibaud-Nissen F, Prakash T, Buell CR, Shiu SH (2009) Evolutionary and expression signatures of pseudogenes in Arabidopsis and rice. Plant Physiol **151:** 3–15

Downloaded from www.plantphysiol.org on December 17, 2014 - Published by www.plant.org
Copyright © 2014 American Society of Plant Biologists. All rights reserved.

CHAPTER 3

GENOME OF THE LONG-LIVING SACRED LOTUS (NELUMBO NUCIFERA GAERTN)

The following is a reprint of an article coauthored by myself, Ray Ming, Robert VanBuren, Yanling Liu, Mei Yang, Yuepeng Han, Lei-Ting Li, Qiong Zhang, Min-Jeong Kim, Michael C Schatz, Jingping Li, John E Bowers, Haibao Tang, Eric Lyon, Ann A Ferguso, Giuseppe Narzis, David R Nelso, Crysten E Blaby-Haa, Andrea R Gschwen, Yuannian Jiao, Joshua P De, Fanchang Zeng, Jennifer Han, Xiang Jia Mi, Karen A Hudso, Ratnesh Sing, Aleel K Grenna, Steven J Karpowicz, Jennifer R Watlin, Kikukatsu Ito, Sharon A Robinson, Matthew E Hudson, Qingyi Yu, Todd C Mockler, Andrew Carroll, Yun Zheng, Ramanjulu Sunkar, Ruizong Jia, Nancy Chen, Jie Arro, Ching Man Wai, Eric Wafula, Ashley Spence, Yanni Han1, Liming Xu1, Jisen Zhang, Rhiannon Peery, Miranda J Haus, Wenwei Xiong, James A Walsh, Jun Wu, Ming-Li Wang, Yun J Zhu, Robert E Paull, Anne B Britt, Chunguang Du, Stephen R Downie, Mary A Schuler, Todd P Michael, Steve P Long, Donald R Ort, J William Schopf, David R Gang, Ning Jiang, Mark Yandell, Claude W dePamphilis, Sabeeha S Merchant, Andrew H Paterson, Bob B Buchanan, Shaohua Li, and Jane Shen-Miller. This article was originally published in *Genome Biology* 2013, 14:R41 and is used with permission.

Personal contribution

I performed the structural and functional annotation of the genome and generated the summary statistics. I wrote the genome annotation sections of the manuscript. I critically read and edited the manuscript.

Ming et al. Genome Biology 2013, 14:R41
http://genomebiology.com/2013/14/5/R41

Genome Biology

RESEARCH Open Access

Genome of the long-living sacred lotus (*Nelumbo nucifera* Gaertn.)

Ray Ming[1,2*†], Robert VanBuren[2†], Yanling Liu[1†], Mei Yang[1†], Yuepeng Han[1], Lei-Ting Li[2,3], Qiong Zhang[1,2], Min-Jeong Kim[4], Michael C Schatz[5], Michael Campbell[6], Jingping Li[7], John E Bowers[8], Haibao Tang[9], Eric Lyons[10], Ann A Ferguson[11], Giuseppe Narzisi[5], David R Nelson[12], Crysten E Blaby-Haas[13], Andrea R Gschwend[2], Yuannian Jiao[7,14], Joshua P Der[14], Fanchang Zeng[2], Jennifer Han[2], Xiang Jia Min[15], Karen A Hudson[16], Ratnesh Singh[17], Aleel K Grennan[2], Steven J Karpowicz[18], Jennifer R Watling[19], Kikukatsu Ito[20], Sharon A Robinson[21], Matthew E Hudson[22], Qingyi Yu[17], Todd C Mockler[23], Andrew Carroll[24], Yun Zheng[25], Ramanjulu Sunkar[26], Ruizong Jia[27], Nancy Chen[28], Jie Arro[2], Ching Man Wai[2], Eric Wafula[14], Ashley Spence[2], Yanni Han[1], Liming Xu[1], Jisen Zhang[29], Rhiannon Peery[2], Miranda J Haus[2], Wenwei Xiong[30], James A Walsh[2], Jun Wu[3], Ming-Li Wang[27], Yun J Zhu[27,31], Robert E Paull[28], Anne B Britt[32], Chunguang Du[30], Stephen R Downie[2], Mary A Schuler[2,33], Todd P Michael[34], Steve P Long[2], Donald R Ort[2,35], J William Schopf[36], David R Gang[4], Ning Jiang[11], Mark Yandell[6], Claude W dePamphilis[14], Sabeeha S Merchant[13], Andrew H Paterson[7], Bob B Buchanan[37], Shaohua Li[1*] and Jane Shen-Miller[36*]

Abstract

Background: Sacred lotus is a basal eudicot with agricultural, medicinal, cultural and religious importance. It was domesticated in Asia about 7,000 years ago, and cultivated for its rhizomes and seeds as a food crop. It is particularly noted for its 1,300-year seed longevity and exceptional water repellency, known as the lotus effect. The latter property is due to the nanoscopic closely packed protuberances of its self-cleaning leaf surface, which have been adapted for the manufacture of a self-cleaning industrial paint, Lotusan.

Results: The genome of the China Antique variety of the sacred lotus was sequenced with Illumina and 454 technologies, at respective depths of 101× and 5.2×. The final assembly has a contig N50 of 38.8 kbp and a scaffold N50 of 3.4 Mbp, and covers 86.5% of the estimated 929 Mbp total genome size. The genome notably lacks the paleo-triplication observed in other eudicots, but reveals a lineage-specific duplication. The genome has evidence of slow evolution, with a 30% slower nucleotide mutation rate than observed in grape. Comparisons of the available sequenced genomes suggest a minimum gene set for vascular plants of 4,223 genes. Strikingly, the sacred lotus has 16 COG2132 multi-copper oxidase family proteins with root-specific expression; these are involved in root meristem phosphate starvation, reflecting adaptation to limited nutrient availability in an aquatic environment.

Conclusions: The slow nucleotide substitution rate makes the sacred lotus a better resource than the current standard, grape, for reconstructing the pan-eudicot genome, and should therefore accelerate comparative analysis between eudicots and monocots.

* Correspondence: rming@life.uiuc.edu; shhli@wbgcas.cn; shenmiller@lifesci.ucla.edu
† Contributed equally
[1]Key Laboratory of Plant Germplasm Enhancement and Specialty Agriculture, Wuhan Botanical Garden, The Chinese Academy of Sciences, Lumo Road, Wuhan 430074, China
[36]IGPP Center for the Study of Evolution and Origin of Life, Geology Building, Room 5676, University of California, Los Angeles, 595 Charles E Young Drive East, Los Angeles, CA 90095-1567, USA
Full list of author information is available at the end of the article

BioMed Central

© 2013 Ming et al.; licensee BioMed Central Ltd. This is an open access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Background

Sacred lotus, so named because of its religious significance in both Buddhism and Hinduism, belongs to the small plant family Nelumbonaceae, with only one genus, *Nelumbo*, and two species: *N. nucifera* (Asia, Australia, Russia) and *N. lutea* (eastern and southern North America) [1]. Lotus is in the eudicot order Proteales, which lies outside of the core eudicots (Figure S1 in Additional file 1); its closest relatives are shrubs or trees belonging to the families Proteaceae and Platanaceae. Lotus was a land plant that has adapted to aquatic environments.

Used as a food for over 7,000 years in Asia, lotus is cultivated for its edible rhizomes, seeds and leaves. Its buds, flowers, anthers, stamens, fruits, leaves, stalks, rhizomes and roots have been used as herbal medicines for treatment of cancer, depression, diarrhea, heart problems, hypertension and insomnia [2,3]. Its seeds have exceptional longevity, remaining viable for as long as 1,300 years, and its vegetative rhizomes remain healthy for more than 50 years [1,2]. The nanoscopic closely packed protuberances of its self-cleaning leaf surface have been adapted in Europe for the manufacture of a 'self-cleaning' industrial paint, Lotusan. The use of this paint results in the so-called lotus effect that is now widely advertised for self-cleaning automobiles, buildings and fabrics.

Here, we report the sequencing and analysis of the sacred lotus genome, which descends from the most ancient lineage of angiosperms. We have studied the evolutionary history of the genome and genes involved in relevant processes governing the unique features of this ancient land plant, including its adaptation to aquatic environments.

Results

Genome sequencing and assembly

We sequenced the genome of the sacred lotus variety 'China Antique' with 94.2 Gb (101×) Illumina and 4.8 Gb (5.2×) 454 sequences. The final assembly includes 804 Mb, 86.5% of the estimated 929 Mb lotus genome [4]. The contig N50 is 38.8 kbp and the scaffold N50 is 3.4 Mbp (Table S1 in Additional file 1). The largest 429 scaffolds account for 94.8% of the assembled genome and 98.0% of the annotated genes. Among the 39 plant genomes published to date, the median N50 scaffold length is about 1.3 Mb, making lotus the eighth best assembled genome (Table S2 in Additional file 1). We constructed a high-density genetic map using 3,895 sequence-based restriction-associated DNA sequencing markers and 156 simple sequence repeat markers [5]. The former were sorted into 562 co-segregating bins and a total of 698 informative markers were mapped into nine linkage groups for the eight lotus chromosomes, with one gap remaining between two linkage groups (Table S3 in Additional file 1). The nine anchored megascaffolds have a combined size of 543.4 Mb, accounting for 67.6% of the genome assembly, and they are mostly proportional to the karyotype of the lotus chromosomes (Figure S2 and S3 in Additional file 1). The high quality of the lotus genome assembly is largely due to the unexpected homozygosity of the 'China Antique' variety. Although lotus is an out-crossing plant, its cultivation and vegetative propagation via rhizomes over the past 7,000 years may have imposed a narrow genetic bottleneck. This could be partly the consequence of its unique feature, seed longevity, which might have further reduced the number of generations in its evolutionary history in addition to vegetative propagation. The estimated heterozygosity in 'China Antique' is 0.03%, lower than the 0.06% of the sequenced papaya cultivar 'SunUp' after 25 generations of inbreeding [6]. The estimated heterozygosity in the American lotus *N. lutea* 'AL1' variety is 0.37%, also low.

Repeat content of the sacred lotus genome

Repetitive sequences account for 57% of the assembled genome, including 47.7% recognizable transposable elements (Table S4 in Additional file 1). Unlike most plants, which exhibit relatively inconsequential non-long terminal repeat retrotransposons (approximately 1% of the genome) [7-9], such non-long terminal repeat retrotransposons contribute 6.4% to the lotus genome. Differing from other plants that usually have more Gypsy-like elements [9,10], Copia and Gypsy-like elements are comparable in copy number and genomic fraction in lotus. Most major DNA transposon families are detected in sacred lotus (occupying 16% of the lotus genome), albeit with more than 10-fold variation in relative abundance. An exception, the *Tc1/Mariner* super-family, is absent from both the lotus and grape genomes [7], suggesting the frequent loss of this family of elements. Surprisingly, *hAT* (*Ac/Ds*-like) elements contribute to nearly 7% of the lotus genome, represented by more than 100,000 copies, more than in any other sequenced plant genome. Of these, CACTA elements are least abundant (0.4%) while *MULE*, *PIF* and *Helitron* elements have amplified to a moderate degree (2.5%, 2.7% and 3.6%, respectively). The lotus genome further includes 1,447 Pack-mutator-like elements that carry genes or gene fragments [11]. Analysis using expressed sequence tags (ESTs) indicated that at least 10 Pack-mutator-like elements are expressed, suggesting that they may play functional roles.

Genome annotation and gene expression

Following repeat-masking and annotation, we inferred 26,685 protein-coding genes in lotus, including all 458 core eukaryotic proteins [12]; 82% of the genes have similarity to proteins in SwissProt as identified by Basic

Local Alignment Search Tool (E < 0.0001). The average gene length is 6,561 bp with median exon and intron lengths of 153 bp and 283 bp, respectively (Table S1 in Additional file 1). The average gene density is one gene per 30 kb, with genes spread more evenly over the assembled genome than in many other plant genomes (Figure S2 in Additional file 1), which are characterized by gene-rich regions often found at the distal regions of chromosomes arms. A total of 12,344 ESTs were aligned to 11,741 gene models, and 174 alternative splicing events were identified from 164 genes involving 380 EST contigs (Table S5 in Additional file 1). Of the annotated genes in lotus, 22,803 (85.5%) show expression in rhizomes, roots, leaves or petioles based on RNAseq data (Figure S4 in Additional file 1). Expression of the remaining genes is likely confined to seeds, flowers and other unsurveyed tissues. Expression of 3,094 protein-coding genes was tissue-specific, including 1,910 genes showing expression only in rhizomes and 841 only in roots; 14,477 genes are expressed across all tissues surveyed. Of the 1,910 rhizome-specific genes, we found several AP2-like ethylene-responsive transcription factors, BTB/POZ domain-containing proteins, heat shock proteins, homeobox transcription factors, kinesins and pentatricopeptide repeat-containing proteins (PPRs) (Table S6 in Additional file 1). In lotus, 544 genes were annotated as PPRs, with 201 of these expressed in the four tissues tested, and 199 only expressed in the rhizome. PPRs have been identified as a group of RNA-binding proteins involved in RNA processing, stability, editing, maturation and translation in plants. Although the molecular mechanism of their function has not yet been elucidated, their broad expression in lotus rhizome is notable.

Ortholog classification and ancestral gene content in eudicots

The protein-coding gene sets from lotus and 16 other sequenced angiosperm species were used to identify putative orthologous gene clusters with Proteinortho v4.20 [13]. A total of 529,816 non-redundant genes were classified into 39,649 orthologous gene clusters (orthogroups) containing at least two genes (Table S7 in Additional file 1). Of the 26,685 protein-coding genes in lotus, 21,427 (80.3%) were classified into 10,360 orthogroups, of which 317 contained only lotus genes.

From this gene classification, we estimate a minimum gene set of 7,165 genes in 4,585 orthogroups for eudicots (Table S7 in Additional file 1). The minimum gene set for core eudicots (7,559 genes in 4,798 orthogroups) is only slightly larger than the eudicot-wide set, suggesting that the minimal gene set of the eudicot-monocot ancestor (6,423 genes in 4,095 orthogroups) would add at least 490 orthogroups associated with the eudicots as a whole.

We reconstructed the ancestral gene content at key nodes of the evolutionary series, as well as the adaptational changes occurring along the branches leading to these nodes: the greatest changes observed in orthogroup presence and absence are specific to terminal lineages (Tables S8 and S9 in Additional file 1 and Figure 1). More than three times as many orthogroup gains occur in the lineage leading to all eudicots, as compared to core eudicots (Figure S5 in Additional file 1), an increase second only to that of the grasses.

Synteny and genome evolution

A major evolutionary force shaping genome architecture in angiosperms is whole genome duplication (WGD) [14,15]. This process is followed by the 'diploidization' of genome organization through rearrangement, and of gene content through 'fractionation,' or homeologous gene loss. Intragenomic analysis of lotus indicates that it has experienced at least one WGD (paleotetraploidy, see Figure S6 in Additional file 1), named λ, but implies that the *Nelumbo* lineage did not experience γ, the paleohexaploidy (triplication) event around 125 million years ago detected in all other sequenced eudicot genomes [6,16-20]. Using lotus as a reference, as many as three post-γ grape subgenomic copies are equally evident, the syntenic regions of which show extensive collinearity of homologous genes (Figure 2). Among the 87.1% of the lotus genic regions retained from this duplication, 5,279 (33.3%) are singletons, 8,578 (54.1%) are duplicated, and 2,007 (12.6%) have more than three homeologs, implying there may have been additional paleo-duplications (Table S10 in Additional file 1).

Based on three lines of evidence, the lineage nucleotide substitution rate in lotus is about 30% slower than that of grape, widely used in angiosperm comparative genomics due to its basal phylogenetic position in rosids, slow mutation rate, and lack of reduplication. First, while phylogenetic evidence firmly dates the lotus-grape divergence before the pan-eudicot γ triplication affecting only grape, synonymous substitution rates (Ks) between genome-wide lotus-grape syntelog pairs (Figure S7 in Additional file 1) are smaller than those among triplicated grape genes. Second, the lotus lineage mutation rate also appears slower (about 29.26% slower) than that of *Vitis* based on a maximum-likelihood tree of 83 plastid genes [21] and expert dating of the respective speciation events [22] using the r8s program [23] with penalized likelihood. Third, the lotus genome has retained more ancestral loci following its lineage-specific WGD. Lotus is a basal eudicot, and its genome is the one from the most ancient lineage of angiosperm sequenced to date (Figure S1 in Additional file 1). Lotus represents an even better model than grape for inferences about the common ancestor of eudicots.

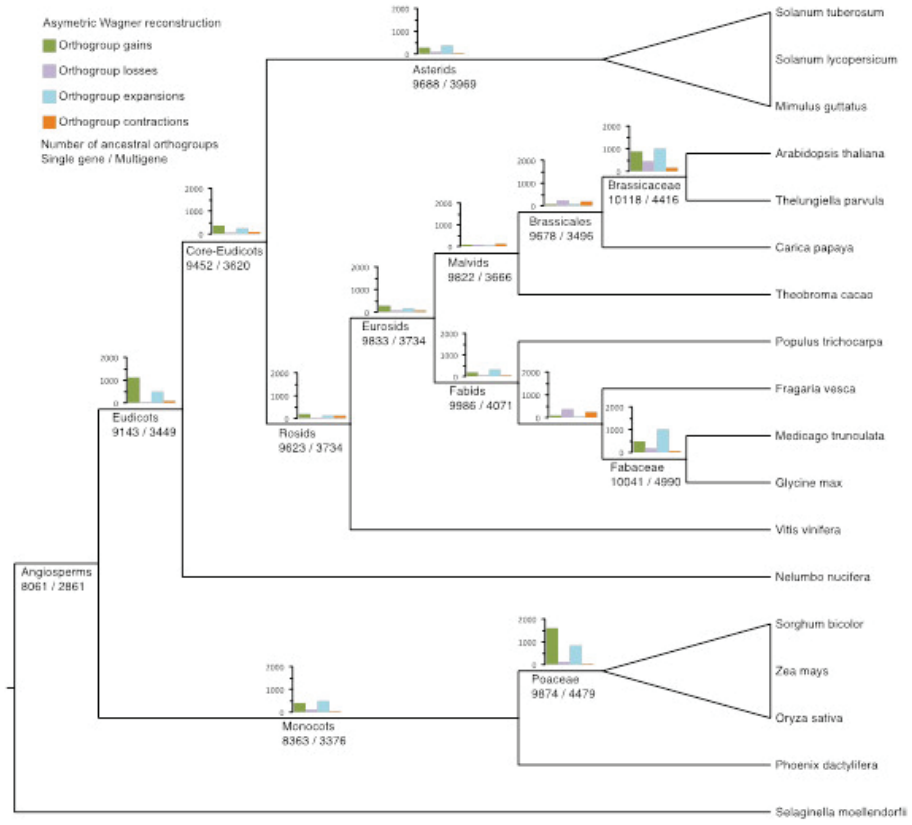

Figure 1 Orthogroup dynamics in lotus and other angiosperm genomes. Ancestral gene content and gene family (orthogroup) dynamics in lotus and other eudicot and monocot genomes identify expansion of the number of gene families and gene content associated with the ancestral eudicot.

The remarkably slow mutation rate in lotus complicates the dating of the λ duplication. λ-duplicated lotus genes have a median synonymous substitution rate (Ks) of 0.5428, corresponding to an age of 27 million years ago (MYA) on the basis of average rates in plants [24] or 54 MYA on the basis of the grape lineage rate (Figure S7 in Additional file 1). Because lotus diverged from its closest sister lineage approximately 135 to 125 MYA [21], before the γ triplication, this suggests that the mutation rate in lotus is much lower than that in grape, and that the lotus-specific WGD event occurred about 65 MYA with a range between 76 and 54 MYA. This date coincides with the Cretaceous-Tertiary mass extinction that led to the loss of roughly 60% of plant species [25]. Polyploidization has been associated with increased adaptation and survivability, and the numerous plant species inferred to have undergone polyploidy within this timeframe suggests a possible advantage to polyploid lineages during the Cretaceous-Paleogene transition, an interpretation supported by the λ duplication in lotus.

By tracing the phylogenetic histories of 688 pairs of grape genes in 528 orthogroups from each of the γ duplication blocks [26], we tested the timing of the γ paleohexaploid event that has been observed in the genomes of *Vitis* [7], papaya [6], *Populus* [20] and other core eudicots [14,17]. About 50% of the resolved trees support the

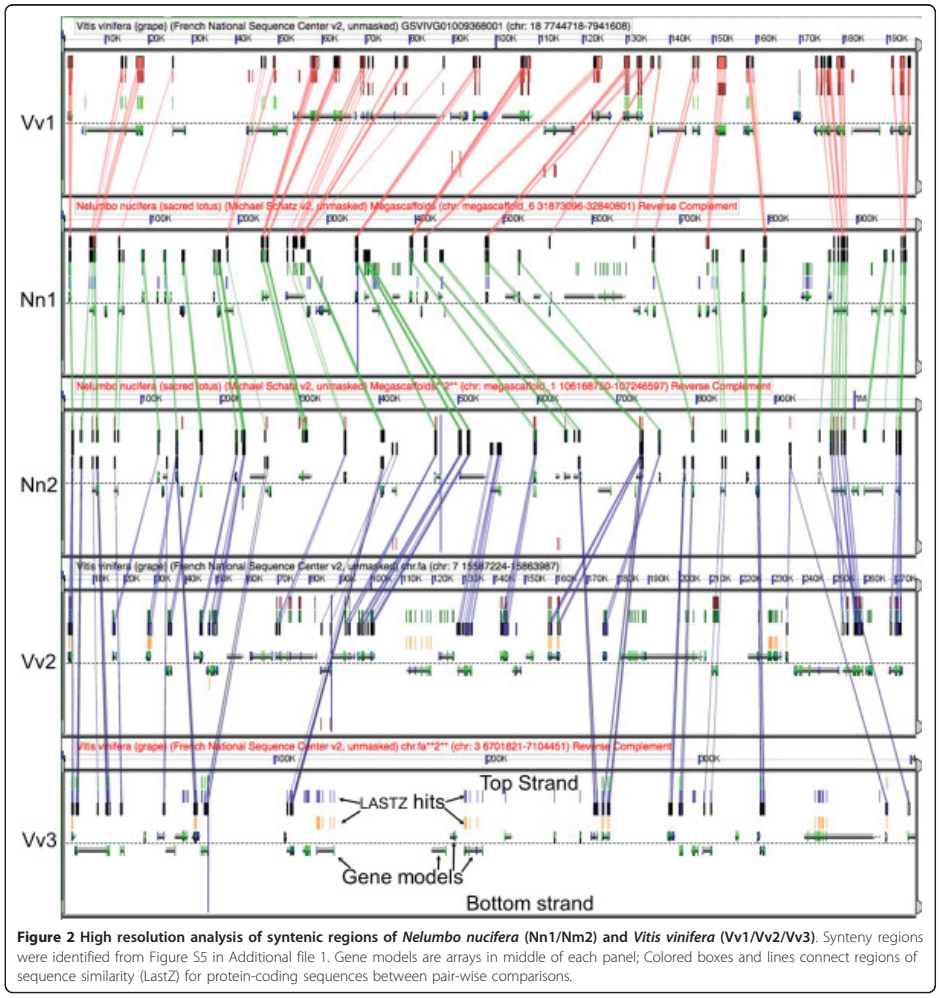

Figure 2 High resolution analysis of syntenic regions of *Nelumbo nucifera* (Nn1/Nm2) and *Vitis vinifera* (Vv1/Vv2/Vv3). Synteny regions were identified from Figure S5 in Additional file 1. Gene models are arrays in middle of each panel; Colored boxes and lines connect regions of sequence similarity (LastZ) for protein-coding sequences between pair-wise comparisons.

timing of the γ event to have occurred 'core-eudicot-wide' after the divergence of lotus, consistent with synteny analysis. By contrast, gene family phylogenies for about half of the γ block duplications include lotus genes (Table S11 in Additional file 1), although, in rare cases, duplicated monophyletic groups contain both lotus and eudicot-wide genes. This is consistent with an earlier phylogenomic analysis using data from numerous plant genomes and basal eudicot transcriptomes, suggesting that 18% to 28% of γ block duplications were eudicot-wide [26], even though the signal is primarily observed in core eudicots (Figure 3).

Such data suggest that a relatively large amount of genetic novelty is specifically associated with eudicots as a whole, even though the core eudicots shared a genome-triplication after divergence from the basal eudicots. By contrast, in monocots it appears that the evolution of the grass family specifically, rather than the earlier node

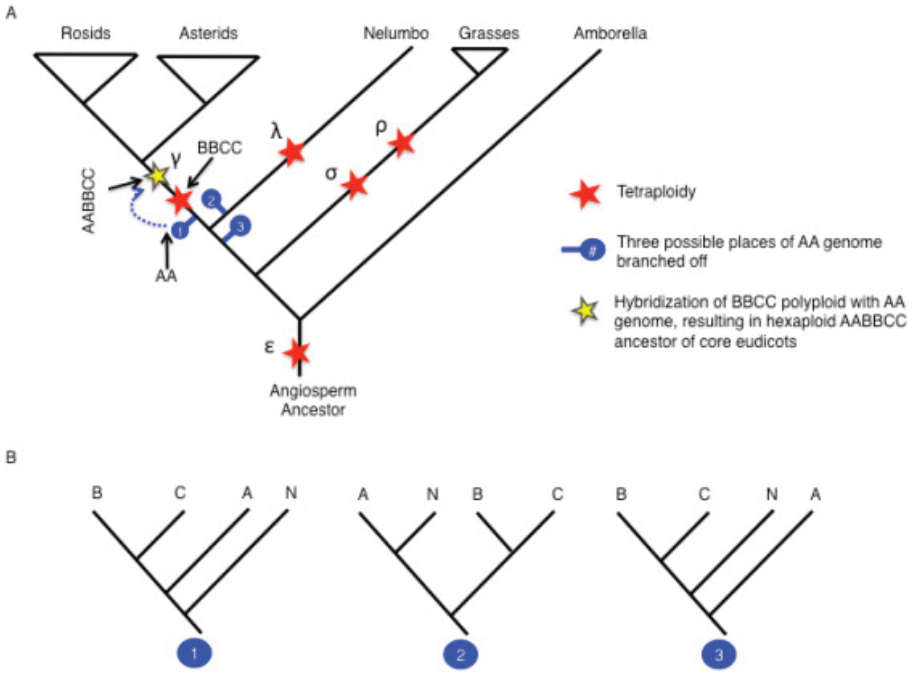

Figure 3 Polyploidy events in the history of angiosperm evolution. (A) Summary of polyploidy events in the history of angiosperm evolution, with a focus on the possible phylogenetic origins of the three subgenomes comprising the gamma paleohexaploidy event in core eudicots. Synteny analysis of the *Nelumbo* genome indicates that gamma is shared only within the core eudicots; however, phylogenomic analysis suggests a more complex history since around half of the gamma pairs were duplicated core-eudicot-wide and the other half eudicot-wide (See Table S10 in Additional file 1). AA, BB, and CC are three subgenomes of the ancestral hexaploidy. Three possible phylogenetic origins of the ancestral AA genome involved in gamma are denoted by 1, 2 and 3. Lamda is defined as the most recent polyploidy event in the evolutionary history of *Nelumbo*. All the other Greek symbols are well-known polyploidy events in the evolutionary history of angiosperms. Gamma: genome-triplication (hexaploid) event in core eudicot genomes [7,23]; Sigma and rho: genome duplications detected in grass genomes [8]; Epsilon: angiosperm-wide duplication detected in large-scale gene family phylogenies. Based on gene tree phylogenomics, we hypothesize that the triplication event involved a tetraploid event (BBCC red star) first, then subgenome AA combined with BBCC to form hexaploidy AABBCC (blue dashed line). **(B)** Predicted gene tree topologies of hypothetical origins of the AA subgenome of the gamma paleohexaploidy. A, B, C indicate surviving genes inherited from AA, BB, CC subgenomes of the AABBCC ancestral hexaploidy. N indicates genes of *Nelumbo*.

comprised of grasses (Poales) and palms (Arecales), was associated with relatively large gains in gene family number and size.

Adaptation to an aquatic environment

Submersed plant growth presents unique physiological challenges. Lotus has had to evolve novel features to cope with its aquatic lifestyle. Possible adaptations include an astonishing number of putative copper-dependent proteins, of which 63 proteins contain at least one COX2 domain, 55 contain a 'copper-binding-like' domain, and 4 contain polyphenol oxidases. The abundance of copper proteins in lotus compared to other plants is attributed to expansions in COG2132, a family of multi-copper oxidases. Most plant genomes encode one or two members of COG2132, whereas lotus has at least 16 members due to WGD and repeated tandem duplications (Figure 4, and see Figure S8 in Additional file 1). The only COG2132 members in Arabidopsis, LPR1 and LPR2, are involved in phosphate starvation signaling in root meristems. Similarly, in lotus, expression of COG2132 family members is confined largely to the roots (Figure 4). The lotus-specific expansion appears to form a separate phylogenetic clade from the LPR1 and 2-like proteins, suggesting a novel

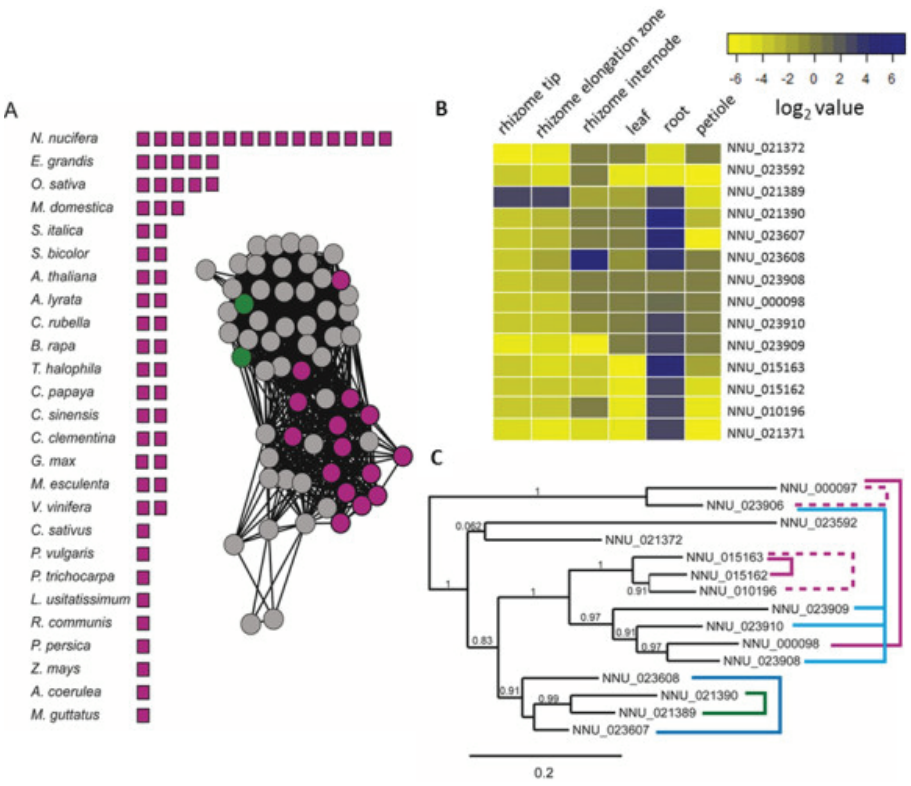

Figure 4 Lotus-specific expansion in LPR1/LPR2 proteins. (A) The number of LPR1/LPR2 homologs in land plants. Homologs detected by Basic Local Alignment Search Tool against the genomes of land plants are represented by a box. A protein similarity network of those proteins is also shown; lotus proteins are represented as purple nodes, Arabidopsis proteins (LPR1 and LPR2) are represented as green nodes and other land plant proteins are represented as grey nodes. **(B)** Heatmap of COG2132 gene family member expression in lotus. Reads per kilo base per million (RPKM) values were $\log_2$ transformed, where blue correlates to high expression, and yellow to low expression. **(C)** A maximum-likelihood tree of LPR1/LPR2-like lotus proteins. Branch support was calculated using an Approximate Likelihood-Ratio Test. Lotus homologs are connected with a dashed bracket, whereas proteins whose genes are found in tandem on the genome are connected with a solid bracket. A detailed phylogeny of COG2132 members can be found in Figure S8 in Additional file 1.

function not found in Arabidopsis (Figure 4, and see Figure S8 in Additional file 1).

Adaptation to phosphate starvation in lotus is also evidenced by expansion of the UBC24 family and the miR399 family that regulates it (Table S12 in Additional file 1). The miR169 family, implicated in adaptation to drought stress in Arabidopsis [27], also shows expansion in lotus, totaling 22 members. The fact that lotus grows aquatically and may rarely be subjected to drought suggests that the miR169 family is involved in other physiological processes.

Several other gene families also show unusual compositions that may reflect adaptation to aquatic lifestyles. The basic helix loop helix (bHLH) family, implicated in light responses including germination, control of flowering and de-etiolation, and root and flower development, lacks three of its 20 subfamilies in lotus: Va, implicated in brassinosteroid signaling; VIIIc2, implicated in root hair development; and XIII, implicated in root meristem development [28]. The largest families of bHLH factors in lotus are XII, involved in developmental processes including control of petal size, brassinosteroid signaling

and floral initiation, and Ia, implicated in stomatal development and patterning.

The PRR1/TOC1 circadian clock family, which coordinates internal biology with daily light/dark cycles and is highly conserved across many plant species, includes three predicted members in lotus compared to the one or two present in other plant genomes. The fact that PRR proteins have key roles in modulating light and temperature input into the circadian clock suggests that lotus may require more sensitive adjustments to its environment than other plants. Consistent with this, the cryptochrome (CRY) family of blue light photoreceptors is also increased with five (two CRY1, two CRY2, one CRY3) compared to three in Arabidopsis and four in poplar (Additional file 1, Table S13). Similar expansion in the CRY family was also noted in another aquatic organism, *Ostreococcus*, a micro green algae. Lotus is adapted to both temperate and tropical climates and day lengths with a wide range of flowering times, perhaps associated with increased numbers of flowering time and circadian clock-associated genes.

Discussion

Paleopolyploids are widespread among eukaryotes and particularly common in angiosperms [14,15]. Lotus diverged from other eudicots early in eudicot history, prior to the γ genome-triplication characteristic of most members of the group [14,15,17,26], and provides insight into the timing and nature of this event associated with a rapid radiation of the large eudicot lineages. When plant genomes of high paleopolyploidy levels are compared, differentiated gene loss (fractionation) among several homologous subgenomes tends to diminish the signals of synteny. In such cases, genomes with few paleopolyploidy events (such as those of grape or papaya) can be used to take advantage of the smaller evolutionary distances between orthologous segments. Extensive collinearity within itself, as well as with other plant genomes such as those of Arabidopsis, grape, rice and sorghum, makes the lotus genome not only a eudicot evo-genomic reference (Figure S9 in Additional file 1), but also a better resource for reconstructing the pan-eudicot genome and facilitating comparative analysis between eudicots and monocots.

Surprisingly, the phylogenomic analysis of gene families associated with the γ include a substantial fraction of eudicot-wide duplications, suggesting the possibility of a two-step model that involved genetic material from a lineage that branched off earlier than the core eudicots (Figure 3A). A substantial fraction of eudicot-wide gene duplications was also observed in phylogenomic analyses that contained large collections of transcriptome data from early branching basal eudicots such as *Platanus*, *Aquilegia* and poppies [26]. Eudicot-wide duplications were detected only rarely in another phylogenomic analysis that introduced transcriptome data from the basal eudicots *Gunnera* and *Pachysandra* [29]. The 34 unigenes available from that study were used to populate five MADS box orthogroups with larger taxon sampling in this study. Phylogenies of these orthogroups identify (at boostrap >50%) one eudicot-wide and three core-eudicot-wide duplications (Table S11 in Additional file 1), consistent with the rest of the findings in the present study.

In contrast to the phylogenomic results, syntenic comparison showed one lotus region matched with up to three *Vitis* homologous regions, indicating that the lotus genome did not share the γ event. We propose that the γ event occurred after the separation of the lotus lineage (Proteales), and involved hybridization with a now extinct species that branched off around the same time (Figure 3A, AA at position #2), or even earlier than lotus (Figure 3A, AA at position #3). This model explains why the phylogenomic analyses could identify some γ duplications occurring before the divergence of lotus, but not observable as a triplication in the lotus genome structure. A similar two-step model was suggested by Lyons *et al.* [30] on the basis of fractionation patterns seen in *Vitis*, and evidence for a two-step hexaploid process is clearly observed in the much more recent paleohexaploid *Brassica rapa* [31]. Additional whole plant genome sequences from lineages close to the γ event, especially ones without the confounding effects of lineage-specific genome duplications, may also help to clarify genome-wide patterns of fractionation among the three γ subgenomes, which could provide further evidence bearing on the timing and event(s) associated with the γ paleohexaploidy event that is associated with what is arguably one of the most important radiations in angiosperm history.

The higher homeolog retention rate in lotus compared with most other genomes studied provided an opportunity to study subfunctionalization [32], a major driving force affecting fates of duplicated genes following paleopolyploidy. Most pairs of lotus homeologs have no difference in PFAM domain families, whereas 453 pairs (11.6%) differ by up to five domains. The unshared domains have mean length 17 amino acids with a range of 0 to 890 amino acids. Between homeologous lotus gene pairs, mRNA length (excluding 5′ and 3′ untranslated regions), coding sequence length, and intron length differences all follow geometric-like distributions (Figure S10 in Additional file 1), consistent with independent accumulation of small insertions and deletions. The changes of length in exonic and intronic regions seem uncorrelated, implying that subfunctionalization affects gene regulation at multiple transcriptional and post-transcriptional levels.

When divergence of lineages is followed by WGD, one predicts similar divergence of the paralogs in one species' genome from a shared ortholog in the other species, confirmed in previous studies [16,33]. Comparison of paired λ paralogs and their grape ortholog generally fit this prediction (Figure S11 in Additional file 1); however, comparisons to cereal (sorghum) orthologs show consistent differentiation in branch lengths. This discrepancy in the lotus-cereal comparison could be explained by fast evolutionary rates in cereal genomes and/or λ being older than it appears, due to the slow *Nelumbo* evolutionary rate. Alternatively, this is also consistent with structural compartmentalization, with genes within the same genome undergoing different evolutionary trajectories [33]. Wider taxa sampling at neighboring branches will help better distinguish the possibilities.

The extraordinary seed longevity and vegetative propagation via rhizomes are likely the causes of the slow evolutionary rate in lotus. The 'China Antique' has a highly homozygous genome, yielding arguably the best assembled genome using next-generation sequencing technologies with pseudo-molecules proportional to its karyotype. The lotus genome provides the foundation for revealing the molecular basis of its many distinguishing biological properties, including seed longevity, adaptation to aquatic environment, the distinctive superhydrophobicity and self-cleaning property of its leaves, and the thermogenesis that is thought to enhance its pollination success.

Sacred lotus is the first true aquatic plant to be sequenced and comparative genomics reveal unique gene family expansions that may have contributed to its adaptations to an aquatic environment. Submersed soils are largely hypoxic and have a decreased reduction-oxidation potential, causing heavy metal precipitation and reduced nutrient availability. Lotus has a dramatic expansion of the COG2132 family, a group of multi-copper oxidases involved in phosphate starvation in root meristems. A role in root-specific processes is supported by the expression of these unique genes in root tissue. Adaptation to phosphate starvation can also be seen in an expansion of the UBC24 family and the miR399 family that regulates it. Lotus lacks four bHLH subfamilies involved in iron uptake and root hair and root meristem development, suggesting novel root growth and iron regulation. These gene family expansions and preferential retention of duplicated genes reflect the challenges of aquatic growth.

Conclusions

Sacred lotus has many unique biological features, most noticeable seed longevity and the lotus effect, in addition to its agricultural and medicinal importance. The purpose of sequencing the lotus genome is to facilitate research in these areas and on agronomic and horticultural traits such as rhizome development and flowering time. The assembly of the lotus genome is surprisingly high quality, largely due to the high level of homozygosity resulting from domestication and vegetative propagation. The lotus genome has a lineage-specific WGD event that occurred about 65 MYA, but shows no structural evidence for the γ hexaploid event shared among core eudicot species. The lotus genome has a 30% slower nucleotide mutation rate than that of grape, contributing in part to the outstanding genome assembly using next-generation sequencing technologies. Analysis of sequenced plant genomes yielded a minimum gene set for vascular plants of 4,223 genes. Strikingly, lotus has 16 COG2132 multi-copper oxidase family proteins with root-specific expression. COG2132 members are involved in root meristem phosphate starvation, reflecting lotus' adaptation to limited nutrient availability in an aquatic environment. The slow nucleotide substitution rate and the lack of the triplication event make lotus genome an excellent reference for reconstructing the pan-eudicot genome and for accelerating comparative analysis between eudicots and monocots. The lotus genome will accelerate the identification of genes controlling rhizome yield and quality, seed size and nutritional profile, flower morphology, and flowering time for crop improvement.

Materials and methods

Illumina (Illumina HiSeq 2000) libraries were generated from purified *N. nucifera* 'China Antique' nuclear DNA with inserts of 180 bp, 500 bp, 3.8 kb and 8 kb and assembled using ALLPATHS-LG. 454/Roche (GSFLX pyrosequencing platform) 20 kb mate pair reads were used for scaffolding. RNAseq data generated from various lotus tissues were used for annotation and RNAseq differential gene expression analysis using CLC Genomics Workbench 5.0 (CLC Bio, Aarhus, Denmark). MAKER version 2.22 was used in combination with the assembled RNAseq data to annotate 26,685 genes in the lotus genome. Detailed methods for genome assembly, annotation and analyses are provided in Additional file 1.

Data access

The assembled *N. nucifera* genome was submitted to GenBank (AQOG00000000; PID PRJNA168000, http://www.ncbi.nlm.nih.gov/Traces/wgs/?val=AQOG01). Whole genome shotgun raw reads are deposited under SRA study: SRP021228 (http://trace.ncbi.nlm.nih.gov/Traces/sra/?study=SRP021228). The raw RNAseq data are deposited under BioProject 196884 (http://www.ncbi.nlm.nih.gov/bioproject/196884).

Additional material

Additional file 1: Supplementary data, including detailed materials and methods, and supplementary tables S1-S13, and figures S1-S14.

Abbreviations
bHLH: basic helix loop helix; bp: base pair; CRY: cryptochrome; EST: expressed sequence tags; MYA: million years ago; PPR: pentatricopeptide repeat-containing proteins; WGD: whole genome duplication.

Authors' contributions
RM, RV, YL, MY, YH and S L designed research; RM, RV, YL, MY, YH, LTL, QZ, JEB, HT, EL, AAF, GN, DRN, CEBH, ARG, YJ, JPD, FZ, JH, XM, KAH, KI, SAR, MEH, QY, TCM, AC, YZ, RS, RJ, NC, JA, CMW, EW, AS, YH, LX, JZ, RP, MJH, WX, JAW, JW, MLW, YJZ, REP, ABB, CD, SRD, MAS, TPM, SPL, DRO, JWS, DRG, NJ, MY, CWD, SSM, AHP, BBB, SL and JSM performed research and analyzed data; RM, RV, JL, AHP, CEBH, JRW, KI, SAR, CWD, SSM and BBB wrote the paper. All authors read and approved the final manuscript.

Acknowledgements
We thank K. Hasenstein for collection of the fruits of *Nelumbo lutea*. This project was supported by the University of California, Los Angeles (JSM); Wuhan Botanical Garden, Chinese Academy of Sciences, P.R. China (SL); and the University of Illinois at Urbana-Champaign (RM).

Authors' details
[1]Key Laboratory of Plant Germplasm Enhancement and Specialty Agriculture, Wuhan Botanical Garden, The Chinese Academy of Sciences, Lumo Road, Wuhan 430074, China. [2]Department of Plant Biology, University of Illinois at Urbana-Champaign, 1201 West Gregory Drive, Urbana, IL 61801, USA. [3]College of Horticulture, Nanjing Agricultural University, 1 Weigang Road, Nanjing 210095, China. [4]Institute of Biological Chemistry, Washington State University, Clark Hall, 100 Dairy Road, Pullman, WA 99164, USA. [5]Simons Center for Quantitative Biology, Cold Spring Harbor Laboratory, One Bungtown Road, Cold Spring Harbor, NY 11724, USA. [6]Eccles Institute of Human Genetics, University of Utah, 15 North 2030 East, Salt Lake City, UT 84112, USA. [7]Plant Genome Mapping Laboratory, University of Georgia, 111 Riverbend Road, Athens, GA 30602, USA. [8]Department of Crop and Soil Sciences, University of Georgia, 120 Carlton Street, Athens, GA 30602, USA. [9]J Craig Venter Institute, 9704 Medical Center Drive, 20850 Rockville, MD, USA. [10]School of Plant Sciences, iPlant Collaborative Bio5 Institute, University of Arizona, 1657 East Helen Street, Tucson, AZ 85745, USA. [11]Department of Horticulture, Michigan State University, A288 Plant and Soil Sciences Building, 1066 Bogue Street, East Lansing, MI 48824, USA. [12]Department of Microbiology, Immunology and Biochemistry, University of Tennessee Health Science Center, 858 Madison Avenue Suite G01, Memphis, TN 38163, USA. [13]Department of Chemistry and Biochemistry and Institute for Genomics and Proteomics, University of California, Los Angeles, 607 Charles E Young Drive East, CA 90095, USA. [14]Department of Biology and Intercollege Graduate Program in Plant Biology, The Pennsylvania State University, 201 Life Sciences Building, University Park, PA 16802, USA. [15]Center for Applied Chemical Biology, Department of Biological Sciences, Youngstown State University, 1 University Plaza, Youngstown, OH, 44555, USA. [16]USDA-ARS, Purdue University, 915 West State Street, West Lafayette, IN 47907, USA. [17]Texas A&M AgriLife Research, Department of Plant Pathology & Microbiology, Texas A&M University System, 17360 Coit Road, Dallas, TX 75252, USA. [18]Department of Biology, University of Central Oklahoma, 100 North University Drive, Edmond, OK 73034, USA. [19]School of Earth and Environmental Sciences, University of Adelaide, North Terrace, Adelaide, 5005, Australia. [20]Cryobiofrontier Research Center, Faculty of Agriculture, Iwate University, Ueda 3-18-8, Morioka, Iwate 020-8550, Japan. [21]Institute for Conservation Biology, The University of Wollongong, Northfields Avenue, Wollongong, NSW 2522, Australia. [22]Department of Crop Sciences, University of Illinois at Urbana-Champaign, 1101 West Peabody Drive, Urbana, IL 61801, USA. [23]Donald Danforth Plant Science Center, 975 North Warson Road, St Louis, MO 63132, USA. [24]Lawrence Berkeley National Laboratory, 1 Cyclotron Road Berkeley, Emeryville, CA 94720, USA. [25]Institute of Developmental Biology and Molecular Medicine & School of Life Sciences, Fudan University, 220 Handan Road, Shanghai, 200433, China. [26]Department of Biochemistry and Molecular Biology, 246 Noble Research Center, Oklahoma State University, Stillwater, OK 74078, USA. [27]Hawaii Agriculture Research Center, 94-340 Kunia Road, Waipahu, HI 96797, USA. [28]Department of Tropical Plant and Soil Sciences, University of Hawaii at Manoa, 3190 Maile Way, Honolulu, HI 96822, USA. [29]Fujian Normal University, Qishan Campus, Minhou, Fuzhou, 350117, China. [30]Department of Biology and Molecular Biology, Montclair State University, 1 Normal Avenue, Montclair, NJ 07043, USA. [31]Institute of Tropical Biosciences and Biotechnology, China Academy of Tropical Agricultural Sciences, 4 Xueyuan Road, Haikou, Hainan 571101, China. [32]Department of Plant and Microbial Biology, University of California, 1 Shields Avenue, Davis CA, 95161, USA. [33]Department of Cell and Developmental Biology, University of Illinois, 1201 West Gregory Drive, Urbana IL, 61801, USA. [34]The Genome Analysis Center, Monsanto, St Louis, MO 63167, USA. [35]Global Change and Photosynthesis Research Unit, Agricultural Research Service, United States Department of Agriculture, 1206 West Gregory Drive, Urbana, IL, USA. [36]IGPP Center for the Study of Evolution and Origin of Life, Geology Building, Room 5676, University of California, Los Angeles, 595 Charles E Young Drive East, Los Angeles, CA 90095-1567, USA. [37]Department of Plant and Microbial Biology, University of California, 411 Koshland Hall, Berkeley, CA 94720, USA.

Received: 4 January 2013 Revised: 19 April 2013
Accepted: 10 May 2013 Published: 10 May 2013

References

1. Shen-Miller J: **Sacred lotus, the long-living fruits of China Antique.** *Seed Sci Res* 2002, **12**:131-143.
2. Shen-Miller J, Schopf JW, Harbottle G, Cao RJ, Ouyang S, Zhou KS, Southon JR, Liu GH: **Long-living lotus: germination and soil g-irradiation of centuries-old fruits, and cultivation, growth, and phenotypic abnormalities of offspring.** *Am J Bot* 2002, **89**:236-247.
3. Duke JA, Bogenschutz-Godwin MJ, duCellier J, Duke AK: *Handbook of Medicinal Herbs* 2002, Boca Raton: CRC Press.
4. Diao Y, Chen L, Yang G, Zhou M, Song Y, Hu Z, Lin JY: **Nuclear DNA C-values in 12 species in Nymphaeales.** *Caryologia* 2006, **59**:25-30.
5. Yang M, Han Y, VanBuren R, Ming R, Xu L, Han Y, Liu Y: **Genetic linkage maps for Asian and American lotus constructed using novel SSR markers derived from the genome of sequenced cultivar.** *BMC Genomics* 2012, **13**:653.
6. Ming R, Hou S, Feng Y, Yu Q, Dionne-Laporte A, Saw JH, Senin P, Wang W, Ly BV, Lewis KL, Salzberg SL, Feng L, Jones MR, Skelton RL, Murray JE, Chen C, Qian W, Shen J, Du P, Eustice M, Tong E, Tang H, Lyons E, Paull RE, Michael TP, Wall K, Rice DW, Albert H, Wang ML, Zhu YJ, *et al*: **The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus).** *Nature* 2008, **452**:991-996.
7. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, Vezzi A, Legeai F, Hugueney P, Dasilva C, Horner D, Mica E, Jublot D, Poulain J, Bruyere C, Billault A, Segurens B, Gouyvenoux M, Ugarte E, Cattonaro F, Anthouard V, Vico V, Del Fabbro C, Alaux M, Di Gaspero G, Dumas V, *et al*: **The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla.** *Nature* 2007, **449**:463-467.
8. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, Schmutz J, Spannagl M, Tang H, Wang X, Wicker T, Bharti AK, Chapman J, Feltus FA, Gowik U, Grigoriev IV, Lyons E, Maher CA, Martis M, Narechania A, Otillar RP, Penning BW, Salamov AA, Wang Y, Zhang L, Carpita NC, *et al*: **The Sorghum bicolor genome and the diversification of grasses.** *Nature* 2009, **457**:551-556.
9. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, *et al*: **The B73 maize genome: complexity, diversity, and dynamics.** *Science* 2009, **326**:1112-1115.
10. Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, Xu D, Hellsten U, May GD, Yu Y, Sakurai T, Umezawa T, Bhattacharyya MK, Sandhu D, Valliyodan B, Lindquist E, Peto M, Grant D, Shu S, Goodstein D, Barry K, Futrell-Griggs M, Abernathy B, Du J, Tian Z, Zhu L, *et al*: **Genome sequence of the palaeopolyploid soybean.** *Nature* 2010, **463**:178-183.
11. Jiang N, Bao Z, Zhang X, Eddy S-R, Wessler S-R: **Pack-MULE transposable elements mediate gene evolution in plants.** *Nature* 2007, **431**:569-573.
12. Parra G, Bradnam K, Korf I: **CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes.** *Bioinformatics* 2007, **23**:1061-1067.
13. Lechner M, Findeiss S, Steiner L, Marz M, Stadler PF, Prohaska SJ: **Proteinortho: Detection of (Co-)orthologs in large-scale analysis.** *BMC Bioinformatics* 2011, **12**:124.

14. Paterson AH, Freeling M, Tang H, Wang X: **Insights from the comparison of plant genome sequences.** *Annu Rev Plant Biol* 2010, **61**:349-372.
15. Soltis DE, Albert VA, Leebens-Mack J, Bell CD, Paterson AH, Zheng C, Sankoff D, dePamphilis CW, Wall PK, Soltis PS: **Polyploidy and angiosperm diversification.** *Am J Bot* 2009, **96**:336-348.
16. Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH: **Synteny and collinearity in plant genomes.** *Science* 2008, **320**:486-488.
17. Tang H, Wang X, Bowers JE, Ming R, Alam M, Paterson AH: **Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps.** *Genome Res* 2008, **18**:1944-1954.
18. The Arabidopsis Genome Initiative: **Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*.** *Nature* 2000, **408**:796-815.
19. International Rice Genome Sequencing Project: **The map-based sequence of the rice genome.** *Nature* 2005, **436**:793-800.
20. Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, Schein J, Sterck L, Aerts A, Bhalerao RR, Bhalerao RP, Blaudez D, Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson J, Chalot M, Chapman J, Chen GL, Cooper D, Coutinho PM, Couturier J, Covert S, Cronk Q, *et al*: **The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray).** *Science* 2006, **313**:1596-1604.
21. Moore M, Soltis PS, Bell CD, Burleigh JG, Soltis DE: **Phylogenetic analysis of 83 plastid genes further resolves the early diversification of eudicots.** *Proc Natl Acad Sci USA* 2010, **107**:4623-4628.
22. Hedges SB, Dudley J, Kumar S: **TimeTree: a public knowledge-base of divergence times among organisms.** *Bioinformatics* 2006, **22**:2971-2972.
23. Sanderson MJ: **r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock.** *Bioinformatics* 2003, **19**:301-302.
24. Wolfe KH, Sharp PM, Li WH: **Rates of synonymous substitution in plant nuclear genes.** *J Mol Evol* 1989, **29**:208-211.
25. Fawcett JA, Maere S, van de Peer Y: **Plants with double genomes might have had a better chance to survive the Cretaceous-Tertiary extinction event.** *Proc Natl Acad Sci USA* 2009, **106**:5737-5742.
26. Jiao Y, Leebens-Mack J, Ayyampalayam S, Bowers JE, McKain MR, McNeal J, Rolf M, Ruzicka DR, Wafula E, Wickett NJ, Wu X, Zhang Y, Wang J, Zhang Y, Carpenter EJ, Deyholos MK, Kutchan TM, Chanderbali AS, Soltis PS, Stevenson DW, McCombie R, Pires JC, Wong GK, Soltis DE, dePamphilis CW: **A genome triplication associated with early diversification of the core eudicots.** *Genome Biol* 2012, **13**:R3.
27. Li W-X, Oono Y, Zhu J, He XJ, Wu JM, Iida K, Lu XY, Cui X, Jin H, Zhu JK: **The Arabidopsis NFYA5 transcription factor is regulated transcriptionally and post transcriptionally to promote drought resistance.** *Plant Cell* 2008, **20**:2238-2251.
28. Pires N, Dolan L: **Origin and diversification of basic-helix-loop-helix proteins in plants.** *Mol Biol Evol* 2010, **27**:862-874.
29. Vekemans D, Proost S, Vanneste K, Coenen H, Viaene T, Ruelens P, Maere S, van de Peer Y, Geuten K: **Gamma paleohexaploidy in the stem lineage of core eudicots: significance for MADS-box gene and species diversification.** *Mol Biol Evol* 2012, **29**:3793-3806.
30. Lyons E, Pedersen B, Kane J, Freeling M: **The value of nonmodel genomes and an example using SynMap within CoGe to dissect the hexaploidy that predates the rosids.** *Tropical Plant Biol* 2008, , **1**: 181-190.
31. Tang H, Woodhouse MR, Cheng F, Schnable JC, Pedersen BS, Conant G, Wang X, Freeling M, Pires JC: **Altered patterns of fractionation and exon deletions in *Brassica rapa* support a two-step model of paleohexaploidy.** *Genetics* 2012, **90**:1563-1574.
32. Lynch M, Force AG: **The origin of interspecific genomic incompatibility via gene duplication.** *Am Nat* 2000, **156**:590-605.
33. Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, Bai Y, Mun JH, Bancroft I, Cheng F, Huang S, Li X, Hua W, Freeling M, Pires JC, Paterson AH, Chalhoub B, Wang B, Hayward A, Sharpe AG, Park BS, Weisshaar B, Liu B, Li B, Tong C, Song C, Duran C, Peng C, Geng C, Koh C, *et al*: **The genome of the mesopolyploid crop *species Brassica rapa*.** *Nat Genet* 2011, **43**:1035-1039.

doi:10.1186/gb-2013-14-5-r41
Cite this article as: Ming *et al.*: Genome of the long-living sacred lotus (*Nelumbo nucifera* Gaertn.). *Genome Biology* 2013 **14**:R41.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central

CHAPTER 4

AUTOMATED UPDATE, REVISION, AND QUALITY CONTROL OF THE MAIZE GENOME ANNOTATIONS USING MAKER-P IMPROVES THE B73 REFGEN_V3 GENE MODELS AND IDENTIFIES NEW GENES

The following is a reprint of an article coauthored by myself, MeiYee Law, Kevin L. Childs, Joshua C. Stein, Andrew J. Olson,Carson Holt, Nicholas Panchy, Jikai Lei, Dian Jiao, Carson M. Andorf, Carolyn J. Lawrence, Doreen Ware, Shin-Han Shiu, Yanni Sun, Ning Jiang, and Mark Yandell. This article was originaly published in *Plant Physiology* 2014 and is used with permission.

Personal Contribution

I helped develop the benchmarking and update strategy for the maize reference genome. I wrote the code for generating summary statistics for the benchmarks. I integrated snoscan and tRNA-scan into MAKER and used these tools to annotate tRNAs and snoRNAs in the new maize assembly. I wrote the ncRNA sections of the manuscript. I critically read and edited the manuscript.

Breakthrough Technologies

Automated Update, Revision, and Quality Control of the Maize Genome Annotations Using MAKER-P Improves the B73 RefGen_v3 Gene Models and Identifies New Genes[1][OPEN]

MeiYee Law, Kevin L. Childs, Michael S. Campbell, Joshua C. Stein, Andrew J. Olson, Carson Holt, Nicholas Panchy, Jikai Lei, Dian Jiao, Carson M. Andorf, Carolyn J. Lawrence, Doreen Ware, Shin-Han Shiu, Yanni Sun, Ning Jiang, and Mark Yandell*

The Jackson Laboratory, Bar Harbor, Maine 04609 (M.L.); Eccles Institute of Human Genetics (M.L., M.S.C., M.Y.), Department of Biomedical Informatics (M.L.), and USTAR Center for Genetic Discovery (C.H., M.Y.), University of Utah, Salt Lake City, Utah 84112; Genetics Program (N.P., S.-H.S., N.J.), Department of Plant Biology (K.L.C., S.-H.S.), Department of Computer Science and Engineering (J.L., Y.S.), and Department of Horticulture (N.J.), Michigan State University, East Lansing, Michigan 48824; iPlant Collaborative, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724 (J.C.S., A.J.O., D.W.); Ontario Institute for Cancer Research, Toronto, Ontario, Canada M5G 1L7 (C.H.); Texas Advanced Computing Center, University of Texas, Austin, Texas 78758 (D.J.); Department of Genetics, Development, and Cell Biology and Department of Agronomy (C.J.L.), and United States Department of Agriculture-Agricultural Research Service Corn Insects and Crop Genetics Research (C.M.A.), Iowa State University, Ames, Iowa 50011; and United States Department of Agriculture-Agricultural Research Service Northeast Area, Robert W. Holley Center for Agriculture and Health, Ithaca, New York 14853 (D.W.)

The large size and relative complexity of many plant genomes make creation, quality control, and dissemination of high-quality gene structure annotations challenging. In response, we have developed MAKER-P, a fast and easy-to-use genome annotation engine for plants. Here, we report the use of MAKER-P to update and revise the maize (*Zea mays*) B73 RefGen_v3 annotation build (5b+) in less than 3 h using the iPlant Cyberinfrastructure. MAKER-P identified and annotated 4,466 additional, well-supported protein-coding genes not present in the 5b+ annotation build, added additional untranslated regions to 1,393 5b+ gene models, identified 2,647 5b+ gene models that lack any supporting evidence (despite the use of large and diverse evidence data sets), identified 104,215 pseudogene fragments, and created an additional 2,522 noncoding gene annotations. We also describe a method for de novo training of MAKER-P for the annotation of newly sequenced grass genomes. Collectively, these results lead to the 6a maize genome annotation and demonstrate the utility of MAKER-P for rapid annotation, management, and quality control of grasses and other difficult-to-annotate plant genomes.

Plant genomes, especially grass genomes, are difficult substrates for genome annotation due to regional and whole-genome duplication events and often contain large numbers of pseudogenes. These factors impact every aspect of gene structure annotation, from revision of existing annotations in light of new data to annotation of newly sequenced plant genomes. These aspects of plant genomes also dramatically lengthen compute times, because the many repeated genes and other sequences result in commensurately more sequence alignments and gene predictions. In many ways, annotation of the maize genome epitomizes these problems.

In 2005, the National Science Foundation, U.S. Department of Agriculture, and Department of Energy announced that the approximately 2.3-Gb genome of the maize (*Zea mays*) inbred line B73, a major contributor to much of the germplasm used for U.S. grain production, would be sequenced using a bacterial artificial chromosome (BAC)-by-BAC approach. The plan was to sequence BACs from a minimal tiling path to approximately 6× coverage and to further improve only the unique genic regions. These sequences would be labeled Phase 1 HTGS_IMPROVED at GenBank, and the GenBank record for each BAC was to include information on the improved regions as well as order and orientation, where available, as comments. The Maize Genome Sequencing Consortium planned to release all data via

[1] This work was supported by the National Science Foundation (grant no. IOS–1126998 to S.-H.S., Y.S., N.J., K.L.C., and M.Y. and grant no. MCB–1119778 to S.-H.S.), the U.S. Department of Agriculture-Agricultural Research Service, and Iowa State University (MaizeGDB and contributions by C.M.A. and C.J.L.).

* Address correspondence to myandell@genetics.utah.edu.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Mark Yandell (myandell@genetics.utah.edu).

[OPEN] Articles can be viewed without a subscription.

www.plantphysiol.org/cgi/doi/10.1104/pp.114.245027

 © 2014 American Society of Plant Biologists. All Rights Reserved.

MaizeSequence.org, a project database, with a plan to transition all data into MaizeGDB (Sen et al., 2009) and Gramene (Monaco et al., 2014), a comparative resource for plant genomics (Youens-Clark et al., 2011), at project close.

Not only did the Maize Genome Sequencing Consortium produce these sequences, they created reference assemblies for each chromosome (the first assembly was named B73 RefGen_v1) as well as structural and functional annotations to genes (Liang et al., 2009; Schnable et al., 2009). The published B73 reference genome (RefGen_v1) available from GenBank consisted of 2,048 Mb in 125,325 sequence contigs (N50 of 40 kb), forming 61,161 scaffolds (N50 of 76 kb) anchored to a high-resolution genetic map (Wei et al., 2009). After predicting transposable elements (TEs), a combination of evidence-based, ab initio approaches and stringent TE filtering resulted in a set of 32,540 high-confidence, predicted protein-encoding genes (the Filtered Gene Set). Due to incomplete sampling of the genome, the B73 reference genome is estimated to be missing approximately 5% to 10% of genes that are physically present in the B73 genome.

Following the release of the first draft, B73 RefGen_v2 improved v1 by the addition of fosmid reads as well as by integrating genetic and optical map information. For B73 RefGen_v2, approximately 80% of the maize genome is ordered and oriented, and optical map and genetic map comparisons suggest that only 2% to 2.5% of the sequences are likely to be misplaced in the assembly (Fusheng Wei, Jeff Glaubitz, and Mike McMullen, personal communication). The set of gene predictions for RefGen_v2 included 110,028 transcript models in the Working Gene Set (5a) with a subset of 39,656 high-confidence structures identified as the Filtered Gene Set (5b). (Note that here we use the naming conventions imposed by the MaizeSequence.org data generators, although alternative naming conventions have been used in some cases for these data sets; e.g. at Phytozome [http://www.phytozome.net/maize.php], the Working Gene Set is called the unfiltered working set.)

In the last year of the project, Roche/454 whole-genome shotgun (WGS) reads were made available to improve the coverage of the gene space not included in the BAC minimal tiling path (and thereby identifying some of the estimated 5%–10% of genes that were missed). Improvements for B73 RefGen_v3 included refinements to contig placement supported by recent improvements to the IBM genetic map and inclusion of 1,844 gene space contigs. These 1,844 contigs were produced from a WGS sequencing library to fill in missing gene space both within and between original BAC sequences. In addition, approximately 65,000 full-length complementary DNAs (cDNAs) were aligned to the RefGen_v2 assembly and the new WGS contigs. The new 5b+ annotation build included 251 new gene models and 213 improved models. The number of protein-coding genes (including all nuclear chromosomes, mitochondrial DNA, chloroplast DNA, and unknown chromosome) actually decreased to 39,475 models due to merging and additional quality control. The annotation consists of 137,208 gene transcripts and 316 short noncoding genes. The maize B73 assemblies and various annotations are represented at Gramene, MaizeGDB, EnsemblPlants, and GenBank.

MaizeGDB, the Maize Genetics and Genomics Database (http://www.maizegdb.org), is the U.S. Department of Agriculture Agricultural Research Service's long-term model organism database and the maize research community's data portal. MaizeGDB makes accessible genetic and genomic data and data analysis tools that are used by researchers to investigate basic biological concepts and translate findings into technology that is deployed in farmers' fields. During the period from 2013 through 2018, the MaizeGDB team is tasked to make accessible high-quality, actively curated, and reliable genetic, genomic, and phenotypic data sets. At the root of a high-quality genome lies a well-supported assembly and annotation. For this reason, the deployment of an automated high-quality genome annotation system is of the utmost importance. As we demonstrate here, MAKER-P will fulfill this need.

Updating a genome's annotations over time is a complex task, and the rapidly changing data landscape can render annotations obsolete almost as they are created. Continuity is another major issue. Many genome projects have annotations that embody years of manual curation and revision. Simply throwing old annotations away and substituting new ones created by another pipeline is hardly desirable. To be truly effective, any revision process must build upon the foundation of existing annotations and provide incremental means to move forward in light of new data.

Next-generation sequencing data, especially RNA sequencing (RNA-seq) data, also hold great potential for the annotation of newly sequenced plant genomes. But again, making use of them is no easy task. For example, using transcriptome data to train gene finders for use on a newly assembled genome can be a difficult, frustrating task, so much so that many genome projects attempt to leverage gene finders trained for other genomes. As we have demonstrated previously (Holt and Yandell, 2011), both approaches are challenging and fraught with difficulties, and gene model accuracy suffers when gene finders are trained with unmatched species parameters.

Moreover, gene space is not limited to protein-coding genes; increasingly, noncoding RNA (ncRNA) annotations are coming to be considered an essential component of every genome's annotations. Pseudogenes are also an issue, especially for plant genomes, due to frequent whole-genome duplication and subsequent degeneration of paralogs (Zou et al., 2009). Consider the rice (*Oryza sativa*) genome, for example, which has approximately 39,000 annotated protein-coding genes and 28,330 pseudogenes (Zou et al., 2009); clearly, means to annotate pseudogenes are needed.

MAKER-P (Campbell et al., 2014) is an easy to use genome annotation pipeline with great software portability, based upon the widely used MAKER genome annotation pipeline (Holt and Yandell, 2011). Designed to address the needs of the plant genomes community, MAKER-P provides means for the annotation of newly

sequenced plant genomes and for automated revision, quality control, and management of existing genome annotations. MAKER-P also extends MAKER to include means for pseudogene annotation and noncoding gene finding. MAKER-P provides the plant genomics community early access to new functionalities prior to their later, general release in the MAKER package. Moreover, MAKER-P is dramatically faster than other genome annotation pipelines, allowing it to scale to even the largest plant genomes. MAKER-P is designed to run on Unix-like operating systems, including Linux and Apple OS X. It can run on laptop and desktop machines, but it also has extensions to take advantage of capabilities offered by high-performance computer clusters. Recent work, for example, has shown that the version of MAKER-P available within the iPlant Cyberinfrastructure can reannotate the entire maize genome in less than 3 h (Campbell et al., 2014) and that it can carry out the complete de novo annotation of the 17.83-Gb draft loblolly pine (*Pinus taeda*) genome in less than 24 h (Neale et al., 2014; Wegrzyn et al., 2014).

Our previous work using the Arabidopsis (*Arabidopsis thaliana*) genome demonstrated MAKER-P's effectiveness for the management and quality control of existing annotations and for de novo annotation using this relatively simple plant genome (Campbell et al., 2014). Here, we apply MAKER-P to the much less tractable maize genome, using it for analysis and quality control of the 5b+ annotation build, to systematically compare the 5b and 5b+ annotation builds with one another, for revision of the 5b+ annotations in light of 96 different RNA-seq data sets, and for de novo annotation of the maize genome. Also presented is maize genome annotation build 6a, which is demonstrably superior to the existing 5b+ build, thereby demonstrating MAKER-P's utility for management and quality control of the maize genome annotations.

RESULTS AND DISCUSSION

Overview of the 5b and 5b+ Builds

Our overarching goal in these analyses was to systematically compare the 5b and 5b+ annotation builds with one another using MAKER-P's management functions, to update and reevaluate the 5b+ annotation build in light of additional RNA-seq evidence, and to determine if MAKER-P was capable of automatically producing an annotation build of comparable quality. Table I summarizes the 5b and 5b+ RefGen builds. The Arabidopsis Information Resource (TAIR) 10 annotations are also included for purposes of comparison. As can be seen, the 5b and 5b+ builds are very similar to one another, differing primarily by 251 new and 213 improved genes in 5b+ (160 new models in chromosomes 1–10). In addition, a higher percentage of 5b+ models have annotated start and stop codons. In what follows, we present a detailed analysis of the relationship of the 5b+ annotation build to its supporting evidence, subjecting it to a series of quality-control analyses. We will also describe three additional annotation builds: a MAKER-P updated version of 5b+; a MAKER-P de novo annotation build; and a new 6a annotation build. The 6a build is a consensus build composed of the MAKER-P updated 5b+ gene models minus a set of 2,647 poorly supported 5b+ gene models. The 6a annotation build also includes 4,466 additional new, but well-supported, gene annotations derived from the MAKER-P de novo build; 102,370 pseudogene fragments; and an additional 2,522 ncRNA gene annotations. Each of these annotation data sets is described in detail below.

Use of RNA-seq Data

RNA-seq data provide means for the independent confirmation and improvement of genome annotations. MAKER-P (Campbell et al., 2014), like its parent pipeline MAKER2 (Holt and Yandell, 2011), provides integrated means for employing RNA-seq data for de novo annotation, for revising existing annotation data sets in light of new RNA-seq data, and for quality-control purposes. MAKER-P uses these data to add additional untranslated region (UTR) and exon sequences to existing gene models and for the creation of new gene models where none existed previously (Holt and Yandell, 2011).

Extensive RNA-seq resources exist for maize, and our goal here was 2-fold: to use these data for purposes of quality control and to determine if MAKER-P could employ them to improve the quality of the 5b+ annotations. For these analyses, we used 96 different RNA-seq data sets downloaded from the Sequence Read Archive repository (Benson et al., 2013). The data sets are derived from various maize genotypes, developmental stages, and plant tissues. The data sets are composed of various read lengths, ranging from single-end 35 bp to 2 × 100 bp (for details, see Supplemental Table S1). Assembly of these data using Trinity (Grabherr et al., 2011; see "Materials and Methods") produced 5,116,586 different transcripts, all of which were used in the analyses described below.

After assembly with Trinity, we ranked the RNA-seq data sets according to their number of assembled transcripts, our assumption being that data sets with the most transcripts would have the greatest value for annotation and quality control. We also sought to determine if there was a constant or perhaps diminishing benefit of using ever-greater numbers of RNA-seq data sets in the annotation process. Table II documents the power of pooling ever-larger numbers of RNA-seq data sets for discovery and quality-control purposes. Column 2 of Table II tallies the number of all 5b+ annotations on maize chromosome 5 that were overlapped, at least by 1 bp, by one or more transcripts using top one, five, 10, 15, 20, and finally all 96 transcript assemblies. The third column tallies the percentage of 5b+ annotations encoding a protein with a Pfam domain (Finn et al., 2014) but without transcript support, as annotations containing known protein domains are less likely

Table I. *Overview of maize annotation builds*

5b and 5b+ refer to nuclear chromosomes 1 to 10 only in versions 5b and 5b+ of Maize Genome Sequencing Project annotation builds, respectively. Also included is a de novo annotation data set generated by MAKER-P. 5b+ update is a MAKER-P updated version of the 5b+ annotation build. 6a is the final, combined data set consisting of the updated 5b+ gene models with evidence support plus an additional 4,964 new gene models derived from the MAKER-P de novo build. TAIR 10 annotations are included for purposes of comparison.

Parameter	5b	5b+	MAKER-P	5b+ Update	6a	TAIR 10
Protein-coding genes	39,024	39,155	44,200	38,783	40,602	27,206
Average gene length	4,100	4,014	3,600	4,203	4,190	1,488
Average protein length per gene	375	366	327	371	366	410
Average exons per mRNA	4.8	4.8	4.6	5	5.1	5.3
Percentage of genes with UTRs	81	81	59	85	86	77
Average UTR length	397	422	284	515	507	259
Average 5′ UTR length	137	161	107	202	199	94
Average 3′ UTR length	260	261	177	313	308	165
Percentage of models with start and stop codons	84	97	86	98	94	96
Percentage of genes with a Pfam domain	64	65	62	65	69	79

to be false positives. As can be seen, the number of additional confirmed annotations begins to plateau beyond 10 transcript assemblies, with only modest improvements thereafter. These results provide two important facts: first, they place an approximate upper bound on the expected percentage of gene models that can be confirmed using the available RNA-seq data: about 91%; second, they provide some guidance regarding the minimum number of transcript assemblies to employ in quality-control and future reannotation efforts. Properties of RNA-seq data sets such as read depth and heterogeneity make generalizations for other genomes and their RNA-seq data sets problematic, but for these data, it appears that it would be advisable to use at least 10 of the RNA-seq data sets. In the interest of performing as near exhaustive analysis as possible, we employed all available maize RNA-seq transcript assemblies as well as an additional 136,673 maize EST and full-length cDNA sequences from the National Center for Biotechnology Information (NCBI) and 33,635 nonmaize SwissProt plant protein sequences in the analyses that follow.

Accuracy of Intron-Exon Structures

MAKER-P provides automated means to assess the accuracy of a genome's annotations in the context of the evidence used to produce them (Campbell et al., 2014). To do so, it uses a performance measure called annotation edit distance (AED; for review, see Yandell and Ence, 2012). AED measures the goodness of fit of an annotation to the evidence supporting it. AED is a number between 0 and 1, with an AED of 0 denoting perfect concordance with the available evidence and a value of 1 indicating a complete absence of support for the annotated gene model. 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 an annotation's congruency with a particular type or types of evidence. By plotting the cumulative distribution function (CDF) of AED across all annotations, a genome-wide perspective can be obtained of how well the annotations reflect the EST, protein, and RNA-seq evidence. Importantly, this can be done even in the absence of a gold-standard set of reference annotations. AED also makes it possible to compare the annotations of different genomes with one another, making possible many new sorts of cross-genome quality-control analyses (Eilbeck et al., 2009; Holt and Yandell, 2011; Yandell and Ence, 2012). For additional information on AED, see Yandell and Ence (2012).

The top of Figure 1 presents AED CFD curves for the 5b and 5b+ annotation builds. For reference purposes,

Table II. *Impact of using increasing numbers of RNA-seq data sets for annotation*

Ninety-six different RNA-seq data sets were ranked according to the number of Trinity-assembled transcripts they produced. The number (and percentage) of maize chromosome 5 5b+ genes supported by the top one, five, 10, 15, 20, or all transcript collections was calculated (column 2). Column 3 shows the number (and percentage) of 5b+ genes containing a Pfam domain but not supported by any transcript evidence.

RNA-seq Data Sets	Transcript-Supported 5b+ Annotations on Chromosome 5	5b+ Annotations with Pfam Domains But without Transcript Support
Best 1	2,670 (59.7%)	886 (19.8%)
Best 5	3,624 (81.0%)	314 (7.0%)
Best 10	3,924 (87.7%)	159 (3.6%)
Best 15	4,015 (89.8%)	130 (2.9%)
Best 20	4,066 (90.9%)	115 (2.6%)
All assemblies	4,082 (91.3%)	121 (2.7%)

also included is the TAIR 10 annotation build, presented previously (Campbell et al., 2014). The bottom of Figure 1 summarizes the same AED CFD curves as stack plots, wherein the AED data have been binned into quartiles. In previous work, we advocated that an AED CDF curve wherein more than 90% of genome annotations have an AED score of less than 0.5 is evidence that that genome is well annotated (Yandell and Ence, 2012). The Arabidopsis, human, and mouse genome annotations, for example, all satisfy this criterion (Eilbeck et al., 2009; Holt and Yandell, 2011; Campbell et al., 2014). As can be seen, approximately 90% of maize annotations have AED scores of less than 0.5, indicating that maize is a relatively well-annotated genome, but less so compared with the TAIR 10 reference annotations. Thus, Figure 1 serves to highlight an essential point regarding the maize genome annotations. Despite the complexity of the maize genome, the quality of its existing gene models as measured by their congruency with the available evidence is reasonably high, but nowhere near that of Arabidopsis. Figure 1 also makes it clear that the 5b+ and 5b builds are of very similar quality as judged by AED. This result, taken together with the data presented in Table I, which demonstrate the similarity of the two builds with regard to gene numbers, lengths, exons, and intron content, makes it clear that the two data sets are globally very similar to one another. Also presented in Figure 1 is an AED curve and stack plot for the 160 new gene models present in the 5b+ build. These new genes, on average, are less well supported.

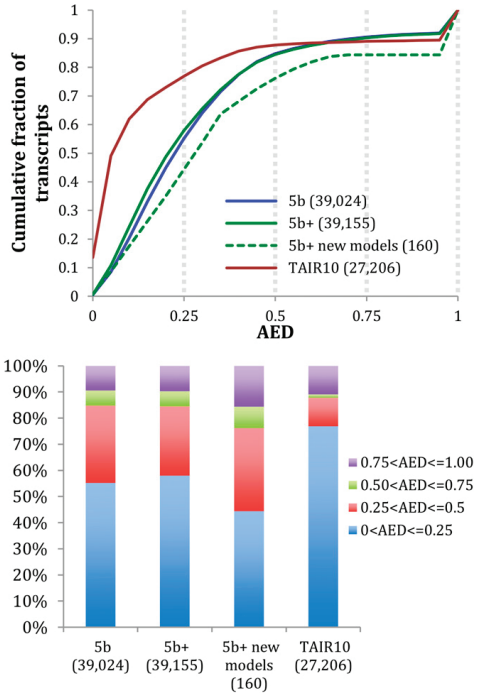

Figure 1. AED analyses of the 5b, 5b+, and TAIR 10 annotation builds. Top, AED CDF curves; bottom, stack plots with the same data broken down into quartiles. 5b+ new models are those models that are not present in 5b.

AED and Gene Category

Closer inspection of Figure 1 reveals that the maize 5b and 5b+ annotation builds, as well as the TAIR 10 build, contain a significant fraction of gene models with very little or no evidence supporting them. These models, with an AED score of 1 or nearly so, produce the sudden ramp present at the far right end of their AED curves. These models are shown in purple in the stack plots.

The TAIR 10 annotation for Arabidopsis can be used to better understand this ramp. TAIR employs a five-star ranking system for quality control of its genome annotations (ftp://ftp.arabidopsis.org/home/tair/Genes/TAIR_gene_confidence_ranking/DOCUMENTATION_TAIR_Gene_Confidence.pdf). In the TAIR schema, the best-supported transcripts are afforded five stars or four stars, with less supported annotations assigned three-, two-, and one-star status. Annotations with no support are assigned to the no-star category. In previous work (Campbell et al., 2014), we cross-validated MAKER-P's AED and TAIR 10's star ratings. For five-star TAIR 10 transcripts, 94% have AED scores of less than 0.5, whereas only 33% of one-star transcripts have an AED less than 0.5. All of the 604 TAIR 10 no-star annotations have AED's of one, indicating that they have no evidence support.

In order to better understand the characteristics of the poorly supported gene models in the maize v3 build, we divided the 5b+ maize annotations into five categories based upon the following categories of homologous relationships: Syntelogs, Orthologs, Conserved, Species-specific, and Other. We term Syntelogs as those gene annotations with syntenic orthologs in rice and/or sorghum (*Sorghum bicolor*). We classified as Orthologs those models with an ortholog in rice and/or sorghum that is not syntenic. Conserved are those gene models that are identified in a multispecies tree but where no orthologous relationships were found. Species-specific are those annotations encoding proteins with one or more paralogs in maize but not found elsewhere. And by Other, we mean gene models not meeting any of the above criteria. The results of this process are shown in Figure 2. As can be seen, the overall level of support and the congruency of the 5b+ gene models' intron-exon structures with their supporting evidence differ in a consistent fashion across the categories. Syntelogs, for example, are characterized by much lower (better) AED scores than the other categories. The 160 new genes in the current 5b+ build are

distributed across these five categories as follows: 68 in the Syntelog category, 23 in the Ortholog category, 11 in the Conserved category, three in the Species-specific category, and 55 in the Other category.

Poorly Supported Annotations in 5b+

Of the five categories, presented in Figure 2, Other is clearly the most problematic. Over 30% of these annotations have AED scores of greater than 0.75. By comparison, less than 1% of Syntelogs fall into this AED quartile. Given that the Other category comprises almost 4% of the 5b+ annotation build, the question naturally arises whether these are real maize genes, but inaccurately annotated, or false positives (i.e. not actually protein-coding genes). Our analyses call into question a considerable portion of genes in the Other category as well as unsupported annotations present in the rest of the categories. Using our evidence data sets (see "Materials and Methods"), a total of 3,141 (8%) of the 5b+ annotations have no supporting experimental evidence (e.g. RNA-seq, protein, and EST or encode Pfam domains). The results from Table II suggest that we should expect around 3% of the 5b+ annotations with protein support or containing a domain to lack transcript support. Although there may have been support for these annotations in prior annotation builds, 3,141 5b+ models have no support (transcript, protein, or domain) in our analysis. These facts suggest that these 3,141 5b+ annotations should be considered questionable and, in turn, that the 5b+ gene build contains 36,014 supported gene models.

MAKER-P Updates to the 5b+ Build

MAKER-P has the capacity to automatically revise an annotation build using new evidence (Campbell et al., 2014). This functionality is especially useful for updating annotations in light of new RNA-seq data. When run in update mode, MAKER-P revises the intron-exon structures of a reference annotation data set, adding additional 5′ and 3′ exons and UTRs to the reference annotations as suggested by the new evidence; reference annotations are split and merged in order to improve their fit to the supporting evidence; and new gene models are created in regions of the genome where experimental evidence supports the existence of a gene but where the reference build has no annotation. Importantly, when run in update mode, MAKER-P will not delete a reference gene model, even when MAKER-P fails to find evidence to support it.

The MAKER-P revision process for 5b+ merged 31 annotations, slightly decreasing the 5b+ gene set from 39,155 (nuclear chromosomes 1–10 only) to 38,783 annotations (for additional details, see Table I). Figure 3 illustrates the impact of revision upon the maize chromosome 10 5b+ gene models. Points along the diagonal line denote models unchanged by the revision process. Note that with MAKER-P revision, AED only improves; it never worsens. This is because MAKER-P defaults to the original reference annotation whenever it is unable to improve upon it. Note too that most changes are to those models having the lowest (best) AED scores in the reference set. This is because it is often the best-annotated models that have the richest supporting evidence: with 96 different RNA-seq data sets and 5,116,586 different assembled transcripts, highly expressed genes are often overlapped by such a superabundance of evidence, some supportive, some not, that human annotators are simply stymied. MAKER-P, in contrast, is able to effectively revise the gene models regardless of the complexity or quantity of evidence. For more on this point, see Campbell et al. (2014).

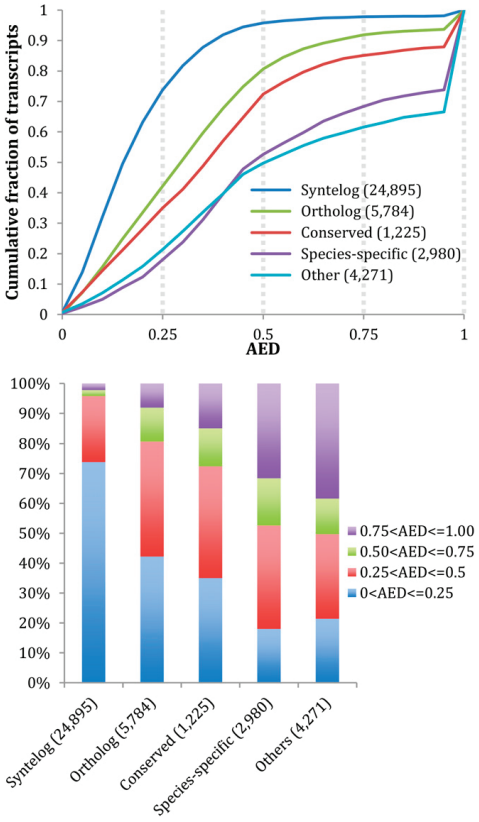

Figure 2. 5b+ annotations with stronger evidence of conservation have correspondingly better AED values. 5b+ maize annotations are broken into five categories: Syntelog, Ortholog, Conserved, Species-specific, and Other. For details of the classification system, see text. Note the extreme AED ramp of the Other category due to a lack of supporting evidence for these gene models. Top, AED curves; bottom, stack plots for the same data broken down into quartiles.

Figure 4 presents the AED CDF curves for the MAKER-P update in the context of both the 5b+ annotations and a MAKER-P de novo annotation build (discussed below). As can be seen, revision of the 5b+ build by MAKER-P shifts its AED CDF curve toward lower AEDs, indicating that the revision process has brought the 5b+ build into still better congruence with the available evidence. Note, however, that the AED ramp at the right side of the curve is unaffected; this is because the MAKER-P revision process has retained every gene model in the 5b+ build for which there was no supporting evidence. As shown, overall, the MAKER-P revised gene models have the highest proportion of genes with AEDs of less than 0.2. Table I summarizes the global differences between the 5b+ build and the MAKER-P 5b+ updated build. As can be seen, the MAKER-P revised models on average have more exons (five versus 4.8), contain additional UTR sequence (515 versus 422 bases of UTR), and the percentage of genes having any UTR at all increases from 81% to 85%. Collectively, these facts demonstrate the power of MAKER-P's update functionality to revise and improve even high-quality maize 5b+ gene models.

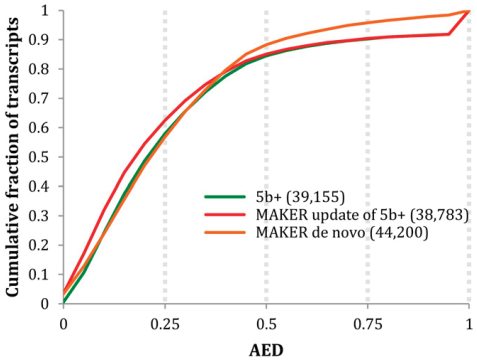

Figure 4. AED analyses of the MAKER-P updated 5b+ gene models. For ease of reference, also included are the MAKER-P de novo annotations and the original 5b+ annotations.

The MAKER-P de Novo Annotations

We also generated a MAKER-P de novo annotation build for the maize genome, using the same evidence data sets as the analyses presented in Table I and Figures 1 to 4 (for details, see "Materials and Methods"). Our goal here was to 2-fold: (1) to measure the performance of MAKER-P on the maize genome by comparing its annotations with the 5b+ annotation build in order to gain an indication of what to expect when using MAKER-P on other difficult-to-annotate plant genomes; and (2) to determine if MAKER-P might identify additional maize genes absent from the 5b+ annotation build.

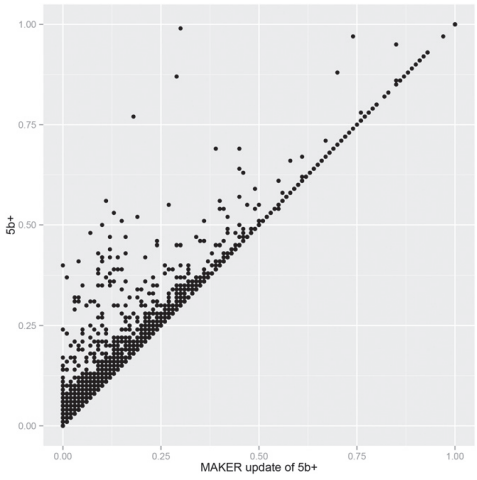

Figure 3. AED-based comparison of the 5b+ and 5b+ updated gene models for maize chromosome 10. Circles represent annotations with physical overlap between a 5b+ and its corresponding updated MAKER-P gene model. *x* axis, AED of the corresponding MAKER-P updated 5b+ gene model; *y* axis, AED of 5b+ models.

Training MAKER-P

Given sufficient training data (i.e. gold-standard gene models), ab initio gene predictors can deliver very accurate gene models (Guigó et al., 2006; Yandell and Ence, 2012). However, for newly sequenced genomes, no training data are usually available. In previous work (Holt and Yandell, 2011; Campbell et al., 2014), we described a procedure whereby MAKER-P can be used to train Augustus (Stanke and Waack, 2003; Stanke et al., 2008) and SNAP (Korf, 2004), two widely used ab initio gene finders. This training process uses RNA-seq data and ESTs in lieu of a preexisting gold-standard set of gene models. These data are aligned to the genome using the splice-aware aligner Exonerate (http://www.ebi.ac.uk/~guy/exonerate/), and an automatically identified postprocessed subset of high-quality alignments is used for gene-finder training.

Grass genomes are generally repeat rich and harbor the results of multiple polyploidization events, making them difficult substrates for annotation. It seemed likely that these same features of grass genomes might negatively impact the effectiveness of MAKER-P's gene-finder training procedures. Maize thus provides an opportunity to examine this problem. The genome is typical of grass genomes: there is a preexisting gold standard of reference annotations (e.g. the conserved Syntelogs of the 5b+ build), and there exist a plethora of maize RNA-seq and

EST data. Equally important, the popular and very accurate gene finder Augustus (Stanke and Waack, 2003; Stanke et al., 2008) comes pretrained for maize, providing an opportunity to benchmark the performance of a version of Augustus trained by MAKER-P using maize RNA-seq and EST data to one trained by the authors of Augustus using the maize reference annotations. Supplemental Figure S1 shows the AED CDF curves for these two versions of Augustus. As expected, the version trained by the Augustus group using the 5b gene models is more accurate than the MAKER-P version trained using the noisy RNA-seq and EST data, but not greatly so. The MAKER-P-trained version of Augustus, for example, calls about 5% more genes, and 87%, as opposed to 91%, of its models have an AED of less than 0.5, indicating that the intron-exon structures of the MAKER-P-trained version of Augustus are nearly as accurate. These results demonstrate that MAKER-P's training procedure is effective even for difficult-to-annotate grass genomes. We used the MAKER-P-trained version of Augustus for the de novo annotation run described below.

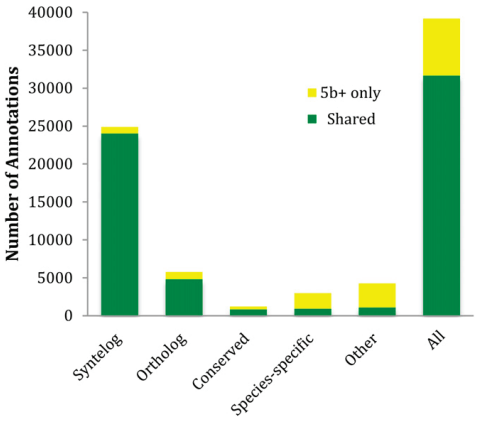

Figure 5. Shared and unique gene models in the 5b+ and the MAKER-P gene de novo gene sets. To facilitate comparison, both builds were broken down into the same five gene categories described for Figure 2. Intersecting genes are shown in green, and gene models unique to the MAKER-P de novo build are shown in yellow.

MAKER-P de Novo Results

AED curves and stack plots comparing the MAKER-P de novo build with the 5b+ and updated 5b+ builds are presented in Figure 4. As can be seen, overall, its models are nearly as congruent with the evidence as the updated 5b+ build. Figure 5 summarizes the intersections between the 5b+ build and the MAKER-P gene set, broken down by gene category. As shown, there is almost perfect agreement among the Syntelog gene set, with less, but still considerable, congruence for the Ortholog and Conserved categories. However, of the 5,401 models comprising the 5b+ Other category, only 1,347 have supporting evidence and are also called by MAKER-P, again suggesting that many of 5b+ genes belonging to the Other category should be considered provisional.

Table I summarizes the relevant statistics of the MAKER-P de novo gene models. Globally, the MAKER-P de novo build is quite similar to the 5b+ build, but it differs in three regards: (1) fewer of its gene models contain UTRs; (2) its gene models are shorter; and (3) it contains 5,045 additional annotations that do not overlap 5b+ gene models. Point 2 is largely a consequence of the additional gene models not present in the 5b+ build. The 5,045 additional gene models tend to be short and are predominantly single-exon genes. In these respects, they are quite similar to the majority of 5b+ genes in the Other category. But they differ in one vital regard: every MAKER-P gene is supported by transcript, protein, and/or domain evidence, whereas the majority of the 5b+ Other genes are supported only by ab initio gene predictions, a point we return to in "Conclusion." Collectively, analyses presented in Figures 4 and 5 and Table I indicate that, globally, the MAKER-P de novo build is slightly inferior to the curated 5b+ build with regard to protein-coding genes, but not dramatically so, demonstrating that MAKER-P is capable of producing a high-quality de novo gene build for a grass genome, one that is a suitable starting point for further manual and automated curation. Moreover, as we document below, the MAKER-P de novo build has no unsupported models and contains additional pseudogene, ncRNA, and well-supported protein-coding gene models not present in the curated 5b+ build.

Nonprotein-Coding Genes

MAKER-P's annotations are not limited to protein-coding genes alone. The MAKER-P toolkit provides a process for the annotation of pseudogenes. The ability to annotate and identify pseudogenes is particularly important for grass genomes, given their abundance. MAKER-P also provides means for the identification of known and new classes of ncRNAs.

Pseudogenes

In total, 102,370 putative partial or complete pseudogenes were identified in maize with MAKER-P. These pseudogenes have a mean length of 191 bp, similar to what was found in Arabidopsis and rice (Zou et al., 2009b; Campbell et al., 2014), with a significant positive skew, indicating that the majority of pseudogenes were on the shorter end of the spectrum. This can be a consequence of the inability to connect pseudoexons of a pseudogene together. Nonetheless, the same MAKER-P pipeline identified only 4,204 pseudogenes in Arabidopsis, far less than what we have recovered in maize.

One explanation is that the gene deletion rate was higher in the Arabidopsis lineage, consistent with the finding that genome size differences between Arabidopsis (150 Mb) and *Arabidopsis lyrata* (207 Mb) is due to extensive DNA loss (Hu et al., 2011). Another possibility is that pseudogenes were generated or retained at a greater rate in the maize lineage. This is consistent with a much more recent whole-genome duplication in the maize lineage (approximately 11 million years ago; Gaut and Doebley, 1997) compared with that in Arabidopsis (α-genome duplication, approximately 50 million years ago; Bowers et al., 2003). In addition, in maize, there is an overabundance of *Helitrons* carrying gene fragments (Du et al., 2009; Yang and Bennetzen, 2009). Among 272 manually annotated *Helitrons*, 94% of them carry captured sequences from 376 genes (Du et al., 2009). There is also evidence suggesting that more than 20,000 gene fragments in the B73 genome are transduplicated and reshuffled due to *Helitron* activities (Yang and Bennetzen, 2009). Together with the suggestion that *Helitrons* are involved in exon shuffling (Feschotte and Wessler, 2001), these findings are consistent with the possibility that *Helitrons* have contributed significantly to the high pseudogene fragment number observed.

To better understand what kinds of duplicates tend to become pseudogenes, MapMan (Thimm et al., 2004) annotations were assigned to pseudogenes based on the maize protein sequences used to identify them. As a result, 54.6% of pseudogenes have one or more MapMan annotations. Fisher's exact test was used to identify MapMan annotations associated with overrepresented and underrepresented numbers of pseudogenes (Figure 6). Overrepresented terms include stress, protein degradation (via ubiquitin), and secondary metabolism (unspecified), which are also known to be overrepresented in Arabidopsis (Zou et al., 2009). Similarly, the *Argonaute* gene family involved in small RNA biogenesis has 43 annotated, presumably functional, members and 127 pseudogenes (Figure 6). *Argonaute* genes are important for viral defense in plants (Qu et al., 2008). In addition, genes involved in external stimulus responses tend not only to experience lineage-specific duplication (Hanada et al., 2008) but also to pseudogenize at a higher rate (Zou et al., 2009). Taken together, the significant overrepresentation of *Argonaute* pseudogenes may be the product of viral defense genes that were no longer useful. We also found that most transcriptional regulators are among the underrepresented class of pseudogenes, except the Homeobox and APETALA2/ethylene response element binding protein families (Figure 6). The underrepresentation of transcription factor pseudogenes is consistent with higher retention rates among plant transcription factor duplicates (Schnable et al., 2009), particularly those derived from whole-genome duplications (Blanc and Wolfe, 2004; Seoighe and Gehring, 2004; Shiu et al., 2005). Therefore, in spite of differences in the number of pseudogenes identified, the pseudogenization of duplicates in Arabidopsis and maize follows similar trends.

ncRNA Genes

The MAKER-P toolkit identified 2,192 total tRNA genes. Of these annotated tRNA genes, 1,398 decode the standard amino acids, four decode seleno-Cys, seven are possible suppressor tRNAs, 12 are undetermined, and 771 appear to have been pseudogenized (Table III). Ultimately, these data contain slight differences from tRNA analyses of previous maize genome assemblies in maize secondary to changes in the v3 assembly (http://lowelab.ucsc.edu/GtRNAdb/Zmays/Zmays-stats.html). Using 12 small RNA-seq experiments, the MAKER-P toolkit also identified 183 microRNAs (miRNAs). As mentioned previously (Campbell et al., 2014), the number of miRNAs predicted by the MAKER-P toolkit is dependent on the small RNA evidence; thus, this number represents a lower bound of miRNAs in the v3 assembly. Most of the predicted mature miRNAs are of length 21, which is the typical plant miRNA length. Of the 183 predictions, 87 of them overlap with the existing 5b+ annotation of miRNAs and others are new predictions. The discrepancy mainly stems from the different methods used for miRNA annotation by MAKER-P and the existing maize miRNA identification method (Zhang et al., 2009). While the miRNA prediction pipeline miR-PREFeR of MAKER-P follows the criteria for plant miRNA annotation (Meyers et al., 2008), 5b+ miRNA annotations were created by aligning genomic sequences against miRBase (Griffiths-Jones et al., 2008) sequences using BLASTN (Altschul et al., 1990). Thus, the reliability of 5b+ miRNA annotation relies heavily on the quality of miRBase collections. Although the underlying annotations in miRBase are generally experimentally determined or experimentally verified, errors have been detected in miRBase annotations (Kozomara and Griffiths-Jones, 2014). In addition, many 5b+ miRNA annotations lack expression evidence in our 12 small RNA-seq samples. Finally, the homology search-based annotation method we adopted may miss miRNAs that are specific to maize. Using the same small RNA-seq data sets, the MAKER-P toolkit identified 727 small nucleolar RNAs (snoRNAs) with AEDs less than 0.5. (See Supplemental Text S1 for the link to the GFF file containing the tRNA, miRNA, and snoRNA predictions.)

The 6a Gene Annotation Build

Table I also provides a summary of an annotation build termed 6a. Our goal in creating the 6a build was to provide the maize community with a single annotation build comprising the best-possible annotated gene models drawn from the 5b+, 5b+ updated, and MAKER-P de novo annotation builds. Thus, the 6a build is a synthetic data set composed of the MAKER-P updated 5b+ gene models, which contain additional 5′ and 3′ exons and UTR sequences, together with additional new, but well-supported, genes derived from the MAKER-P de novo build. We also excluded from 5b+ 2,647 5b+ gene models for which we could find no supporting evidence and 249 models that overlapped with our predicted

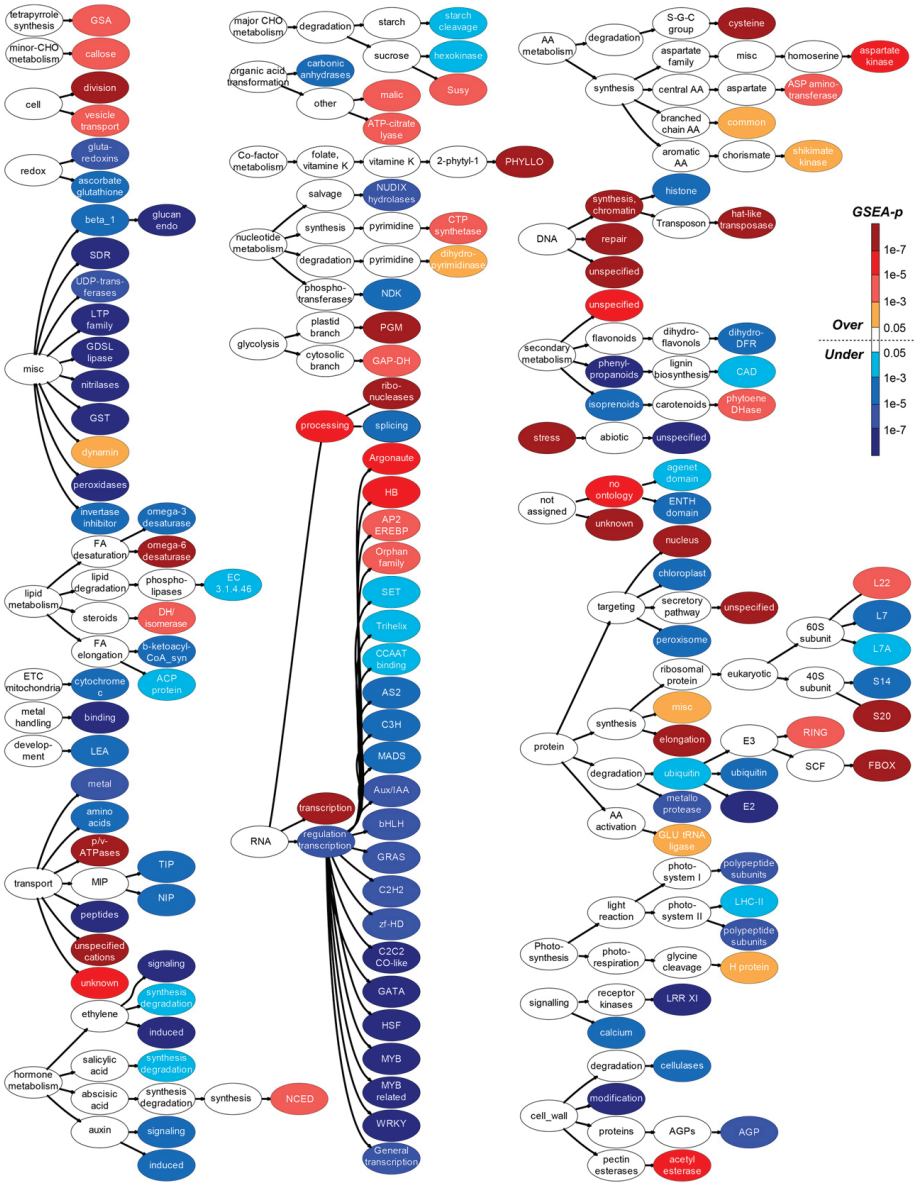

Figure 6. MapMan terms with overrepresented or underrepresented numbers of maize pseudogenes. The ovals indicate overrepresented (shades of red) and underrepresented (shades of blue) terms and their parent terms (white). Some terms are truncated or abbreviated. For full terms and associated statistics, see Supplemental Table S4.

Table III. *Summary of ncRNA annotations*

Numbers of ncRNAs are broken down by type for 5b, 5b+, and 6a annotation builds. The last column gives corresponding numbers in the TAIR 10 annotation of Arabidopsis for reference. NA denotes classes of annotations not present in the non-MAKER-P-derived builds.

ncRNA Type	5b	5b+	6a	Common to 5b+ and 6a	TAIR 10
miRNA	NA	316	183	87	180
tRNA	NA	NA	2,192	NA	631
snoRNA	NA	NA	727	NA	71

ncRNA models. These gene models are included in a separate file (Supplemental Table S2) under the title Provisional v3 Gene Models.

The 44,200 MAKER-P de novo protein-coding genes (Table I; Fig. 5) comprised the starting point for our attempt to identify a core set of additional high-quality gene models for inclusion in the 6a build. To identify these models, we first removed any unique MAKER-P de novo gene models that resided within transposons, as these might represent gene fragments carried by transposons; this reduced the number by about 10%. We then broke the remaining MAKER-P unique protein-coding gene models into two classes: (1) multiexon models with at least one splice site perfectly confirmed by RNA-seq or EST alignments; and (2) single-exon models that encode a domain and have annotated start and stop codons. Our reasoning was that models supported by spliced transcript data and having canonical splice sites were reasonable candidates for additional genes. We also enforced an additional criterion on these genes: they must have at least one coding exon predicted by a gene finder. With regard to the unique MAKER-P single-exon gene models, because single-exon genes are often spuriously overlapped by transcript data, we did not consider transcript support as proof of a single-exon gene's existence. Thus, enforcing the additional criteria that these single-exon genes encode a known domain, their single exon be predicted by a gene finder, and they have annotated start and stop codons should diminish the proportion of the models that constitute a common form of false-positive annotation: random open reading frames fortuitously overlapped by RNA-seq data from noisy transcription data. Likewise, the requirement for start and stop codons should avoid false positives where the supposed single-exon gene consists of portions of a pseudogene with a partial open reading frame encoding a remnant portion of a protein domain. Of course, none of these criteria can guarantee that every one of the additional new genes is truly a new maize protein-coding gene, but what is true is that each of the new gene models identified in the analysis meets a stringent set of criteria for inclusion in the 6a build. Certainly, they are better candidates than the 2,647 provisional gene models we identified in our analyses of the 5b+ build, none of which meet any of these criteria; hence, replacing those provisional models with these additional MAKER-P-derived new models seems reasonable.

Table IV summarizes the results of this analysis. In total, 4,049 of the new MAKER-P gene models encode multiexon transcripts with at least one confirmed splice site. Note that the average number of exons is 4.9, and 45% of these putative genes encode a Pfam domain. Thus, although they are shorter than the average 5b+ annotation (2,836 versus 4,014), many are sizable, multiexon gene models that contain domains. All 417 of the single-exon models encode a domain, have transcript support, and have annotated start and stop codons. In addition, all of the new models have gene-finder support. Figure 7 presents AED stack plots for the 6a build and various portions thereof. Also included for reference purposes are the 5b+ reference build and the subset of models that we identified as provisional and, thus, that are not included in the 6a build. As can be seen from an inspection of Tables I, III, and IV, the 6a build contains more supported gene models and more models with 5′ and 3′ UTRs, and its gene models have longer UTRs compared with the original 5b+ build, contain more exons, and encode longer proteins. The 6a models are also more congruent with the available evidence as judged by AED. Also included are an additional 3,006 ncRNA genes and 102,370 pseudogene annotations not present in the 5b+ build.

CONCLUSION

We have carried out systematic analyses of the maize 5b+ annotation build using MAKER-P's management and quality-control functions. This work has allowed us to reevaluate the 5b+ annotation build in light of additional RNA-seq evidence and to update the 5b+ build using these same data. We have also compared MAKER-P de novo annotations with those of the 5b+ reference build in order to gain an indication of what to expect when using MAKER-P on other difficult-to-annotate plant genomes. These same analyses have identified additional maize genes absent from the 5b+ annotation build.

As we have shown, MAKER-P can further improve an existing genome annotation build. The MAKER-P 5b+ update, for example, contains every model present in the 5b+ build but adds additional exons and UTR sequences. It also contains a number of gene splits and merges suggested by the RNA-seq data. The result is an updated 5b+ build that is demonstrably in better agreement with the available evidence. Importantly, these results also show how using MAKER-P for the management of a genome's annotations does not necessitate a switch from one pipeline's annotations to another. MAKER-P can improve an existing community annotation resource without introducing any break in continuity (i.e. the existing models are kept but brought forward incrementally to reflect additional evidence).

Our de novo training results demonstrate that MAKER-P also can be used to train a widely used gene finder such as Augustus for employment on newly sequenced plant genomes and that the resulting performance is a close match to that obtained using a gold-standard training set. This is important because previous work by our group and others has made it clear that attempts to leverage gene finders trained from other genomes rarely produce accurate gene predictions. Our analysis of the MAKER-P de

Table IV. *Summary of new gene models included in the 6a build*

Parameter	Multiexon MAKER-P de Novo	Single-Exon MAKER-P de Novo	6a
Protein-coding genes	4,049	417	40,602
Average gene length	2,836	676	4,190
Average exons per mRNA	4.9	1	5.1
Average exon length	195	648	315
Average protein length	216	221	366
Percentage of genes with a Pfam domain	45	100	68

novo annotations demonstrates that, although the MAKER-P de novo models are slightly inferior with regard to the accuracy of its intron-exon structures, it is demonstrably superior in its relationship to the available evidence (i.e. the average model is more congruent with its overlapping evidence, and importantly, every one of its annotations has supporting evidence). Collectively these results make clear that MAKER-P provides an effective means for de novo annotation of even difficult-to-annotate grass genomes.

The 6a annotation build provides the maize community a genome annotation data set that is notably superior to both the 5b+ and MAKER-P de novo builds. Informed by new expression evidence assembled from an extensive collection of RNA-seq studies, the 6a build contains the MAKER-P updated 5b+ gene models together with an additional 4,466 new genes not contained in the 5b+ annotation build.

The 6a build also lacks 2,647 5b+ genes for which we could find no support, despite the number and diversity of evidence data sets used. Thus, the improvements offered by the 6a build are not limited solely to new contents. Considering these 2,647 5b+ genes as provisional has important consequences for future work: first, these poorly supported gene models, for example, will no longer introduce biases into comparative studies with regard to statistics such as domain content, UTR lengths, and exon number sets; second, knowledge that these 5b+ genes are provisional will provide a starting point for focused experimental follow-up studies aimed at confirming or denying their existence.

Collectively, the 6a build is a demonstrable improvement upon the 5b+ build. Its genes have more exons, have longer UTRs, and are more congruent with the evidence. Furthermore, the 6a build also supplements the 5b+ build with 102,370 pseudogene and 3,006 ncRNA annotations.

Recent work has shown that the version of MAKER-P available within the iPlant Cyberinfrastructure can reannotate the entire maize genome using the same evidence data sets described here in less than 3 h (Campbell et al., 2014) and that it can carry out a complete de novo annotation of the 20-Gb draft loblolly pine genome in less than 24 h (Neale et al., 2014; Wegrzyn et al., 2014).

These facts have important implications for the future of plant genome annotation. First, they show that MAKER-P provides effective means for the annotation of plant genomes; second, its update mode provides a means to refresh the annotations of established plant genomes to reflect new data; and third, these updates can be carried out much more rapidly and frequently than has heretofore been possible. Perhaps even more important is that MAKER-P's speed and flexibility will enable individual iPlant users to generate their own custom genome annotation data sets using public annotation builds as starting points but embodying their own data. The 6a annotations and related documents are available for download at http://documents.maizegdb.org/makerp/. The latest version of MAKER-P is available as part of the

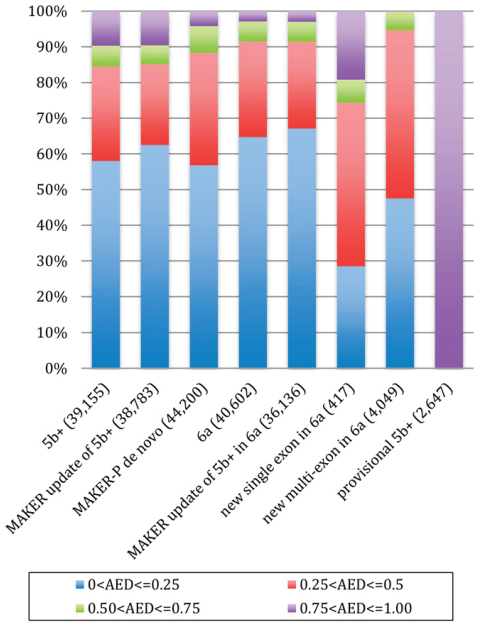

Figure 7. AED analyses of the 6a build. AED stack plots are broken down into quartiles: 5b+ build, MAKER update of 5b+, MAKER-P de novo, 6a build, 5b+ models in 6a, new MAKER de novo multiexons and single exon in 6a, and provisional 5b+ models. Numbers in parentheses indicate the number of annotations in each gene set.

MAKER package download at http://www.yandell-lab.org/software/maker-p.html

MATERIALS AND METHODS

Transcripts and Protein Evidence

Transcripts and transcript assemblies were used as evidence for gene predictions and MAKER updates. Maize (*Zea mays*) ESTs and full-length cDNAs were downloaded from the NCBI GenBank. Ninety-five RNA-seq data sets were downloaded from NCBI's Sequence Read Archive (Supplemental Table S1). One additional RNA-seq data set was described by Takacs et al. (2012) and can be obtained from the authors (Supplemental Table S1). The RNA-seq reads from these data sets were cleaned using tools from the FASTX toolkit (version 0.0.13; http://hannonlab.cshl.edu/fastx_toolkit/). The fastx-clipper program removed adapter sequences from all reads, and the fastx-artifacts-filter was used to remove aberrant reads. These steps were followed by running the fastx-trimmer program, which removed bases with quality scores less than 20 and discarded reads that were less than 30 bases in length. Cleaned RNA-seq reads from individual studies (Supplemental Table S1) were assembled using the Trinity transcript assembly package (Grabherr et al., 2011) and used for annotation. SwissProt plant protein sequences were downloaded from UniProt. Maize protein sequences were removed, and the remaining plant protein sequences were used as annotation evidence. The maize genome (Zea_mays.AGPv3.21.dna.genome.fa.gz) was downloaded from ftp://ftp.ensemblgenomes.org/pub/release-21/plants/fasta/zea_mays/dna/. MAKER-P analyses focused on all nuclear chromosomes 1 to 10 unless specified otherwise.

Classification of the 5b+ Annotation Set Using Comparative Genomics Criteria

We utilized the output of Ensembl Compara Gene Trees and associated synteny builds available from Gramene release 39 (October 2013), currently archived at http://archive.gramene.org/. The Ensembl method identifies ortholog and paralog relationships between genes using phylogenetic inference (Vilella, et al., 2009; see also http://useast.ensembl.org/info/genome/compara/homology_method.html). The Gramene project subsequently maps collinear and near-collinear orthologous genes between related species (Youens-Clark et al., 2011), adapting a protocol originally developed for the analysis of synteny in maize (Schnable et al., 2009; for details, see supporting online materials: http://www.sciencemag.org/content/suppl/2009/11/18/326.5956.1112.DC1/Schnable.SOM.pdf), which uses DAGChainer (Haas et al., 2004). The Compara Gene Trees in Gramene release 39 incorporated gene sets for 25 plant and five nonplant species. This release also included synteny maps for maize-sorghum (*Sorghum bicolor*) and maize-rice (*Oryza sativa*). From these data, we classified the maize 5b+ annotation set as follows: Syntelog, having orthologs in rice and/or sorghum that are arranged in a collinear or near-collinear fashion; Ortholog, having a called ortholog in rice and/or sorghum that is not a Syntelog; Conserved, found in a multispecies tree but lacking an identified ortholog; Species-specific, found in a maize-specific gene tree (i.e. having paralogs in maize but without homology to other species); and Other, not found in a tree (thus having no detectable homology with other species in the set).

Repeat Library and Examination of New Genes for Transposons

The repeat library used in this study was derived from the following two sources. First, 1,526 transposon exemplar sequences were downloaded from the maize TE database (http://maizetedb.org/~maize/). Second, 10,619 maize Sirevirus sequences were downloaded from MASiVEdb (Bousios et al., 2012) and masked by the 1,526 transposon sequences from the maize TE database. For a Sirevirus sequence, if 90% of the length was masked with a similarity of 80% or higher, it was excluded, since it was considered to be already present in the 1,526 sequences. Exemplar sequences were chosen from the remainder of the Sirevirus sequences to reduce the redundancy as follows: all sequences were compared using BLASTN. The element with the most matches (cutoff at 80% identity in 90% of the element length) was considered as the first exemplar. Thereafter, this element and its matches were excluded from the group and a second-round BLASTN search was conducted with the remainder of the elements, leading to the generation of the second exemplar. This process was repeated until all elements were excluded. These exemplar sequences were combined with the 1,526 transposon sequences from the maize TE database, and the combined library was used in this study.

Since the combined library only contains true transposon sequences, gene fragments that are carried with transposons such as those in Pack-Mutator-like transposable elements (MULEs) were not included in the library. To test whether the new MAKER-P genes identified in this study were actually gene fragments inside transposons, the relevant gene coordinates were first compared with previously identified Pack-MULEs in maize (Jiang et al., 2011). If over 50% of the mRNA sequence of a gene was located inside a Pack-MULE, this gene was considered a transposon and excluded from the 6a build. For the remainder of the genes, the gene and the 5-kb flanking sequence on both sides of the gene were retrieved and the transposons in the entire fragment were annotated using RepeatMasker with the library mentioned above. If the gene was flanked by two transposons from the same superfamily of transposon and both transposons were truncated by 30 bp or more on the side facing the gene, this gene was considered to reside inside a transposon and excluded from 6a. If only part of the gene was inside the transposon, a 50% cutoff of the transcribed sequences was taken for consideration. In summary, if 50% or more of the mRNA of a gene is inside a transposon, the gene is considered a transposon.

MAKER-P de Novo Annotation and Update of 5b+

RNA-seq data sets from public repositories (Supplemental Table S1) were assembled and used as evidence in MAKER-P 2.31 r1081, along with Uniprot/SwissProt protein evidence and a set of traditional full-length cDNAs. A custom repeat library (see above) was used to mask the repetitive regions (for details, see preceding paragraph). Genes were predicted using Augustus (Stanke and Waack, 2003; Stanke et al., 2008) trained in an iterative fashion in MAKER-P as described before (Campbell et al., 2014). The MAKER de novo annotation set represents those predictions that are supported by evidence or contained a Pfam domain. To obtain a set of MAKER-P revised annotations, maize 5b+ models are passed to MAKER-P as gene predictions, together with the same evidence set and RepeatMasker as above.

Utility of Transcript Assembly Evidence for Gene Predictions

Our Trinity-derived transcript assemblies from 96 different RNA-seq data sets were ranked by the number of sequences in each assembly. While this approach may not recover the best RNA-seq data sets in all cases (e.g. a data set might contain genomic contamination, resulting in large numbers of spurious transcripts), we found that this simple procedure provided a practical means to select subsets of RNA-seq data when many different data sets are available. Collections of the top one, five, 10, 15, 20, or all transcript assemblies were used as evidence in MAKER-P runs. MAKER-P was run in pass-through mode using the 5b+ gene predictions and the different collections of transcript assemblies as evidence. The 5b+ gene models were unmodified but were assigned AED scores based on the transcript support for each model. Genes with AED scores less than 1 were scored as being supported by the given transcript evidence set.

6a Annotations

MAKER de novo annotations that were not overlapped by MAKER updated 5b+ gene models were retained when (1) single-exon models encoded a domain and contained annotated start and stop codons and (2) multiexon models with at least one splice site was confirmed by EST alignment. Maize 5b+ updated models with domain support or RNA-seq evidence support were combined, along with MAKER-P ncRNA annotations with these two classes of MAKER de novo annotations, to generate the final 6a build. 5b+ models without evidence support (AED = 1.00) and/or encoded Pfam domains were classified as provisional. MAKER de novo annotations residing within transposons were also excluded.

ncRNA Annotation

tRNAs were identified using tRNAscan-SE (Lowe and Eddy, 1997) within the parallelized MAKER-P framework. The snoRNAs were predicted using snoscan (Lowe and Eddy, 1999) also within the parallelized MAKER-P framework. To limit the inevitable false positives resulting from the

genome-scale use of stochastic context-free grammars in snoscan, we limited our results to snoscan predictions that matched a ribosomal RNA (rRNA) *O*-methylation site and had an AED of less than 0.5. rRNA *O*-methylation sites for maize 26S (Refseq accession no. NR_028022 version NR_028022.2) and 17S (Refseq accession no. NR_036655 version NR_036655.1) rRNAs were inferred based on homology to known rRNA methylation sites (http://lowelab.ucsc.edu/snoscan/default-files/Hu-meth.sites) in human 28S (GenBank accession no. M11167 version M11167.1) and 18S (GenBank accession no. NR_003286 version NR003286.2) rRNA, respectively.

The miRNAs were identified using miR-PREFeR pipeline (Lei and Sun, 2014), which is an improved version of the miRNA annotation pipeline described previously (Campbell et al., 2014). Expression of these miRNAs was confirmed within the miR-PREFeR pipeline using 12 small RNA sequencing experiments from seven tissues (Supplemental Table S3). miR-PREFeR utilizes expression patterns of miRNAs and follows the criteria for plant miRNA annotation (Meyers et al., 2008) to accurately predict plant miRNAs from one or more small RNA-seq samples. The primary criterion is that the small RNA-seq data should provide evidence of precise miRNA/passenger miRNA (miRNA*) excision. Specifically, there should exist abundant reads corresponding to the mature miRNA sequence, and there should be at least one read that can be precisely mapped back to the miRNA* sequence. The miRNA and miRNA* sequences should form a duplex with two-nucleotide 3′ overhangs. In addition, the miRNA/miRNA* duplex needs to present the following structural characteristics: there are typically four or fewer unpaired bases in the miRNA/miRNA* duplex, and asymmetric bulges are rare and small in size.

As the expression of miRNAs can be tissue or condition specific, we aimed to provide a comprehensive miRNA annotation by using multiple RNA-seq samples from different tissues/conditions/developmental stages. There are two advantages of predicting miRNAs from multiple RNA-seq samples. First, some miRNAs are poorly expressed and cannot be identified in a single RNA-seq sample. miR-PREFeR can predict poorly expressed miRNAs by combining all reads from multiple samples. Second, due to fast degradation, some miRNAs lack reads mapping to their miRNA* region and will not satisfy the strict plant miRNA annotation criteria. In our method, if the corresponding miRNA loci from multiple samples demonstrate other typical miRNA characteristics, including high expression, the existence of a well-formed stem loop, and precise miRNA/miRNA* excision in the predicted stem loop, we conclude that this locus contains a true miRNA gene by dropping the requirement for the presence of the star sequence. In this implementation, when there is no read corresponding to the star sequence, we require that there should be at least 1,000 reads in all samples and at least 100 reads in each sample.

Pseudogene Identification

Pseudogenes were identified by MAKER-P according to the method described previously (Campbell et al., 2014). Annotated protein sequences were searched against a version of the genome masked for 6a annotations and filtered using four criteria: e value ($<1e^{-5}$), identity (greater than 40%), length (more than 30 amino acids), and coverage of the query sequence (5%). Using a maximum interval of 2,032 bp (95th percentile intron length), 510,259 pseudoexons were combined into putative pseudogenes, which were subsequently filtered if they overlapped with annotated gene regions and/or known Viridiplantae repeats. Note that some of these putative pseudogenes are substantially shorter than their annotated, presumably functional, paralogs but do not have disabling mutations (stop or frame shift). In addition, some pseudogenes may be functional genes that are split between contigs or scaffolds. Thus, we only examined putative pseudogenes with one or more disabling mutations or those located distantly from the ends of contigs based on a threshold distance. This threshold distance is defined as the sum of the 95th percentile intron length and a consideration of functional paralog length. Suppose a functional paralog to a pseudogene has length L and the pseudogene match is from $M1$ and $M2$, functional paralog length is defined as the larger of $M1$ or $L - M2$.

Supplemental Data

The following supplemental materials are available.

Supplemental Figure S1. Comparing two versions of trained Augustus within MAKER-P on Chromosome 10.

Supplemental Table S1. RNA-seq data sources used for transcript assemblies.

Supplemental Table S2. Provisional 5b+ gene models.

Supplemental Table S3. Small RNA-seq experiments used in miRNA identification.

Supplemental Table S4. MapMan terms and statistics.

Supplemental Text S1. GFF file containing the tRNA, miRNA, and snoRNA predictions.

ACKNOWLEDGMENTS

We thank iPlant, Texas Advanced Computing Center, and MaizeGDB support personnel for their efforts.

Received June 19, 2014; accepted November 2, 2014; published November 10, 2014.

LITERATURE CITED

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol **215:** 403–410

Benson DA, Cavanaugh M, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2013) GenBank. Nucleic Acids Res **41:** D36–D42

Blanc G, Wolfe KH (2004) Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. Plant Cell **16:** 1679–1691

Bousios A, Minga E, Kalitsou N, Pantermali M, Tsaballa A, Darzentas N (2012) MASiVEdb: the Sirevirus Plant Retrotransposon Database. BMC Genomics **13:** 158

Bowers JE, Chapman BA, Rong J, Paterson AH (2003) Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. Nature **422:** 433–438

Campbell MS, Law M, Holt C, Stein JC, Moghe GD, Hufnagel DE, Lei J, Achawanantakun R, Jiao D, Lawrence CJ, et al (2014) MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. Plant Physiol **164:** 513–524

Du C, Fefelova N, Caronna J, He L, Dooner HK (2009) The polychromatic Helitron landscape of the maize genome. Proc Natl Acad Sci USA **106:** 19916–19921

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

Feschotte C, Wessler SR (2001) Treasures in the attic: rolling circle transposons discovered in eukaryotic genomes. Proc Natl Acad Sci USA **98:** 8923–8924

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, et al (2014) Pfam: the protein families database. Nucleic Acids Res **42:** D222–D230

Gaut BS, Doebley JF (1997) DNA sequence evidence for the segmental allotetraploid origin of maize. Proc Natl Acad Sci USA **94:** 6809–6814

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol **29:** 644–652

Griffiths-Jones S, Saini HK, van Dongen S, Enright AJ (2008) miRBase: tools for microRNA genomics. Nucleic Acids Res **36:** D154–D158

Guigó R, Flicek P, Abril JF, Reymond A, Lagarde J, Denoeud F, Antonarakis S, Ashburner M, Bajic VB, Birney E, et al (2006) EGASP: the human ENCODE Genome Annotation Assessment Project. Genome Biol (Suppl 1) **7:** S2

Haas BJ, Delcher AL, Wortman JR, Salzberg SL (2004) DAGchainer: a tool for mining segmental genome duplications and synteny. Bioinformatics **20:** 3643–3646

Hanada K, Zou C, Lehti-Shiu MD, Shinozaki K, Shiu SH (2008) Importance of lineage-specific expansion of plant tandem duplicates in the adaptive response to environmental stimuli. Plant Physiol **148:** 993–1003

Holt C, Yandell M (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics **12:** 491

Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H, et al (2011) The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. Nat Genet **43:** 476–481

Jiang N, Ferguson AA, Slotkin RK, Lisch D (2011) Pack-Mutator-like transposable elements (Pack-MULEs) induce directional modification

of genes through biased insertion and DNA acquisition. Proc Natl Acad Sci USA **108:** 1537–1542

Korf I (2004) Gene finding in novel genomes. BMC Bioinformatics **5:** 59

Kozomara A, Griffiths-Jones S (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res **42:** D68–D73

Lei J, Sun Y (2014) miR-PREFeR: an accurate, fast and easy-to-use plant miRNA prediction tool using small RNA-Seq data. Bioinformatics **30:** 2837–2839

Liang C, Mao L, Ware D, Stein L (2009) Evidence-based gene predictions in plant genomes. Genome Res **19:** 1912–1923

Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res **25:** 955–964

Lowe TM, Eddy SR (1999) A computational screen for methylation guide snoRNAs in yeast. Science **283:** 1168–1171

Meyers BC, Axtell MJ, Bartel B, Bartel DP, Baulcombe D, Bowman JL, Cao X, Carrington JC, Chen X, Green PJ, et al (2008) Criteria for annotation of plant microRNAs. Plant Cell **20:** 3186–3190

Monaco MK, Stein J, Naithani S, Wei S, Dharmawardhana P, Kumari S, Amarasinghe V, Youens-Clark K, Thomason J, Preece J, et al (2014) Gramene 2013: comparative plant genomics resources. Nucleic Acids Res **42:** D1193–D1199

Neale DB, Wegrzyn JL, Stevens KA, Zimin AV, Puiu D, Crepeau MW, Cardeno C, Koriabine M, Holtz-Morris AE, Liechty JD, et al (2014) Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. Genome Biol **15:** R59

Qu F, Ye X, Morris TJ (2008) Arabidopsis DRB4, AGO1, AGO7, and RDR6 participate in a DCL4-initiated antiviral RNA silencing pathway negatively regulated by DCL1. Proc Natl Acad Sci USA **105:** 14732–14737

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, et al (2009) The B73 maize genome: complexity, diversity, and dynamics. Science **326:** 1112–1115

Sen TZ, Andorf CM, Schaeffer ML, Harper LC, Sparks ME, Duvick J, Brendel VP, Cannon E, Campbell DA, Lawrence CJ (2009) MaizeGDB becomes "sequence-centric." Database (Oxford) **2009:** bap020

Seoighe C, Gehring C (2004) Genome duplication led to highly selective expansion of the Arabidopsis thaliana proteome. Trends Genet **20:** 461–464

Shiu SH, Shih MC, Li WH (2005) Transcription factor families have much higher expansion rates in plants than in animals. Plant Physiol **139:** 18–26

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

Stanke M, Waack S (2003) Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics (Suppl 2) **19:** ii215–ii225

Takacs EM, Li J, Du C, Ponnala L, Janick-Buckner D, Yu J, Muehlbauer GJ, Schnable PS, Timmermans MCP, Sun Q, et al (2012) Ontogeny of the maize shoot apical meristem. Plant Cell **24:** 3219–3234

Thimm O, Bläsing O, Gibon Y, Nagel A, Meyer S, Krüger P, Selbig J, Müller LA, Rhee SY, Stitt M (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J **37:** 914–939

Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E (2009) EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. Genome Res **19:** 327–335

Wegrzyn JL, Liechty JD, Stevens KA, Wu LS, Loopstra CA, Vasquez-Gross HA, Dougherty WM, Lin BY, Zieve JJ, Martínez-García PJ, et al (2014) Unique features of the loblolly pine (Pinus taeda L.) megagenome revealed through sequence annotation. Genetics **196:** 891–909

Wei F, Zhang J, Zhou S, He R, Schaeffer M, Collura K, Kudrna D, Faga BP, Wissotski M, Golser W, et al (2009) The physical and genetic framework of the maize B73 genome. PLoS Genet **5:** e1000715

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

Yang L, Bennetzen JL (2009) Distribution, diversity, evolution, and survival of Helitrons in the maize genome. Proc Natl Acad Sci USA **106:** 19922–19927

Youens-Clark K, Buckler E, Casstevens T, Chen C, Declerck G, Derwent P, Dharmawardhana P, Jaiswal P, Kersey P, Karthikeyan AS, et al (2011) Gramene database in 2010: updates and extensions. Nucleic Acids Res **39:** D1085–D1094

Zhang L, Chia JM, Kumari S, Stein JC, Liu Z, Narechania A, Maher CA, Guill K, McMullen MD, Ware D (2009) A genome-wide characterization of microRNA genes in maize. PLoS Genet **5:** e1000716

Zou C, Lehti-Shiu MD, Thibaud-Nissen F, Prakash T, Buell CR, Shiu SH (2009) Evolutionary and expression signatures of pseudogenes in Arabidopsis and rice. Plant Physiol **151:** 3–15

CHAPTER 5

ADDING THE FUNCTIONALITY OF A COMBINER TO THE MAKER GENOME ANNOTATION PIPELINE

Introduction

I have incorporated EVidence Modeler (EVM)[1] into the MAKER genome annotation pipeline[2]. Previously the MAKER genome annotation pipeline produced structural genome annotations by running multiple gene finders across a genome and giving them hints derived from aligned evidence when possible. MAKER would then choose the gene model that best matched the evidence and add three- and five-prime untranslated regions (UTRs) based on expression data to serve as the final structural annotation[2]. Though MAKER can modify the three and five prime end of genes to better match the evidence, it cannot change the internal exon structure. EVM belongs to a class of gene prediction tools called Combiners. Combiners chose a combination of exons for each gene model based on aligned evidence and *ab initio* gene predictions[3]. EVM combines evidence types based on user supplied weights and known error profiles in a manner that minimizes error[1]. In a recent gene prediction competition, combiners were found to outperform all of the other gene prediction algorithms[4]. I hypothesized that incorporating EVM into MAKER would improve the exon level sensitivity of the MAKER annotation pipeline. I chose to use the right arm of *Drosophila melanogaster*

chromosome III for these comparisons. The protein-coding gene annotations for *Drosophila melanogaster* have gone trough extensive manual curation[5], making them suitable for use as a truth set, and the right arm of chromosome III contains enough genes to train gene finders.

There are several places that EVM could be employed in the MAKER pipeline. EVidence Modeler could be employed upstream of the MAKER annotation and the outputs of EVidence Modeler could be given to MAKER as gene predictions to compete against models produced by other gene finders, such as SNAP[6] and Augustus[7,8]. This approach was used independently of the Yandell Lab in the reannotation of the *anolis* lizard genome, and led to significantly improved annotations[9]. EVidence Modeler could also be used downstream of MAKER. In this scenario, EVidence Modeler would be given the MAKER output in GFF3[10] format and allowed to make new models based on the aligned evidence and MAKER annotations. These options require users to run EVM independently of MAKER, failing to capitalize on MAKER's advanced parallel computing abilities[11]. To maintain efficient parallel computing, EVM can be employed within MAKER. Currently MAKER aligns protein and mRNA-seq/EST evidence to the genome using BLAST and Exonerate. MAKER then runs a series of *ab initio* gene predictors[2]. These alignments and gene predictions are the required inputs for EVM[1]. For this experiment MAKER passes aligned evidence and gene predictions to EVM and runs EVM internally. EVM output gene models reenter the MAKER pipeline as gene predictions to compete with the *ab initio* gene predictions. See Figure 5.1 for a graphical representation of the MAKER workflow with EVM running internally.

Results and discussion

Running EVM within MAKER resulted in minimal improvement in exon and nucleotide accuracy, and slightly worse gene accuracy (see Table 5.1). Interestingly, the average gene length for Flybase genes is much larger than the average gene length for any of the other annotation sets, while the median lengths are comparable. This observation suggests a small number of very large genes are responsible for the difference in gene length. The average and median exon lengths are similar between all of the annotation sets, but the average intron is much longer in the flybase genes while the median length is similar to that of the other annotation sets. This suggests that a small number of genes with very large introns are responsible for the observed differences in average gene length. (See Table 5.2.) Introns larger than 10kb are challenging to annotate, using computational methods in the *Drosophila melanogaster* genome because they are rare. In these experiments, the settings used to run MAKER would rarely annotate a gene with an intron larger than 10kb. The BLAST parameters in MAKER could be relaxed to support larger intron identification, but would likely have deleterious effects on the other gene models. It is most likely that the largest genes in the *Drosophila* genome were manually annotated.

Methods

Incorporating EVidence Modeler into MAKER

Subroutines were added to the GI.pm, MpiChunk.pm, auto_annotator.pm perl modules, and the MAKER executable. A new widget module was written to run EVM (Widget::evm), and PhatHit and PhatHSP modules were written to convert EVM output

to a MAKER-usable format (Bio::Search::Hit::PhatHit::evm and Bio::Search::HSP:: PhatHSP::evm).

Genomic DNA and gold standard gene model acquisition

The right arm of chromosome three was downloaded from NCBI (ftp://ftp.flybase.net/genomes/Drosophila_melanogaster/dmel_r6.03_FB2014_06/fasta/dmel-all-chromosome-r6.03.fasta.gz). The Flybase models were downloaded from Flybase in GFF3 format (ftp://ftp.flybase.net/releases/FB2014_06/dmel_r6.03/gff/dmel-all-no-analysis-r6.03.gff.gz). The transcript with the longest coding sequence was chosen as a representative transcript for calculating sensitivity, specificity, and accuracy for alternatively spliced genes.

mRNA-Seq acquisition, alignment, and assembly

Publicly available mRNA-seq reads were downloaded from the sequence read archive (ftp://ftp-trace.ncbi.nlm.nih.gov/sra/sra-instant/reads/ByRun/sra/SRR/SRR191/SRR1914096/SRR1914096.sra). Adapters and low-quality bases were removed from the reads using SeqyClean[12]. Transcripts were assembled using reference-guided and reference-free methods. For the reference-guided approach, reads were aligned to the genome using Tophat2[13] and assembled using StringTie[14]. For the reference-free approach, reads were assembled using Trinity[15].

Protein evidence

Whole proteomes from *Drosphila simulans* and *C. elegans* were downloaded using the NCBI taxonomy browser (txid7240 limited to refseq proteins) and from wormbase, respectively (ftp://ftp.wormbase.org/pub/wormbase/species/c_elegans/sequen

ce/protein/c_elegans.current.protein.fa.gz). Uniprot-swissprot was downloaded from the Uniprot website, and all *Drosophila melanogaster* sequences were removed (ftp://ftp.uniprot.org/pub/databases/uniprot/previous_major_releases/release-2015_01 /knowledgebase/uniprot_sprot-only2015_01.tar.gz).

Repeat masking

The *Drosophila melanogaster* repeats in Repbase[16] and a collection of known transposable element proteins distributed with MAKER were used to mask the genome[2]. Further soft masking of low-complexity sequence was done by BLAST prior to evidence alignment[17].

Training gene finders

The recently introduced BRAKER pipeline was used to train Augustus using the aligned mRNA-seq reads described above[18]. The gene models produced by BRAKER were then used to train SNAP which was then further trained using the iterative process described previously[2].

Running MAKER

MAKER was run five times: once using Basic Protocol 3 from[19] to add quality metrics to the gene models from Flybase, and four times using Basic Protocol 1 from[19] to generate the SNAP only; Augustus Only; SNAP and Augustus; and SNAP, Augustus, and EVM annotation sets. For each run protein evidence and Trinity-assembled transcripts were passed to MAKER in FASTA format. StringTie transcripts were passed to MAKER in GFF3 format. Repeats were masked by Repeatmasker using the inputs described above.

Calculating sensitivity, specificity, and accuracy

Sensitivity and specificity were calculated using the evaluate_gff.pl script in the eval software package[20]. Sensitivity is defined as the true positives divided by the true positives plus the false negatives. Specificity is defined as the true positives divided by the true positives plus the false positives. Accuracy was reported as the sensitivity plus the specificity divided by two[3]. Calculations were limited to the coding sequence (CDS) of each gene. To qualify as a true positive, a given feature in the predictions must match a feature in the Flybase annotations perfectly. A false positive is defined as a feature present in the predicted set, but not in the Flybase annotations, and a false negative is a feature present in the Flybase annotations and not present in the predictions.

Conclusion

Incorporating EVM into the MAKER pipeline did not appreciably improve the gene, transcript, exon, or nucleotide level sensitivity, specificity, or accuracy of MAKER annotations above that of the default MAKER annotation pipeline for the *Drosophila melanogaster* genome. Though the *Drosophila melanogaster* genome is a popular genome for benchmarking gene predictors, it may not be the best genome for these experiments. Future benchmarking of MAKER with EVM on more complex genomes, such as those found in mammals or conifers, may show greater improvements in annotation accuracy.

References

1. Haas, B. J. *et al.* Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9,** R7 (2008).

2. Cantarel, B. L. *et al.* MAKER : An easy-to-use annotation pipeline designed for emerging model organism genomes. 188–196 (2008). doi:10.1101/gr.6743907.1

3. Yandell, M. & Ence, D. A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.* **13,** 329–42 (2012).

4. Coghlan, A. *et al.* nGASP--the nematode genome annotation assessment project. *BMC Bioinformatics* **9,** 549 (2008).

5. flybase. at <flybase.org>

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

7. Stanke, M. & Waack, S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics* **19,** ii215–ii225 (2003).

8. Stanke, M., Diekhans, M., Baertsch, R. & Haussler, D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24,** 637–44 (2008).

9. Eckalbar, W. L. *et al.* Genome reannotation of the lizard Anolis carolinensis based on 14 adult and embryonic deep transcriptomes. *BMC Genomics* **14,** 49 (2013).

10. Generic Feature Format version 3. at <http://www.sequenceontology.org/gff3.shtml>

11. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12,** 491 (2011).

12. Zhbannikov, I. SeqyClean. (2015). at <https://bitbucket.org/izhbannikov/seqyclean>

13. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14,** R36 (2013).

14. Pertea, M. *et al.* StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33,** (2015).

15. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29,** 644–52 (2011).

16. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110,** 462–7 (2005).

17. Korf, I., Yandell, M. & Bedell, J. *Blast*. (O'Reilly, 2003).

18. Lange, S., Hoff, K. J., Lomsadze, A., Borodovsky, M. & Stanke, M. BRAKER. (2015). at <http://exon.gatech.edu/GeneMark/braker1.html>

19. Campbell, M. S., Holt, C., Moore, B. & Yandell, M. *Genome Annotation and Curation Using MAKER and MAKER-P. Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* **48,** (2014).

20. Keibler, E. & Brent, M. R. Eval: a software package for analysis of genome annotations. *BMC Bioinformatics* **4,** 50 (2003).

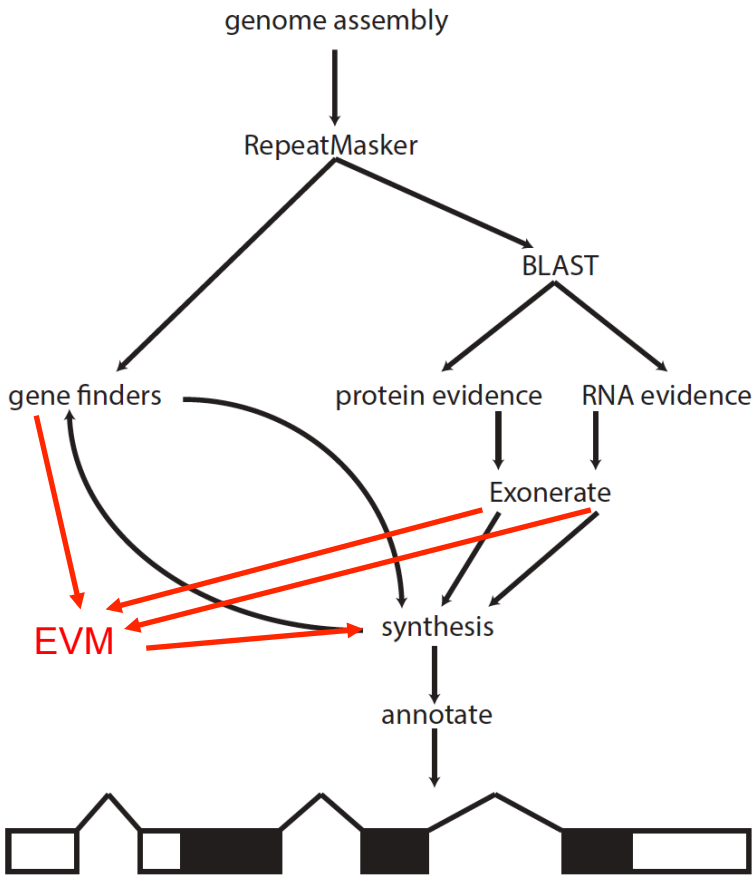

Figure 5.1. Position of EVM in the MAKER annotation pipleline. EVM receives aligned evidence and gene predictions as input. EVM-generated gene models are then included in the super set of gene prediction available to MAKER. MAKER then choses the gene prediction that best matches the evidence in the synthesis step to continue to the annotation step, where three and five prime UTR are added and quality metrics are calculated. Figure adapted from [19].

Table 5.1.

Sensitivity, specificity, and accuracy of the MAKER annotation pipeline with SNAP, Augustus, and EVM as gene predictors

SNAP	+	-	+	+
Augustus	-	+	+	+
EVM	-	-	-	+
Gene Sensitivity	42.10%	54.77%	52.84%	52.56%
Gene Specificity	44.13%	54.26%	52.63%	52.23%
Gene Accuracy	**43.12%**	**54.52%**	**52.74%**	**52.40%**
Exon Sensitivity	75.47%	78.94%	79.58%	79.37%
Exon Specificity	71.25%	76.43%	75.04%	75.26%
Exon Accuracy	**73.36%**	**77.69%**	**77.31%**	**77.32%**
Nucleotide Sensitivity	90.27%	93.85%	94.22%	94.16%
Nucleotide Specificity	91.35%	91.97%	91.84%	91.97%
Nucleotide Accuracy	**90.81%**	**92.91%**	**93.03%**	**93.07%**

Table 5.2.

Basic annotation metrics

SNAP	+	-	+	+	Flybase
Augustus	-	+	+	+	
EVM	-	-	-	+	
Protein coding genes	3,102	3,284	3,268	3,276	3,261
Gene length average (median)	4,782 (2,651) bp	3,667 (2,044) bp	4,050 (2,226) bp	3,970 (2,207) bp	6,020 (2,142) bp
Exons per mRNA average (median)	4.75 (4)	4.39 (3)	4.53 (3)	4.50 (3)	4.51 (3)
Exon length average (median)	413 (237) bp	439 (247) bp	436 (246) bp	438 (246) bp	496 (280) bp
Intron length average (median)	752 (76) bp	504 (70) bp	580 (71) bp	565 (71) bp	1,076 (70) bp

CHAPTER 6

COMMUNITY OUTREACH: BROADENING THE SCIENTIFIC IMPACT OF THE MAKER GENOME ANNOTATION PIPELINE

My software development efforts involving the MAKER genome annotation pipeline have been funded by the National Science Foundation (NSF). NSF grants and progress reports are evaluated for intellectual merit and broader impacts. Intellectual merit is well understood, but broader impacts can be more difficult to interpret. Broader impacts include activities that promote teaching, training, and learning, and especially those that encourage the involvement of underrepresented groups and enhance the infrastructure for research and education[1]. To meet these broader impact criteria, I have relied on community outreach activities.

To be useful in advancing science, a software package must be used. For a tool to be used, a user must first know of the tool, and second, know how to use the tool. Publishing a paper describing the tool is a good start, but is ultimately a passive approach to establishing a user base because it requires users to find the tool on their own. Community outreach efforts actively introduce users to a tool and teach them how to use it. Community outreach began early in the development of the MAKER annotation pipeline and was accelerated when MAKER became a part of the Generic Model Organism Database (GMOD) project[2]. These early outreach efforts were lead by Carson

Holt during his graduate training in the Yandell lab[3] (University of Utah[4]) and included speaking and conducting hands-on training sessions at conferences, teaching genome annotation courses at the GMOD summer school, and maintaining a MAKER developers mailing list where MAKER users could get help with their annotation projects[5].

These outreach efforts led to the worldwide adoption of the MAKER genome annotation pipeline (see Figure 6.1), with a combined total of 212 citations for the MAKER and MAKER2 publications. As the number of MAKER users has grown, so have the help requests sent to the MAKER developers list. These requests placed an additional burden on MAKER developers' time, thus impairing software development, data analysis, and other research-related activities. Motivated by the belief that improving the MAKER documentation would mitigate many of these requests, I created a media wiki called MAKER wiki (http://weatherby.genetics.utah.edu/MAKER/wiki/index .php/Main_Page) that was linked to the MAKER download page, and the MAKER developers list was made searchable. The wiki has had over 11,000 hits and the MAKER development team has observed a substantial decrease in help requests through the developers list. I have also written a unit on genome annotation using MAKER for *Current Protocols in Bioinformatics*[6]. The protocols and data from this publication served as the foundation for a genome annotation class taught by Marvin B. Moore, Director of Science and Research at the USTAR Center for Genetic Discovery[7] at the University of Utah[4], to researchers at the University of Johannesburg[8] in 2014.

To reach users through nonprint media I have spoken at conferences, including Plant and Animal Genomes XXI[9] and the Society for Molecular Biology and Evolution 2014 meeting[10], and taught hands-on genome annotation courses, including GMOD

Summer School 2013[11], and GMOD Malaysia 2014[12]. While in Malaysia I taught sessions on the GFF3 format and the WebApollo annotation viewer and editor, in addition to genome annotation using MAKER. My efforts at GMOD summer school 2013 resulted in an acknowledgement on the golden eagle genome publication[13].

The wiki, protocols paper, and my personal teaching/training efforts have clearly contributed to teaching, training, and learning as outlined by the NSF. These materials and activities have also aided Malaysian and South African genomics research, encouraging the involvement of underrepresented groups and enhancing the infrastructure for genomics research in these countries.

References

1. March, P. Broader Impacts Review Criterion. *National Science Foundation* (2008). at <http://www.nsf.gov/pubs/2007/nsf07046/nsf07046.jsp>
2. GMOD. at <www.gmod.org>
3. Yandell Lab. at <http://www.yandell-lab.org/>
4. University of Utah. at <http://www.utah.edu/>
5. Holt, C. H. Tools and Techniques for Genome Annotation Analysis. (2011).
6. Campbell, M. S., Holt, C., Moore, B. & Yandell, M. *Genome Annotation and Curation Using MAKER and MAKER-P. Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.]* **48,** (2014).
7. USTAR Center for Genetic Discovery. at <http://www.innovationutah.com/>
8. University of Johannesburg. at <http://www.uj.ac.za/EN/Pages/Home.aspx>
9. Campbell, M. MAKER. in *PAG XXI* (2013). at <https://pag.confex.com/pag/xxi/webprogram/Paper7243.html>
10. Society for Molecular Biology and Evolution. at <www.smbe.org>

11. GMOD Summer School 2013. (2013). at <http://www.gmod.org/wiki/2013_GMOD_Summer_School>

12. GMOD Malaysia. (2014). at <http://gmod.org/wiki/GMOD_Malaysia_2014>

13. Doyle, J. M. *et al.* The genome sequence of a widespread apex predator, the golden eagle (Aquila chrysaetos). *PLoS One* **9,** 20–22 (2014).

Figure 6.1. World-wide adoption of MAKER. Each red balloon represents a registered download of the MAKER pipeline. MAKER users are located throughout the world on every continent excepting Antarctica. Image courtesy of Moore.

CHAPTER 7

GENOME ANNOTATION AND CURATION USING MAKER AND MAKER-P

The following is a reprint of an article coauthored by myself, Carson Holt, Barry Moore, and Mark Yandell. This article was originally published in *Current Protocols in Bioinformatics* 2014, 48:4.11.1-4.11.39 and is used with permission.

Personal contribution

I developed the protocols, assembled the test data, and wrote the manuscript.

Genome Annotation and Curation Using MAKER and MAKER-P

Michael S. Campbell,[1] Carson Holt,[1,2] Barry Moore,[1,2] and Mark Yandell[1,2]

[1]Eccles Institute of Human Genetics, University of Utah, Salt Lake City, Utah
[2]USTAR Center for Genetic Discovery, University of Utah, Salt Lake City, Utah

This unit describes how to use the genome annotation and curation tools MAKER and MAKER-P to annotate protein-coding and noncoding RNA genes in newly assembled genomes, update/combine legacy annotations in light of new evidence, add quality metrics to annotations from other pipelines, and map existing annotations to a new assembly. MAKER and MAKER-P can rapidly annotate genomes of any size, and scale to match available computational resources. © 2014 by John Wiley & Sons, Inc.

Keywords: genome annotation • comparative genomics • gene finding • plants

How to cite this article:
Campbell, M. S., Holt, C., Moore, B. and Yandell, M. 2014.
Genome Annotation and Curation Using MAKER and MAKER-P.
Curr. Protoc. Bioinform. 48:4.11.1-4.11.39.
doi: 10.1002/0471250953.bi0411s48

INTRODUCTION

In this unit, we describe the MAKER genome annotation and curation pipeline. All of the input files used in the following protocols are found in `CPB_MAKER.tar.gz`, available for download at *http://weatherby.genetics.utah.edu/CPB_MAKER/CPB_MAKER.tar.gz*. Also described is MAKER-P, a version of MAKER optimized for plant genome annotation efforts that offers a number of new functionalities such as ncRNA annotation capabilities and support for pseudogene identification (Zou et al., 2009; Campbell et al., 2014). Both MAKER and MAKER-P are available for download from *http://www.yandell-lab.org*. MAKER-P is also installed in the Texas Advanced Computing Center as part of the iPlant Cyberinfrastructure (Goff et al., 2011); see *https://pods.iplantcollaborative.org/wiki/display/sciplant/MAKER-P+at+iPlant* and *UNIT 1.22* in this manual.

MAKER and MAKER-P annotate and mask repetitive elements in the genome, and align protein and RNA evidence to the assembly, in a splice-aware fashion to accurately identify splice sites. They also run multiple ab initio gene predictors, compare all predicted gene models to RNA and protein alignment evidence, and then revise the ab initio gene models in light of this evidence. The best supported gene models are chosen using a quality metric called Annotation Edit Distance (AED), developed by the Sequence Ontology (Eilbeck et al., 2009). MAKER and MAKER-P's outputs include FASTA files (Lipman and Pearson, 1985; see *APPENDIX 1B* for description of FASTA format) of transcripts and proteins for each annotated gene, and GFF3 (Generic Feature Format version 3; see Internet Resources) files that describe the gene models and their supporting evidence. These GFF3 files also provide a number of quality metrics (including AED) for each gene model. This basic workflow is visually represented in Figure 4.11.1.

Though MAKER was originally developed for de novo annotation of emerging model organisms, it has expanded into a multiuse genome annotation and curation tool (Holt

Copyright © 2014 John Wiley & Sons, Inc.

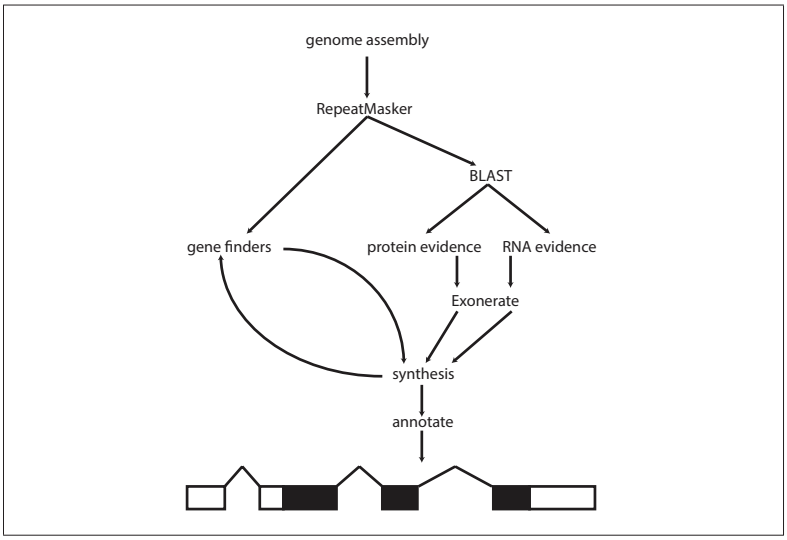

Figure 4.11.1 MAKER annotation workflow. MAKER masks repeats with RepeatMasker, aligns evidence to the genome with BLAST, polishes those alignments around splice sites using Exonerate, and runs a number of gene finders. MAKER also feeds evidence-based hints to the gene finders in order to improve their accuracy. These data are then synthesized into gene annotations.

and Yandell, 2011). In addition to de novo annotation, MAKER and MAKER-P can also be used to update existing annotations in the light of new experimental evidence and for quality control of gene models produced by other annotation pipelines (Campbell et al., 2013; Law et al., 2014)

MAKER and MAKER-P are both highly parallelized applications with support for the Message Passing Interface (MPI); this allows them to efficiently utilize multiple CPUs. Given enough CPUs, MAKER can annotate large mammalian and plant genomes in hours (Campbell et al., 2013). MAKER-P, which is available in massively parallel mode as part of the iPlant project (Goff et al., 2011), is even more powerful. For example, it was recently used to annotate the entire 22-GB loblolly pine genome assembly in less than 24 hr using over 8000 CPUs (Neale et al., 2014; Zimin et al., 2014). The highly parallelized architectures of MAKER and MAKER-P mean that users can experiment with alternate parameters and datasets to optimize annotation quality. It also makes it trivial to regularly update annotations as new evidence and assemblies become available.

STRATEGIC PLANNING

Know your organism

Knowledge of your organism's phylogenetic relationships and any previously annotated close relatives is crucial. The NCBI taxonomy browser can help identify closely related organisms and help find corresponding transcript and protein sequences to use as evidence while annotating your genome. UniProt/Swiss-Prot, NCBI genomes, and Ensembl are good places to look for protein data, while the sequence read archive (SRA), Genbank, and Ensembl are good places to look for RNA evidence.

Get the best assembly you can

MAKER has been used successfully on genomes derived from many different sequencing platforms and assemblers. For a comparison of assemblers, see the Assemblathon 2 paper

(Bradnam et al., 2013). As a rule of thumb, if the scaffold N50 of your assembly is less than the expected average gene length (including introns and UTR), the assembly should be improved before attempting to annotate it with MAKER (Yandell and Ence, 2012). You should also consider evaluating the "completeness" of the assembly using tools like CEGMA (Parra et al., 2007), which can indicate the upper limit of recoverable gene content from draft assemblies.

Sequence the genome with its eventual annotation in mind

Use a portion of your genome-sequencing budget to produce expression data. mRNA-seq data from multiple tissue types and stages of development helps greatly with gene annotation. Likewise, small RNA-seq data sets provide evidence to support ncRNA annotations.

DE NOVO GENOME ANNOTATION USING MAKER

BASIC PROTOCOL 1

Identifying the protein-coding genes in a newly assembled genome is a common first step in genome analysis. These protein-coding gene annotations enable further computational analyses and serve as the basis for diverse molecular biology experiments. Successful downstream analyses and experiments are contingent upon the quality of the underlying gene annotations.

Necessary Resources

Hardware

Computer with a Unix-based operating system (e.g., Linux, Mac OS X)

Software

MAKER and MAKER-P are available for download at *yandell-lab.org*. Installation instructions are included in the tarball. For brevity's sake, the following protocols describe MAKER, but apply to MAKER-P as well.

MAKER will identify and download all of its necessary external dependencies including BLAST, Exonerate, RepeatMasker, and a number of Perl modules [automatic download and installation of Perl modules requires CPAN (*https://metacpan.org/pod/CPAN*) to be installed]. MAKER will also install a number of additional programs such as SNAP, Augustus, and MPICH2. This example uses a version of MAKER installed with NCBI BLAST+, Exonerate, RepeatMasker, with optional RepBase libraries, and SNAP.

Files

Genome assembly to be annotated in FASTA format

Protein evidence in FASTA format

Assembled mRNA-seq transcripts from the species of interest in FASTA format

Optional: a species parameter/HMM file for SNAP generated for the organism of interest or a closely related species. The process used to create a species parameter/HMM file is described in SNAP's internal documentation (Korf, 2004).

1. From the Unix command line (using the "bash" shell), generate the MAKER control files (in the text below, lines that start with `%` show the command prompt; the `%` should not be typed; lines starting with `#` are comments and should not be typed):

```
% maker -CTL
```

This command generates three files: `maker_opts.ctl`, `maker_bopts.ctl`, *and* `maker_exe.ctl`. *User input is given to MAKER through these three files. For a detailed explanation of the options and parameters in the* `*.ctl`, *files, please see Critical Parameters and Advanced Parameters sections, below.*

2. Edit the `maker_opts.ctl` file to specify the genome assembly sequence, experimental alignment evidence, and which gene-finding method to use. Any text editor will work, but for purposes of this protocol we will use 'emacs':

```
% emacs maker_opts.ctl
#-----Genome (these are always required)
genome=$PATH_TO_CBP_maker_inputs/dpp_data/dpp_contig
  .fasta
#genome sequence (fasta file or fasta embeded in GFF3
  file)
organism_type=eukaryotic #eukaryotic or prokaryotic.
  Default is eukaryotic
#-----EST Evidence (for best results provide a file for
  at least one)
est=$PATH_TO_CBP_makerinputs/dpp_data/dpp_est.fasta
#set of ESTs or assembled mRNA-seq in fasta format
#-----Protein Homology Evidence (for best results
  provide a file for at least one)
protein=$PATH_TO_CBP_maker_inputs/dpp_data/dpp
  _protein.fasta
#protein sequence file in fasta format (i.e., from
  multiple organisms)
#-----Gene Prediction
snaphmm=$PATH_TO_CBP_maker_inputs/dpp_data/
  D.melanogaster.hmm #SNAP HMM file
```

Relative or absolute paths can be used in all of the `*.ctl` files. To ensure proper parsing of these files, make sure that there are no spaces between the equal sign and the path to the files. With the exception of the `genome=` parameter, multiple files can be given to MAKER as a comma-separated list of paths. Protein evidence and mRNA-seq data are commonly given to MAKER in multiple files to better keep track of evidence sources in the final outputs (see Support Protocol 4).

3. Run MAKER:

```
% maker 2> maker.error
```

The locations of the control files for a MAKER run can be specified on the command line. If they are not specified, the control files in the current working directory are used. As MAKER runs, it will output a number of progress messages to the screen along with any error messages (you can reduce the volume of messages by running MAKER with a `-q` (quiet mode) to limit the status messages, or `-qq` (very quiet mode) to eliminate everything but errors). It is often helpful to save these status and error messages to a file for future reference, which is what was done on the above command line with the `2>` redirect).

In addition to status and warning messages, MAKER creates an output directory named after the input genome FASTA file (if you would rather specify the name of the output directory you can do that on the command line by using the `-base` option). In this example, the name of the output directory is `dpp_contig.maker.output`. After MAKER runs, you will find a number of additional files and directories inside this output directory. Of primary interest are the `datastore` directory and the `datastore` index log (both of which are named after the base name if given on the command line, or by default using the name of the genome FASTA file).

Because genome annotation can produce hundreds of files for each of tens of thousands of contigs in the assembly, the MAKER datastore directory uses a hashed directory tree structure to separate the outputs for individual contigs/scaffolds from your assembly. Inside each contig directory you can find all of the genome annotation results that pertain to that

contig, together with a number of intermediate files that are saved to speed up subsequent MAKER runs. The datastore index log is the key to easily locating results in the datastore directory. It provides the final path to output for every annotated contig. The datastore index also indicates the run status of each contig processed (whether a contig has started, finished, failed, or was skipped).

4. Check the standard error output and datastore index file to see if MAKER is finished:

```
% tail --n 2 maker.error
maker is now finished!!!
% cat\
dpp_contig.maker.output/dpp_contig_master_datastore
  _index.log
contig-dpp-500-500
  dpp_contig_datastore/05/1F/contig-dpp-500-500/ STARTED
contig-dpp-500-500
  dpp_contig_datastore/05/1F/contig-dpp-500-500/
  FINISHED
```

If everything went well, the last line of the `MAKER.error` *file will read* `MAKER is now finished!!!`*, and the datastore index log will have an entry for when MAKER started each entry in the genome FASTA file and when it finished or failed that entry. Since we have only one entry in our genome FASTA file, we have only two entries in our datastore index log. In this case, MAKER finished running and successfully finished our contig.*

5. Collect the results from all individual contigs into genome wide annotations:

```
% gff3_merge -d
dpp_contig.maker.output/dpp_contig_master_datastore
  _index.log
dpp_contig.all.gff
% fasta_merge -d
dpp_contig.maker.output/dpp_contig_master_datastore
  _index.log
dpp_contig.all.maker.proteins.fasta
dpp_contig.all.maker.transcripts.fasta
dpp_contig.all.maker.snap_masked.proteins.fasta
dpp_contig.all.maker.snap_masked.transcripts.fasta
dpp_contig.all.maker.non_overlapping_ab_initio
  .proteins.fasta
dpp_contig.all.maker.non_overlapping_ab_initio
  .transcripts.fasta
```

MAKER uses two output formats, GFF3 and FASTA. Gene predictions, evidence alignments, repetitive elements, and the final gene models are output in GFF3 format, while transcript and protein sequences are output in FASTA format. Here we used two of the accessory scripts distributed with MAKER to collect the GFF3 and FASTA results from individual contigs and merge them to provide genome-wide results. These scripts use the directory paths present in the datastore index log to find the relevant files for each contig.

After merging, you will have a single GFF3 file, together with protein and transcript sequences of the MAKER annotations. Depending upon runtime parameters, MAKER's outputs may also include additional FASTA files for the ab initio gene predictions, and/or rejected gene predictions with no evidence support that do not overlap a MAKER annotated gene. These additional files are given to the user for reference and evaluation purposes, and their presence depends on the user defined setting in the `MAKER_opts.ctl` *file. For example, if the* `keep_preds` *parameter in the* `maker_opts.ctl` *file is set to* `1`*, there will not be FASTA output for non-overlapping ab initio predictions because they will all be contained in the* `maker-transcripts` *and* `maker-protein` *files. These files may*

also be absent if every locus with a gene prediction was supported by evidence and thus annotated by MAKER.

Once the GFF3 and FASTA files are merged together, the structural protein-coding gene annotation is complete. Subsequent protocols document post-processing options and functional annotation protocols available to MAKER/MAKER-P users.

De novo genome annotation using MAKER-P

Building on MAKER, MAKER-P adds noncoding RNA and pseudogene annotation functionality as well as protocols for generating species specific repeat libraries. Mi-RPREFer was developed as part of the MAKER-P tool kit to annotate miRNAs, and can be found at *https://github.com/hangelwen/miR-PREFeR*. tRNAscan-SE (Lowe and Eddy, 1997) and snoscan (Lowe, 1999) are also integrated into the MAKER-P framework, and are run by using `trna=` and `snoscan_rrna=` in the `maker_opts.ctl` file (see `trna=` and `snoscan_rrna=` in Table 4.11.2 at the end of this unit). Pseudogenes are annotated using the method described here (Zou et al., 2009; Campbell et al., 2014). A protocol for annotating pseudogenes can be found at *http://shiulab.plantbiology.msu.edu/wiki/index.php/Protocol:Pseudogene*. See Campbell et al. (2014) for benchmarking results for MAKER-P annotated ncRNAs and pseudogenes on the Arabidopsis genome.

Adequate repeat masking is critical for accurate gene annotations. Basic and advanced protocols for generating species-specific repeat libraries can be found at *http://weatherby.genetics.utah.edu/MAKER/wiki/index.php/Repeat_Library_Construction--Basic* and *http://weatherby.genetics.utah.edu/MAKER/wiki/index.php/Repeat_Library_Construction--Advanced*. See Campbell et al. (2014) for benchmarking of these repeat library generation protocols on the Arabidopsis genome. MAKER-P is available for use on the iPlant infrastructure; see *https://pods.iplantcollaborative.org/wiki/display/sciplant/MAKER-P+at+iPlant* for MAKER-P usage on iPlant as an atmosphere image and on the Texas Advanced Computing Center (TACC) compute clusters.

ALTERNATE PROTOCOL 1

DE NOVO GENOME ANNOTATION USING PRE-EXISTING EVIDENCE ALIGNMENTS AND GENE PREDICTIONS

Aligning evidence to a genome assembly is one of the more time consuming and computationally expensive steps in genome annotation. Using pre-aligned evidence will substantially decrease the time it takes to annotate a genome. MAKER can take protein and mRNA-seq/EST alignments as evidence as well as aligned repetitive elements for masking. It can also accept existing gene predictions. All of these data need to be in GFF3 format.

Necessary Resources

Hardware

Computer with a Unix-based operating system (e.g., Linux, Mac OS X)

Software

MAKER and MAKER-P are available for download at *yandell-lab.org*. Installation instructions are included in the tarball. For brevity's sake, the following protocols describe MAKER, but apply to MAKER-P as well.

MAKER will identify and download all of its necessary external dependencies including BLAST, Exonerate, RepeatMasker, and a number of Perl modules [automatic download and installation of Perl modules requires CPAN (*https://metacpan.org/pod/CPAN*) to be installed]. MAKER will also install a number of additional programs such as SNAP, Augustus, and MPICH2. This

example uses a version of MAKER installed with NCBI BLAST+, Exonerate, RepeatMasker, with optional RepBase libraries, and SNAP.

Files

Genome assembly to be annotated in FASTA format, protein evidence alignments in GFF3 format, assembled mRNA-seq transcript alignments from the species of interest in GFF3 format, gene predictions for the genomic assembly you wish to annotate in GFF3 format, and repetitive elements to be masked in GFF3 format.

1. From the Unix command line, generate the MAKER control files:

```
% maker -CTL
```

This is the same as in Basic Protocol 1.

2. Edit the `maker_opts.ctl` file to add the genomic sequence and evidence, and specify the gene=finding method:

```
% emacs maker_opts.ctl
#-----Genome (these are always required)
genome=$PATH_TO_CBP_maker/maker_inputs/dpp_data/dpp
  _contig.fasta #genome sequence (fasta file or fasta
  embeded in GFF3 file)
organism_type=eukaryotic #eukaryotic or prokaryotic.
  Default is eukaryotic
#-----EST Evidence (for best results provide a file for
  at least one)
est_gff=$PATH_TO_CBP_maker/maker_inputs/dpp_data/
  mRNA_seq_evidence.gff
#aligned ESTs or mRNA-seq from an external GFF3 file
#-----Protein Homology Evidence (for best results
  provide a file for at least one)
protein_gff=$PATH_TO_CBP_maker/maker_inputs/dpp
  _data/protein_evidence.gff
#aligned protein homology evidence from an external GFF3
  file
#-----Repeat Masking (leave values blank to skip repeat
  masking)
rm_gff=$PATH_TO_CBP_maker/maker_inputs/dpp
  _data/repeats.gff
#pre-identified repeat elements from an external GFF3
  file
#-----Gene Prediction
pred_gff=$PATH_TO_CBP_maker/maker_inputs/dpp_data/
  snap_predictions.gff #ab-initio predictions from an
  external GFF3 file
```

MAKER is expecting alignments in the GFF3 file to be represented as `match/match_part` two-level features. Below is an example from the `mRNA_seq_evidence.gff` file. Importantly, MAKER assumes that evidence passed in as GFF3 represents the correct exon boundaries of transcripts; for best results, make sure that precomputed BLAST alignments have been aligned to the genome in a splice-aware fashion before passing them to MAKER in GFF3 format:

```
contig-dpp-500-500 est2genome expressed_sequence_match
  26786
```

```
  31656 14993 +.
  ID=contig-dpp-500-500:hit:53:3.2.0.0;Name=dpp-mRNA-5
contig-dpp-500-500 est2genome match_part 26786 26955
  14993 +. ID=contig-dpp-500-500:
  hsp:62:3.2.0.0;Parent=contig-dpp-500-
  500:hit:53:3.2.0.0;Target=dpp-mRNA-5 1 170
  +;Gap=M170
contig-dpp-500-500 est2genome match_part 27104 27985
  14993 +. ID=contig-dpp-500-500:
  hsp:63:3.2.0.0;Parent=contig-dpp-500-
  500:hit:53:3.2.0.0;Target=dpp-mRNA-5 171 1052
  +;Gap=M882
contig-dpp-500-500 est2genome match_part 29709 31656
  14993 +. ID=contig-dpp-500-500:hsp:
  64:3.2.0.0;Parent=contig-dpp-500-
  500:hit:53:3.2.0.0;Target=dpp-mRNA-5 1053 3000
  +;Gap=M1948
```

3. Run MAKER and check/collect the results as outlined in Basic Protocol 1, steps 3 to 5.

ALTERNATE PROTOCOL 2

PARALLELIZED DE NOVO GENOME ANNOTATION USING MPI

Users can dramatically decrease the time required for annotating a genome by spreading the computation out across multiple compute cores (CPUs). MAKER is fully MPI compliant, allowing users to parallelize their genome annotation efforts.

Necessary Resources

Hardware

Multicore server or cluster with a Linux-based operating system

Software

MAKER and MAKER-P are available for download at *yandell-lab.org*. Installation instructions are included in the tarball. For brevity's sake, the following protocols describe MAKER, but apply to MAKER-P as well.

MAKER will identify and download all of its necessary external dependencies including BLAST, Exonerate, RepeatMasker, and a number of Perl modules [automatic download and installation of Perl modules requires CPAN (*https://metacpan.org/pod/CPAN*) to be installed]. MAKER will also install a number of additional programs such as SNAP, Augustus, and MPICH2. This example uses a version of MAKER installed with NCBI BLAST+, Exonerate, RepeatMasker, with optional RepBase libraries, and SNAP.

OpenMPI or MPICH2

Files

Genome assembly to be annotated in FASTA format, protein evidence alignments in GFF3 format, assembled mRNA-seq transcript alignments from the species of interest in GFF3 format, gene predictions for the genomic assembly you wish to annotate in GFF3 format, and repetitive elements to be masked in GFF3 format.

1. Configure MAKER to run with MPI during the installation step of MAKER:

```
% cd $PATH_TO_MAKER/maker/src
% perl Build.PL
MAKER supports distributed parallelization via MPI.
  Would you like to configure MAKER for MPI (This
```

```
  requires that you have an MPI client installed)? [N]Y
Please specify the path to 'mpicc' on your system:
  [/usr/local/mpich2/bin/mpicc ]
Please specify the path to the directory containing
  'mpi.h': [/usr/local/mpich2/include ]
```

The text below the command lines is generated by MAKER and requires user input. The default input is printed in the brackets and can be accepted by pressing return/enter or changed by entering the requested information and pressing return/enter. These steps can be done when you install MAKER. In the above example, MAKER found `mpicc` and `mpi.h` in the path and gave them as the default response to the specify path request. If you would like to use another version/flavor of MPI, you can specify it at this point. In this example we are using MPICH2.

When installing MPICH2 or OpenMPI, it is important to compile them with shared libraries enabled. For OpenMPI, this may require the addition of a line similar to the one below to your `~/.bash_profile` or equivalent.

```
export LD_PRELOAD=/path/to/openmpi/lib/libmpi.so:$LD
  _PRELOAD
```

OpenMPI and MPICH2 exhibit very similar performance on jobs using less than 100 CPUs. When using more than 100 CPUs, the OpenMPI implementation of MPI is more stable.

2. Generate the MAKER control files and edit `maker_opts.ctl` as outlined in Basic Protocol 1, steps 1 and 2.

3. Run MAKER using mpiexec on the number of CPU cores you wish to utilize:

```
% mpiexec -n 26 maker
```

The first part of this command, `mpiexec`, is a standard way of starting an MPI job regardless of the MPI implementation. The `--n` argument to `mpiexec` is used to specify the number of processors (in this case 26). The next command is the `maker` executable. Please note that according to `mpiexec` documentation, in order to run this same command in the background or under control of `nohup`, you must also attach `/dev/null` to STDIN as demonstrated below:

```
% nohup mpiexec -n 26 maker < /dev/null &
```

Different cluster environments may also require additional command-line arguments for `mpiexec`; check with your cluster administrator and/or MPICH2 and OpenMPI documentation for additional details. For example, disabling OpenFabrics support may be required on Infiniband-based clusters for MAKER to work correctly with OpenMPI:

```
% mpiexec -mca btl ^openib -n 26 maker
```

4. Check/collect the results as outlined in Basic Protocol 1, steps 4 to 5.

PARALLELIZED DE NOVO GENOME ANNOTATION WITHOUT MPI

ALTERNATE PROTOCOL 3

If it is not possible to install MPICH2 or OpenMPI on the server or cluster where you wish to run MAKER, there is still a way to annotate your genome in parallel. This is done by starting multiple MAKER instances in the same directory. Each instance of MAKER will then use file locks together with the datastore index log to coordinate contig processing across multiple MAKER instances. If a datastore index log entry indicates that a contig

is being processed by a separate instance of MAKER, then that instance of MAKER will skip to the next contig in the FASTA. This checking and skipping process will continue until a given instance of MAKER finds an entry that has not been started.

Necessary Resources

Hardware

Multicore server or cluster with a Linux based operating system.

Software

MAKER and MAKER-P are available for download at *yandell-lab.org*. Installation instructions are included in the tarball. For brevity's sake, the following protocols describe MAKER, but apply to MAKER-P as well.

MAKER will identify and download all of its necessary external dependencies including BLAST, Exonerate, RepeatMasker, and a number of Perl modules [automatic download and installation of Perl modules requires CPAN (*https://metacpan.org/pod/CPAN*) to be installed]. MAKER will also install a number of additional programs such as SNAP, Augustus, and MPICH2. This example uses a version of MAKER installed with NCBI BLAST+, Exonerate, RepeatMasker, with optional RepBase libraries, and SNAP.

1. Generate the MAKER control files and edit the `maker_opts.ctl` as outlined in Basic Protocol, steps 1 and 2.
2. Start multiple instances of MAKER in the same directory (started as background processes):

```
% maker 2> maker1.error
% maker 2> maker2.error &
% maker 2> maker3.error &
```

3. Check/collect the results as outlined in Basic Protocol 1, steps 4 to 5.

SUPPORT PROTOCOL 1

TRAINING GENE FINDERS FOR USE WITH MAKER

Ab initio gene finders can achieve very high accuracies when well trained. Training data normally takes the form of a 'gold-standard' set of pre-existing gene annotations. Unfortunately, training data is usually not available for a newly sequenced organism. Here we outline a method for generating training data for a novel, never before annotated genome. The key is using MAKER in an iterative fashion. For more on this topic, see Holt and Yandell (2011). In this example, we train SNAP, but this method can be applied to other gene finders as well.

Necessary Resources

Hardware

Computer with a Unix-based operating system (e.g., Linux, Mac OS X)

Software

MAKER and MAKER-P are available for download at *yandell-lab.org*. Installation instructions are included in the tarball. For brevity's sake, the following protocols describe MAKER, but apply to MAKER-P as well.

MAKER will identify and download all of its necessary external dependencies including BLAST, Exonerate, RepeatMasker, and a number of Perl modules [automatic download and installation of Perl modules requires CPAN (*https://metacpan.org/pod/CPAN*) to be installed]. MAKER will also install a number of additional programs such as SNAP, Augustus, and MPICH2. This example uses a version of MAKER installed with NCBI BLAST+, Exonerate, RepeatMasker, with optional RepBase libraries, and SNAP.

Files

This example uses a larger data set than Basic Protocol 1 so as to generate enough gene models to train the gene finder. The file types are the same, with the exception of the SNAP species parameter/HMM file, which we are going to create. Here we are using a data set from *Pythium ultimum* (Lévesque et al., 2010). Note that the protein and EST evidence could also be given in GFF3 format (see Alternate Protocol 1).

1. Create MAKER control files as outlined in Basic Protocol 1, step 1.
2. Edit the `maker_opts.ctl` file to add the genomic sequence and evidence, and specify the gene finding method:

```
% emacs maker_opts.ctl
#-----Genome (these are always required)
genome=$PATH_TO_CBP_maker/maker_inputs/pyu_data/pyu-
  contig.fasta #genome sequence (fastafile or fasta
  embeded in GFF3 file)
organism_type=eukaryotic #eukaryotic or prokaryotic.
  Default is eukaryotic
#-----EST Evidence (for best results provide a file
  for at least one)
est=$PATH_TO_CBP_maker/maker_inputs/pyu_data/pyu-
  est.fasta
#set of ESTs or assembled mRNA-seq in fasta format
#-----Protein Homology Evidence (for best results
  provide a file for at least one)
protein=$PATH_TO_CBP_maker/maker_inputs/pyu_data/pyu-
  protein.fasta #protein sequence file in fasta format
  (i.e., from multiple organisms)
#-----Gene Prediction
est2genome=1
#infer gene predictions directly from ESTs, 1 = yes, 0
  = no
```

Note that the configuration shown above differs from that in Basic Protocol 1 in the Gene Prediction section. By setting `est2genome=1`*, MAKER will infer gene models directly from the EST/mRNA-seq evidence. Remember that if these data are given in GFF3 format, they must have been aligned to the genome in a splice-aware fashion. BLAST data will not suffice. If given in FASTA format, as in this example, MAKER will take care of the aligning and polishing. Given the nature of these data, many of the resulting gene models will be partial. However, there is usually enough information in these gene models for first-round training of a gene finder. Alternatively, you could also set* `protein2genome=1` *to derive gene models from splice-aware aligned protein evidence.*

3. Run MAKER with or without MPI (see Basic Protocol 1 and Alternate Protocols 2 and 3).
4. Check/collect the results as outlined in Basic Protocol 1, steps 4 to 5.

   For gene finder training, you only need to collect the GFF3 file for the genome.

5. Make a directory for SNAP training and go to it:

```
% mkdir snap1
% cd snap1
```

6. Run `maker2zff`:

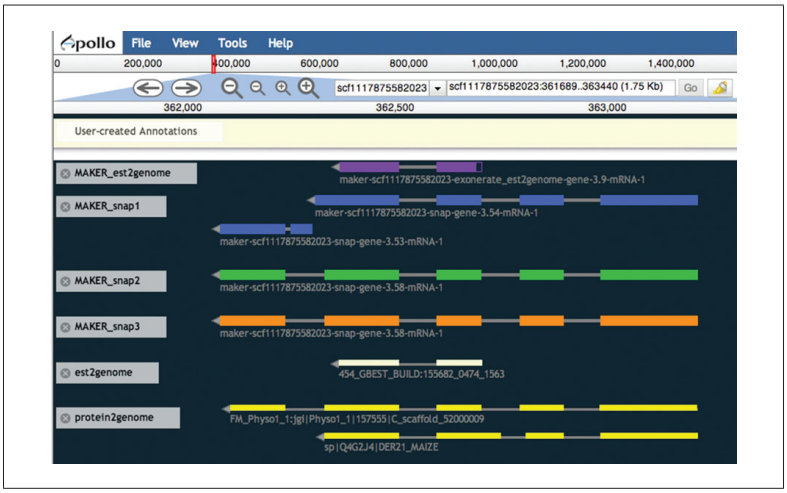

Figure 4.11.2 Iterative gene finder training improves gene annotations. Using only EST data and no gene finders, MAKER annotates a single two-exon gene at the locus `MAKER_est2genome` (purple). This annotation is consistent with the EST alignment (`est2genome`, beige), but is inconsistent with protein evidence data (`protein2genome`, yellow). After one round of SNAP training, MAKER annotates two models at this locus (`MAKER_snap1`, blue); these two models are more consistent with the protein evidence (`protein2genome`, yellow). An additional round of training yields a single MAKER annotation (`MAKER_snap2`, green) that is still more consistent with the protein evidence. Note that SNAP is not improved on with further training (`MAKER_snap3`, orange).

```
% maker2zff ../ pyu-contig.all.gff
genome.ann
genome.dna
```

`maker2zff` is an accessory script that comes with MAKER. It generates a ZFF-formatted file (`genome.ann`) and a FASTA file (`genome.dna`) that are required to train SNAP. To produce these files, the input GFF3 file must contain the genomic FASTA sequence appended to the end according to the GFF3 specification (this is the default used by `gff3_merge`). In order for a gene model to be considered suitable for training, it has to pass several quality filters imposed by the `maker2zff script`. By default, a gene model must have half of its splice sites confirmed by an EST/mRNA-seq alignment; half of its exons must overlap an EST/mRNA-seq alignment; and its annotation edit distance must be less than 0.5. All of these criteria can be modified on the command line.

7. Run `fathom with the categorize option` (part of SNAP package):

```
% fathom -categorize 1000 genome.ann genome.dna
```

8. Run `fathom with the export option`:

```
% fathom -export 1000 -plus uni.ann uni.dna
```

9. 9 Run `forge` (part of SNAP package):

```
% forge export.ann export.dna
```

10. Run `hmm-assembler.pl` (part of SNAP package) to generate the final SNAP species parameter/HMM file and return to the MAKER working directory:

```
% hmm-assembler.pl pyu1 . > pyu1.hmm
% cd ..
```

11. Edit the maker_opts.ctl file to use the newly trained gene finder:

```
%emacs maker_opts.ctl
#-----Gene Prediction
snaphmm=./snap1/pyu1.hmm #SNAP HMM file
est2genome=0
#infer gene predictions directly from ESTs, 1 = yes, 0
  = no
```

12. Optional bootstrap training can be done by now repeating steps 3 to 10 and using the initial SNAP HMM file to seed the next round of SNAP training.

Generally there is little further improvement after two rounds of bootstrap training with the same evidence, and you run the risk of overtraining (which can actually decrease SNAP's accuracy). See Figure 4.11.2. Once SNAP is trained, you can use the SNAP-derived annotations to train other gene finders following this same bootstrap procedure. When all of your gene finders are trained, you are ready to annotate your genome using Basic Protocol 1 or any of the alternate protocols above.

RENAMING GENES FOR GENBANK SUBMISSION

SUPPORT PROTOCOL 2

You can learn a lot about a MAKER gene annotation from the name assigned to the gene. Take for example the gene named `maker-contig-dpp-500-500-snap-gene-0.3`. Since it starts with `maker`, we know that that it is derived from a MAKER 'hint-based' prediction (for more information about how MAKER passes evidence derived 'hints' to the gene predictors, see Cantarel et al., 2008; Holt and Yandell, 2011). We have the name of the scaffold that the gene is on (`contig-dpp-500-500`) followed by the name of the gene finder used to generate the original and hint-based model (`snap`). The numbers following the gene predictor are used to make the ID unique.

Though useful, these IDs are not intended to be permanent. Once you have a registered genome prefix, you can use two of the accessory scripts distributed with MAKER to replace your MAKER gene names with NCBI-style gene IDs.

Necessary Resources

Hardware

Computer with a Unix-based operating system (e.g., Linux, Mac OS X)

Software

`maker_map_ids, map_gff_ids, and map_fasta_ids` distributed with MAKER.

Files

MAKER generated GFF3 and FASTA files

1. Generate an id mapping file using `maker_map_ids`:

```
% maker_map_ids --prefix DMEL_ --justify 6\
dpp_contig.all.gff > dpp_contig.all.map
```

This creates a two-column tab-delimited file with the original id in column 1 and the new id in column 2. The `--prefix` is where you give your registered genome prefix; the value following `--justify` determines the length of the number following the prefix (make sure that you allow adequate places for the number of genes in the annotation set, e.g., if you have 10,000 genes, `--justify` should be set to at least 5).

2. Look at the contents of the `contig-dpp-500-500.map` file.

```
% cat dpp_contig.all.map
maker-contig-dpp-500-500-snap-gene-0.3 DMEL_000001
maker-contig-dpp-500-500-snap-gene-0.3-mRNA-1
  DMEL_000001-RA
```

You will notice that the `.map` files are simply two-column files showing the conversion of the existing gene/transcript ID (column 1) to a new ID (column 2).

3. Use the map file created in step 1 to change the ids in the GFF3 and FASTA file

```
% map_gff_ids dpp_contig.all.map dpp_contig.all.gff
% map_fasta_ids dpp_contig.all.map
  dpp_contig.all.maker.proteins.fasta
% map_fasta_ids dpp_contig.all.map
  dpp_contig.all.maker.transcripts.fasta
% head --n 3 dpp_contig.all.gff
contig-dpp-500-500 . contig 1 32156 . . .
  ID=contig-dpp-500-500;Name=contig-dpp-500-500
contig-dpp-500-500 maker gene 23054 31656 . + .
  ID=DMEL_000001;Name=DMEL_000001;Alias=maker-contig-
  dpp-500-500-snap-gene-0.3;
```

Note that the above command lines do not redirect standard out (STDOUT). These scripts do an in-place edit of the file to save disk space. Therefore, it is important not to interrupt these scripts as they run, or the files can be corrupted/truncated. In this example, our long MAKER-generated gene ID `maker-contig-dpp-500-500-snap-gene-0.3` was changed to `DMEL_000001` in both the GFF3 and FASTA files, with the original MAKER name kept as an alias.

SUPPORT PROTOCOL 3

ASSIGNING PUTATIVE GENE FUNCTION

MAKER also provides support for functional annotation (i.e., identifying putative gene functions, protein domains, etc.). This protocol uses NCBI BLAST+ and the well-curated UniProt/Swiss-Prot set of proteins to assign putative functions to newly annotated genes.

Necessary Resources

Hardware

Computer with a Unix-based operating system (e.g., Linux, Mac OS X)

Software

`NCBI BLAST+, maker_functional_gff, and maker_functional_fasta` (from MAKER)

Files

UniProt/SwissProt multi-FASTA file (*http://www.uniprot.org*), MAKER-generated GFF3 and FASTA files

1. Index the UniProt/Swiss-Prot multi-FASTA file using `makeblastdb`:

```
% makeblastdb -in uniprot_sprot.fasta -input_type fasta
  -dbtype prot
```

2. BLAST the MAKER-generated protein FASTA file to UniProt/SwissProt with BLASTP. Some command lines are longer than a single printed (displayed) line. These long commands include a '\' before the continued line, so that multiple lines are read as a single line:

```
% blastp -db uniprot_sprot.fasta\
-query contig-dpp-500-500.maker.proteins.fasta -out
  maker2uni.blastp -evalue\ .000001 -outfmt 6
  -num_alignments 1 -seg yes -soft_masking true\
-lcase_masking -max_hsps_per_subject 1
```

The key parts of this BLAST command line include the specification of the tabular format (`-outfmt 6`), and the `-num_alignments 1` and `-max_hsps_per_subject 1` flags which limit the hits returned for a given sequence to a single line in the BLAST report. The output for this BLAST search is:

```
DMEL_000001-RA sp|P07713|DECA_DROME 100.00 588 0 0 1 588
  1 588 0.0 1220
```

Tabular-formatted WUBLAST/ABBLAST output works as well for this protocol.

3. Add the protein homology data to the MAKER GFF3 and FASTA files with `maker_functional_gff and maker_functional_fasta`.

```
% maker_functional_gff uniprot_sprot.fasta\
maker2uni.blastp dpp_contig.all.gff\
contig-dpp-500-500_functional_blast.gff
% maker_functional_fasta uniprot_sprot.fasta\
  maker2uni.blastp\ dpp_contig.all.maker.proteins.fasta\
dpp_contig.all.maker.proteins_functional_blast.fasta
% maker_functional_fasta uniprot_sprot.fasta\
  maker2uni.blastp\
dpp_contig.all.maker.transcripts.fasta\
dpp_contig.all.maker.transcripts_functional_blast
  .fasta
```

This procedure added `Note=Similar to dpp: Protein decapentaplegic (Drosophila melanogaster);` to column 9 of the gene and mRNA feature lines in the MAKER GFF3 file:

```
% head -n 4 dpp_contig.all.functional_blast.gff
##gff-version 3
contig-dpp-500-500 . contig 1 32156 . . .
  ID=contig-dpp-500-500;Name=contig-dpp-500-500
contig-dpp-500-500 maker gene 23054 31656 . + .
  ID=DMEL_000001;Name=DMEL_000001;Alias=maker-contig-
  dpp-500-500-snap-gene-0.3;Note=Similar to dpp:
  Protein decapentaplegic (Drosophila melanogaster);
contig-dpp-500-500 maker mRNA 23054 31656 . + .
  ID=DMEL_000001-
  RA;Parent=DMEL_000001;Name=DMEL_000001-
  RA;Alias=maker-contig-dpp-500-500-snap-gene-0.3-mRNA-
  1;_AED=0.13;_QI=1422|1|1|1|0.5|0.33|3|1049|588;
  _eAED=0.13;Note=Similar to dpp: Protein
  decapentaplegic (Drosophila melanogaster);
```

This also added “Name:” Similar to dpp Protein decapentaplegic (Drosophila melanogaster)“” to the definition lines of the FASTA entries.

```
% head -n 1
  dpp_contig.all.maker.proteins_functional_blast.fasta
```

```
DMEL_000001-RA protein Name:"Similar to dpp Protein
  decapentaplegic (Drosophila melanogaster)" AED:0.13
  eAED:0.13 QI:1422|1|1|1|0.5|0.33|3|1049|588
% head -n 1
  dpp_contig.all.maker.transcripts_functional_blast
  .fasta
DMEL_000001-RA transcript Name:"Similar to dpp Protein
  decapentaplegic (Drosophila melanogaster)" offset:1422
  AED:0.13 eAED:0.13 QI:1422|1|1|1|0.5|0.33|3|1049|588
```

A similar tool called `ipr_update_gff` *is also distributed with MAKER. This tool allows users to add functional annotations from InterProScan (Quevillon et al., 2005), including protein family domains to the MAKER GFF3 file. See Basic Protocol 5 for more on this point.*

SUPPORT PROTOCOL 4

LABELING EVIDENCE SOURCES FOR DISPLAY IN GENOME BROWSERS

Many genome annotation projects entail the use of multiple RNA-seq and protein datasets. For example, the RNA-seq datasets might come from multiple tissue types, stages of life, strains, accessions, and treatments, and the protein datasets might comprise the proteomes of related species. All of these data can be passed to MAKER as evidence in the form of a comma-separated list added to the `maker_opts.ctl` file. Additionally, each file can be given a tag that is moved forward to the MAKER GFF3 output to identify the source of any given evidence alignment. This tag can be very helpful when visualizing your data in a genome browser or when mining data from the MAKER-generated GFF3 file to use in other applications/protocols.

Necessary Resources

Hardware

Computer with a Unix-based operating system (e.g., Linux, Mac OS X)

Software

MAKER and MAKER-P are available for download at *yandell-lab.org*. Installation instructions are included in the tarball. For brevity's sake, the following protocols describe MAKER, but apply to MAKER-P as well.

MAKER will identify and download all of its necessary external dependencies including BLAST, Exonerate, RepeatMasker, and a number of Perl modules (automatic download and installation of Perl modules requires CPAN (*https://metacpan.org/pod/CPAN*) to be installed. MAKER will also install a number of additional programs such as SNAP, Augustus, and MPICH2. This example uses a version of MAKER installed with NCBI BLAST+, Exonerate, RepeatMasker, with optional RepBase libraries, and SNAP.

Files

Genome assembly to be annotated in FASTA format

Protein evidence in FASTA format

Assembled mRNA-seq transcripts from the species of interest in FASTA format

Optional: a species parameter/HMM file for SNAP generated for the organism of interest or a closely related species. The process used to create a species parameter/HMM file is described in SNAP's internal documentation (Korf, 2004).

1. Create MAKER control files as outlined in Basic Protocol 1, step 1.
2. Edit the `maker_opts.ctl` file to add the genomic sequence and evidence, and specify the gene-finding method. Add the tags to the evidence at this time:

```
% emacs maker_opts.ctl
#-----Genome (these are always required)
genome=$PATH_TO_CBP_maker/maker_inputs/dpp_data/dpp
  _contig.fasta #genome sequence (fasta file or fasta
  embeded in GFF3 file)
organism_type=eukaryotic #eukaryotic or prokaryotic.
  Default is eukaryotic
#-----EST Evidence (for best results provide a file for
  at least one)
est=$PATH_TO_CBP_maker/maker_inputs/dpp_data/dpp_est
  .fasta:3instar
#set of ESTs or assembled mRNA-seq in fasta format
#-----Protein Homology Evidence (for best results
  provide a file for at least one)
protein=$PATH_TO_CBP_maker/maker_inputs/dpp_data/dpp
  _protein.fasta:Dsim #protein sequence file in fasta
  format (i.e., from multiple organisms)
#-----Gene Prediction
snaphmm=$PATH_TO_CBP_maker/maker_inputs/dpp_data/
  D.melanogaster.hmm #SNAP HMM file
```

3. Run MAKER and check/collect the results as outlined in Basic Protocol 1, steps 3 to 5.

The tags are added after the evidence dataset file name as a suffix consisting of a colon, followed by the identification tag. In the above example, the tag `3instar` was added to the `est` file and the tag `Dsim` was added to the protein file. In the final GFF3 output, the source column (column 2 in bold below) for the BLASTN alignments from the `dpp_est.fasta` file is changed from `blastn` to `blastn:3instar`. Similarly, as sources the GFF3 file contains `est2genome:3instar`, `blastx:Dsim`, and `protein2genome:Dsim`.

```
% grep blastn dpp_contig.all.gff | head -n 1
contig-dpp-500-500 blastn:3instar
  expressed_sequence_match 26786 31656 170 + .
  ID=contig-dpp-500-500:hit:48:3.2.0.0;Name=dpp-mRNA-5
% grep est2genome dpp_contig.all.gff | head -n 1
contig-dpp-500-500 est2genome:3instar
  expressed_sequence_match 26786 31656 14993 + .
  ID=contig-dpp-500-500:hit:53:3.2.0.0;Name=dpp-mRNA-5
% grep blastx dpp_contig.all.gff | head -n 1
contig-dpp-500-500 blastx:Dsim protein_match 27118
  30604 1482 + .
  ID=contig-dpp-500-500:hit:58:3.10.0.0;Name=dpp-CDS-5
% grep protein2genome dpp_contig.all.gff | head -n 1
contig-dpp-500-500 protein2genome:Dsim protein_match
  27118 30604 3062 + .
  ID=contig-dpp-500-500:hit:63:3.10.0.0;Name=dpp-CDS-5
```

This feature simplifies loading different lines of evidence into a genome browser as separate tracks.

BASIC PROTOCOL 2

UPDATING/COMBINING LEGACY ANNOTATION DATASETS IN LIGHT OF NEW EVIDENCE

MAKER provides means to employ new evidence to improve the accuracy of existing genome annotations without completely reannotating the genome. This allows MAKER users to rapidly update existing annotations in light of new mRNA-seq data sets and protein evidence. Note that the starting annotations need not have been produced using MAKER. The protocol outlined below assumes that a starting dataset of annotations is available in GFF3 format. If this is not available, see Basic Protocol 4, "Mapping annotations to a new assembly," which explains how to map pre-existing transcripts (produced by any annotation pipeline) to a genome assembly and produce a GFF3 file for later use with MAKER.

Necessary Resources

Hardware

Computer with a Unix-based operating system (e.g., Linux, Mac OS X)

Software

MAKER and MAKER-P are available for download at *yandell-lab.org*. Installation instructions are included in the tarball. For brevity's sake, the following protocols describe MAKER, but apply to MAKER-P as well.

MAKER will identify and download all of its necessary external dependencies including BLAST, Exonerate, RepeatMasker, and a number of Perl modules (automatic download and installation of Perl modules requires CPAN (*https://metacpan.org/pod/CPAN*) to be installed. MAKER will also install a number of additional programs such as SNAP, Augustus, and MPICH2. This example uses a version of MAKER installed with NCBI BLAST+, Exonerate, RepeatMasker, with optional RepBase libraries, and SNAP.

Files

Genome assembly from the original annotation in FASTA format, new protein evidence in FASTA format, new assembled mRNA-seq transcripts from the species of interest in FASTA format, annotations to be updated/combined in GFF3 format

1. Create MAKER control files as outlined in Basic Protocol 1, step 1.
2. Edit the `maker_opts.ctl` file to add the genomic sequence, evidence, and gene models you wish to update:

```
% emacs maker_opts.ctl
#-----Genome (these are always required)
genome=$PATH_TO_CBP_maker/maker_inputs/legacy_data/
  legacy-contig.fasta #genome sequence (fasta file or
  fasta embeded in GFF3 file)
organism_type=eukaryotic #eukaryotic or prokaryotic.
  Default is eukaryotic
#-----EST Evidence (for best results provide a file for
  at least one)
est=$PATH_TO_CBP_maker/maker_inputs/legacy_data/
  legecy-new-mRNAseq.fasta #set of ESTs or assembled
  mRNA-seq in fasta format
#-----Protein Homology Evidence (for best results
  provide a file for at least one)
```

```
protein=$PATH_TO_CBP_maker/maker_inputs/legacy_data/
  legacy-new-protein.fasta #protein sequence file in
  fasta format (i.e., from multiple organisms)
#-----Gene Prediction
pred_gff=$PATH_TO_CBP_maker/maker_inputs/legacy
  _data/legacy-set1.gff,
  $PATH_TO_CBP_maker/maker_inputs/legacy_data/
legacy-set2.gff #ab-initio predictions from an external
  GFF3 file
#-----MAKER Behavior Options
keep_preds=0
#Concordance threshold to add unsupported gene
  prediction (bound by 0 and 1)
```

Passing gene models in GFF3 format to MAKER as `pred_gff` *allows MAKER to update the models in the light of new evidence by adding new 3′ and 5′ exons, additional UTR, and merging split models. This method will not change internal exons, nor will it entirely delete any existing gene model. When run in this mode with an additional gene finder turned on, MAKER will also create new annotations where new evidence suggest a gene but no corresponding model was previously present. In this example, two annotation sets (*`legacy-set1.gff`, `legacy-set2.gff`*) are being merged and updated. When two models are annotated at the same locus, MAKER will chose the model that best matches the evidence for inclusion in the final annotation set. Setting* `keep_preds=1` *will ensure that no gene models are lost from the legacy annotations. If* `keep_preds=0` *is set, gene models that are not supported by the evidence will not be included in the final MAKER annotation build. For this example, we have set* `keep_pres=0` *because we are concerned about false positives in the legacy annotations.*

3. Run MAKER and check/collect the results as outlined in Basic Protocol 1, steps 3 to 5.

   The original legacy annotation sets contained 237 and 203 gene annotations:

```
% grep -cP '\tgene\t' \
$PATH_TO_CBP_maker/maker_inputs/legacy_data/legacy-
  set1.gff
237
% grep -cP '\tgene\t' \
$PATH_TO_CBP_maker/maker_inputs/legacy_data/legacy-
  set2.gff
203
```

The combined annotation set contains 180 genes. This number of genes in the combined set is a result of adding genes from one set that were not annotated in the other set, merging or splitting genes, and discarding genes that are not supported by the evidence.

```
% grep -cP '\tgene\t' legacy-contig.all.gff
180
```

In addition to reconciling these two annotation sets based on the evidence, MAKER also added 3′ and 5′ UTR features that were supported by the new RNA evidence. Neither of the legacy annotation sets contained three or five prime UTR features.

```
% grep -cP '\t(three|five)_prime_UTR\t'
  legacy-contig.all.gff
47
```

BASIC PROTOCOL 3

ADDING MAKER'S QUALITY-CONTROL METRICS TO ANNOTATIONS FROM ANOTHER PIPELINE

The MAKER annotation pipeline strives to be transparent in its use of evidence for each gene annotation. To accomplish this transparency, MAKER does two things. First, all of the evidence alignments, repeat masked regions, ab initio gene predictions, etc, are included in MAKER's GFF3 output with its annotations. Second, MAKER generates a series of quality metrics for each annotated gene model. These metrics include (1) the MAKER mRNA Quality index (QI), and (2) an Annotation Edit Distance (AED). Both of these data types are attached to each MAKER transcript.

The MAKER mRNA Quality index (QI) is a nine-dimensional summary of a transcript's key features and how they are supported by the data gathered by MAKER's compute pipeline. A typical QI might look as follows: `QI:0|0.77|0.68|1|0.77|0.78|19|462|824`. Table 4.11.2 provides a key for the QI data fields. Values are delimited by pipe symbols. Interpretation is easy. For example, the transcript with the QI string above has no 5′ UTR; 77% of its splice sites are confirmed by transcript data; 68% of its exons overlap transcript evidence; all of its exons overlap transcript or protein alignments; 77% of its splice sites are precisely confirmed by an ab initio gene prediction; 78% of its exons overlap an ab initio prediction; the transcript has 19 exons; the 3′ UTR is 462 base pairs long; and the protein it encodes is 824 amino acids in length. QI strings are easily parsed, and thus provide a good starting point for MAKER users seeking to write their own scripts for genome annotation quality control.

Also included in MAKER's GFF3 outputs is a second quality control measure called Annotation Edit Distance (AED; Eilbeck et al., 2009; Holt and Yandell, 2011; Yandell and Ence, 2012). MAKER and MAKER-P use AED 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). In other words, the AED score provides a measure of each annotated transcript's congruency with its supporting evidence. See (Yandell and Ence, 2012) for a further discussion of AED.

The protocol below adds QI tags and AED scores to gene models produced by other pipelines.

Table 4.11.1 MAKER Quality Index Summary (adapted from Cantarel et al., 2008)

Position	Definition
1	Length of the 5′ UTR
2	Fraction of splice sites confirmed by an EST/mRNA-seq alignment
3	Fraction of exons that match an EST/mRNA-seq alignment
4	Fraction of exons that overlap EST/mRNA-seq or protein alignments
5	Fraction of splice sites confirmed by ab initio gene prediction
6	Fraction of exons that overlap an ab initio gene prediction
7	Number of exons in the mRNA
8	Length of the 3′ UTR
9	Length of the protein sequence produced by the mRNA

Necessary Resources

Hardware

Computer with a Unix-based operating system (e.g., Linux, Mac OS X)

Software

MAKER and MAKER-P are available for download at *yandell-lab.org*. Installation instructions are included in the tarball. For brevity's sake, the following protocols describe MAKER, but apply to MAKER-P as well.

MAKER will identify and download all of its necessary external dependencies including BLAST, Exonerate, RepeatMasker, and a number of Perl modules [automatic download and installation of Perl modules requires CPAN (*https://metacpan.org/pod/CPAN*) to be installed]. MAKER will also install a number of additional programs such as SNAP, Augustus, and MPICH2. This example uses a version of MAKER installed with NCBI BLAST+, Exonerate, RepeatMasker, with optional RepBase libraries, and SNAP.

Files

Genome assembly from the original annotation in FASTA format, protein evidence in FASTA format, assembled mRNA-seq transcripts from the species of interest in FASTA format, gene annotations in GFF3 format

1. Create MAKER control files as outlined in Basic Protocol 1, step 1.

2. Edit the `maker_opts.ctl` file to add the genomic sequence, evidence, and gene models you wish to add quality metrics to:

```
% emacs maker_opts.ctl
#-----Genome (these are always required)
genome=$PATH_TO_CBP_maker/maker_inputs/legacy_data/
  legacy-contig.fasta #genome sequence (fasta file or
  fasta embeded in GFF3 file)
organism_type=eukaryotic #eukaryotic or prokaryotic.
  Default is eukaryotic
#-----EST Evidence (for best results provide a file for
  at least one)
est=$PATH_TO_CBP_maker/maker_inputs/legacy_data/
  legecy-new-mRNAseq.fasta #set of ESTs or assembled
  mRNA-seq in fasta format
#-----Protein Homology Evidence (for best results
  provide a file for at least one)
protein=$PATH_TO_CBP_maker/maker_inputs/legacy_data/
  legacy-new-protein.fasta #protein sequence file in
  fasta format (i.e., from multiple organisms)
#-----Gene Prediction
model_gff=/home/mcampbell/project_links/CPB_maker/
  maker_inputs/legacy_data/legacy-set1.gff
#annotated gene models from an external GFF3 file
  (annotation pass-through)
```

Annotations given to MAKER as `model_gff` *remain unchanged in the final MAKER output and are kept regardless of evidence support.*

3. Run MAKER and check/collect the results as outlined in Basic Protocol 1, steps 3 to 5. Then repeat this procedure for `legacy-set2.gff`.

This gives us a GFF3 for each annotation set. They are renamed below for simplicity:

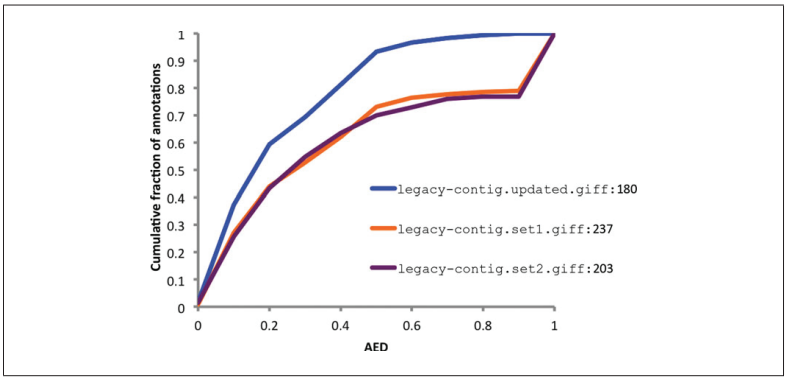

Figure 4.11.3 Adding quality metrics to legacy annotations facilitates comparison between annotation sets. Shown on the *y* axis is the cumulative distribution of AED for each dataset. The two legacy annotation sets are of comparable quality, with approximately 70% of their annotations having AEDs of less than 0.5 (orange and purple lines). Combining and updating legacy annotations results in a much improved annotation build (blue line), in which greater than 90% of the annotations have and AED less than 0.5.

```
legacy-contig.set1.gff
legacy-contig.set2.gff
```

Now that we have quality metrics for all of the annotations, we can compare them. A cumulative distribution function curve based on AED is a simple way to visually compare annotation sets. An example using the above annotation sets as well as the results from Basic Protocol 2, where we combined and updated these legacy annotations, is shown in Figure 4.11.3.

BASIC PROTOCOL 4

MAPPING ANNOTATIONS TO A NEW ASSEMBLY

Genome assemblies can change over time for a variety of reasons. Removing contaminants and improving assemblies with new genomic sequence data are two common reasons. Changes in the reference sequence make it necessary to also alter the beginning and ending coordinates of annotated genes. The simplest way to fix this is to use MAKER to map existing annotations forward onto the new assembly. The protocol below explains this process. Assembly changes can also invalidate pre-existing gene models, requiring structural revisions. In cases where the assembly has changed substantially. Basic Protocol 2, "Updating/combining legacy annotation datasets in light of new evidence," provides an easy means to simultaneously remap and update the existing gene annotations.

Necessary Resources

Hardware

Computer with a Unix-based operating system (e.g., Linux, Mac OS X)

Software

MAKER and MAKER-P are available for download at *yandell-lab.org*. Installation instructions are included in the tarball. For brevity's sake, the following protocols describe MAKER, but apply to MAKER-P as well.

MAKER will identify and download all of its necessary external dependencies including BLAST, Exonerate, RepeatMasker, and a number of Perl modules [automatic download and installation of Perl modules requires CPAN (*https://metacpan.org/pod/CPAN*) to be installed]. MAKER will also install a number of additional programs such as SNAP, Augustus, and MPICH2. This

example uses a version of MAKER installed with NCBI BLAST+, Exonerate, RepeatMasker, with optional RepBase libraries, and SNAP.

Files

The new genome assembly in FASTA format, and the manually curated transcripts in FASTA format. This example uses the transcripts generated by MAKER in Basic Protocol 1, and a version of the genomic sequence with 60 bases removed from the first intron of the MAKER annotated gene.

1. Create MAKER control files as outlined in Basic Protocol 1, step 1.
2. Edit the `maker_opts.ctl` file to add the changed genomic sequence and the transcripts you wish to map forward:

```
% emacs maker_opts.ctl
#-----Genome (these are always required)
genome=$PATH_TO_CBP_maker/maker_inputs/new_assembly/
  new_assembly.fasta #genome sequence (fasta file or
  fasta embeded in GFF3 file)
organism_type=eukaryotic #eukaryotic or prokaryotic.
  Default is eukaryotic
#-----EST Evidence (for best results provide a file for
  at least one)
est=$PATH_TO_CBP_maker/maker_inputs/new_assembly/
  manualy_curated_transcript.fasta #set of ESTs or
  assembled mRNA-seq in fasta format
#-----Gene Prediction
est2genome=1 #infer gene predictions directly from
  ESTs, 1 = yes, 0 = no
```

MAKER will align the manually curated transcripts to the genome. By setting `est2genome=1`*, MAKER will create gene models directly from those alignments.*

3. Manually add the following line to the `maker_opts.ctl` file:

```
est_forward=1
```

By setting this hidden option in MAKER, the sequence id from the FASTA header will maintained as part of the gene name in the GFF3 output.

4. Run MAKER and check/collect the results as outlined in Basic Protocol 1, steps 3 to 5.

Shown below are the lines for exon two in the GFF3 file. The coordinates for exon 2 have shifted by 60 bp. Further exploration of the MAKER outputs will also show that the final transcript and protein outputs have not changed between the two assemblies.

```
Original assembly exon 2
contig-dpp-500-500 maker exon 27104 27985
Updated assembly exon 2
contig-dpp-500-500 maker exon 27044 27925
```

THE MAKER GENE BUILD/ RESCUING REJECTED GENE MODELS

BASIC PROTOCOL 5

MAKER users can decide which gene models to include in their final annotation build. This is accomplished using the MAKER tools and procedures described below. The resulting datasets are termed either *default*, *standard*, or *max*. The MAKER *default build* includes only those gene models that are supported by the evidence (i.e., AED <1.0). The MAKER-P *standard build* includes every gene model in the default build, plus

every ab initio gene prediction that encodes a Pfam domain as detected by InterProScan (Quevillon et al., 2005), and does not overlap an annotation in the MAKER default set. The MAKER *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. We recommend that users choose the standard build, as previous work (Holt and Yandell, 2011; Campbell et al., 2014) has shown that this build procedure has the best overall accuracy. Nevertheless some users may prefer specificity to sensitivity, choosing the default build, whereas others may wish to include every possible gene model by using the max build procedure.

Necessary Resources

Hardware

Computer with a Unix-based operating system (e.g., Linux, Mac OS X)

Software

MAKER and MAKER-P are available for download at *yandell-lab.org*. Installation instructions are included in the tarball. For brevity's sake, the following protocols describe MAKER, but apply to MAKER-P as well.

MAKER will identify and download all of its necessary external dependencies including BLAST, Exonerate, RepeatMasker, and a number of Perl modules [automatic download and installation of Perl modules requires CPAN (*https://metacpan.org/pod/CPAN*) to be installed]. MAKER will also install a number of additional programs such as SNAP, Augustus, and MPICH2. This example uses a version of MAKER installed with NCBI BLAST+, Exonerate, RepeatMasker, with optional RepBase libraries, and SNAP.

InterProScan

Files

Genome assembly to be annotated in FASTA format

Protein evidence in FASTA format

Assembled mRNA-seq transcripts from the species of interest in FASTA format

Optional: a species parameter/HMM file for SNAP generated for the organism of interest or a closely related species. The process used to create a species parameter/HMM file is described in SNAP's internal documentation (Korf, 2004).

1. Create MAKER control files as outlined in Basic Protocol 1, step 1.
2. Edit the `maker_opts.ctl` file to add the genomic sequence, evidence, and specify the gene finding method(s):

```
% emacs maker_opts.ctl
#-----Genome (these are always required)
genome=$PATH_TO_CBP_maker/maker_inputs/pyu_data/pyu-
  contig.fasta #genome sequence (fasta file or fasta
  embeded in GFF3 file)
organism_type=eukaryotic #eukaryotic or prokaryotic.
  Default is eukaryotic
#-----EST Evidence (for best results provide a file for
  at least one)
est=$PATH_TO_CBP_maker/maker_inputs/pyu_data/pyu-
  est.fasta
#set of ESTs or assembled mRNA-seq in fasta format
#-----Protein Homology Evidence (for best results
  provide a file for at least one)
```

```
protein=$PATH_TO_CBP_maker/maker_inputs/pyu_data/pyu-
  protein.fasta #protein sequence file in fasta format
  (i.e., from multiple organisms)
#-----Gene Prediction
snaphmm=./pyu3.hmm #SNAP HMM file
#-----MAKER Behavior Options
keep_preds=1
#Concordance threshold to add unsupported gene
  prediction (bound by 0 and 1)
```

This step is very similar to step 2 in Basic Protocol 1. The key difference is setting `keep_preds=1` *in the MAKER behavior options section. Setting* `keep_preds=1` *prevents MAKER from rejecting unsupported gene models. The* `pyu3.hmm` *file was made using Support Protocol 1 and is found in the* `CPB_maker` *tarball described previously.*

3. Follow steps 3 to 5 in Basic Protocol 1.

Instructing MAKER to retain unsupported gene models trades specificity for sensitivity. See Yandell and Ence (2012) for discussion of annotation specificity and sensitivity issues. For small, very compact genomes such as those of many fungi, this approach often works quite well; i.e., the sensitivity/specificity trade-off is minimal. However, for larger eukaryotic genomes, such as large animal and plant genomes, setting `keep_preds=1` *can result in thousands of false-positive gene models, so further filtering is necessary. One of the simplest ways to identify true positives is to run InterProScan (Quevillon et al., 2005) on the MAKER annotations. The idea is that a gene model without EST or protein homology that encodes a known protein domain is likely to be a true positive.*

4. Run InterProScan on the MAKER generated proteins to identify proteins with known functional domains:

```
% interproscan.sh -appl PfamA -iprlookup -goterms -f
  tsv\
-i pyu-contig.all.fasta
```

The above example uses the stand-alone version of InterProScan and limits the search to Pfam domains. InterProScan can be run multiple ways and any of them that output a `.tsv` *file will work.*

5. Update the MAKER generated GFF3 file with the InterProScan results using `ipr_update_gff`:

```
% ipr_update_gff contig-dpp-500-500.gff\
pyu-contig.all.maker.proteins.fasta.tsv\
pyu-contig.max.functional_ipr.gff
```

This procedure added a `Dbxref` *tag to column nine of the gene and mRNA features that have Pfam domains identified by InterProScan in the GFF3 file. The value for this tag contains InterPro and Pfam ids as well as the Gene Ontology ids associated with the identified domains, and looks like this:* `Dbxref=InterPro:IPR001300,Pfam:PF00648;Ontology_term=GO:0004198,GO:0005622,GO:0006508`. *The resulting GFF3 file from this command serves as the MAKER max build containing all gene models regardless of evidence support.*

6. Use the `quality_filter.pl` script distributed with MAKER to filter the gene models based on domain content and evidence support. Start by running `quality_filter.pl` without any options to see the usage:

```
% quality_filter.pl
quality_filter.pl: generates default and standard
```

```
 gene builds from a maker generated gff3_file with
 iprscan data pushed onto column 9 using
 ipr_update_gff.
USAGE: quality_filter.pl -[options] <gff3_file>
OPTIONS: -d Prints transcripts with an AED <1 (MAKER
 default)
-s Prints transcripts with an AED <1 and/or Pfam
 domain if in gff3 (MAKER Standard)
-a <number between 0 and 1> Prints transcripts
 with an AED < the given value
```

We can generate the MAKER default build and the MAKER standard build using the `-d` and `-s` options respectively:

```
% quality_filter.pl -d pyu-
  contig.max.functional_ipr.gff\
> pyu-contig.default.functional_ipr.gff
% quality_filter.pl -s pyu-
  contig.max.functional_ipr.gff\
> pyu-contig.standard.functional_ipr.gff
```

When we count the number of genes in these two files, we can see that we were able to rescue 161 genes that were not annotated due to lack of evidence but are supported by Pfam domain content:

```
% grep -cP '\tgene\t'\
pyu-contig.default.functional_ipr.gff
404
% grep -cP '\tgene\t'\
pyu-contig.standard.functional_ipr.gff
565
```

This procedure was used in the MAKER-P paper for benchmarking MAKER-P on the Arabidopsis genome. When the gene models with Pfam domain support were included, sensitivity improved at the expense of specificity, but the best accuracy was obtained using the TAIR10 annotations as truth (Campbell et al., 2014).

GUIDELINES FOR UNDERSTANDING RESULTS

MAKER and MAKER-P are designed with three general use-case scenarios in mind. These are (1) de novo annotation of new genomes; (2) updating annotations to reflect assembly changes and/or new evidence; and (3) quality control of genome annotations. Classic model-organism genomes such mouse (Waterston et al., 2002), *C. elegans* (Press et al., 1998), and *Drosophila melanogaster* (Adams et al., 2000) benefited from pre-existing gold-standard gene annotations. These were used to train gene finders and to evaluate the accuracy of genome annotations. In contrast, the genomes being sequenced today are novel, and their contents are unknown. Thus, evidence, in the form of transcript and protein alignments, must be used as a surrogate for gold-standard annotations. Accordingly, MAKER and MAKER-P provide means for employing transcript and protein alignments to train gene finders and for evaluating the accuracy of the genome annotations, i.e., quality control. These operations are primarily accomplished using Annotation Edit Distance (AED). AED is a distance measure that summarizes the congruency of each annotation with its supporting evidence. A value of 0 indicates that the annotation matches the evidence perfectly, while a value of 1 indicates that the annotation has no evidence support. See Yandell and Ence (2012) for more discussion on this topic; also see Cantarel et al. (2008); Eilbeck et al. (2009; Holt and Yandell (2011).

Protein domain content provides another means to judge the quality of de novo protein coding annotations. Previous work (Holt and Yandell, 2011) has shown that somewhere between 55% to 65% of the proteins comprising a well annotated eukaryotic proteome will contain a recognizable domain. See Basic Protocol 5 for more on how to employ MAKER and MAKER-P to carry out domain-based analyses of annotations.

Together, AED and proteome domain content provide two simple summary statistics with which to globally compare one genome's annotations to another's (Holt and Yandell, 2011). As a rule of thumb, a genome annotation build where 90% of the annotations have an AED less than 0.5, and over 50% of its proteome contains a recognizable domain, can be considered well annotated (Holt and Yandell, 2011; Yandell and Ence, 2012; Campbell et al., 2014).

Gene number is a third important summary statistic for evaluating the overall quality of a genome annotation build. Clearly, a build comprising only a handful of genes is hardly a satisfactory result, no matter how domain-rich their proteins, or how well they agree with the transcript and protein alignment evidence. Unfortunately, there is no sure way to determine gene number for a genome. Some guidance, however, can be had from considering gene numbers from model organism genomes. Generally—and biology is full of exceptions—MAKER users should expect to see somewhere around 10,000 protein-coding annotations for fungal genome, between 12,000 and 20,000 for an invertebrate genome, and around 20,000 to 30,000 for a vertebrate genome. Plant gene numbers are even more difficult estimate because whole-genome duplications are common in plant evolution, but somewhere between 20,000 and 40,000 protein-coding genes are a good first guess. Consider too that fragmented assemblies will inflate these numbers, as a gene will often be split across multiple scaffolds. Again, keep in mind that these are ballpark figures. Biology is all about exceptions to the rule. This is one reason that MAKER and MAKER-P offer three different annotation build protocols: default, standard, and max. Generally, the MAKER default build provides a useful lower bound of well annotated genes with which to estimate gene numbers, the max build an upper bound, and the standard build a best first estimate for gene number.

COMMENTARY

Background Information

MAKER was developed as an easy-to-use annotation pipeline for emerging model organism genomes (Cantarel et al., 2008). The overarching goal of MAKER was to enable small, independent research groups without extensive bioinformatics expertise or resources to annotate genomes.

MAKER 2 is a backwardly compatible extension of MAKER (Holt and Yandell, 2011). MAKER2 improved MAKER's gene-finding capabilities, offering improved, dynamic means to inform gene predictors, and provided new means for quality control using AED, as well as means for updating legacy annotations in light of new transcript and protein evidence.

MAKER-P is designed to address the needs of the plant genome community. MAKER-P provides means for annotation of complex plant genomes, and for automated revision, quality control, and management of existing genome annotations. MAKER-P also provides means for annotation of ncRNA genes and pseudogene annotation. MAKER-P is dramatically faster than other genome-annotation pipelines, including the original MAKER2, allowing it to scale to even the largest plant genomes. Recent work, for example, has shown that the version of MAKER-P available within the iPlant Cyberinfrastructure can re-annotate the entire maize genome in less than 3 hr (Campbell et al., 2014), and that it can carry out the complete de novo annotation of the 17.83-GB draft loblolly pine genome in less than 24 hr (Neale et al., 2014; Wegrzyn et al., 2014). MAKER-P can be used to annotate any genome, not just plants, and is now the main production release of the MAKER pipeline.

MAKER, MAKER2, and MAKER-P are available for download at *http://www.yandell-lab.org*. In addition, MAKER-P is available on the iPlant Cyberinfrastructure. Instructions for using MAKER-P on iPlant can be found at *https://pods.iplantcollaborative.org/wiki/display/sciplant/MAKER-P+at+iPlant*.

Critical Parameters

Critical parameters are defined here as parameters that will have global effects on the de novo annotation of a genome. These fall into four broad classes: repeat masking, evidence alignment, gene prediction, and MAKER behavior.

Repeat masking. Good repeat masking is essential in producing high-quality gene annotations. When not adequately masked, portions of transposable elements can be erroneously included in annotations of neighboring protein-coding genes. Species-specific repeat libraries will provide the best masking, especially for organisms that are phylogenetically distant from those currently found in RepBase (Jurka et al., 2005). See *http://weatherby.genetics.utah.edu/MAKER/wiki/index.php/Repeat_Library_Construction--Basic* and *http://weatherby.genetics.utah.edu/MAKER/wiki/index.php/Repeat_Library_Construction--Advanced* for basic and advanced protocols for generating species specific repeat libraries, respectively.

Evidence. It is crucial that MAKER has access to as complete an evidence dataset as possible. Ideally these data will include assembled RNA-seq transcripts from several tissues and developmental time points, as well as the complete proteomes of both a closely related organism and of an outgroup to account for lineage-specific gene loss. It is also advisable to include an omnibus protein dataset such as UniProt/Swiss-Prot. If RNA-seq data is not available, high-quality gene annotations can still be obtained from protein data alone, but they will lack untranslated regions (UTR), and MAKER may miss genes specific to the organism at hand. Remember that, by default, MAKER will not annotate genes that have no evidence support, so incomplete evidence datasets can lead to lower overall gene counts.

Gene prediction. It is important to understand that MAKER does not predict genes; rather, the gene finders you select in the control files predict the genes (SNAP, Augustus, etc.). Poorly trained gene finders will result in lower-quality final annotations. The gene finders will perform better inside of MAKER than they would have on their own because of evidence-derived hints being passed to them by MAKER (see Holt and Yandell, 2011, for more on this point), but the better trained the gene finder, the better this process will work. See Support Protocol 1 for directions for using MAKER to train gene finders. Consider too that not all gene finders perform well on every organism. A gene finder that performs well on fungi may not perform as well on plants and animals. Don't be afraid to remove a poorly performing gene finder from your analysis. Poor performance from multiple gene predictors likely indicates other problems such as insufficient repeat masking. You may want to build a species-specific repeat library for use with MAKER. A Web tutorial outlining this process is available at *http://weatherby.genetics.utah.edu/MAKER/wiki/index.php/Repeat_Library_Construction--Advanced*. There may also be widespread assembly errors or assembly fragmentation problems (these break open reading frames and erase potential splice sites, making it impossible to generate accurate annotations). Programs such as CEGMA (Parra et al., 2007) can be used to estimate what fraction of gene content will be recoverable from your genome assembly.

MAKER behavior. Important MAKER options that should be kept in mind include `split_hit`, `max_dna_len`, and `single_exon`. These options are set in the `maker_opts.ctl` file. The value of `split_hit` can be thought of as the longest intron that you expect in your genome. As a rule of thumb, 20 kb for vertebrates and 40 kb for mammals are reasonable values to try first. The default 10 kb works for many plants and most invertebrates and fungi. However, you may want to set it even lower for gene-dense genomes with short introns. Setting this value too low will result in truncated annotations, while setting it too high can result in concatenated genes (as evidence alignments will extend across neighboring paralogs). The `max_dna_len` parameter controls the window size for the genomic blocks MAKER will operate on at a time (larger values increase memory usage). This value must be set to at least three times the `split_hit` value to avoid issues with very large genes extending across multiple windows.

The `single_exon` parameter controls whether or not MAKER will consider single-exon EST alignments when generating hints for gene predictors. It is turned off by default. Setting `single_exon=1` will allow MAKER to annotate single-exon genes based on unspliced EST/mRNA-seq data, but will also greatly increase the false-positive rate for gene annotation. Single-exon alignments often result from spurious alignments, library contamination, background transcription of the genome, pseudogenes, and repeat elements. These facts should be considered carefully before enabling the `single_exon` parameter.

Nevertheless, for intron-poor genomes, you may want to turn this option on. If you choose to do so, the `single_length` parameter can be used to set a minimum size for single-exon alignments to accept. Shorter alignments are more likely to be spurious than longer alignments; 250 base pairs is a good minimum value for this parameter.

Troubleshooting

MAKER users should subscribe to the `MAKER_dev` mailing list (*http://yandell-lab.org/mailman/listinfo/maker-devel_yandell-lab.org*). Answers to common MAKER use errors can be found by searching the archived posts from the MAKER mailing list found at *https://groups.google.com/forum/#!forum/maker-devel*.

Advanced Parameters

MAKER has a large number of options and parameters. For a full list of the MAKER control file options including descriptions, see Table 4.11.2.

Acknowledgments

This work was supported by NSF IOS-1126998 to MY. A portion of M.C.'s efforts were supported by NIH 1R01GM104390-01 to MY.

Literature Cited

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J. D., Amanatides, P. G., and Venter, J.C. 2000. The genome sequence of *Drosophila melanogaster*. *Science* 287:2185-2195.

Bradnam, K.R., Fass, J.N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., Boisvert, S., Chapman, J.A., Chapuis, G., Chikhi, R., Chitsaz, H., Chou, W.C., et al. 2013. Assemblathon 2: Evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience* 2:10. doi:10.1186/2047-217X-2-10.

Campbell, M., Law, M., Holt, C., Stein, J., Moghe, G., Hufnagel, D., Lei, J., Achawanantakun, R., Jiao, D., Lawrence, C.J., Ware, D., Shiu, S.H., Childs, K.L., Sun, Y., Jiang, N., and Yandell, M. 2013. MAKER-P: A tool-kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiol.* 164:513-524.

Cantarel, B.L., Korf, I., Robb, S.M.C., Parra, G., Ross, E., Moore, B., Holt, C., Sanchez Alvarado, A., and Yandell, M. 2008. MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. 18:188-196.

Eilbeck, K., Moore, B., Holt, C., and Yandell, M. 2009. Quantitative measures for the management and comparison of annotated genomes. *BMC Bioinformatics* 10:67.

Goff, S.A., Vaughn, M., McKay, S., Lyons, E., Stapleton, A.E., Gessler, D., Matasci, N., Wang, L., Hanlon, M., Lenards, A., Muir, A., Merchant, N., et al. 2011. The iPlant collaborative: Cyberinfrastructure for plant biology. *Front. Plant Sci.* 2:34.

Holt, C. and Yandell, M. 2011. MAKER2: An annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* 12:491.

Jurka, J., Kapitonov, V.V, Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* 110:462-467.

Korf, I. 2004. Gene finding in novel genomes. *BMC Bioinformatics* 5:59.

Law, M., Childs, K.L., Campbell, M.S., Stein, J.C., Holt, C., Olson, A.J., Holt, C., Lei, J., Jiao, D. Andorf, C.M., Ware, D. Shiu, S.-H., Sun, Y. Jiang, N., and Yandell, M. 2014. Automated update, revision and quality control of the *Zea mays* genome annotations using MAKER-P improves the B73 RefGen_v3 gene models and identifies new genes. *Plant Physiol.* In press.

Lévesque, C.A., Brouwer, H., Cano, L., Hamilton, J.P., Holt, C., Huitema, E., Raffaele, S., Robideau, G.P., Thines, M., Win, J., Zerillo, M.M. Beakes, G.W., et al. 2010. Genome sequence of the necrotrophic plant pathogen *Pythium ultimum* reveals original pathogenicity mechanisms and effector repertoire. *Genome Biol.* 11:R73.

Lipman, D.J. and Pearson, W.R. 1985. Rapid and sensitive protein similarity searches. *Science* 227:1435-1441.

Lowe, T.M. 1999. A computational screen for methylation guide snoRNAs in yeast. *Science* 283:1168-1171.

Lowe, T.M. and Eddy, S.R. 1997. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25:955-964.

Campbell, M.S., Law, M., Holt, C., Stein, J.C., Moghe, G.D., Hufnagel, D.E., Lei, J., Achawanantakun, R., Jiao, D., Lawrence, C.J., Ware, D., Shiu, S.H., Childs, K.L., Sun, Y., Jiang, N., and Yandell, M. 2014. MAKER-P: A tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiol.* 164:513-524.

Neale, D.B., Wegrzyn, J.L., Stevens, K.A., Zimin, A.V., Puiu, D., Crepeau, M.W., Cardeno, C., Koriabine, M., Holtz-Morris, A.E., Liechty, J.D., Martínez-García, P.J., Vasquez-Gross, H.A., et al. 2014. Decoding the massive genome of loblolly pine using haploid DNA and novel assembly strategies. *Genome Biol.* 15:R59.

Parra, G., Bradnam, K., and Korf, I. 2007. CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23:1061-1067.

Press, H., York, N., and Nw, A. 1998. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* 282:2012-2018.

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., and Lopez, R. 2005. InterProScan: Protein domains identifier. *Nucleic Acids Res.* 33:W116-W120.

Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., Antonarakis, S.E., Attwood, J., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420:520-562.

Wegrzyn, J.L., Liechty, J.D., Stevens, K.A., Wu, L.-S., Loopstra, C.A., Vasquez-Gross, H.A., Dougherty, W.M., Lin, B.Y., Zieve, J.J., Martínez-García, P.J., Holt, C., Yandell, M., Zimin, A.V., Yorke, J.A., Crepeau, M.W., Puiu, D., Salzberg, S.L., Dejong, P.J., Mockaitis, K., Main, D., Langley, C.H., and Neale, D.B. 2014. Unique features of the loblolly pine (*Pinus taeda* L.) megagenome revealed through sequence annotation. *Genetics* 196:891-909.

Yandell, M. and Ence, D. 2012. A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.* 13:329-342.

Zimin, A., Stevens, K.A., Crepeau, M.W., Holtz-Morris, A., Koriabine, M., Marçais, G., Puiu, D., Roberts, M., Wegrzyn, J.L., de Jong, P.J., Neale, D.B., Salzberg, S.L., Yorke, J.A., and Langley, C.H. 2014. Sequencing and assembly of the 22-gb loblolly pine genome. *Genetics* 196:875-890.

Zou, C., Lehti-Shiu, M.D., Thibaud-Nissen, F., Prakash, T., Buell, C.R., and Shiu, S.-H. 2009. Evolutionary and expression signatures of pseudogenes in Arabidopsis and rice. *Plant Physiol.* 151:3-15.

Internet Resources

http://www.sequenceontology.org/gff3.shtml

Generic Feature Format version 3.

https://metacpan.org/pod/CPAN

CPAN Web site (A. König, developer).

Table 4.11.2 MAKER Control File Options Found in the `maker_opts.ctl` and `maker_bopts.ctl` Files

Option	Comments
`maker_opts.ctl`	
`#-----Genome (these are always required)`	Headings for sections in the control files are marked by a pound sign and five dashes. These headings are not actually used by MAKER but are helpful when trying to find a specific option or parameter.
`genome= #genome sequence (fasta file or fasta embeded in GFF3 file)`	This is a single multifasta file that contains the assembled genome. Both absolute and relative file paths are allowed. It is also important to note that though there are a large number of characters accepted by FASTA format to represent nucleotides, many of them are not supported by some of the tools MAKER calls, so make sure that FASTA sequence contains only A, T, C, G, and N.
`organism_type=eukaryotic #eukaryotic or prokaryotic. Default is eukaryotic`	MAKER's default is eukaryotic. Setting this to prokaryotic changes some of MAKER's behavior options automatically (such as turning off repeat masking).
`#-----Re-annotation Using MAKER Derived GFF3`	This section was developed as a convenience method for using the output of a previous MAKER run as the evidence to a new MAKER run.
`maker_gff= #MAKER derived GFF3 file`	Path to the MAKER generated GFF3 file.
`est_pass=0 #use ESTs in maker_gff: 1 = yes, 0 = no`	Set to `1` to use the EST/mRNA-Seq alignments from the MAKER file. See `est=` below for details.
`altest_pass=0 #use alternate organism ESTs in maker_gff: 1 = yes, 0 = no`	Set to `1` to use the alternative EST/mRNA-seq alignments from the MAKER file. See `altest=` below for details.

continued

Table 4.11.2 MAKER Control File Options, *continued*

Option	Comments
`maker_opts.ctl`	
`protein_pass=0 #use protein alignments in maker_gff: 1 = yes, 0 = no`	Set to 1 to use the protein alignments from the MAKER file. See `protein=` below for details.
`rm_pass=0 #use repeats in maker_gff: 1 = yes, 0 = no`	Set to 1 to use the repeat masking data from the MAKER file. See the `#-----Repeat Masking` section below for details.
`model_pass=0 #use gene models in maker_gff: 1 = yes, 0 = no`	Set to 1 to use the gene models from the MAKER file. See `model_gff=` below for details.
`pred_pass=0 #use ab-initio predictions in maker_gff: 1 = yes, 0 = no`	Set to 1 to use the gene predictions from the MAKER file. See `pred_gff=` below for details.
`other_pass=0 #passthrough anyything else in maker_gff: 1 = yes, 0 = no`	Set to 1 to pass any other features through from a previous MAKER file. See other_`gff=`, below for details.
`#-----EST Evidence (for best results provide a file for at least one)`	This section contains options pertaining to Transcript Evidence., e.g., EST, mRNA-seq and assembled full length cDNAs. These are assumed to be correctly assembled and they will be aligned in a splice aware fashion (MAKER uses Exonerate to do this). MAKER can use these alignments to infer gene models directly when the est2genome option is turned on. MAKER also uses them as support for intron/exon boundaries in hints sent to the gene finders, and for AED calculations. MAKER also uses these data to infer alternate splice forms and UTR regions. How these alignments cluster with other evidence (protein, for example) will help MAKER infer gene boundaries in some cases.
`est= #set of ESTs or assembled mRNA-seq in fasta format`	Specifies files containing assembled mRNA-Seq transcripts, ESTs, or full-length cDNAs. You may provide multiple files in a comma-separated list.
`altest= #EST/cDNA sequence file in fasta format from an alternate organism`	Specifies files containing assembled mRNA-Seq transcripts, ESTs, or full-length cDNAs from *another* related organism. This option is useful when there is no transcript evidence available for the genome at hand, but this data is available for a closely related species. However, these alignments are done using tblastx, which makes these data very expensive computationally. Use protein evidence from a relate species if at all possible before using transcript evidence. You may provide multiple files in a comma-separated list.
`est_gff= #aligned ESTs or mRNA-seq from an external GFF3 file`	These are prealigned transcripts from the organism being annotated in GFF3 format. The most common sources of these kinds of data are alignment based transcript assemblers such as cufflinks, or outputs from a previous MAKER run. You may provide multiple files in a comma-separated list.

continued

Table 4.11.2 MAKER Control File Options, *continued*

Option	Comments
`maker_opts.ctl`	
`altest_gff= #aligned ESTs from a closly relate species in GFF3 format`	These are prealigned transcripts from a related species in GFF3 format. The most common source of these kinds of data is output from a previous MAKER run. You may provide multiple files in a comma-separated list.
`#-----Protein Homology Evidence (for best results provide a file for at least one)`	This section of the control file covers options controlling the use of protein homology evidence. Protein homology evidence helps MAKER locate coding regions and gene boundaries. These alignments will also be used to generate hints for the gene finders and as part of the AED calculation.
`protein= #protein sequence file in fasta format (i.e., from mutiple oransisms)`	This is a collection of protein sequences (usually from related species) in FASTA format. A minimum of one full proteome from a related species should be used. Multiple files in a comma-separated list are allowed.
`protein_gff= #aligned protein homology evidence from an external GFF3 file`	These are pre-aligned proteins in GFF3 format. The most common source of these data is a previous MAKER run. Multiple files in a comma-separated list are allowed.
`#-----Repeat Masking (leave values blank to skip repeat masking)`	Repeats will be masked to stop EST and proteins from aligning to repetitive regions and to keep gene prediction algorithms from being allowed to call exons in those regions.
`model_org=all #select a model organism for RepBase masking in RepeatMasker`	Specifies the model organism to use for RepeatMasker when RepBase libraries are installed. Common values are mammal, grass, primate, fungi, etc. The genus and species can also be used so long as they are bound by double quotes e.g., "`drosophila melanogaster`". See RepeatMasker documentation for valid entries. You can also use the value `simple` to specify only low complexity repeats (this option is MAKER specific). If you have gone to the trouble of making a custom repeat library (i.e., `rmlib=`) you do not need to use RepBase and may leave this option blank.
`rmlib= #provide an organism specific repeat library in fasta format for RepeatMasker`	Specifies the location of a custom repeat library. The file should be in FASTA format.
`repeat_protein= #provide a fasta file of transposable element proteins for RepeatRunner`	Specify transposable element proteins in FASTA format. A default file comes packaged with MAKER. These are aligned in protein space to help mask known transposable elements that have diverged over time.
`rm_gff= #pre-identified repeat elements from an external GFF3 file`	These are pre-aligned repeats in GFF3 format. The most common source of these data is a previous MAKER run, but these are also available from some organism databases.

continued

Table 4.11.2 MAKER Control File Options, *continued*

Option	Comments
`maker_opts.ctl`	
`prok_rm=0 #forces MAKER to repeatmask prokaryotes (no reason to change this), 1 = yes, 0 = no`	As a general rule, masking a prokaryotic genome is unnecessary and can lead to truncated gene models.
`softmask=1 #use soft-masking rather than hard-masking in BLAST (i.e., seg and dust filtering)`	Soft-masking in BLAST prevents alignments from seeding in regions of low complexity but allows alignments to extend through these regions.
`#-----Gene Prediction`	This section covers the gene finders used by MAKER. Unless gene finders are specified, MAKER will not annotate any genes. MAKER will run each gene predictor without hints once (ab initio predictions) and once with hints. Models produced by the gene finder will only be maintained in the final annotation set if there is some form of evidence supporting their structure. If multiple models overlap, only the one with the lowest AED (best evidence match) will be maintained in the final annotation set.
`snaphmm= #SNAP HMM file`	Specifies the location of the HMM file required to run SNAP. Always use an HMM specific for the genome at hand if at all possible, although a related species can be used to generate models that can then be used for training for the genome at hand. Multiple files in a comma-separated list are allowed.
`gmhmm= #GeneMark HMM file`	Specify an HMM file for GeneMark. Multiple files in a comma-separated list are allowed.
`augustus_species= #Augustus gene prediction species model`	Specify the species model to use for Augustus. This is just a name and not a file path. To get a list of valid options, look in the `../augustus/config/species` directory. Multiple files in a comma- separated list are allowed.
`fgenesh_par_file= #FGENESH parameter file`	Location of an FGENESH parameter file. Multiple files in a comma- separated list are allowed.
`pred_gff= #ab-initio predictions from an external GFF3 file`	Predictions from any gene finder can be used in MAKER, so long as the gene finder's output has been converted to GFF3 format. Multiple files in a comma- separated list are allowed.
`model_gff= #annotated gene models from an external GFF3 file (annotation pass-through)`	These are assumed to be high-confidence gene models usually from a previous annotation of the genome. Because these models are considered high confidence, they will be used to merge evidence clusters around existing loci. This clustering will slightly bias MAKER towards keeping rather than replacing previous models for borderline cases. MAKER is only allowed to keep or replace these models and cannot modify them, although if `map_forward=1` is set, their names will be mapped forward onto whatever model replaces them. If no evidence supports these models,

continued

Table 4.11.2 MAKER Control File Options, *continued*

Option	Comments
maker_opts.ctl	
	MAKER will still keep them because they are assumed to be high confidence (but MAKER will tag them with an AED score of 1). Multiple files in a comma- separated list are allowed.
est2genome=0 #infer gene predictions directly from ESTs, 1 = yes, 0 = no	This option is used to create gene models directly from the transcript evidence. This option is useful when no gene predictor is trained on your organism or there is not a training file available from a closely related organism. The gene models from this option will be fragmented and incomplete because of the nature of transcript data (especially mRNA-Seq). These gene models are most useful for first round training of gene finders. Once there is a trained gene predictor, turn this option off.
protein2genome=0 #infer predictions from protein homology, 1 = yes, 0 = no	Similar to est2genome. This option will make gene models directly from protein alignments. Like est2genome this option is most useful for training gene predictors and should be turned off afterwards.
trna=0 #find tRNAs with tRNAscan, 1 = yes, 0 = no	Set to 1 to use tRNAscan-SE to annotate tRNAs.
snoscan_rrna= #rRNA file to have Snoscan find snoRNAs	Specify a FASTA file containing rRNAs that will be used by snoscan to annotate snoRNAs.
unmask=0 #also run ab-initio prediction	This option lets the gene finders run on the unmasked sequence as well as the masked sequence.
programs on unmasked sequence, 1 = yes, 0 = no	This allows repetitive regions to be included in gene models (does not affect evidence alignment), which may be useful in cases where over masking of repeats is suspected.
#-----Other Annotation Feature Types (features MAKER doesn't recognize)	This section covers parameters that allow users to add additional annotations to MAKER's set.
other_gff= #extra features to pass-through to final MAKER generated GFF3 file	These are GFF3 lines you just want MAKER to add to your files. These are things MAKER does not annotate: promoter/enhancer regions, CpG islands, restrictions sites, etc. MAKER will not attempt to validate the features, but will just pass them through "as is" into the final GFF3 file. Multiple files in a comma-separated list are allowed.
#-----External Application Behavior Options	These options are passed to external programs like BLAST and can usually be left as default, especially if you are running MAKER with MPI.
alt_peptide=C #amino acid used to replace non-standard amino acids in BLAST databases	This option allows the user to specify amino acid codes that will be used to replace non-standard amino acids in protein alignment databases used by BLAST and Exonerate. Cysteine (C) is the default because it has the lowest overall substitution penalty of all of the amino acids in the BLOSUM matrix.

continued

Table 4.11.2 MAKER Control File Options, *continued*

Option	Comments
`maker_opts.ctl`	
`cpus=1 #max number of CPUs to use in BLAST and RepeatMasker (not for MPI, leave 1 when using MPI)`	Specifies the number of CPUs to use when running BLAST. If using MAKER with MPI, leave this as 1 or it will act like a multiplier to the CPUs already specified by mpiexec and can overburden your job.
`#-----MAKER Behavior Options`	These options affect internal MAKER behavior. They can be tuned to help MAKER run more effectively.
`max_dna_len=100000 #length for dividing up contigs into chunks (increases/decreases memory usage)`	Affects the window size used by MAKER for looking at blocks of the genome. Larger values use more memory. It is important that this parameter be at least three times the expected maximum intron size, or genes can bridge multiple windows and performance will suffer. 300,000 is a good `max_dna_len` on large vertebrate genomes if memory is not a limiting factor.
`min_contig=1 #skip genome contigs below this length (under 10kb are often useless)`	Causes MAKER to skip short contigs without attempting to annotate them. For large, repeat-rich genomes, setting this option to 10,000 can decrease run time without sacrificing annotation quality because contigs shorter than this are usually un-annotatable (they are too short to contain a full gene).
`pred_flank=200 #flank for extending evidence clusters sent to gene predictors`	Gene finders require flanking sequence on either side of a gene to correctly find start and stop locations. This parameter adds flanking sequence to evidence clusters to ensure the required flanking sequence is there. This option also affects how close evidence islands must be before clustering together. If you are annotating a genome with a sparse/fragmented evidence set increasing this value can capture exons missing from your evidence. Decreasing this value can help decrease gene mergers in organisms with high gene density.
`pred_stats=0 #report AED and QI statistics for all predictions as well as models`	Adds AED and QI statistics to the reference ab initio models in the GFF3. This can be computationally expensive, but can be useful when evaluating rejected gene models.
`AED_threshold=1 #Maximum Annotation Edit Distance allowed (bound by 0 and 1)`	Restricts the final gene models to have at least a given threshold of evidence support. Setting this option to a value lower than 1 will result in a final annotation set with fewer gene models but they will be better supported by the evidence.
`min_protein=0 #require at least this many amino acids in predicted proteins`	Sometimes gene predictors can generate very short predictions, especially on fragmented genomes with very short contigs. Setting this option can filter them out from the final annotation set.
`alt_splice=0 #Take extra steps to try and find alternative splicing, 1 = yes, 0 = no`	When this parameter is set to `0` MAKER will generate a single transcript for each gene that best matches the evidence. When set to `1`, MAKER will separate the transcript evidence into mutually

continued

Table 4.11.2 MAKER Control File Options, *continued*

Option	Comments
`maker_opts.ctl`	
	exclusive intron/exon sets. The information from each evidence set is then independently given to the gene finders as hints. If the gene finder predicts an alternative transcript using the alternate evidence set, then it is kept as an isoform in the final GFF3 output. Be careful when using this feature of MAKER in conjunction with noisy RNA-seq data, as this can result in an excess of alternative transcripts being predicted.
`always_complete=0 #extra steps to force start and stop codons, 1 = yes, 0 = no`	Will extend or truncate gene models to try and force canonical start/stop codons even if they are not biologically correct.
`map_forward=0 #map names and attributes forward from old GFF3 genes, 1 = yes, 0 = no`	When a gene from `model_gff` input is replaced with a new model, that new model will inherit the name from the model it replaced. Allows for naming conservation when re-annotating a genome.
`keep_preds=0 #Concordance threshold to add unsupported gene prediction (bound by 0 and 1)`	This is used when you want an annotation set with maximum sensitivity. As a general rule, gene finders tend to over-predict on novel genomes, so MAKER rejects models that do not have at least some form of evidence support. This flag removes the evidence support requirement. On some genomes with high gene density where over-prediction is modest (fungi, oomycetes, etc.), setting this parameter to `1` can be beneficial. However, doing so on larger plant and animal genomes can lead to false-positive gene calls, outnumbering true gene models by an order of magnitude or more.
`split_hit=10000 #length for the splitting of hits (expected max intron size for evidence alignments)`	This option is currently used to keep BLAST from aligning transcripts and proteins with exons unreasonably far apart, which can causes false merging of neighboring paralogs or spurious alignment of terminal exons.
`single_exon=0 #consider single exon EST evidence when generating annotations, 1 = yes, 0 = no`	By default MAKER does not use single exon transcript alignments as supporting evidence for gene models. Single exon alignments overwhelmingly represent spurious alignments, library contamination, background transcription of the genome, pseudogenes, and repeat elements. This somewhat decreases the sensitivity of MAKER, but greatly improves the specificity and overall accuracy. Turn this parameter on if the genome contains many single exon genes.
`single_length=250 #min length required for single exon ESTs if 'single_exon is enabled'`	If `single_exon` is set to `1`, this option filters out the shortest alignments, because spurious alignments and contamination tend to be biased toward shorter sequences.

continued

Table 4.11.2 MAKER Control File Options, *continued*

Option	Comments
`maker_opts.ctl`	
`correct_est_fusion=0 #limits use of ESTs in annotation to avoid fusion genes`	This option helps prevent merging of gene models because of overlapping UTRs (common in fungal genomes) or because of falsely merged RNA-seq assemblies (e.g., you did not turn on the Jaccardian clip option when running Trinity). If you see gene models where transcript evidence is causing a neighboring gene model to be merged into the UTR, or you see gene models that are being rejected only because they slightly overlap the UTR of a neighboring gene, then turn this option on. It will trim back the low confidence UTRs on both genes to allow both models into the final annotation set.
`tries=2 #number of times to try a contig if there is a failure for some reason`	Sets the maximum number of retries before MAKER considers an assembly contig to have failed. Large computes especially in cluster environments can be hindered by random failures caused by the network or I/O performance. This option gets past such failures by just trying again. It will not, however, get around systematic failures caused by errors in your dataset.
`clean_try=0 #remove all data from previous run before retrying, 1 = yes, 0 = no`	MAKER tries to recover from failures before trying a contig again, and it starts off where it left off in the analysis. However, some failures can result in irrecoverable file corruption that MAKER cannot fix. In those cases, it is better to just delete all files from the contig and start again from scratch. This is the best way to get around stubborn random failures caused by slow or unreliable NFS file system implementations.
`clean_up=0 #removes theVoid directory with individual analysis files, 1 = yes, 0 = no`	This option will help save disk space by deleting individual raw results files (such as BLAST, Exonerate, and gene predictor outputs) once they are no longer needed. If you have the disk space it is usually best to keep this set to `0`. Having those files around will make rerunning MAKER much faster if it's ever necessary.
`TMP= #specify a directory other than the system default temporary directory for temporary files`	Many programs MAKER uses create temporary files, and some programs need fast I/O performance or non-NFS storage to run correctly. MAKER uses `/tmp` or whatever your system's temporary directory is by default; however, you may specify an alternate location. Never specify an NFS-mounted location, however, or MAKER will fail in a very ugly way.
`maker_bopts.ctl`	
`blast_type=ncbi+ #set to 'ncbi+', 'ncbi' or 'wublast'`	MAKER can use three of the major BLAST engines. Choosing a BLAST engine is more likely to be influenced by what flavor of BLAST is installed on the system rather than performance of one over the other.

continued

Table 4.11.2 MAKER Control File Options, *continued*

Option	Comments
`maker_opts.ctl`	
`pcov_blastn=0.8 #Blastn Percent Coverage Threhold EST-Genome Alignments`	Sets the required percent coverage (end-to-end) for an EST/mRNA-seq alignment to be maintained as evidence.
`pid_blastn=0.85 #Blastn Percent Identity Threshold EST-Genome Aligments`	Sets the required percent identity for an EST/mRNA-seq alignment to be maintained as evidence.
`eval_blastn=1e-10 #Blastn eval cutoff`	Sets the required BLAST e-value cutoff for an EST/mRNA-seq alignment to be maintained as evidence.
`bit_blastn=40 #Blastn bit cutoff`	Sets the required BLAST bit value cutoff for an EST/mRNA-seq alignment to be maintained as evidence.
`depth_blastn=0 #Blastn depth cutoff (0 to disable cutoff)`	Allows the user to limit the number of BLAST alignments that are kept and used for annotation. Setting this to a non-zero number will save memory and improve runtime for large evidence datasets.
`pcov_blastx=0.5 #Blastx Percent Coverage Threhold Protein-Genome Alignments`	These options are analogous to the BLASTN options above but are applied to protein evidence.
`pid_blastx=0.4 #Blastx Percent Identity Threshold Protein-Genome Aligments`	
`eval_blastx=1e-06 #Blastx eval cutoff`	
`bit_blastx=30 #Blastx bit cutoff`	
`depth_blastx=0 #Blastx depth cutoff (0 to disable cutoff)`	
`pcov_tblastx=0.8 #tBlastx Percent Coverage Threhold alt-EST-Genome Alignments`	These options are analogous to the BLASTN options above but are applied to `alt_ests` (EST/mRNA-seq data from a closely related species) evidence.
`pid_tblastx=0.85 #tBlastx Percent Identity Threshold alt-EST-Genome Aligments`	
`eval_tblastx=1e-10 #tBlastx eval cutoff`	
`bit_tblastx=40 #tBlastx bit cutoff`	
`depth_tblastx=0 #tBlastx depth cutoff (0 to disable cutoff)`	
`pcov_rm_blastx=0.5 #Blastx Percent Coverage Threhold For Transposable Element Masking`	These options are analogous to the BLASTN options above but are applied to transposon protein alignments used in repeat masking.

continued

Table 4.11.2 MAKER Control File Options, *continued*

Option	Comments
`maker_opts.ctl`	
`pid_rm_blastx=0.4 #Blastx Percent Identity Threshold For Transposbale Element Masking`	
`eval_rm_blastx=1e-06 #Blastx eval cutoff for transposable element masking`	
`bit_rm_blastx=30 #Blastx bit cutoff for transposable element masking`	
`ep_score_limit=20 #Exonerate protein percent of maximal score threshold`	Setting this higher will require polished protein to have a higher exonerate score to be maintained as evidence.
`en_score_limit=20 #Exonerate nucleotide percent of maximal score threshold`	Same as above but for Polished EST/mRNA-seq alignments.

CHAPTER 8

CONCLUSION

Over the course of my graduate studies, I have gained a depth of knowledge in the field of genome annotation that spans the technical details of computer science, genome biology, and genetics essential for working in today's genome science. My efforts in software development have facilitated the adoption of the MAKER genome annotation pipeline by the plant community, thus facilitating the rapid annotation of organisms in all domains of life. I clearly demonstrated the utility of MAKER in plants in previous chapters (2, 3, and 4). I added ncRNA annotations to the maize genome (Chapter 4), increasing the body of genomic knowledge of the dominant agricultural product of the United States. Adding the functionality of a combiner to MAKER (Chapter 5) did not improve the final gene annotations produced by the pipeline based on the benchmarks presented for the *Drosophila melanogaster* genome; however, incorporating EVidence Modeler into MAKER advanced my knowledge of, and skill associated with, the Perl programming language tremendously. These advanced skills in computer programing will allow me to approach the biological questions of the future with confidence and competence. Chapters 6 and 7 demonstrate my commitment to globally advancing the field of genome biology. "If you give a man a fish he is hungry in an hour. If you teach him to catch a fish you do him a good turn"[1]. Distributing skills in addition to knowledge is thus an investment in the future. More minds equipped to approach today's and

tomorrow's questions in genome biology will advance the body of science further than if I spent the time and effort represented in Chapters 6 and 7 annotating newly sequenced and assembled genomes myself.

The future of genome annotation

The genome project paradigm has shifted in the course of my graduate career from analysis of a single genome to analysis of multiple species and whole populations. Early on, I was involved in a number of single genome analyses, including the genomes of the African coelacanth[2], the sea lamprey[3], and the algae *Nannochloropsis*[4]. As the field progressed, I progressed with it, with projects that involved multiple species and populations, such as the gibbon[5] and pigeon[6] genome projects, respectively. As genome science moves to population-level analyses, it is becoming painfully clear that the linear reference genome is inadequate. The limitations of the linear reference genome are most clearly demonstrated in agricultural crops.

Well-assembled and annotated reference genomes have revolutionized agricultural crop breeding, allowing for the identification of gene networks responsible for important agricultural traits, such as yield and drought tolerance[7,8]. Recent sequencing efforts directed at three divergent strains of rice identified large structural differences between strains using *de novo* genome assembly. Many of these structural variations are megabases in length and contain protein-coding genes unique to a given strain[9]. This *de novo* assembly approach is an improvement over a single reference genome, but we are still left with three genomes representing more than 124,000 rice accessions held in the International Rice Genebank[10]. Not only would sequencing and *de novo* assembly of more than 124,000 accessions of rice be cost prohibitive, it would also generate a huge

amount of redundant data, as many of these accessions will differ only slightly from the reference. Here I propose to use a variant graph approach to create a pan-genome annotation for rice by leveraging the vast amount of available rice sequencing data, including the newly available *de novo* genome assemblies of the *indica*, *aus*, and *temperate japonica* rice genomes[9,11]. I also propose to put in place the infrastructure to replicate this resource for any population of agricultural plants using iPlant[12] resources.

Variant graphs provide an elegant solution to the high cost of *de novo* assembly and storage of redundant data, while accurately representing populations. A variant graph is generated by aligning re-sequenced individuals from the population to a starting reference sequence. Each base in the genome is represented as a node and the edges are the observed path through the genome. Variant sites result in an alternative path or bubble in the graph. Figure 8.1 shows an example of an alignment and the resulting variant graph. This graph can then be used to inform downstream analyses such as gene annotation, primer design, and future variant calling. The deletion in Figure 8.1 might well result in an alternative start codon and a frame shift resulting in a premature stop for Indv1. The insertion in Indv2 might create an alternative start and a frame shift that changes the amino acid sequence; and the single nucleotide variant (SNV) will change the amino acid sequence to a Q in Indv2 and a T in Indv3. These changes are only clear when all of the variants are considered at once, and all have the potential to result in functionally different proteins. Variants are problematic in nongenic regions as well. The changes in the nuclear sequence have the potential to affect transcription factor binding or, in molecular biology experiments, primer specificity and affinity. Importantly, with the right software, any annotatable feature on the reference genome can be automatically

identified on the variant graph and 'projected' forward with appropriate revisions to a subpopulation or single individual. Over the next three years I am going to work in the domain of agricultural genomics, focusing on nonlinear representations of genomic sequence.

References

1. Thackeray, A. I. *Mrs. Dymond*. (Smith, Elder, and CO., 1885).

2. Amemiya, C. T. *et al.* The African coelacanth genome provides insights into tetrapod evolution. *Nature* **496,** 311–6 (2013).

3. Smith, J. J. *et al.* Sequencing of the sea lamprey (Petromyzon marinus) genome provides insights into vertebrate evolution. *Nat. Genet.* **45,** 415–21, 421e1–2 (2013).

4. Vieler, A. *et al.* Genome, functional gene annotation, and nuclear transformation of the heterokont oleaginous alga Nannochloropsis oceanica CCMP1779. *PLoS Genet.* **8,** e1003064 (2012).

5. Carbone, L. *et al.* Gibbon genome and the fast karyotype evolution of small apes. *Nature* **513,** 195–201 (2014).

6. Shapiro, M. D. *et al.* Genomic diversity and evolution of the head crest in the rock pigeon. *Science* **339,** 1063–7 (2013).

7. Huang, X. *et al.* Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* **42,** 961–7 (2010).

8. Xu, X. *et al.* Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat. Biotechnol.* **30,** 105–11 (2012).

9. Schatz, M. C. *et al.* Whole genome de novo assemblies of three divergent strains of rice, Oryza sativa, document novel gene space of aus and indica. *Genome Biol.* **15,** 506 (2014).

10. Jackson, M. T. Conservation of rice genetic resources: the role of the International Rice Genebank at IRRI. *Plant Mol. Biol.* **35,** 61–7 (1997).

11. Gao, Z. *et al.* Dissecting yield-associated loci in super hybrid rice by resequencing recombinant inbred lines and improving parental genome sequences. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 14492–7 (2013).

12. Goff, S. a *et al.* The iPlant Collaborative: Cyberinfrastructure for Plant Biology. *Front. Plant Sci.* **2,** 34 (2011).

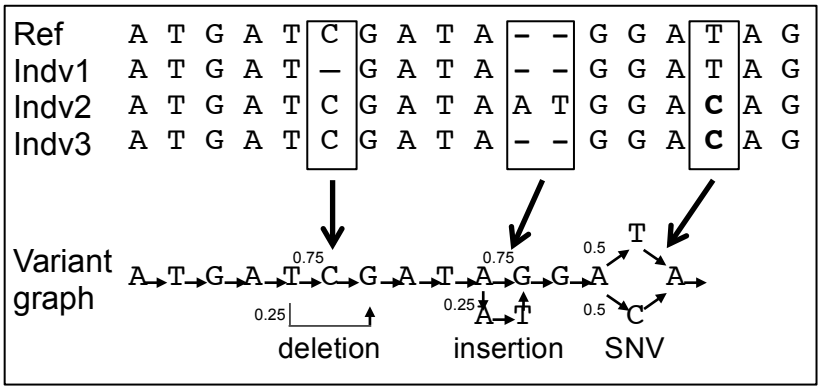

Figure 8.1. From alignment to graph. Sequences from three individuals are aligned to a reference. Variant positions in the individuals relative to the reference are boxed above. Invariable nucleotide sites are represented as a single node in the variant graph. Variants are depicted as bubbles (i.e. observed alternate paths) through the graph. Depicted here are a deletion, an insertion, and an SNV. The observed frequencies of edges (i.e. variant frequency) are recorded in the graph, with unlabeled edges having a frequency of 1.0.