

TOOLS AND TECHNIQUES FOR GENOME
ANNOTATION AND ANALYSIS

by

Carson Hinton Holt

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

August 2011

Copyright © Carson Hinton Holt 2011

All Rights Reserved

The University of Utah Graduate School

STATEMENT OF DISSERTATION APPROVAL

The dissertation of Carson Hinton Holt

has been approved by the following supervisory committee members:

Mark Yandell , Chair 5/16/2011
Date Approved

Carl Thummel , Member 5/16/2011
Date Approved

Alejandro Sanchez , Member 5/16/2011
Date Approved

Karen Eilbeck , Member 5/16/2011
Date Approved

Lewis Frey , Member 5/16/2011
Date Approved

and by Lynn Jorde , Chair of

the Department of Human Genetics

and by Charles A. Wight, Dean of The Graduate School.

ABSTRACT

Whole genome sequencing projects have expanded our understanding of evolution, organism development, and human disease. Now advances in second-generation technologies are making whole genome sequencing routine even for small laboratories. However, advances in annotation technology have not kept pace with genome sequencing, and annotation has become the major bottleneck for many genome projects (especially those with limited bioinformatics expertise). At the same time, challenges associated with genomics research extend beyond merely annotating genomes, as annotations must be subjected to diverse downstream analyses, the complexities of which can confound smaller research groups. Additionally, with improvements in genome assembly and the wide availability of next generation transcriptome data (mRNA-seq), researchers have the opportunity to re-annotate previously published genomes, which creates new difficulties for data integration and management that are not well addressed by existing tools.

In response to the challenges facing second-generation genome projects, I have developed the annotation pipeline MAKER2 together with accessory software for downstream analysis and data management. The MAKER2 annotation pipeline finds repeats within a genome, aligns ESTs and cDNAs, identifies sites of protein homology, and produces database-ready gene annotations in association with supporting evidence. However MAKER2 can go beyond structural annotation to identify and integrate functional annotations. MAKER2 also provides researchers

with the capability to re-annotate legacy genome datasets and to incorporate mRNA-seq. Additionally, MAKER2 supports distributed parallelization on computer clusters, thus providing a scalable solution for datasets of any size.

Annotations produced by MAKER2 can be directly loaded into many popular downstream annotation analysis and management tools from the Generic Model Organism Database Project. By using MAKER2 with these tools, research groups can quickly build genome annotations, perform analyses, and distribute their data to the wider scientific community.

Here I describe the internal architecture of MAKER2, and document its computational capabilities. I also describe my work to annotate and analyze eight emerging model organism genomes in collaboration with their associated genome projects. Thus, in the course of my thesis work, I have addressed a specific need within the scientific community for easy-to-use annotation and analysis tools while also expanding our understanding of evolution and biology.

For my beautiful wife, Marlene, and my four wonderful
children who are the inspiration to everything I do

TABLE OF CONTENTS

ABSTRACT	iii
LIST OF TABLES	viii
LIST OF FIGURES	ix
ACKNOWLEDGMENTS.....	x
Chapter	
1 INTRODUCTION	1
Background.....	1
MAKER2, an easy-to-use annotation solution.....	6
Summary of findings	9
References.....	9
2 MAKER2: AN ANNOTATION PIPELINE AND GENOME-DATABASE MANAGEMENT TOOL FOR SECOND-GENERATION GENOME PROJECTS ..	13
Abstract.....	14
Background.....	14
Results and discussion	17
Conclusions	32
Methods.....	34
List of abbreviations.....	42
Authors' contributions.....	42
Acknowledgements.....	43
References.....	43
3 THE PROOF-OF-PRINCIPLE ANNOTATION OF EIGHT EMERGING MODEL ORGANISM GENOMES	58
Introduction	58
Results and discussion	58
Conclusions	73
Methods.....	74
References.....	84
4 EDUCATIONAL OUTREACH	94
References.....	96

5	MAKER WEB ANNOTATION SERVICE: AN ONLINE PORTAL FOR GENOME ANNOTATION AND ANALYSIS	97
	Abstract.....	97
	Rationale.....	97
	Algorithm/website design	98
	Annotation of <i>Pinus taeda</i> BAC clones	103
	Conclusions.....	105
	Availability and requirements.....	106
	Methods.....	106
	References.....	107
6	ALGORITHM DESIGN	115
	Supported system architectures	116
	Input to MAKER2	116
	MAKER2's output	117
	Step-by-step overview of MAKER2	117
	Algorithm stability and error handling.....	127
	References.....	128
7	SUMMARY AND CONCLUSIONS	132
	References.....	132

LIST OF TABLES

Table

2.1	Gene model accuracy for gene prediction/annotation programs	49
2.2	Sensitivity and specificity for prediction/annotation programs	49
2.3	Gene model accuracy using unmatched species parameters.....	50
2.4	Sensitivity and specificity using unmatched species parameters	50
2.5	Pfam domain content and GO molecular functions in six reference genomes .	51

LIST OF FIGURES

Figure

1.1	View of MAKER2 annotations and evidence.....	12
2.1	MAKER2 vs. <i>ab initio</i> predictors on second-generation genomes.....	52
2.2	Evaluating AED as a metric for annotation quality control.....	53
2.3	AED evaluation of <i>Homo sapiens</i> reference annotations	54
2.4	Re-annotation of a portion of the maize genome using MAKER2.....	55
2.5	MAKER2 as a management tool for existing genome annotations.....	56
2.6	MAKER2 scales to even the largest genomes	57
3.1	Comparison of domain content in Oomycetes	89
3.2	Domain content in reference eukaryotes	90
3.3	Comparison of intron/exon structure across eukaryotes	91
3.4	Phylogenetic analysis of <i>Schmidtea mediterranea</i>	92
3.5	Genes shared by <i>Schmidtea mediterranea</i> with other metazoans.....	93
5.1	MWAS login screen.....	109
5.2	The MWAS job submission screen	110
5.3	View of MWAS annotations in Apollo.....	111
5.4	Annotation summary statistics	112
5.5	View of MWAS functional annotation integration.....	113
5.6	Association of results and analyses with gene annotations	114
6.1	Flowchart of MAKER2's design and operation	130
6.2	Summary of MAKER2's Quality Indices	131

ACKNOWLEDGMENTS

I'd like to thank and recognize all contributions from my advisor Mark Yandell, lab member Hao Hu, and former lab members Brandi Cantarel and Hadi Islam in developing MAKER2. I would also like to recognize our collaborator Ian Korf at UC Davis and his lab member Genis Parra for their contributions. Dave Clements at Emory University deserves special recognition for his work in promoting MAKER2 and maintaining MAKER2 documentation on the GMOD wiki page. In addition, I want to thank and recognize the involvement and contributions made by the many collaborators who have been integrally involved in annotation and analysis of the genomes processed by MAKER2: Alejandro Sánchez Alvarado (University of Utah), Robin Buell (Michigan State University), Weiming Li (Michigan State University), Allen Kovach (UC Davis), Brenda Wingfield (University of Pretoria, South Africa), Christopher D. Smith (San Francisco State University), Christopher R. Smith (Earlham College), Jürgen Gadau (Arizona State University), Neil Tsutsui (UC Berkeley), and Cameron Currie (University of Wisconsin-Madison).

This work was supported by the NIH/NHGRI grant R01-HG004694 and partially supported by the NIH Genetics Training Grant T32-GM007464. An allocation of computer time from the Center for High Performance Computing at the University of Utah is also gratefully acknowledged.

CHAPTER 1

INTRODUCTION

MAKER2 is an automated genome annotation pipeline that identifies repetitive elements in a genome, aligns Expressed Sequence Tags (ESTs) and protein homology evidence to a genome, and synthesizes these data into database-ready genome annotations. MAKER2 is based on the annotation pipeline MAKER[1]. Here I provide an introduction to the process of genome annotation, the rationale behind MAKER2's development, and an explanation of important design considerations for the pipeline.

Background

While second-generation sequencing technologies are making great strides in bringing down sequencing costs, focus on these achievements tends to overlook the fact that raw DNA sequence in and of itself really isn't that useful. Given a newly sequenced genome, what researchers most want to know is, "where are the genes and what do they do"? The process of identifying genes within a genome sequence, documenting their intron-exon structures with supporting evidence, and assigning them putative functions is referred to as genome annotation.

What are genome annotations?

Annotations are essentially models describing a gene's intron-exon structure, alternate splice forms, UTR locations, coding regions, etc. Gene annotations can also describe other features such as gene expression profiles, a gene's molecular function, the biological pathway a gene is involved in, or the orthologous relationship of a gene to another species. Gene annotations are therefore the nuclei around which our electronic knowledge of a genome and an organism grows.

The annotation process

Because of the size of most genomes, experimental identification and verification of all genes and annotations within them is impossible (at least in the short run). Annotations are therefore the result of logical deductions based on evidence from EST and cDNA alignments, protein homology, and *ab initio* gene predictions.

Eukaryotes pose a significant challenge for gene annotation because of their large intron-containing genes, relatively low gene densities, alternative splice forms, and high concentrations of transposons and repetitive elements. Finding and describing genes in these organisms can be difficult, and building genome annotations for them requires an exhaustive process in which *ab initio* gene predictions, EST and cDNA alignments, and homology to known proteins must be taken into consideration. Of course, managing these data on a genomewide scale also requires sophisticated computational methods.

To a large degree, the first-generation of genome annotations were created manually (these were classic model organisms like *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Saccharomyces cerevisiae*). Manual annotation is and

was a very complex, time-consuming endeavor, especially considering the thousands of genes present in the average eukaryotic organism. Despite it being so laborious, for the first generation of genome projects, the process of annotation was still far less expensive and labor intensive than genome sequencing. For this reason, organizations associated with early genome projects, like FlyBase[2] and Celera, were willing to employ hundreds of scientists as manual curators. Even today manual curation still remains the ‘gold-standard’[3, 4] by which to judge the quality of automated annotation tools.

At the end of the annotation process, gene annotations, the evidence supporting those annotations, and genome features such as repeat elements are combined into a database in order to produce a complete model of the genome (including structure and features). Researchers can query the database for information they need to design experiments such as protein domain analyses, gene knock-out/knock-in type experiments, etc. The genome database also serves as a substrate for improving gene annotations via manual and automated review of experimental evidence (ESTs, protein homology, etc.).

Automatically generated annotations are far from perfect, and manual gene annotation remains the ‘gold-standard’ for evaluating annotation confidence and quality. However, manual annotation is an expensive and time-consuming process. For this reason, classification and prioritization of annotations for later manual review is critical for the maintenance of a genome database. A classification scheme requires that each annotation be tagged with information describing the type of evidence that supports each gene model. The requirement for evidence trails, therefore, further complicates genome annotation.

The genome annotation bottleneck

As second-generation sequencing technology has improved, annotation (not sequencing) has become the major bottleneck to genomics research. For example, as of January 2011, 463 genomes were fully sequenced yet unpublished (an ever-growing backlog of un-annotated genomes), and over 1,600 eukaryotic genome projects were underway[5]. With a conservative estimate of just 10,000 genes per genome, these projects alone will produce over 16,000,000 new gene annotations that must be generated, maintained, and updated to reflect new research. With even more genome projects just over the horizon, it is obvious that manual curation is no longer feasible, and better fully-automated solutions are required. Unfortunately, second-generation genome projects are overwhelmingly associated with small research communities that often lack the bioinformatics skills necessary to implement their own automated annotation pipelines or easily build and maintain an annotation database.

Alternative solutions for annotation

Some research groups have taken an alternate approach to building and maintaining their own genome annotation pipeline by outsourcing annotations to major databases such as ENSEMBL[6] or VectorBase[7]. However, the number of un-annotated genomes far exceeds the capacity and stated purview of these databases. ENSEMBL, for example, has traditionally been focused on vertebrate genomes, and VectorBase is limited to insect vectors of human disease. Many research groups are also unwilling to give up control of their genome datasets to third-party organizations and would rather attempt genome annotation independently.

In recognition of the difficulties confronted by research groups trying to annotate their respective organism's genome, some sequencing centers and major genome databases have made parts of their own in-house annotation pipelines available to the public. ENSEMBL, for example, provides an extensive suite of annotation, alignment, and data management tools[8]. However, distributing annotation tools is not part of the primary mission of large genome databases, and what is made available constitutes only a subset of their internal systems. For large research groups with extensive bioinformatics experience, these tools can be sufficient when supplemented with other in-house pipeline development. Small research groups with little bioinformatics experience, however, are left with few options when trying to build and manage gene annotations and evidence.

The Generic Model Organism Database project

Because of the limited availability of software for genome annotation and analysis, the Generic Model Organism Database project (GMOD)[9] was formed. The goal of GMOD is to provide open source software tools for managing genome-scale biological databases as well as facilitating downstream analyses. GMOD also has the goal of having all its software tools be interoperable via a standard file format, Generic Feature Format version 3 (GFF3)[10], which is used for describing genome features and annotations. While the GMOD suite of components contains a large selection of software for manipulating and analyzing annotations, they fall short of providing a tool that can build and combine annotations and evidence into a new database (at least until MAKER2 was developed). Once again leaving small research projects with no easy annotation solution.

MAKER2, an easy-to-use annotation solution

To confront the many difficulties faces by second-generation genome projects, I have developed MAKER2, a genome annotation pipeline designed to be easy-to-use for small research groups with little bioinformatics experience. MAKER2 joined GMOD in 2008 and fills a “big hole” in that project (as there was no tool capable of producing *de novo* annotations in GMOD before MAKER2). The MAKER2 annotation pipeline is based on the earlier genome annotation pipeline MAKER[1], but MAKER2 greatly expands on the earlier program’s capabilities and functionality. MAKER2 integrates existing software tools into a package that produces database-ready genome annotations together with associated evidence and quality control statistics. It identifies and masks repetitive elements in the genome, aligns ESTs and protein homology evidence, produces *ab initio* gene predictions, infers five and three prime UTRs, and integrates all these data to produce final gene annotations with quality control statistics that help prioritize genes for downstream review and manual curation. MAKER2 does not produce gene predictions by itself; rather, it integrates existing *ab initio* gene prediction programs like SNAP[11], Augustus[12], and GeneMark[13] for this purpose. However, rather than just accepting the raw *ab initio* gene predictions produced by these algorithms, MAKER2 uses evidence alignments to provide ‘hints’ to the prediction programs as to the location of probable introns, exons, and coding regions. MAKER2 also actively modifies the resulting predictions to include features like UTR that can be inferred from EST alignments. In this way, MAKER2 guides the behavior of *ab initio* prediction programs using experimental evidence to produce improved models. The final output of MAKER2 is in GFF3 format, which is the common file format used by GMOD tools. Because of the common format, MAKER2’s output can be directly

loaded into GMOD tools like Apollo[14] (an annotation viewing and curation tool), Chado (a database schema)[15], GBrowse[16] (an online annotation viewing and distribution tool), and Galaxy[17] (an analysis pipeline). MAKER2's output can thus be easily utilized for downstream curation, data distribution, and experimental analysis.

Design considerations for MAKER2

Creating an easy-to-use annotation pipeline forces one to confront several software design challenges. First, the pipeline should be simple to install, configure, and run; this means using it should require only basic bioinformatics skills and it must be runnable with the types of computational resources that would be encountered in an average laboratory. However, because of the wide variation in genome size and content for different organisms, the pipeline must also be scalable – able to handle datasets both large and small by taking full advantage of all computational resources that may be available to it. MAKER2 meets these requirements by being compatible with UNIX-like operating systems such as Linux and Mac OS X (machines found in most laboratories) and by integrating support for Message Passing Interface (MPI), which is a distributed parallelization protocol used in computer clusters. This means MAKER2 can either run on a laptop computer, or if needed, expand its analyses on a computer cluster to thousands of CPUs and process datasets of virtually any size.

An easy-to-use pipeline must perform basic tasks of evidence alignment and interpretation. Therefore not only is MAKER2 required to identify repetitive elements, produce *ab initio* gene predictions, and align ESTs and proteins to the genome, but the pipeline must also integrate those data to synthesize feature-rich

gene annotations that include three and five prime UTRs, alternative splice forms, and an evidence trail that can be used for downstream analysis and quality control. MAKER2 achieves this by leveraging existing tools such as RepeatMasker[18] (for repeat identification), BLAST[19] (for evidence alignment), Exonerate[20] (to polish alignments), and SNAP[11] (a gene-predictor). These are all programs that have been highly optimized to do a specific task very well. By integrating and interpreting their outputs, MAKER2 takes advantage of these tools' many combined years of research and development, to produce gene annotations that could not have been generated by any of the programs individually. Figure 1.1 shows how evidence alignments from programs used by MAKER2 correlate with the resulting genome annotations, thus both suggesting and confirming each aspect of the final gene model.

Another design consideration for an easy-to-use annotation pipeline is trainability. Because every genome is different, an annotation pipeline must be easily trained, thus maximizing the accuracy of gene models produced for each new organism. MAKER2 takes advantage of aligned evidence to identify organism specific patterns in intron/exon structures that can then be conveyed to *ab initio* gene-predictors like SNAP and Augustus. These programs then produce improved evidence-based gene models that can become the basis for further training of *ab initio* gene-predictors. In this way, MAKER2 can be trained on new organisms via a bootstrap-like procedure.

The final essential feature of an easy-to-use annotation pipeline is that its output must rigorously describe all aspects of the gene annotations and their associated evidence in a machine-readable fashion. The output format must also be compatible with other commonly used tools and applications, and users must be able

to view and edit individual contigs with only minimal computational resources. These tasks have been simplified for us by the Generic Model Organism Database project (GMOD), which provides a suite of tools for viewing, curating, distributing, and analyzing genome annotations via a single common file format, GFF3. Thus by producing genome annotations in GFF3 format, MAKER2 gives its users access to extensive software resources that are already freely available.

Summary of findings

MAKER2 is a simple solution to the genome annotation needs of projects associated with smaller research groups. It provides an efficient mechanism to produce annotations using evidence derived from ESTs, protein alignments, and *ab initio* gene predictions. MAKER2's output also facilitates downstream analysis via integration with GMOD tools.

While MAKER2, by design, is meant to facilitate the generation and analysis of genome annotations, it is not meant to be an exhaustive algorithm. The pipeline does not identify noncoding RNA genes, nor does it provide comprehensive solutions to every problem in genome annotation. MAKER2 does, however, produce database-ready protein-coding gene annotations that serve as the substrate for further analysis and experimentation, thus helping to jump-start research in newly sequenced organisms.

References

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

2. Tweedie S, Ashburner M, Falls K, Leyland P, McQuilton P, Marygold S, Millburn G, Osumi-Sutherland D, Schroeder A, Seal R, et al: **FlyBase: enhancing drosophila gene ontology annotations.** *Nucleic acids research* 2009, **37**:D555-D559.
3. Guigo R, Reese M: **EGASP: collaboration through competition to find human genes.** *Nature methods* 2005, **2**:575 - 577.
4. Coghlan A, Fiedler T, McKay S, Flicek P, Harris T, Blasiar D, the n GC, Stein L: **nGASP - the nematode genome annotation assessment project.** *BMC Bioinformatics* 2008, **9**:549.
5. Liolios K, Tavernarakis N, Hugenholtz P, Kyrpides N: **The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide.** *Nucleic acids research* 2006:D332 - 334.
6. Hubbard TJ, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, et al: **Ensembl 2007.** *Nucleic Acids Res* 2007:D610 - 617.
7. Lawson D, Arensburger P, Atkinson P, Besansky NJ, Bruggner RV, Butler R, Campbell KS, Christophides GK, Christley S, Dialynas E, et al: **VectorBase: a home for invertebrate vectors of human pathogens.** *Nucl Acids Res* 2007, **35**:D503-505.
8. Stabenau A, McVicker G, Melsopp C, Proctor G, Clamp M, Birney E: **The Ensembl core software libraries.** *Genome Res* 2004, **14**:929-933.
9. **GMOD** [<http://www.gmod.org>]
10. **GFF3** [<http://www.sequenceontology.org/gff3.shtml>]
11. Korf I: **Gene finding in novel genomes.** *BMC Bioinformatics* 2004, **5**:59.
12. Stanke M, Waack S: **Gene prediction with a hidden Markov model and a new intron submodel.** *Bioinformatics* 2003, **19**:ii215-225.
13. Lomsadze A, Ter-Hovhannisyanyan V, Chernoff YO, Borodovsky M: **Gene identification in novel eukaryotic genomes by self-training algorithm.** *Nucl Acids Res* 2005, **33**:6494-6506.
14. Lewis SE, Searle SMJ, Harris N, Gibson M, Iyer V, Richter J, Wiel C, Bayraktaroglu L, Birney E, Crosby MA, et al: **Apollo: a sequence annotation editor.** *Genome Biology* 2002, **3**:research0082.0081 - 0082.0014.
15. Mungall C, Emmert D: **A Chado case study: an ontology-based modular schema for representing genome-associated biological information.** *Bioinformatics (Oxford, England)* 2007, **23**:i337 - 346.

16. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S: **The Generic Genome Browser: a building block for a model organism system database.** *Genome Res* 2002, **12**:1599-1610.
17. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, et al: **Galaxy: a platform for interactive large-scale genome analysis.** *Genome research* 2005, **15**:1451-1455.
18. **RepeatMasker** [<http://repeatmasker.org>]
19. Altschul SF, Gish W, Miller W, Meyers EW, Lipman DJ: **Basic Local Alignment Search Tool.** *Journal of Molecular Biology* 1990, **215**:403-410.
20. Slater G, Birney E: **Automated generation of heuristics for biological sequence comparison.** *BMC Bioinformatics* 2005, **6**:31.

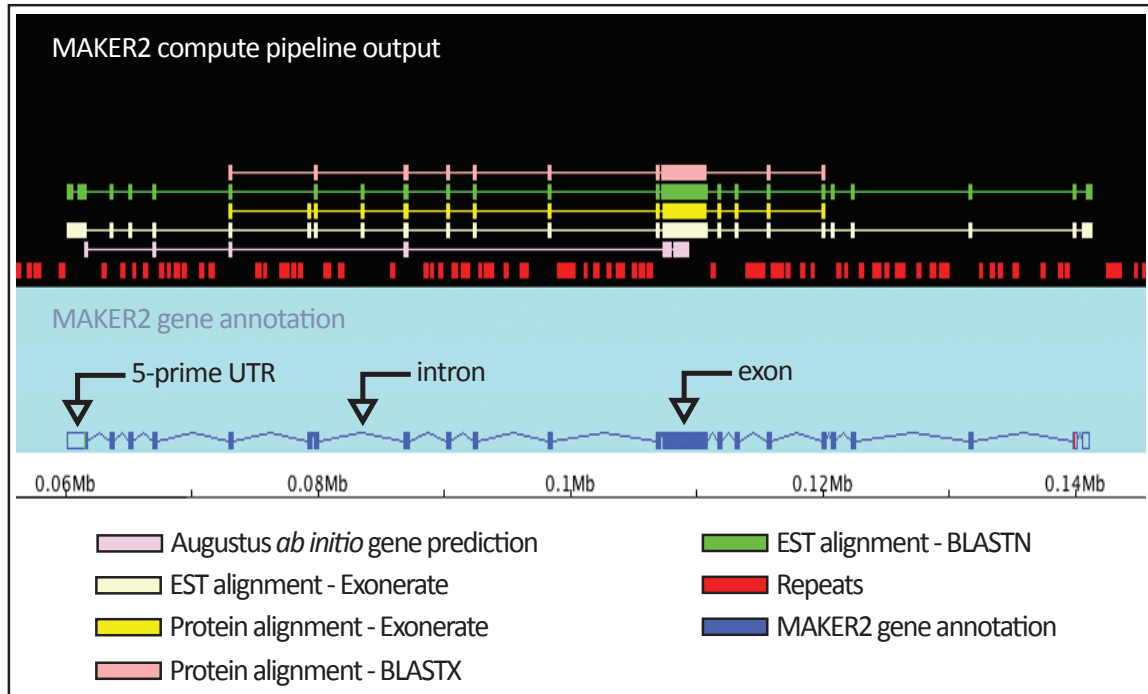


Figure 1.1. View of MAKER2 annotations and evidence.

MAKER2 produced genome annotations and aligned experimental evidence are displayed in the Apollo annotation curation tool. The pattern of experimental alignments (upper dark panel) correlates with the intron/exon structure of the genome annotations (lower blue panel), thus both suggesting and supporting different features of the final gene model.

CHAPTER 2

MAKER2: AN ANNOTATION PIPELINE AND GENOME-DATABASE MANAGEMENT TOOL FOR SECOND-GENERATION GENOME PROJECTS

From “MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects,” by Carson Holt and Mark Yandell, Submitted to Genome Biology. Copyright 2011 Carson Holt and Mark Yandell. 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.

My contributions to this study include conceiving and carrying out the experiments, developing the software for the analyses, and writing the manuscript.

Abstract

Second-generation sequencing technologies are precipitating major shifts with regards to what kinds of genomes are being sequenced and how they are annotated. While the first generation of genome projects focused on well-studied model organisms, today's genome projects are focused on new, often exotic, organisms whose genomes are largely *terra incognita*. This complicates their annotation, because unlike first-generation projects, there are no pre-existing gene-models with which to train gene-finders. Today's second-generation genome projects are also faced with challenges that extend beyond *de novo* annotation. Improvements in genome assembly and the wide availability of mRNA-seq data are creating opportunities to re-annotate previously published genomes and to update their legacy annotations. This, in turn, creates data management problems not encountered by the first wave of genome projects. A better understanding of the impact of these issues on today's genome projects is essential for successful annotation and on-going annotation management. Here, we provide in-depth analyses of these challenges and explain how MAKER2 solves them.

Background

Second-generation sequencing technologies are creating new opportunities as well as new challenges for genomics research. While first-generation genome projects focused primarily on established model organisms such as *Drosophila melanogaster*[1], *Caenorhabditis elegans*[2], and *Mus musculus*[3], falling sequencing costs allow second-generation genome projects to focus on less well-understood, often exotic and phylogenetically isolated organisms. These genomes

present researchers with formidable challenges, as their novel contents can make them difficult substrates for annotation. The large volumes of data produced by second-generation sequencing technologies also create difficulties for data management not encountered by first-generation projects. Together, these two factors can spell disaster for the small research communities in charge of many second-generation genome projects, as they generally lack the bioinformatics experience and resources necessary to manually revise, curate, and manage annotations.

A major challenge facing many second-generation projects is the exotic nature of their organism's genomes. The first-generation of genome projects benefited greatly from large bodies of pre-existing knowledge regarding their organisms' genomes. For *D. melanogaster*, *C. elegans*, and *H. sapiens*[4, 5], for example, hundreds of gene models already existed before the genomes were sequenced. These pre-existing gene models were critical for annotation and analysis, because they allowed researchers to train and optimize gene prediction tools for each genome. They also provided a standard by which to judge the accuracy of annotations. Second-generation projects rarely have access to such information. This severely limits their ability to train *ab initio* gene-finders, the result being low-quality gene predictions. The lack of pre-existing gene models also leaves many second-generation projects with no objective standards with which to gauge annotation accuracy. Quality control is thus a significant issue for these projects. Data management is another.

New techniques such as high-throughput transcriptome sequencing (mRNA-seq) have great potential to improve annotation quality, but they produce enormous amounts of data; likewise, the existence of legacy annotations for a large number of

both first and second-generation genomes is also creating data management challenges. How exactly does one go about updating a genome's worth of annotations to take into account new mRNA-seq data? How does one know that the annotations are improved?

Demographic and economic issues further exacerbate these difficulties. The small research communities driving second-generation genome projects often lack significant bioinformatics experience. They also often lack the economic resources needed to manage and distribute the results of the genome annotation projects. Thus, it is essential that the output of an annotation pipeline be easily converted into a genome database — and there is a great need for automated means to manage them.

MAKER2 builds upon MAKER[6], an easy-to-use genome annotation pipeline that has seen wide adoption[7-19]. MAKER2 provides straightforward solutions to the problems facing today's second-generation genome projects. Here, we demonstrate its ability to overcome handicaps resulting from a lack of pre-existing gene models with which to train gene-predictors for use on novel genomes; its ability to use mRNA-seq data to improve annotation quality; and its ability to update legacy annotations (significantly improving their quality).

Throughout these analyses, we measure MAKER2's performance using an integrated annotation quality control measure, Annotation Edit Distance (AED)[20], developed by the Sequence Ontology project[21]. Thus, MAKER2 is not only an improved annotation engine; it is also an annotation management tool. We show that MAKER2 can both evaluate the global quality of genome annotations, and identify and prioritize individually problematic annotations for manual review; these are functionalities offered by no other annotation tool.

Results and discussion

Genome annotation in classic model organism genomes

The performance of *de novo* annotation tools such as HMM based *ab initio* gene-predictors and evidence based annotation pipelines have previously been explored in competitions such as EGASP[22] and NGASP[23], which looked at gene prediction and annotation accuracy in the human and *C. elegans* genomes, respectively. From these competitions, the metrics sensitivity, specificity, and accuracy have emerged as the standard methods for evaluating the quality of gene predictions[24]. Both measurements require a set of reference gene models that are assumed to be correct; gene predictions are then compared to the reference model to generate sensitivity, specificity, and accuracy values (see Methods).

Using these metrics, we compared the performance of MAKER2 to the *ab initio* gene prediction programs SNAP[25], GeneMark[26], and Augustus[27, 28]. We used the organism specific parameter files that come bundled with each of these algorithms to produce *ab initio* gene predictions for *D. melanogaster* chromosome 3R, *C. elegans* chromosome V, and *Arabidopsis thaliana*[29] chromosome 4. For comparison, we then produced evidence-based genome annotations by running the same three algorithms (SNAP, GeneMark, and Augustus) inside of the MAKER2 genome annotation pipeline. Sensitivity, specificity, and accuracy values were then calculated against the respective reference genome using the program Eval[30] (Table 2.1 and Table 2.2).

As seen in Table 2.1, the base pair and exon level accuracy values for *ab initio* predictions produced by SNAP, Augustus, and GeneMark are very similar, generally within a few percentage points of each other. In *C. elegans*, for example, the difference between low and high base pair level accuracies is only 3.19% (85.10% for

SNAP vs. 88.29% for Augustus). The corresponding MAKER2 annotations have similar accuracies relative to the *ab initio* gene predictions, and more often than not, they are slightly improved over the *ab initio* gene predictions (but the difference is small). In *C. elegans*, for example, base pair level accuracies in MAKER2 range from 86.29% to 88.48% which is comparable to the 85.10% to 88.29% range for the *ab initio* gene predictions. This is not the first time that this trend has been observed[23] — given large enough training sets, *ab initio* gene prediction programs can match or even outperform annotation pipelines. Augustus, for example, achieved an exon-level accuracy in *C. elegans* of 74.62%, compared to MAKER2’s 68.60% (Table 2.1).

The relative similarity of accuracy measurements for *ab initio* prediction methods vs. MAKER2 suggests that MAKER2 is performing on par with these *ab initio* tools (but not greatly improving accuracy). However, as we show below, such comparisons can be quite misleading from a second-generation genome perspective. The key to understanding why is grasping that Table 2.1 reports the performance of the *ab initio* predictors after they have been trained using each genome’s existing annotations — datasets containing tens of thousands of often hand-curated gene models. Data such as those shown in Table 2.1 thus represent the upper bounds for performance of the *ab initio* prediction algorithms. As we demonstrate below, when training sets decrease in quality and/or size, the accuracy of *ab initio* tools drops dramatically; MAKER2’s accuracy, however, remains high. This feature of MAKER2 makes it especially useful for second-generation genome projects as these projects generally lack large enough training datasets for *ab initio* predictors to achieve accuracies comparable to those shown in Table 2.1.

Genome annotation using unmatched species parameters

To better understand how these algorithms perform using limited or poor quality training data, we repeated our analysis shown in Table 2.1 using the same portions of *D. melanogaster* chromosome 3R, *C. elegans* chromosome V, and *A. thaliana* chromosome 4; but this time we ran the gene-predictors using the wrong species file for each organism. *D. melanogaster* and *C. elegans* were analyzed using the species file from *A. thaliana*, and *A. thaliana* was analyzed using the species file from *C. elegans*. Each *ab initio* gene prediction program was then run inside of the MAKER2 annotation pipeline using the same incorrect species file for comparison.

As expected, the accuracy of the *ab initio* prediction algorithms is reduced substantially (Table 2.3 and Table 2.4). The reduction in accuracy is most notable at the exon level where all accuracies were approximately half of what was seen in the previous analysis. However, when each *ab initio* prediction program was run inside of MAKER2, accuracies dramatically improved for every organism at both the base pair and exon levels. The degree of improvement was most notable for SNAP, where exon level accuracies for *A. thaliana* increased from 18.58% to 60.11% (an over three-fold increase in accuracy). In fact, SNAP's performance inside of MAKER2 using the incorrect species file seems to match or even beat levels of performance experienced by all three *ab initio* gene-predictors when run using the correct species files. For example in *D. melanogaster*, when using the incorrect SNAP parameter file, MAKER2 produces exon level accuracies of 53.69%; whereas when using the correct parameter files outside of MAKER2, the programs GeneMark, SNAP, and Augustus produce exon level accuracies of 47.31%, 47.01%, and 61.37%, respectively. The level of improvement seen suggests that for second-generation genome projects where training data may be limited or of poor quality, running *ab initio* gene-

predictors as part of an evidence based pipeline like MAKER2 can provide substantial benefits in performance.

Gene prediction/annotation in second-generation genomes

When analyzing the performance of gene-predictors in sequenced second-generation genomes, the same metrics of sensitivity, specificity, and accuracy used for first-generation genomes cannot be applied (Table 2.1). This is because second-generation genomes lack the high-quality reference gene models required to calculate these values (accuracy measures the overlap between a prediction and the supposed correct reference). Therefore we are forced to apply a different metric.

In the experiments below, we use Pfam[31] domain content (mapped using InterProScan[32]) as a proxy metric for annotation quality. Although expansion and contraction of gene families can be an important mode of organism evolution, the high level of domain content of eukaryotic proteomes is relatively invariant[33]; this fact can be clearly seen in Table 2.5, which documents the high-level Pfam domain frequencies for six different well annotated eukaryotic model organisms (*H. sapiens*, *M. musculus*, *D. melanogaster*, *C. elegans*, *A. thaliana*, and *Saccharomyces cerevisiae*[34]). This relative invariance in the domain content of eukaryotic proteomes can be used as the basis for a rough standard with which to assess the accuracy of a second-generation genome's annotations.

At the grossest level of resolution, the percentage of annotations containing one or more Pfam domains provides an indication of accuracy. Alternatively the relative frequencies of different types of domains can be used to provide an even more precise measure, but such analyses must be considered with caution as there is always the possibility that some new, exotic second-generation genome may exhibit

divergent relative domain frequencies – even though these are grossly static at the whole-proteome scale for every published genome we have examined (Table 2.5).

To compare the performance of *ab initio* gene prediction algorithms to that of MAKER2 on second-generation genomes, we performed the proof-of-principle genome annotation of *Schmidtea mediterranea* (flatworm) and *Linepithema humile* (Argentine ant). For this analysis, we used the *ab initio* gene-predictor SNAP because it can be easily trained for new genomes using CEGMA[35] (an HMM-based program that identifies and annotates a subset of highly conserved, universal eukaryotic genes). The gene models produced by CEGMA then serve as the initial training set for SNAP. MAKER2 annotation of *S. mediterranea* also integrated mRNA-seq reads from that organism into the pipeline thus demonstrating how next-generation transcriptome data can be used to improve gene models (see Methods).

Figure 2.1a gives the high level breakdown of domain contents for six reference genomes. On average, 68% of annotations in these six genomes contain a Pfam domain. For individual proteomes, the percent enrichment ranges from a low of 57% for *C. elegans* to a high of 78% for *M. musculus*. In contrast, only 15% of SNAP produced *ab initio* gene predictions in *L. humile* contain a Pfam domain (Figure 2.1b), and in *S. mediterranea* the percent enrichment is even lower at just 6% (Figure 2.1c). The MAKER2-based proteomes, by comparison, are highly enriched for domains (Figure 2.1c). In total, 56% of the *L. humile* and 52% of *S. mediterranea* MAKER2-supervised SNAP predictions contain Pfam domains (values far more similar to the 68% enrichment seen in the reference proteomes).

Interestingly, not only are domain counts low for the SNAP *ab initio* predictions, but the gene counts produced for both species are well above what is expected for both organism. Approximately 15,000 genes are expected for *S.*

mediterranea and approximately 17,000 are expected for *L. humile* [6, 9], both values well below the 63,622 and 420,224 gene predictions produced (respectively) when running SNAP on its own (outside of MAKER2). *Ab initio* gene-predictors have a recognized tendency to over predict[23], and as these results demonstrate, this tendency can be greatly exacerbated by limited data. In contrast, MAKER2's supervised SNAP-based gene counts are dramatically more consistent with the published expected counts. MAKER2 produce 13,785 gene annotations for *L. humile* and 17,883 for *S. mediterranea* (Note this is even without further optimization and training of the gene-predictor SNAP).

These results stand in stark contrast to the great accuracy obtained by SNAP on model organism genomes presented in Table 2.1. They also make it clear that when training data are limited or of low quality, *ab initio* gene-predictors produce much more reliable results when supervised by an evidence based pipeline like MAKER2. This conclusion is also consistent with our earlier analyses where we annotated three model organism genomes using unmatched species parameter files (Table 2.3). Additionally MAKER2's use of mRNA-seq reads for annotating *S. mediterranea* demonstrates that these types of next-generation techniques can be effectively utilized to improve final gene models.

Annotation Edit Distance as a quality control metric

As the number of published genomes continues to expand, the resources and manpower dedicated to the maintenance of existing genome annotations is being spread thinner and thinner. Manual curation and validation of every annotation in every genome is therefore simply infeasible. A more practical approach would be to dedicate limited resources and manpower to curation and validation of only those

gene annotations most in need of improvement. As we demonstrate below, MAKER2 provides an effective means for automated quality control of genome annotations. Even in cases where the administrators of genome databases have no plans to undertake manual curation, quality control measures are still desirable, as they provide a means for downstream users to judge the quality of an annotation before proceeding with experiments that depend upon the annotation's accuracy for success. Identifying low quality gene annotations however is a challenge not well addressed by existing tools. While quality metrics such as sensitivity, specificity, and accuracy are convenient for evaluating the performance of gene-predictors, they presuppose the existence of reference gene models against which to compare each annotation; this precludes their use in quality control of second-generation genome annotations. Researchers working with second-generation annotations are thus in need of new quality control measures and annotation management tools.

To address this issue, we have adapted the Annotation Edit Distance (AED)[20] measurement, developed by the Sequence Ontology, for use in MAKER2 as an annotation quality control metric. The AED metric is an extension of the sensitivity and specificity measures used to judge gene-finder performance[24], but it differs in that no reference is used. Instead AED measures the distance between two annotations (each from a different releases of the same genome), and it makes no assumptions as to which one is the more correct. As originally formulated, AED provides a means to measures changes to a gene annotation from release to release. We have adapted AED for use in MAKER2 as a means to quantify the distance between a gene annotation and its supporting evidence – EST, protein, and mRNA-seq alignments (see Methods for details). As we show in the analyses presented

below, MAKER2's AED values provide a useful measure for annotation quality control.

AED values are bounded between 0 and 1, with a value of 0 indicating an exact match between an annotation and its evidence and 1 indicating no evidence support. Thus, AED can be used as a rational basis for how much faith should put in an annotation before proceeding with downstream bench experiments, and database managers can use AED to sort gene models from best supported to worst in order to prioritize them for downstream manual review.

As proof-of-principle, we compared MAKER2 produced AED scores for every annotation in release 30 of the *M. musculus* reference annotations (2003) to those of release 37.1 (2007) (Figure 2.2). We also performed the same analysis using reference annotations from human releases 33 (2003) compared to human release 37.2 (2010) (Figure 2.3). In order to perform this analysis, we used MAKER2 to align EST and protein homology evidence against reference annotations from mouse releases 30 and 37.1 and human releases 33 and 37.2 (thus producing AED scores for all annotations in the two datasets). We then plotted the cumulative distribution of AED for each dataset (Figure 2.2c and 2.3c). As can be seen in mouse release 30, there exists an abundance of genes with limited evidence support (large portion of genes with high AED values). In contrast, for the more recent mouse release 37.1, the AED distribution is shifted toward lower AED (better) values. These two curves thus provide a high-level quantitative overview of the genome-wide improvements to the mouse gene-annotations between 2003 and 2007.

Notably, many of the release 33 mouse annotations nearly or completely lack support from EST and protein homology (as indicated by a spike of genes distributed around the AED value of 1). In contrast, for the more recent mouse release 37.1,

there is nearly a complete elimination of the spike of gene around AED score 1; its absence suggests that the earlier releases contained an abundance of false positive gene predictions that were deleted by release 37.1.

To further explore the extent to which AED scores are indicative of annotation quality, we also investigated the AED distribution of the highest quality subset of reference GenBank annotations from each of the mouse and human genome releases (the highest quality genes are those with NM prefixes assigned by RefSeq[36, 37]) (dotted lines in Figure 2.2c and Figure 2.3c). The RefSeq NM prefix provides us with an independently identified ‘gold-standard’ dataset of best quality annotations for comparison. For all releases, we see that the ‘gold-standard’ NM annotation datasets produce cumulative AED distributions that are shifted toward lower AED scores than the reference sets they are derived from. This indicates that MAKER2 is able to both quantify and verify the higher quality of these genes, providing further support for its use as a tool for annotation quality control.

We also investigated how well AED scores agreed with Pfam domain content. As can be seen in Figure 2.2a and Figure 2.3a, AED scores accord well with domain content. In mouse release 30, for example, 87% of genes with AED scores from 0 to 0.25 contain a known domain, whereas only 44% of genes with an AED score ranging from 0.75 to 1.0 contain a domain. The trend is even more striking in human release 33 where only 15% of annotations with AED scores between 0.75 and 1.0 contain a domain, suggesting the abundance of false positive gene predictions in that subset of genes (Figure 2.3a). Tracking these annotations across releases supports this hypothesis: 86% of genes from human release 33 with AED scores between 0.75 and 1.0 are absent by release 37.2 (Figure 2.3b). The same trend is observed in mouse: 59% of annotations in release 30 with AED scores between 0.75 and 1.0 were deleted

by release 37.1 (Figure 2.2b). In comparison, only 14% of genes with AED scores between 0 and 0.25 were deleted between mouse release 30 and release 37.1. Collectively, these results show that gene annotations judged to be of low quality by MAKER2 were also judged to be of low quality by GenBank and preferentially deleted (demonstrating that AED scores mirror the independent curation decisions made by the mouse and human research communities). These facts demonstrate the utility of MAKER2 as an annotation management tool.

Re-annotation of existing genomes and legacy annotation sets

While there are a large number of second-generation genome projects underway, falling sequencing costs are also leading many researchers to revisit published genomes to improve gene models in light of new evidence (such as mRNA-seq) or to take advantage of newer, more complete genome assemblies. There are also instances where researchers are sequencing individual strains/mutants of organisms where a published reference genome is already available or where multiple sets of legacy annotations exist and they wish to carry over annotations from the reference genome and merge them into a nonredundant consensus dataset. MAKER2 provides a simple method to perform these tasks via an external annotation pass-through mechanism that accepts as input any pre-existing genome annotations as well as aligned experimental evidence provided in a GFF3 formatted file (i.e., if the user supplies MAKER2 with gene models or pre-aligned experimental evidence in GFF3 format, that data will merge seamlessly into the pipelines existing analysis).

When using this GFF3 pass-through mechanism, MAKER2 takes the user provided gene models (from GFF3 files), aligns any additional experimental evidence

against the genome (from standard FASTA files), and then calculates quality control statistics such as AED. If the user supplied MAKER2 with more than one legacy annotation dataset (i.e., multiple GFF3 files of alternate legacy annotations), MAKER2 chooses the one model most consistent with the evidence for each locus and carries it forward into the consensus dataset (nonredundant). Researchers can also select to run *ab initio* gene-predictors (as is done for *de novo* annotation) in addition to providing a GFF3 file of legacy annotations. In this case, MAKER2 can produce new gene models for regions where the evidence suggests the existence of a gene that was not found in the legacy set, and with the help of the gene-finders MAKER2 can try and update/revise the legacy annotations to better account for features suggested by aligned evidence.

As proof-of-principle of MAKER2's model pass-through and re-annotation capabilities, we used the pipeline to process a 22 megabase region of maize inbred line B73 chromosome 4[38] together with version 5a.59 of the MaizeSequence.org Working Gene Set[39]. For maize chromosome 4, we produced a *de novo* annotation gene set, a pass-through dataset (in which all reference annotations were maintained but tagged with evidence associations and AED values), and a re-annotation dataset (wherein MAKER2 was allowed to maintain or update reference annotations based on aligned experimental evidence). The cumulative distribution of AED scores for these three datasets was then graphed and is shown in Figure 2.4c. We also plotted the AED distribution of the high quality subset of reference annotations from the Maize Classical Gene List[40] for comparison as an independently identified 'gold-standard' control dataset (Figure 2.4c, gold curve).

During re-annotation, 304 out of 493 version 5a.59 reference gene models were altered/updated to reflect features suggested by evidence alignments; 88 new

gene models were produced for regions where the evidence suggested the existence of a gene but no model existed; and 189 reference gene models were left unchanged. A total of 89 of the unmodified reference gene models had no evidence support and were prioritized by MAKER2 for manual review as possible false positive annotations. Alterations to gene models during the re-annotation process caused the AED distribution curve for the re-annotation dataset (Figure 2.4c, purple curve) to shift towards lower AED values (better) relative to the reference annotation set (Figure 2.4c, red curve). This shift suggests that re-annotation using MAKER2 successfully brought gene models more in line with experimental evidence, thus improving their quality. A further comparison of both the re-annotation dataset and the unmodified reference dataset to the 'gold-standard' annotation set (Figure 2.4c, gold curve) supports this conclusion as these high quality gene models also tend to be distributed around lower AED values (with more than 80% of 'gold-standard' annotations having AED values of < 0.2 compared to just 40% for the version 5a.59 reference annotation set). The spike in the AED distribution for both the unmodified reference dataset and the re-annotation dataset represents gene models that have little-to-no evidence support and are prioritized by MAKER2 for manual review. In comparison, the *de novo* annotation set (Figure 2.4c, blue curve) has an AED distribution shifted toward lower values than either the re-annotation or reference dataset; this is primarily due to the exclusion of unsupported gene models as the average AED for both the *de novo* and re-annotation datasets is identical when unsupported models are excluded (average AED of 0.17 in both).

Managing existing annotation databases

With the proliferation of existing sequencing data, researchers have access to published genomes of multiple related species that may have been annotated using very different methods and to varying degrees of quality. Here, we evaluate how MAKER2's annotation pass-through option can be used to map cross-species data to multiple related genomes. We also explore how these data can be used to fuel downstream analyses such as cross-species orthology.

We used MAKER2 to map experimental evidence as well as reference annotations to six published ant genomes: *Atta cephalotes*[7], *Pogonomyrmex barbatus*[8], *L. humile*[9], *Harpegnathos saltator*[41], *Camponotus floridanus*[41], and *Solenopsis invicta*[18]. The protein datasets provided to MAKER2 consisted of all proteins from UniProt/Swiss-Prot, *D. melanogaster*, *Nasonia vitripennis* (wasp)[42], *Apis mellifera* (honey bee)[43], and each of the previously mentioned published ant species (the individual species whose genome was being evaluated was always excluded from the protein dataset). We also included all Apocrita and Formicidae ESTs in dbEST[44] with the EST dataset. Resulting cumulative AED distributions were then plotted for each ant species; average percent orthology and domain content were also evaluated for each quartile of the AED distribution (Figure 2.5).

Low AED scores indicate gene models that are more in agreement with evidence alignments while higher values mean less evidence support. The cumulative distribution of AED scores for the six ant species can be seen in Figure 2.5c. For each ant species, there is a spike in the distribution curves around AED score 1. This spike represents genes that MAKER2 has prioritized for manual review. We see in Figure 2.5a that Pfam domain content is well correlated with AED

score, and an average of 63% of genes with scores between 0 and 0.25 contain a Pfam domain compared to only 11% of genes with scores between 0.75 and 1.0. The low domain enrichment suggests that genes prioritized by MAKER2 are most likely false positive gene predictions, but there is also the potential that these represent novel genes with domains that would not be found in the Pfam domain database. If we further expand our analysis to look at orthology among the ant species, we see that percent orthology between the six ant species is well correlated to AED. For example, 94% of genes with AED scores between 0 and 0.25 have orthology to at least one protein in another ant species (on average there are 4.41 orthologous genes in other ant species that associate back to each of these), whereas only 26% of genes with AED scores between 0.75 and 1.0 have at least 1 ortholog in another ant species (for the genes here that have an ortholog there are only 1.85 orthologs that map back to them on average). Together with the domain analysis, the association of AED and orthology suggests that genes with AED scores near 1 are indeed false positive gene predictions, thus supporting the use of the AED statistic for quality control. The correlation of AED to orthology also suggests that AED could be used as a selection tool for identifying genes sets for downstream experimentation with well conserved genes shared across species more likely to group around low AED values. Because the annotations and evidence are all loaded into a database ready output format, researchers can also view MAKER2 cross-species alignments in GMOD tools like GBrowse and Apollo, which would allow them to explore more detailed aspects of cross-species conservation such as presence/absence of exons and introns.

The ability of MAKER2 to align cross-species data to multiple genomes in this way demonstrates how MAKER2 can be used to generate common resources even when genomes are annotated using very different methods. Because all

annotations and experimental evidence have been processed into a common format, they can now be easily loaded into downstream GMOD tools for analysis and data distribution. MAKER2 thus provides an efficient automated mechanism for research communities and organizations to manage shared genome database resources.

High-throughput parallelization

MAKER2 has been optimized to support high-throughput parallelization using Message Passing Interface (MPI), a distributed cluster communication protocol. To explore how data throughput in MAKER2 scales with processor usage, we annotated the 10 megabase NGASP[23] dataset for *C. elegans* using an increasing number of processor cores (Figure 2.6). We see from the analysis that data throughput scales linearly with processor usage, annotating the entire 10 megabase dataset in just under 1 hour on 32 CPU cores; this means MAKER2 should be able to annotate the entire *C. elegans* genome in less than 10 hours using similar settings. Researcher with access to distributed computer clusters (300-3000 CPU cores) could expect to annotate even human-sized genomes (~2-3 gigabases) in less than 24 hours, while smaller fungal sized genomes (~40-80 megabases) could easily be annotated on laptop or desktop machines in the same time period. The scalability of data throughput for MAKER2 therefore allows researchers to process datasets of virtually any size or to process multiple datasets in a timely manner. MAKER2's high-throughput parallelization also provides a potential solution to the problem of annotating ultra large genomes such as pine trees, which have genomes in the 20-30 gigabase range[13].

It is important to note that much of MAKER2's computation time is spent aligning experimental evidence to the genome and analyzing the results. For this

reason, the overall time required for genome annotation is expected to vary not only with genome length but also with the size of the input experimental evidence dataset. This upfront investment in computation time, however, provides enormous benefits downstream as all supplied EST reads, protein homology data, and gene predictions are available as searchable features in the final output. By loading MAKER2's output into GMOD tools like Chado[45], Galaxy[46], and GBrowse[47], researchers can quickly perform downstream analyses such as exploring protein orthology and analyzing sequence conservation. They can also identify cross-species changes in intron exon structures with the advantage of having all the information available directly from MAKER2's output without having to perform any additional computation.

Conclusions

MAKER2's annotation of the classic model systems *D. melanogaster*, *C. elegans*, and *A. thaliana* demonstrates that the accuracy of gene models produced by MAKER2 is comparable to that produced by other existing tools (*ab initio* gene-predictors). We also see that with enough training, HMM based *ab initio* gene prediction algorithms can perform as well, if not better, than more computationally intensive evidence-based annotation pipelines like MAKER2. This has previously been seen in genome annotation competitions like NGASP[23]. However, the performance of *ab initio* gene-predictors is heavily dependent on the availability of extensive training data, and further analysis demonstrates the fallacy of expecting similar outcomes for emerging model organism genomes.

First-generation (classic model organism) genome projects benefitted heavily from the extensive knowledge of genes and gene structure that was already

available before the genome projects even began. Unfortunately, second-generation (emerging model organisms) genomes, which represent the overwhelming majority of new sequencing projects, do not share the advantage of an extensive pre-existing knowledgebase. There are usually few if any pre-existing gene models, often no genetics; in fact, the genome project may be the primary resource to begin future research into these organisms. For genomes such as these, using *ab initio* gene predictors as part of an annotation pipeline, like MAKER2, produces better results (Figure 2.1).

By aligning evidence from ESTs, mRNA-seq, and protein homology, MAKER2 also provide a convenient way to add these types of experimental data to new and existing genome databases for downstream analysis or visualization in GMOD tools such as Apollo and GBrowse. Additionally, the association of evidence to gene models via the AED statistic provides a simple mechanism for focusing limited resources to the subset of genes most in need of review and manual curation. As proof-of-principle, we demonstrated that MAKER2 was able to prioritize genes for review from mouse release 30 and human release 33. This prioritization overwhelmingly correlates with the deletion and revision of the same genes in subsequent mouse release 37.1 and human release 37.2, indicating that AED prioritization closely emulates manual and automatic quality control methods used for these genomes. MAKER2 however provides the advantage of performing this type of quality control prioritization as a fully automated analysis that can be performed by individuals with little bioinformatics experience on both new and existing genome annotations.

It is important to realize that the primary benefit of the aligned evidence and quality control statics produced by MAKER2 is that they allow researchers to make

more informed decisions when designing experiments. Researchers are often unaware of the confidence associated with a gene model, which can have potentially disastrous results as incorrect annotations poison every experiment that uses them.

Methods

De novo annotation of first-generation genomes

D. melanogaster chromosome 3R and GFF3 annotations for release r5.32 were downloaded from FlyBase. *C. elegans* chromosome V and GFF3 annotations for release WS221 were downloaded from WormBase. *A. thaliana* chromosome 4 and GFF3 annotations were downloaded from TAIR. Each set of reference gene annotations were passed to MAKER2's model_gff option with all prediction and evidence alignment options turned off. This has the effect of repackaging the reference gene models into more standardized GFF3 files compatible with downstream analysis scripts.

Ab initio gene predictions were produced by the programs SNAP version 2010-07-28, Augustus 2.5.5, and GeneMark-ES 2.3a, using the *D. melanogaster*, *C. elegans*, and *A. thaliana* parameter files pre-packaged with each algorithm (GeneMark parameter files are packaged with the GeneMark.hmm download). To produce all predictions in standardized GFF3 format, these algorithms were run through MAKER2 with all evidence alignments options turned off and the keep_preds flag set to 1. This has the effect of only producing raw *ab initio* gene predictions in standardized GFF3 format.

Evidence-based gene annotations in MAKER2 were produced using default settings. The species parameter files were the same as those used for the *ab initio* gene-predictors. EST and protein homology datasets were provided for each

organism. For *D. melanogaster*, the EST dataset consisted of all *D. melanogaster* ESTs available from dbEST, and the protein homology input consisted of all *Anopheles gambiae* proteins from NCBI together with all of the UniProt/Swiss-Prot database proteins (minus Drosophila proteins). For *C. elegans*, the EST dataset consisted of all *C. elegans* release WS221 ESTs available from WormBase, and protein homology input consisted of all *Caenorhabditis briggsae* WS221 proteins from WormBase together with all of the UniProt/Swiss-Prot database proteins (minus Caenorhabditis proteins). The EST dataset for *A. thaliana* consisted of all *A. thaliana* ESTs from dbEST, and the protein homology dataset consisted of all *Oryza sativa* release 6.1 proteins from PlantGDB and all of the UniProt/Swiss-Prot database proteins (minus Arabidopsis proteins). For *A. thaliana*, MAKER2 was also provided with the Arabidopsis transposable element FASTA file available from TAIR (assists in repeat masking).

The reference gene models, *ab initio* gene predictions, and evidence-based gene annotations were converted to GTF format using the maker2eval script packaged with MAKER2. Values for sensitivity and specificity were then produced using Eval[30].

De novo annotation using unmatched species parameters

To simulate non-optimal training of the *ab initio* gene-finders, *ab initio* predictions and MAKER2 annotations were produced for *D. melanogaster*, *C. elegans*, and *A. thaliana* using unmatched species parameter files. This was done by running SNAP, Augustus, GeneMark, and MAKER2 on *C. elegans* and *D. melanogaster* using the *A. thaliana* parameter files. These programs were then run

on *A. thaliana* using the *C. elegans* parameter files. All other steps and procedures were identical to the previous analysis.

De novo annotation of second-generation genomes

S. mediterranea assembly 3.1 and *L. humile* assembly 4.0 were used to evaluate the performance of the *ab initio* gene-predictor SNAP and the annotation pipeline MAKER2 on second-generation genome projects. To produce SNAP required parameter files for each species, we first ran CEGMA, which produces gene models that can be used for training SNAP from a core set of universal genes that should be found in all eukaryotes. CEGMA gene models were converted to SNAP required ZFF format using the `cegma2zff` script that comes bundled with MAKER2. SNAP was then trained in accordance with its documentation.

To produce all predictions in standardized GFF3 format, SNAP was ran via MAKER2 with all evidence alignments options turned off and the `keep_preds` flag set to 1. This has the effect of only producing raw *ab initio* gene predictions in standardized GFF3 format.

MAKER2 was run on *S. mediterranea* using an EST dataset consisting of all ESTs available for *S. mediterranea* found in dbEST together with the SmedGD EST dataset[6, 17]. The protein homology dataset consisted of all proteins in the UniProt/Swiss-Prot protein database, all *Schistosoma mansoni* v4.0 proteins from Sanger, and all GenBank proteins for *Nematostella vectensis*, *H. sapiens*, *C. elegans*, and *S. mediterranea*. The SmedGD repeat library was also used[6, 17].

Short read mRNA-seq transcriptome datasets for *S. mediterranea* were downloaded from the NCBI Sequence Read Archive (SRP006000). TopHat[48] v1.2.0 and Cufflinks[49] v0.9.3 were used to align and process these short reads. The script

tophat2gff3 and cufflinks2gff3 were then used to process the results into GFF3 format. The resulting GFF3 files were provided to the est_gff option in MAKER2.

MAKER2 was run on *L. humile* using the published genome project EST[9] dataset together with all Apocrita and Formicidae ESTs available from dbEST. The protein homology dataset consisted of all of the UniProt/Swiss-Prot protein database, *D. melanogaster* r5.32 proteins, *N. vitripennis* OGS 1.2 proteins, *A. mellifera* OGS 2 proteins, and all Formicidae proteins from GenBank. A combined repeat FASTA file from the published *L. humile* and *P. barbatus* genomes was also provided[8, 9].

Pfam domain analysis

InterProScan[32] was used to identify Pfam[31] domains for all gene prediction/annotation datasets. Domains were filtered to remove reverse transcriptase, integrase, and virus related protein domains. Any domain listed as unknown, uncharacterized, or NULL was ignored.

To explore the upper bound of expected Pfam domain content in a newly annotated genome, we used InterProScan to identify Pfam protein domains in *H. sapiens* release 37.2, *M. musculus* release 37.1, *D. melanogaster* r5.32, *C. elegans* WS221, and *S. cerevisiae* (NCBI release). Domains were filtered as before (i.e. remove reverse transcriptase, integrase, and virus related domains). The average domain enrichment for these reference genomes was then calculated for comparison.

Calculating Annotation Edit Distance

Sensitivity, specificity, and accuracy are commonly used metrics for evaluating the performance of gene prediction algorithms by comparing the

resulting gene prediction to a well-supported reference annotation[50]. Sensitivity is defined as the fraction of a reference overlapping a prediction; specificity is defined as the fraction of a prediction overlapping a reference; and accuracy is commonly defined as the average of sensitivity and specificity (although several alternate formulations exist). Both sensitivity and specificity can be calculated for any feature in the genome at different levels of stringency (i.e., base pair level, exon level, etc.).

Given a gene prediction i and a reference j , the base pair level sensitivity can be calculated using the formula $SN = |i \cap j| / |j|$; where $|i \cap j|$ represents the number of overlapping nucleotides between i and j , and $|j|$ represents the total number of nucleotides in the reference j . Alternatively, specificity is calculated using the formula $SP = |i \cap j| / |i|$, and accuracy is the average of the two.

When calculating Annotation-Evidence Distance, we adapt the calculation of sensitivity and specificity to account for the fact we do not have a reference gene model for comparison; instead, we cluster experimental evidence aligned against the genome to approximate the reference. So for $SN = |i \cap j| / |j|$, the value $|i \cap j|$ represents the number of nucleotides in a gene prediction overlapped by experimental evidence, and $|j|$ represents the total base pair count for experimental evidence in that cluster. Because we are not comparing to a high quality reference, it is more correct to refer to the average of sensitivity and specificity as the *congruency* rather than accuracy; where $C = (SN+SP)/2$. The *incongruency*, or distance between i and j , then becomes $D = 1-C$, with a value of 0 indicating complete agreement of an annotation to the evidence, and values at or near 1 indicating disagreement or no evidence support.

AED evaluation for the human and mouse genomes

H. sapiens annotations for releases 33 and 37.2 as well as *M. musculus* annotations for releases 30 and 37.1 were downloaded from NCBI in GenBank file format. They were converted to GFF3 format using the `genbank2gff3` script available in the BioPerl[51] 1.6 distribution. The resulting GFF3 files were passed to MAKER2's `model_gff` option with all prediction and evidence alignment options turned off. This has the effect of repackaging the gene models into more standardized GFF3 files compatible with downstream analysis scripts.

The standardized GFF3 files were then provided to MAKER2's `model_gff` option once again together with protein and EST datasets to produce downstream quality control metrics for each gene model. The human reference gene annotations were processed using all human ESTs from dbEST and a protein dataset consisting of all mouse proteins together with all of UniProt/Swiss-Prot (minus human proteins), and the genome was masked using the mammal subset of repeats from RepBase. The mouse reference gene annotations were processed using all mouse ESTs from dbEST and a protein dataset consisting of all human proteins together with all of UniProt/Swiss-Prot (minus mouse proteins), and the genome was masked using the mammal subset of repeats from RepBase.

The presence/absence of human release 33 genes in release 37.2 and mouse release 30 genes in release 37.1 was determined using BLASTP and reciprocal best hits analysis (where genes from each dataset are each others best hit). A threshold e-value of 1×10^{-6} was required for all hits. Pfam domains were also mapped to all genes using the previously described methodology.

Re-annotation of the maize genome

To demonstrate MAKER2's ability to re-annotate existing genomes with respect to legacy annotations, we re-annotated a 22 megabase region of the *Zea mays* (maize) inbred line B73 chromosome 4, available from MaizeSequence.org. We then used the subset of reference annotations that are also included in the Maize Classical Gene List[40] as a 'gold standard' set to evaluate MAKER2's performance.

We first produced a standardized GFF3 file for the maize reference annotations by using the map2assembly script bundled with MAKER2 to map maize reference transcripts onto the genome. We then provided the resulting GFF3 file to MAKER2 via the model_gff option and provided an EST dataset consisting of all ESTs/cDNAs for maize available from the Maize Full Length cDNA Project[52] and dbEST. The protein homology dataset we used consisted of the *A. thaliana* proteome and all of the UniProt/Swiss-Prot database (minus any maize proteins). Maize specific repeats were acquired from the Maize Transposable Element Database[53]. The resulting MAKER2 output was a GFF3 file containing AED quality control values for all reference transcripts. The AED distribution of the reference was then graphed together with the AED distribution for the 'gold standard' genes identified as overlapping the Maize Classical Gene List.

Next we produced *de novo* annotation and a re-annotation dataset using MAKER2. The *de novo* annotation dataset was produced using the maize prediction parameter file that comes bundled with SNAP. We also provided MAKER2 with the same EST, protein, and repeat datasets used in the previous analysis. To produce the re-annotation dataset, we again used the same EST, protein, repeat, and SNAP files; however, we also passed MAKER2 all legacy annotations by indicating the

location of the reference GFF3 file in the `model_gff` option. We then graphed the AED distributions as was done previously for the reference dataset.

Evidence alignment and analysis of published ant genomes

To demonstrate how MAKER2 can be used to add experimental evidence and quality control statistics to existing genome databases (which can fuel downstream analyses or be used to improve annotations), we used MAKER2 to add cross-species homology data to six published ant genomes. We downloaded annotations for *A. cephalotes* OGS 1.2, *L. humile* OGS 1.2, *P. barbatus* OGS 1.2, *C. floridanus* v3.3, *H. saltator* v3.3, and *S. invicta* v2.2.0 from the Hymenoptera Genome Database[43]. Most species had GFF3 format annotations that were passed to MAKER2's `model_gff` option, with all prediction and evidence alignment options turned off. This has the effect of repackaging the gene models into more standardized GFF3 files compatible with downstream analysis scripts. For *S. invicta*, however, we used the `map2assembly` script bundled with MAKER2 to map transcripts onto the genome assembly (thus producing a standardized GFF3 formatted annotation file).

We next ran MAKER2 on each of the six ant species. Standardized GFF3 files were passed to MAKER2's `model_gff` option. We used an EST dataset consisting of all Apocrita and Formicidae ESTs available from dbEST (this did not include ESTs for any of the six species being analyzed). We used a protein homology dataset consisting of all of the UniProt/Swiss-Prot protein database, *D. melanogaster* r5.32, *N. vitripennis* OGS 1.2, *A. mellifera* OGS 2, and all of the published ant proteomes (always excluding the species being processed at the time). A combined ant repeat FASTA file from the published *L. humile* and *P. barbatus* genomes was also provided.

Orthology of the six ant species was explored using BLASTP and reciprocal best hits analysis. A threshold e-value of 1×10^{-6} was required for all hits. We also used InterProScan to identify Pfam domains for all proteins using the previously described methodology.

Evaluation of high through-put parallelization

The parallelization performance of MAKER2 was evaluated on a server with four, twelve-core AMD Opteron 6174 Processors (48 total CPU cores) running Red Hat Enterprise Linux Server release 5.5. MAKER2 was configured with default settings and the NGASP protein, EST, and genomic sequence datasets available from WormBase. The NGASP genomic sequence is a selected 10 megabase sampling of the *C. elegans* genome (release WS160). We ran MAKER2 (the parallel executable is `mpi_maker`) using 1, 4, 8, 16, and 32 CPU cores under MPICH2 1.3.1. The Linux `time` command was used to evaluate process run time.

List of abbreviations

(AED) Annotation Edit Distance; (GMOD) Generic Model Organism Database project; (GFF3) Generic Feature Format version 3; (MPI) Message Passing Interface

Authors' contributions

CH and MY conceived the study and wrote the manuscript. CH carried out experiments and wrote software for the analyses.

Acknowledgements

This work was supported by the grant NIH/NHGRI-R01-HG004694 to MY and partially supported by the NIH Genetics Training Grant T32-GM007464. An allocation of computer time from the Center for High Performance Computing at the University of Utah is gratefully acknowledged. The authors would like to thank C. Robin Buell and John Hamilton of Michigan State University for their assistance with the maize genome re-annotation, as well as Alejandro Sánchez Alvarado and Eric Ross of the University of Utah for their assistance in collecting the datasets to annotate the *Schmidtea mediterranea* genome. Finally, the assistance provided by Christopher Smith (of San Francisco State University) in gathering the necessary datasets for analysis of all six published ant species is also gratefully recognized.

References

1. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, et al: **The genome sequence of *Drosophila melanogaster***. *Science* 2000, **287**:2185-2195.
2. The *C. elegans* Sequencing Consortium: **Genome sequence of the nematode *C. elegans*: a platform for investigating biology**. *Science* 1998, **282**:2012-2018.
3. Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, et al: **Initial sequencing and comparative analysis of the mouse genome**. *Nature* 2002, **420**:520-562.
4. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al: **The sequence of the human genome**. *Science* 2001, **291**:1304-1351.
5. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**:860-921.

6. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, Holt C, Sanchez Alvarado A, Yandell M: **MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes.** *Genome Res* 2008, **18**:188-196.
7. Suen G, Teiling C, Li L, Holt C, Abouheif E, Bornberg-Bauer E, Bouffard P, Caldera EJ, Cash E, Cavanaugh A, et al: **The genome sequence of the leaf-cutter ant *Atta cephalotes* reveals insights into its obligate symbiotic lifestyle.** *PLoS Genet* 2011, **7**:e1002007.
8. Smith CR, Smith CD, Robertson HM, Helmkampf M, Zimin A, Yandell M, Holt C, Hu H, Abouheif E, Benton R, et al: **Draft genome of the red harvester ant *Pogonomyrmex barbatus*.** *Proceedings of the National Academy of Sciences* 2011, **108**:5667-5672.
9. Smith CD, Zimin A, Holt C, Abouheif E, Benton R, Cash E, Croset V, Currie CR, Elhaik E, Elsik CG, et al: **Draft genome of the globally widespread and invasive Argentine ant (*Linepithema humile*).** *Proceedings of the National Academy of Sciences* 2011, **108**:5673-5678
10. Levesque CA, Brouwer H, Cano L, Hamilton J, Holt C, Huitema E, Raffaele S, Robideau G, Thines M, Win J, et al: **Genome sequence of the necrotrophic plant pathogen *Pythium ultimum* reveals original pathogenicity mechanisms and effector repertoire.** *Genome biology* 2010, **11**:R73.
11. Baxter SW, Nadeau NJ, Maroja LS, Wilkinson P, Counterman BA, Dawson A, Beltran M, Perez-Espona S, Chamberlain N, Ferguson L, et al: **Genomic hotspots for adaptation: the population genetics of Mullerian mimicry in the *Heliconius melpomene* clade.** *PLoS Genet* 2010, **6**:e1000794.
12. Ferguson L, Lee SF, Chamberlain N, Nadeau N, Joron M, Baxter S, Wilkinson P, Papanicolaou A, Kumar S, Kee T-J, et al: **Characterization of a hotspot for mimicry: assembly of a butterfly wing transcriptome to genomic sequence at the HmYb/Sb locus.** *Molecular Ecology* 2010, **19**:240-254.
13. Kovach A, Wegrzyn J, Parra G, Holt C, Bruening G, Loopstra C, Hartigan J, Yandell M, Langley C, Korf I, Neale D: **The *Pinus taeda* genome is characterized by diverse and highly diverged repetitive sequences.** *BMC Genomics* 2010, **11**:420.
14. MacDonald J, Doering M, Canam T, Gong Y, Guttman DS, Campbell MM, Master ER: **Transcriptomic responses of the softwood-degrading white-rot fungus *Phanerochaete carnosae* during growth on coniferous and deciduous wood.** *Appl Environ Microbiol* 2011:AEM.02490-02410.

15. Legeai F, Shigenobu S, Gauthier JP, Colbourne J, Rispe C, Collin O, Richards S, Wilson ACC, Murphy T, Tagu D: **AphidBase: a centralized bioinformatic resource for annotation of the pea aphid genome.** *Insect Molecular Biology* 2010, **19**:5-12.
16. Martin J, Abubucker S, Wylie T, Yin Y, Wang Z, Mitreva M: **Nematode.net update 2008: improvements enabling more efficient data mining and comparative nematode genomics.** *Nucleic acids research* 2009, **37**:D571-D578.
17. Robb S, Ross E, Alvarado A: **SmedGD: the Schmidtea mediterranea genome database.** *Nucleic Acids Res* 2007:D599 - 606.
18. Wurm Y, Wang J, Riba-Grognuz O, Corona M, Nygaard S, Hunt BG, Ingram KK, Falquet L, Nipitwattanaphon M, Gotzek D, et al: **The genome of the fire ant Solenopsis invicta.** *Proceedings of the National Academy of Sciences* 2011, **108**:5679-5684.
19. Hauser PM, Burdet FX, Cisse OH, Keller L, Taffe P, Sanglard D, Pagni M: **Comparative genomics suggests that the fungal pathogen Pneumocystis is an obligate parasite scavenging amino acids from its host's lungs.** *PLoS ONE* 2010, **5**:e15152.
20. Eilbeck K, Moore B, Holt C, Yandell M: **Quantitative measures for the management and comparison of annotated genomes.** *BMC Bioinformatics* 2009, **10**:67.
21. Eilbeck K, Lewis S, Mungall C, Yandell M, Stein L, Durbin R, Ashburner M: **The Sequence Ontology: a tool for the unification of genome annotations.** *Genome biology* 2005, **6**:R44.
22. Guigo R, Flicek P, Abril JF, Reymond A, Lagarde J, Denoeud F, Antonarakis S, Ashburner M, Bajic VB, Birney E, et al: **EGASP: the human ENCODE genome annotation assessment project.** *Genome Biol* 2006, **7**:1 - 31.
23. Coghlan A, Fiedler T, McKay S, Flicek P, Harris T, Blasiar D, the n GC, Stein L: **nGASP - the nematode genome annotation assessment project.** *BMC Bioinformatics* 2008, **9**:549.
24. Burset M, Guigo R: **Evaluation of gene structure prediction programs.** *Genomics* 1996, **34**:353 - 367.
25. Korf I: **Gene finding in novel genomes.** *BMC Bioinformatics* 2004, **5**:59.
26. Lomsadze A, Ter-Hovhannisyanyan V, Chernoff YO, Borodovsky M: **Gene identification in novel eukaryotic genomes by self-training algorithm.** *Nucl Acids Res* 2005, **33**:6494-6506.
27. Stanke M, Waack S: **Gene prediction with a hidden Markov model and a new intron submodel.** *Bioinformatics* 2003, **19**:ii215-225.

28. Stanke M, Diekhans M, Baertsch R, Haussler D: **Using native and syntenically mapped cDNA alignments to improve de novo gene finding.** *Bioinformatics* 2008, **24**:637 - 644.
29. The Arabidopsis Genome Initiative: **Analysis of the genome sequence of the flowering plant Arabidopsis thaliana.** *Nature* 2000, **408**:796-815.
30. Keibler E, Brent M: **Eval: A software package for analysis of genome annotations.** *BMC Bioinformatics* 2003, **4**:50.
31. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, et al: **Pfam: clans, web tools and services.** *Nucl Acids Res* 2006, **34**:D247-251.
32. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R: **InterProScan: protein domains identifier.** *Nucl Acids Res* 2005, **33**:W116-120.
33. Zmasek C, Godzik A: **Strong functional patterns in the evolution of eukaryotic genomes revealed by the reconstruction of ancestral protein domain repertoires.** *Genome biology* 2011, **12**:R4.
34. Cherry JM, Ball C, Weng S, Juvik G, Schmidt R, Adler C, Dunn B, Dwight S, Riles L, Mortimer RK, Botstein D: **Genetic and physical maps of Saccharomyces cerevisiae.** *Nature* 1997, **387**:67-73.
35. Parra G, Bradnam K, Korf I: **CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes.** *Bioinformatics* 2007, **23**:1061-1067.
36. Pruitt K, Tatusova T, Maglott D: **NCBI reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic acids research* 2005:D501 - 504.
37. Pruitt KD, Tatusova T, Maglott DR: **NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic Acids Res* 2007:D61 - 65.
38. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, et al: **The B73 maize genome: complexity, diversity, and dynamics.** *Science* 2009, **326**:1112-1115.
39. Wei F, Stein JC, Liang C, Zhang J, Fulton RS, Baucom RS, De Paoli E, Zhou S, Yang L, Han Y, et al: **Detailed analysis of a contiguous 22-Mb region of the maize genome.** *PLoS Genet* 2009, **5**:e1000728.
40. **Maize Classical Gene List** [http://synteny.cnr.berkeley.edu/wiki/index.php/Classical_Maize_Genes]

41. Bonasio R, Zhang G, Ye C, Mutti NS, Fang X, Qin N, Donahue G, Yang P, Li Q, Li C, et al: **Genomic comparison of the ants *Camponotus floridanus* and *Harpegnathos saltator***. *Science*, **329**:1068-1071.
42. Werren JH, Richards S, Desjardins CA, Niehuis O, Gadau Jr, Colbourne JK, Group TNGW: **Functional and evolutionary insights from the genomes of three parasitoid nasonia species**. *Science*, **327**:343-348.
43. Munoz-Torres MC, Reese JT, Childers CP, Bennett AK, Sundaram JP, Childs KL, Anzola JM, Milshina N, Elisk CG: **Hymenoptera Genome Database: integrated community resources for insect species of the order Hymenoptera**. *Nucleic acids research*, **39**:D658-D662.
44. Boguski MS, Lowe TMJ, Tolstoshev CM: **dbEST - database for expressed sequence tags**. *Nat Genet* 1993, **4**:332-333.
45. Mungall C, Emmert D: **A Chado case study: an ontology-based modular schema for representing genome-associated biological information**. *Bioinformatics (Oxford, England)* 2007, **23**:i337 - 346.
46. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, et al: **Galaxy: a platform for interactive large-scale genome analysis**. *Genome research* 2005, **15**:1451-1455.
47. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S: **The Generic Genome Browser: a building block for a model organism system database**. *Genome Res* 2002, **12**:1599-1610.
48. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-seq**. *Bioinformatics* 2009, **25**:1105-1111.
49. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation**. *Nat Biotech* 2010, **28**:511-515.
50. Guigo R, Dermitzakis ET, Agarwal P, Ponting CP, Parra G, Reymond A, Abril JF, Keibler E, Lyle R, Ucla C, et al: **Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes**. *Proceedings of the National Academy of Sciences* 2003, **100**:1140-1145.
51. **BioPerl** [<http://www.bioperl.org>]

52. Soderlund C, Descour A, Kudrna D, Bomhoff M, Boyd L, Currie J, Angelova A, Collura K, Wissotski M, Ashley E, et al: **Sequencing, mapping, and analysis of 27,455 maize full-length cDNAs.** *PLoS Genet* 2009, **5**:e1000740.
53. **Maize Transposable Element Database** [<http://maizetedb.org/>]

Table 2.1 Gene model accuracy for gene prediction/annotation programs.

Reference Organism	Performance Category	<i>Ab Initio</i> Prediction			MAKER Annotations		
		Augustus	GeneMark	SNAP	Augustus	GeneMark	SNAP
<i>A. thaliana</i>	Nucleotide Accuracy	77.04%	74.68%	69.78%	80.53%	79.39%	80.27%
	Exon Accuracy	67.03%	61.31%	56.40%	67.81%	69.60%	68.78%
<i>D. melanogaster</i>	Nucleotide Accuracy	76.08%	66.54%	69.29%	76.42%	73.66%	74.33%
	Exon Accuracy	61.37%	47.31%	47.01%	58.56%	58.03%	58.49%
<i>C. elegans</i>	Nucleotide Accuracy	88.29%	88.09%	85.10%	87.14%	86.29%	88.48%
	Exon Accuracy	74.62%	68.88%	61.38%	68.60%	65.03%	66.19%

Table 2.2 Sensitivity and specificity for prediction/annotation programs.

	<i>Ab Initio</i> Prediction			MAKER2 Annotations		
	Augustus	GeneMark	SNAP	Augustus	GeneMark	SNAP
<u><i>Arabidopsis thaliana</i></u>						
Gene Sensitivity	35.91%	29.41%	15.67%	23.72%	28.43%	26.33%
Gene Specificity	30.23%	19.05%	13.53%	36.50%	35.07%	34.17%
Transcript Sensitivity	29.85%	24.06%	12.82%	19.61%	23.33%	21.61%
Transcript Specificity	30.23%	19.05%	13.53%	33.41%	35.07%	34.09%
Exon Sensitivity	70.79%	68.06%	58.86%	63.30%	63.39%	62.69%
Exon Specificity	63.27%	54.56%	53.93%	72.31%	75.80%	74.87%
Nucleotide Sensitivity	85.49%	83.08%	76.38%	73.01%	70.91%	71.93%
Nucleotide Specificity	68.58%	66.27%	63.17%	88.05%	87.87%	88.60%
<u><i>Drosophila melanogaster</i></u>						
Gene Sensitivity	19.28%	7.62%	8.30%	17.10%	18.78%	17.60%
Gene Specificity	22.41%	7.61%	9.06%	23.34%	21.72%	21.72%
Transcript Sensitivity	14.07%	5.44%	6.08%	12.42%	13.60%	12.78%
Transcript Specificity	22.41%	7.61%	9.06%	21.23%	21.72%	21.69%
Exon Sensitivity	57.37%	50.18%	49.17%	53.05%	51.88%	52.77%
Exon Specificity	65.36%	44.44%	44.85%	64.07%	64.17%	64.21%
Nucleotide Sensitivity	76.67%	69.07%	70.18%	69.77%	66.33%	66.92%
Nucleotide Specificity	75.49%	64.00%	68.40%	83.07%	80.98%	81.74%
<u><i>Caenorhabditis elegans</i></u>						
Gene Sensitivity	36.48%	29.95%	22.89%	24.57%	24.70%	25.39%
Gene Specificity	41.76%	25.04%	18.01%	34.88%	24.03%	29.87%
Transcript Sensitivity	32.59%	25.68%	20.21%	21.79%	21.76%	22.30%
Transcript Specificity	41.76%	25.04%	18.01%	33.47%	24.03%	29.76%
Exon Sensitivity	71.43%	72.79%	65.41%	62.01%	61.35%	63.48%
Exon Specificity	77.81%	64.96%	57.34%	75.18%	68.70%	68.90%
Nucleotide Sensitivity	87.34%	92.50%	87.31%	80.13%	80.64%	84.64%
Nucleotide Specificity	89.23%	83.67%	82.89%	94.14%	91.94%	92.32%

Table 2.3 Gene model accuracy using unmatched species parameters.

Reference Organism	Performance Category	<i>Ab Initio</i> Prediction			MAKER Annotations		
		Augustus	GeneMark	SNAP	Augustus	GeneMark	SNAP
<i>A. thaliana</i>	Nucleotide Accuracy	57.85%	48.62%	43.84%	68.56%	57.96%	73.77%
	Exon Accuracy	30.71%	16.51%	18.58%	53.31%	28.87%	60.11%
<i>D. melanogaster</i>	Nucleotide Accuracy	67.47%	66.51%	48.92%	73.78%	72.83%	74.44%
	Exon Accuracy	30.62%	26.25%	19.94%	43.10%	39.74%	53.69%
<i>C. elegans</i>	Nucleotide Accuracy	66.18%	67.26%	68.24%	74.32%	71.92%	85.02%
	Exon Accuracy	28.33%	30.01%	35.44%	38.52%	39.42%	63.14%

Table 2.4 Sensitivity and specificity using unmatched species parameters.

	<i>Ab Initio</i> Prediction			MAKER2 Annotations		
	Augustus	GeneMark	SNAP	Augustus	GeneMark	SNAP
<u><i>Arabidopsis thaliana</i></u>						
Gene Sensitivity	3.45%	4.54%	2.72%	9.36%	3.92%	14.34%
Gene Specificity	7.40%	4.27%	3.93%	17.61%	7.90%	22.34%
Transcript Sensitivity	2.94%	3.78%	2.18%	7.82%	3.38%	11.92%
Transcript Specificity	7.40%	4.27%	3.93%	17.61%	7.90%	22.34%
Exon Sensitivity	19.51%	10.82%	9.48%	38.04%	13.54%	46.86%
Exon Specificity	41.91%	22.19%	27.67%	68.57%	44.19%	73.35%
Nucleotide Sensitivity	42.09%	30.99%	19.79%	45.89%	25.45%	54.82%
Nucleotide Specificity	73.60%	66.24%	67.88%	91.22%	90.46%	92.71%
<u><i>Drosophila melanogaster</i></u>						
Gene Sensitivity	18.17%	18.17%	17.60%	13.94%	13.14%	25.03%
Gene Specificity	4.96%	3.94%	3.43%	14.36%	11.98%	23.30%
Transcript Sensitivity	13.37%	13.14%	12.35%	10.29%	9.73%	17.99%
Transcript Specificity	4.96%	3.94%	3.43%	14.36%	11.98%	23.30%
Exon Sensitivity	40.12%	36.00%	24.24%	39.18%	33.34%	44.28%
Exon Specificity	21.12%	16.50%	15.63%	47.02%	46.13%	63.09%
Nucleotide Sensitivity	82.45%	81.32%	53.76%	66.82%	62.39%	61.44%
Nucleotide Specificity	52.49%	51.69%	44.07%	80.74%	83.27%	87.44%
<u><i>Caenorhabditis elegans</i></u>						
Gene Sensitivity	4.39%	8.13%	12.09%	4.80%	8.29%	24.84%
Gene Specificity	5.27%	6.99%	9.29%	7.57%	11.29%	29.94%
Transcript Sensitivity	4.18%	7.97%	11.64%	4.32%	8.10%	22.22%
Transcript Specificity	5.28%	7.00%	9.29%	7.05%	11.29%	29.59%
Exon Sensitivity	20.19%	23.63%	27.98%	28.36%	26.94%	55.44%
Exon Specificity	36.46%	36.38%	42.90%	48.68%	51.89%	70.84%
Nucleotide Sensitivity	52.64%	55.84%	56.64%	56.46%	52.49%	75.92%
Nucleotide Specificity	79.72%	78.67%	79.84%	92.17%	91.35%	94.12%

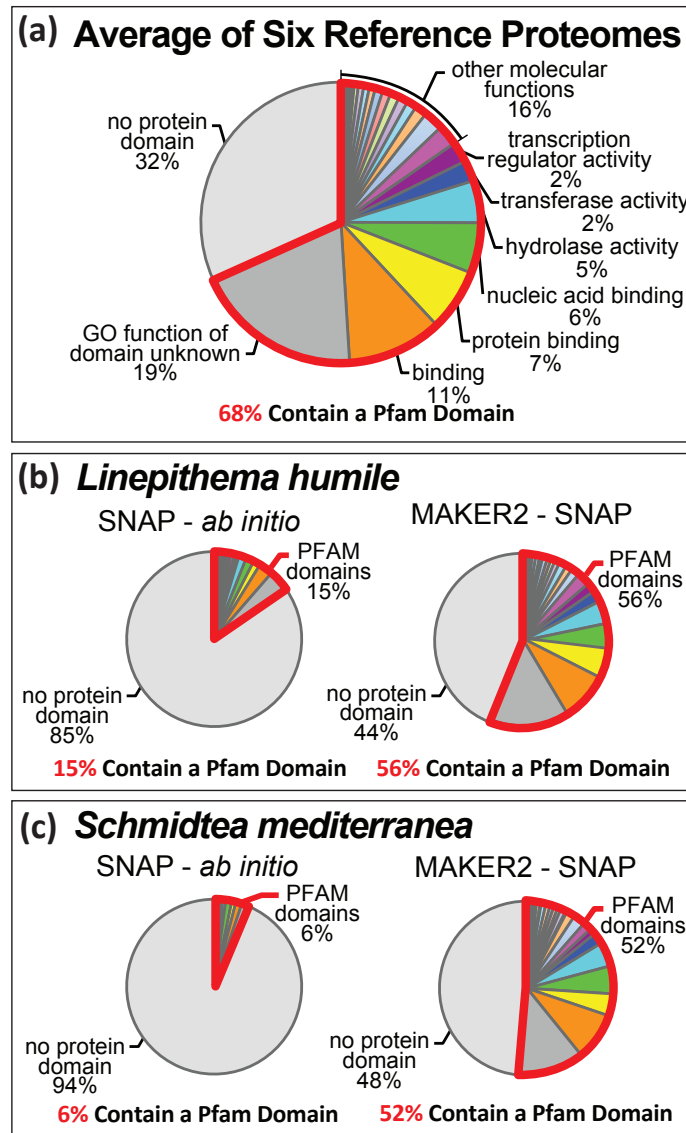


Figure 2.1 MAKER2 vs. *ab initio* predictors on second-generation genomes. We compared the performance of the *ab initio* predictor SNAP to the annotation pipeline MAKER2 on two second-generation genomes: *L. humile* (Argentine ant) and *S. mediterranea* (flatworm). Pfam domain content was used as a means to evaluate the performance of these algorithms, under the assumption that a poorly annotated genome will be globally depleted for domains relative to well-annotated genomes. (A) The average Pfam domain contents for six well annotated eukaryotic reference proteomes: *H. sapiens*, *M. musculus*, *D. melanogaster*, *C. elegans*, *A. thaliana*, and *S. cerevisiae*. These data provide an upper bound for the expected domain content of a newly sequenced genome. The region of the pie chart outlined in red indicates the percentage of genes containing a Pfam domain; these are further subdivided by GO molecular function. (B) The Pfam domain content of SNAP produced *ab initio* predictions compared to MAKER2-SNAP gene annotations for the *L. humile* genome. (C) The Pfam domain content of SNAP *ab initio* gene predictions and MAKER2-SNAP annotations in the *S. mediterranea* genome.

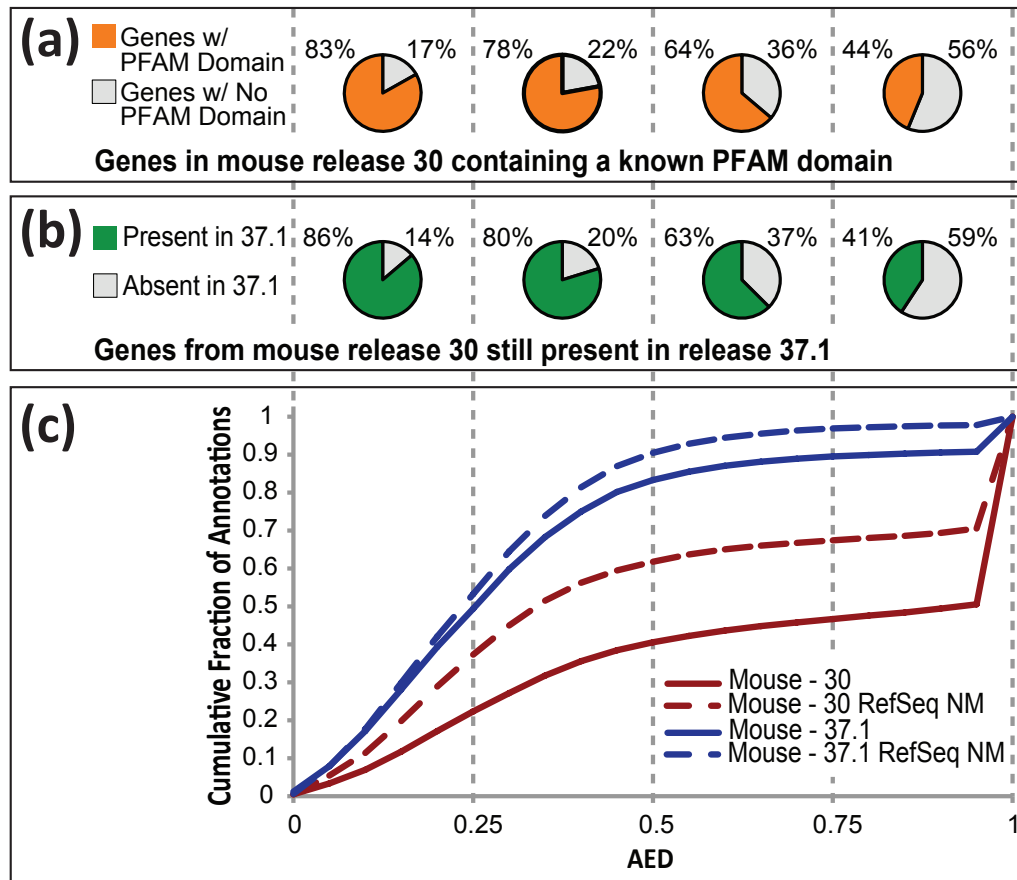


Figure 2.2 Evaluating AED as a metric for annotation quality control.

Annotation Edit Distance (AED) provides a measurement for how well an annotation agrees with overlapping aligned ESTs, mRNA-seq and protein homology data. AED values range from 0 and 1, with 0 denoting perfect agreement of the annotation to aligned evidence, and 1 denoting no evidence support for the annotation. We evaluated the use of AED as a quality control metric by comparing MAKER2 produced AED scores for release 30 (2003) of the *M. musculus* genome to the AEDs for release 37.1 (2007). These data show how AED can be used to quantify improvements to the annotations between each release. (A) The Pfam domain content of *M. musculus* release 30 for genes found in each quartile of the MAKER2 AED distribution. Note that genes with low AEDs are highly enriched for domains. (B) The fraction of *M. musculus* genes from release 30 maintained/removed from subsequent release 37.1 for each MAKER2 AED distribution quartile. These data show how AED mirrors the independent curation decisions made by the mouse research community between 2003 and 2007. (C) The cumulative AED distributions of *M. musculus* release 30 and 37.1 demonstrate how AED quantifies improvements made between releases. The subset of genes with NM prefixes assigned by RefSeq (which indicates the highest level of annotation quality) is plotted separately to show that these independently identified ‘gold-standard’ gene annotations tend to have lower AED values in comparison to the genome as a whole.

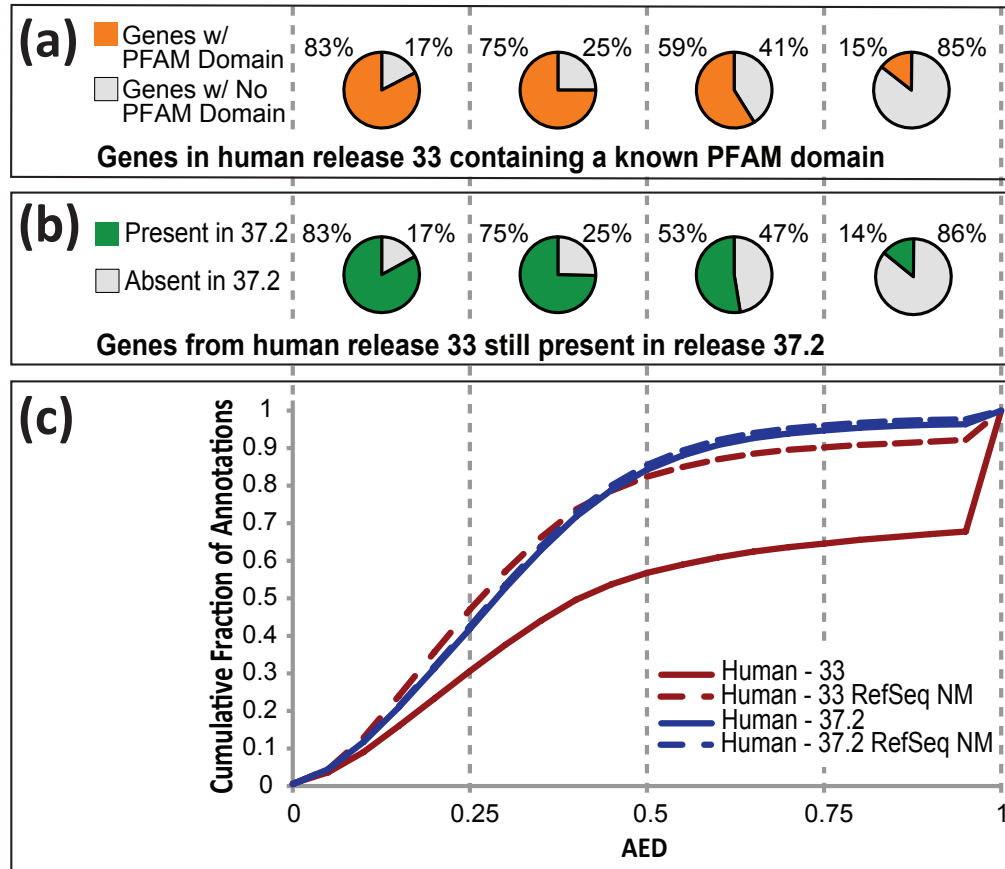


Figure 2.3 AED evaluation of *Homo sapiens* reference annotations.

Annotation Edit Distance (AED) provides a measurement for how well an annotation agrees with its associated evidence, as regards its overlap relative to aligned ESTs, mRNA-seq and protein homology data. AED values range from 0 to 1, with 0 denoting perfect agreement of the annotation to aligned evidence, and 1 denoting no evidence support for an annotation. We evaluated the use of AED as a quality control metric by comparing MAKER2 produced AED scores for release 33 (2003) of the *H. sapiens* genome to the AEDs for release 37.2 (2010). These data show how AED can be used to quantify improvements to annotations between releases. (A) The Pfam domain contents of *H. sapiens* release 33 for genes found in each quartile of the MAKER2 AED distribution. Note that genes with low AEDs are highly enriched for domains. (B) The fraction of *H. sapiens* genes from release 33 maintained/removed from subsequent release 37.2 for each MAKER2 AED distribution quartile. These data show how AED mirrors the independent curation decisions made by the human genetics research community between 2003 and 2010. (C) The cumulative AED distributions of *H. sapiens* release 33 and 37.2 demonstrate how AED quantifies improvements made between releases. The subset of genes with NM prefixes assigned by RefSeq (which indicates the highest level of annotation quality) is plotted separately to show that these independently identified ‘gold-standard’ gene annotations tend to have lower AED values in comparison to all genes as a whole.

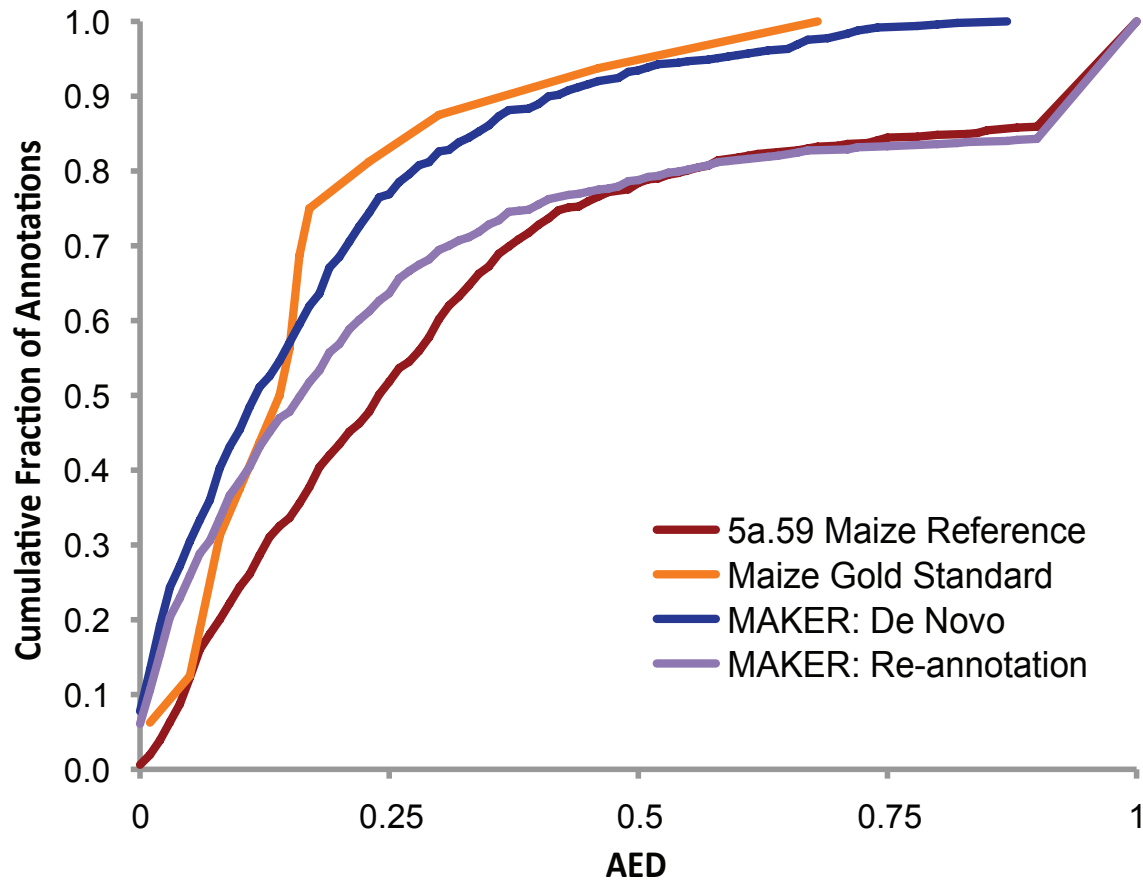


Figure 2.4 Re-annotation of a portion of the maize genome using MAKER2. Annotation Edit Distance (AED) provides a measurement for how well an annotation agrees with its associated evidence (see text and Figure 1 for additional details). Shown are cumulative AED distributions for several maize annotation datasets. Gold curve: AED distribution of high-quality ‘gold standard’ annotations in the benchmark region that are members of the J. Schnable and M. Freeling Maize Classical Genes List; These genes generally have the lowest AEDs. Red curve: all maize gene models from the MaizeSequence.org 5a.59 Working Gene Set in the benchmark region; Blue curve: MAKER2’s first pass, *de novo* annotations for the benchmark region; note that these genes generally have lower AEDs than the 5a.59 Working Gene Set (red curve). Purple curve: automatic MAKER2-based update/revision of the maize 4a.53 Working Gene Set annotations. Note that the revised dataset now exceeds the quality of the 5a.59 Working Gene Set as judged by AED.

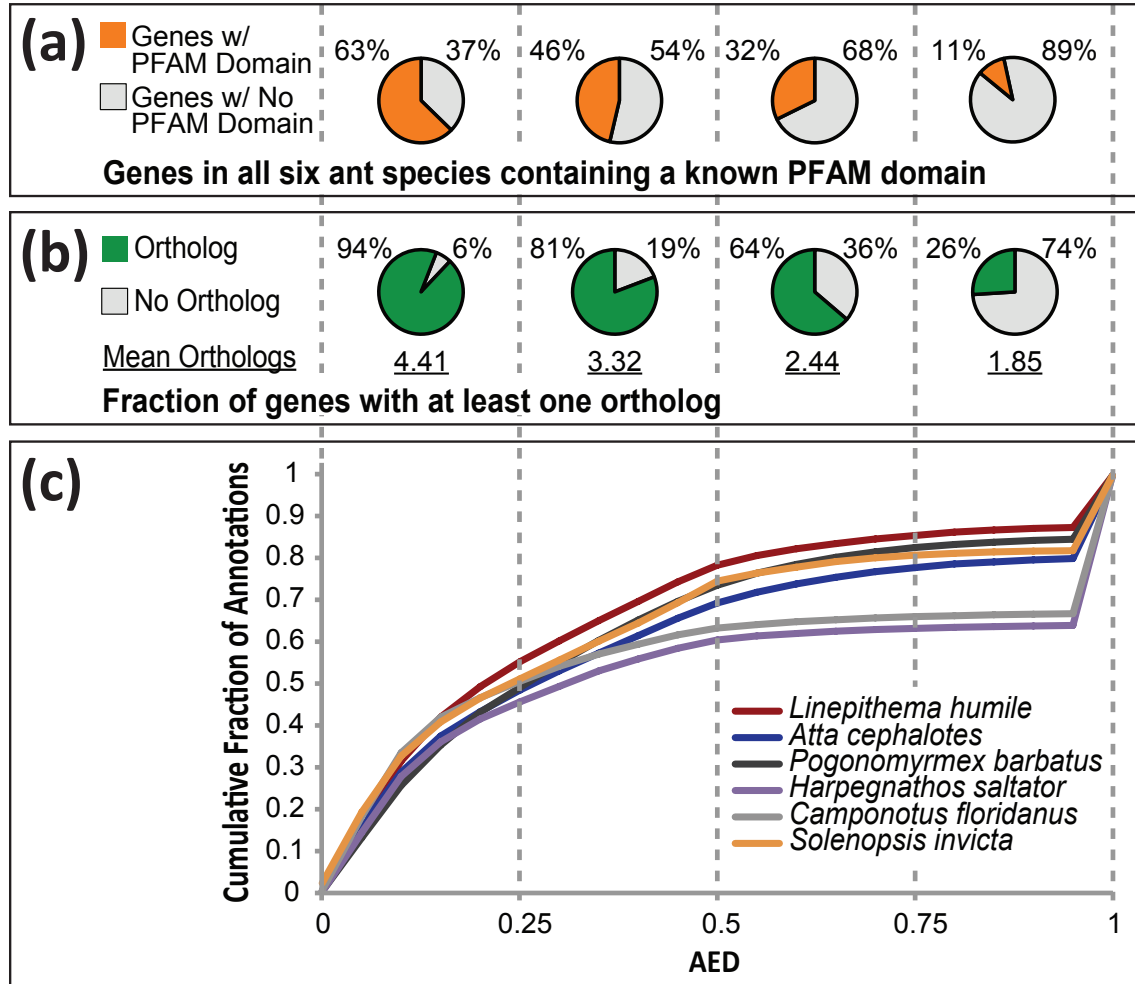


Figure 2.5 MAKER2 as a management tool for existing genome annotations.

MAKER2 was used to add cross species homology evidence and AED values to six published ant species. These data show how MAKER2 can be used both to add new data to existing datasets and for downstream prioritization of genes in those datasets for further analysis and curation. (A) The Pfam domain content in each AED quartile. Genes receiving higher AED scores are less likely to contain a domain, thus prioritizing them as possible false positive gene predictions. (B) The percent of genes in each AED quartile having an orthologous protein in a related ant species with the average number of orthologs per gene (for the subset of orthologous genes) listed at the bottom. AED score is highly correlated with orthology. (C) The cumulative AED distribution for all six ant species. The spike of genes with AED score at or near 1 suggests potential false positive genes predictions rather than species-specific genes, as these annotations also generally lack EST support and Pfam domains; these gene models are first in MAKER2's list for manual review.

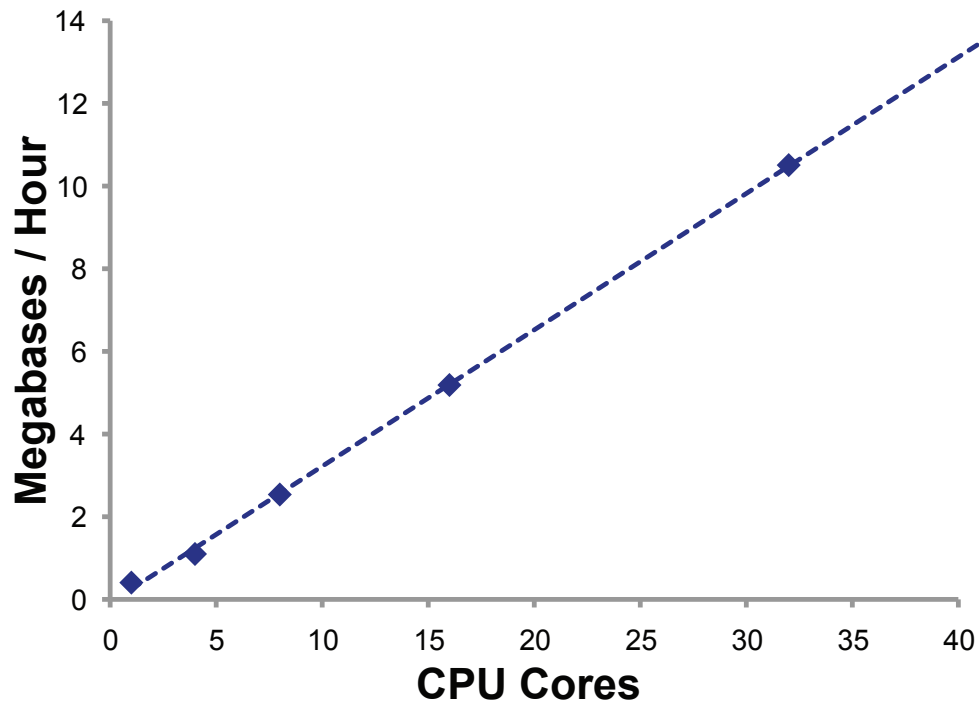


Figure 2.6 MAKER2 scales to even the largest genomes

MAKER2 was used to annotate a 10 megabase section of the *C. elegans* genome (NGASP dataset). The algorithm was parallelized using MPI on an increasing number of CPU cores. This demonstrates how MAKER2 scales almost linearly with CPU number (with a slope of near 1). If we project our results forward to the entire *C. elegans* genome (~100 megabases), MAKER2 should take under 10 hours on 32 CPUs to complete; similarly, the human genome (~3 gigabases) would require fewer than 24 hours on 400 CPUs.

CHAPTER 3

THE PROOF-OF-PRINCIPLE ANNOTATION OF EIGHT EMERGING MODEL ORGANISM GENOMES

Introduction

Genome databases and the annotations they contain are invaluable resources that facilitate research and discovery. While MAKER2 and related tools developed for MAKER2 do not test hypotheses directly, they facilitate and promote the broad acceleration of research by producing feature-rich genome annotations that act as substrates for further computational and experimental work.

Here, I describe the proof-of-principle annotation of eight emerging model organism genomes using MAKER2, thus contributing critically to the understanding of their evolutionary history and functional biology. The annotation of these genomes was done in collaboration with their respective genome projects, and final genome annotations are all publicly available through those projects.

Results and discussion

Annotation of the necrotic plant pathogen *Pythium ultimum*

Pythium ultimum var ultimum is a plant pathogen and a member of the class Oomycota (also referred to as Oomycetes). Oomycetes were at one time believed to be fungi because of their similar morphology and lifestyle; however, based on molecular

evidence, these organisms are now classified as part of the separate kingdom Chromalveolata, which is equidistant evolutionarily to animals, fungi, and plants[1]. Because of their unique position on the tree of life, the study of Oomycetes has the potential to reveal important insights into the history of eukaryotic evolution.

The *Pythium ultimum* genome is not the first Oomycete to be sequenced. In fact there has been significant interest in studying this group of organisms primarily because they represent major pathogens affecting agriculturally important crops. *Phytophthora infestans* for example was the cause of the Irish potato famine, and was among the first Oomycete genomes to be sequenced[68]. Annotation databases for the Oomycetes *Phytophthora sojae*[2] and *Phytophthora ramorum*[2] also exist.

One of the most interesting features of parasitic Oomycetes (and the primary focus of research in these organisms) is their use of secreted effector molecules which manipulate gene expression of a host to facilitate infection[3]. Study of these molecules may help elucidate the mechanisms used by Oomycetes for host infection and could potentially provide new solutions to protect agriculturally important crops from these pathogens. *Pythium ultimum* is basal evolutionarily to other sequenced Oomycetes (all of which belong to the phylum Phytophthora); therefore study of the *Pythium ultimum* genome has the potential to reveal trends in the evolution of this class of organisms especially with respect to their repertoire of secreted effector proteins.

I used the program MAKER2 to annotate the *Pythium ultimum* genome, and produce a final annotation dataset consisting of 15,323 genes[4]. To evaluate the quality of the gene models, I compared the relative domain content of the *Pythium ultimum* proteome to that of other annotated Oomycetes and eukaryotic organisms (Figures 3.1 and 3.2). Previously, I have shown that a comparison of protein domain

content across organisms serves as a proxy metric for gene annotation quality (See the section in Chapter 2 titled ‘Gene prediction/annotation in second-generation genomes’). When comparing *Pythium ultimum* protein domain content to other organisms, I used InterPro domains (as opposed to Pfam domains) because InterPro is a comprehensive dataset and provides great sensitivity for detecting divergent protein domains.

In the analysis of the *Pythium ultimum* proteome, 60% of genes contained at least one known InterPro protein domains (see Figure 3.1). In comparison the proteomes for *Phytophthora sojae*, *Phytophthora ramorum*, and *Phytophthora infestans* (all Oomycetes) contained 61%, 66%, and 55% domain enrichments, respectively (Figure 3.1). The relative similarities in protein domain content among related Oomycetes reflect well on the confidence associated with the *Pythium ultimum* gene annotations. InterPro domain enrichments for the proteomes of *Aspergillus nidulans* (fungi), *Saccharomyces cerevisiae* (fungi), and *Arabidopsis thaliana* (plant) in comparison are 73%, 81%, and 79%, respectively (Figure 3.2). The fact that other eukaryotes have higher domain enrichments in comparison to Oomycetes is not surprising because of the evolutionary isolation of Oomycota.

Global analysis of the intron/exon structure in the *Pythium ultimum* genome reveals many interesting trends in Oomycete evolution. Exons in the *Pythium ultimum* genome tend to be relatively long when compared to other eukaryotes (they have an average length of 498 bp), and many genes are encoded by single-exon transcripts with extremely long open reading frames (the longest being 21,357 bp in length). Organisms like *Drosophila melanogaster* (animal) and *Arabidopsis thaliana* (plant) in comparison have average exon length of 380 bp and 234 bp, respectively. While there are examples of intron-rich genes in *Pythium ultimum*, the majority of

genes tend to have few introns (with an average 1.6 introns occurring per gene). Eukaryotes like *D. melanogaster* (animal) and *A. thaliana* (plant) in comparison contain an average of 5.1 and 5.4 introns per gene, respectively. Introns in *Pythium ultimum* also tend to be short (the average intron being 115 bp in length). Gene structure is thus similar to other Oomycetes like *Phytophthora infestans* (where the average intron is 124 bp in length, and the average gene contains 1.7 introns).

Interestingly, homologs of large single exon *Pythium ultimum* genes tend to be intron rich in other eukaryotes. One example is midasin, a highly conserved ~600 kDa nuclear chaperone protein. The *Pythium ultimum* ortholog of this gene is intronless; yet the same gene contains 72 introns in *A. thaliana* and 101 introns in *Mus musculus* (Figure 3.3). In *Saccharomyces cerevisiae*, however, this gene is encoded by a single exon, a striking similarity to *Pythium ultimum* considering the large evolutionary distance between fungi and Oomycetes.

The tendency of *Pythium ultimum* genes to have few introns that are also short in length may be the result of evolutionary selection for quick production of proteins (since, in theory, short transcripts that undergo little to no splicing could be rapidly processed for translation). Consistent with this hypothesis, generation time has previously been shown to strongly correlate with intron length in many organisms[5]. The fact that large, single exon *Pythium ultimum* genes also tend to be single exon in fungi (despite the long evolutionary distance between these organisms) suggests that similar intron/exon structure may be a convergent adaptation related to a shared life history.

In Oomycetes, the secretome (a collection of all secreted proteins coded by the genome) is a very active area of research due to the importance of these proteins in manipulating a host's response to infection by these parasites. In our analysis, 2,908

proteins were identified as potentially secreted based on SignalP[6] detection of a signal peptide motif in the N-terminus of each protein (SignalP is a signal peptide domain detection program and was run within InterProScan as part of the InterPro domain analyses). While the presence of the signal peptide motif is not exclusive to secreted effector molecules, this list of proteins provides a potentially rich pool of candidates for future research into how *Pythium ultimum* interacts with its host during infection.

Protein domain analyses also identified a number of *Pythium ultimum* DNA methylases. Interestingly, the Oomycete *Phytophthora infestans* lacks DNA methylase genes, the absence of which is believed to be responsible for repeat element expansion within that organism. Repeats make up more than 50% of the *Phytophthora infestans* genome[7]. By comparison, identified repeat elements make up only 7% of the *Pythium ultimum* genome.

Annotation of BAC clones from the loblolly pine *Pinus taeda*

Commonly known as the Loblolly Pine, *Pinus taeda* is a large coniferous tree native to the Southern United States. Because of the large size of the genome (~20-30 gigabases), only a small sample of the genome had been cloned and sequenced at the time of our analysis (~1 megabases). I provided draft annotations and summary statistics for *P. taeda* using MAKER2. I also performed in-depth analysis of the repeat content of the sequence as an estimate for the genome as a whole.

Because of the small amount of sequence available, I was unable to train gene-predictors such as SNAP for *P. taeda*. However, I was able to provide rough draft annotations by using a combination of monocot and dicot parameters from other plant species. In previous analyses, I have shown that evidence informed

MAKER2 annotations using incorrect species parameter files can achieve accuracies comparable to those of *ab initio* gene predictions using correct species parameters (See Table 2.3 in Chapter 2). In total, MAKER2 produced 18 gene models with an average length of 1,178 bp and 2.8 exons per gene under monocot parameters; similarly, MAKER2 produced 18 genes with an average length of 1401 bp and 2.9 exons per gene under dicot parameters[8]. When scanned for InterPro domains, 12 genes contained a known domain, or 67% of genes. However, incomplete protein homology alignments and a lack of consensus start codons suggest that many of these may be pseudogenes, which would not be unexpected in a genome this large (~20-30 gigabases). The presence of pseudo-genes could suggest a historical round of genome duplication or even polyploidy. However, previous analyses have found no evidence supporting whole-genome duplication[9].

Sequencing efforts in the *P. taeda* genome are still underway, and since the MAKER2 analysis of the genome, an additional ~11 megabases of high quality sequence has become available (providing ~12 megabases total). This is enough sequence to train the SNAP *ab initio* gene-predictor specifically for *P. taeda*. I explore annotation of this additional sequence in Chapter 5.

MAKER2's ability to annotate a selection of the *P. taeda* genome demonstrates the feasibility of annotating the entire ~20-30 gigabase genome sometime in the future. It also demonstrates the power of MAKER2 to extract useful information about gene structure, even when given extremely limited datasets for analysis. As more genome sequence becomes available, it will become possible to explore more definitively whether *P. taeda* has indeed undergone ancestral genome duplications as suggested by what appear to be an abundance of pseudogenes.

Annotation of the pitch canker fungus *Fusarium circunatum*

Fusarium circunatum is a fungal pathogen of pine forests, with great ecological as well as economic significance to the forest industry. This genome project was one of a number of projects promoted by the South African government in part to help develop genomics and bioinformatics expertise in that country, and *F. circunatum* is the first eukaryotic genome to be sequenced in Africa.

MAKER2 was used in this genome project to provide high-quality genome annotations. Because one of the primary goals of this project is the generation of bioinformatics expertise within South Africa, I also integrated MAKER2's annotations into the GMOD community annotation system (CAS). In the CAS system, genome annotations are loaded into a Chado[10] database. Researchers then use the program Apollo[11] to access and edit all gene models live, via a remote connection. From there, researchers can either save back to the database or to a local GFF3 file. Saving to the database means all researchers involved in a genome project have instant access to the most up to date annotation set at all times, and changes made to a gene model are updated live for everyone. Using CAS, the *F. circunatum* community was able to provide at least some degree of manual curation to every gene in the genome, a major feat of human parallelization.

The final MAKER2-plus-manual-curation dataset contained a total of 14,973 genes, 67% of which contained known InterPro domains (based on InterProScan analysis). In comparison the fungal proteomes for *Neurospora crassa*, *A. nidulans*, and *S. cerevisiae* have InterPro domain enrichments of 62%, 73%, and 81%, respectively (Figure 3.2). The *F. circunatum* domain enrichment is therefore comparable to what is seen in other annotated fungal genomes. Further analysis of the *F. circunatum* genome is ongoing by the research community and will

undoubtedly provide important insights into the organism's evolution as well as mechanisms it uses for pathogenesis.

Annotation of the red harvester ant *Pogonomyrmex barbatus*

Pogonomyrmex barbatus (the red harvester ant) is a species of ant found in the deserts of North America and is one of several ant species to be recently sequenced. Having a strict social order with biologically defined castes, study of *P. barbatus* could reveal important insights into the evolution of social behavior in these insects, including the pathways of gene expression underlying caste determination. With the additional sequencing of the related organisms *Apis mellifera* (honey bee)[12] and *Nasonia vitripennis* (wasp)[13], comparative analysis of the *P. barbatus* genome annotations could help elucidate the mechanisms of the adaptive radiation of Hymenoptera species (the order of insects that includes bees, wasps, and ants).

MAKER2 was used to provide draft annotations for the *P. barbatus* genome. Gene annotations were then integrated into the GMOD community annotation system (CAS). This permits programs such as Apollo and GBrowse to remotely access a Chado database to rapidly distribute and curate the gene annotations live over the Internet, thus allowing for strong integration of the wider research community into the project.

In total, the MAKER2-plus-manually-curation dataset contained 17,101 genes (OGS 1.2), with more than 1,000 genes receiving some form of manual curation[30]. All final annotations were deposited in the Hymenoptera Genome Database[14] for distribution to the wider research community. The overall gene count for *P. barbatus* was similar to that of other published ant species (*Solenopsis*

invicta with 16,611 genes (v2.2.0)[31], *Harpegnathos saltator* with 18,564 genes (v3.3)[15], and *Camponotus floridanus* with 17,064 genes (v3.3)[15]. Quality control analysis showed InterPro domain enrichment of 53% in *P. barbatus*; this is also comparable to other published ant species (*S. invicta* with 53%, *H. saltator* with 49%, and *C. floridanus* with 54%). The similarity of InterPro domain enrichments among all ant species reflects well on the quality of the *P. barbatus* genome annotations.

Annotation of the Argentine ant *Linepithema humile*

Linepithema humile (the Argentine ant) is a species of ant native to Argentina that has invaded regions in the United States and other countries on six continents (devastating many native species). Research into this organism is important for understanding not only its biology and evolution, but also possible mechanisms for controlling the expansion of *L. humile*.

MAKER2 was used to annotate the *L. humile* genome in conjunction with manual curation and annotation using the GMOD community annotation system (CAS). In total, 16,049 gene models were produced for *L. humile* (OGS 1.2)[29] – a number comparable to that of other ant species (see results for *P. barbatus* annotation above). Additionally, 58% of annotations contained a known InterPro domain, which bodes well for annotation quality in this organism and is the highest domain enrichment among published ant species (see results for *P. barbatus* annotation above).

The final genome annotations for *L. humile* are available through the Hymenoptera Genome Database[14]. The MAKER2 produced annotations for *L. humile* provide a resource not only for exploring questions such as the mechanisms

of sociality in ants, but also may provide ways to counter the ecological damage caused by this organism.

Annotation of the leaf-cutter ant *Atta cephalotes*

Atta cephalotes, more commonly referred to as the leaf-cutter ant, is a species of ant found in the rain forests of Latin America. They are known to harvest leaves that are then use to cultivate fungus. The fungus serves as the colony's primary food source. This behavior is one of the few occurrences of farming exhibited outside of humans. Study of the *A. cephalotes* genome may reveal not only unique insights into the sociality of ants, but also into the effects of the selection imposed by obligate ant-fungus mutualism on the evolution of this organism.

The *A. cephalotes* genome was annotated using MAKER2 together with manual curation via the GMOD community annotation system (CAS). The final annotation set consisted of 18,062 gene models (OGS 1.2)[27], with 52% of genes containing at least one known InterPro domain. This domain content is similar to what is seen in other ant species – which range from 49-58% domain enrichment (see results above for *P. barbatus* and *L. humile* annotations). Gene annotations are publicly available through the Hymenoptera Genome Database[14].

Annotation of the sea lamprey *Petromyzon marinus*

Petromyzon marinus is a species of lamprey, which is a jawless fish. Lampreys exist in mostly coastal and fresh waters. They are also found in, but are not native to, the Great Lakes. The invasion of the Great Lakes by these organisms has had both profound economic and ecological impacts on the region. Lampreys occupy an interesting position on the vertebrate evolutionary tree, just prior to the

evolution of the jaw. Analysis of the lamprey genome can therefore provide valuable insights into vertebrate evolution. Many characteristics of these organisms, such as their unique immune system (with parts unrelated to antibodies found in higher vertebrates)[16] and their ability to regenerate their spinal cord[17, 18], make them important subjects for continued research.

MAKER2 was used to provide an initial set of draft annotations for *P. marinus*, whose genome is comparable in size to the human genome (the lamprey genome contains an estimated 2 gigabases of sequence). The genome assembly for *P. marinus* is relatively fragmented (consisting of ~25,000 super-contigs), which makes this organism a relatively difficult substrate for annotation. We supplemented MAKER2's input evidence dataset with mRNA-seq to help offset the effects of genome fragmentation by increasing sensitivity for partial gene models. MAKER2 produced an annotation set consisting of 24,132 gene models. This number of gene annotations is likely inflated relative to the true gene count for this organism; but this is to be expected given the fragmented nature of the genome assembly, as many genes are likely split across contigs. An analysis of InterPro domain content for quality control purposes showed a 54% enrichment for domain content in this organism, a percentage similar to what is seen in other eukaryotic genomes (see Figures 3.1 and 3.2) which bodes well for the confidence in the current annotation set (especially considering that many genes were likely too fragmented by assembly issues to recover entire domains).

A major point of interest in studying the lamprey genome is the possible existence of whole-genome duplications in the lineage leading to lamprey as well as other vertebrates (commonly referred to as “the vertebrate two genome duplication hypothesis”)[19]. According to this hypothesis, a series of two or more genome

duplications occurred sometime early in vertebrate evolution and was one of the primary mechanisms for rapid expansion and diversification of vertebrate lineages. While there is evidence of these duplications in many vertebrate species, the timing of the duplication events is in dispute[20, 21]. The original proposal hypothesized that these duplications occurred in the stem leading to both lancelets and vertebrates; however, recent work suggests that these duplications may only have occurred in vertebrate lineages[22]. Analysis of the newly published amphioxus (lancelet) genome suggests that rounds of whole-genome duplication may have occurred both before and after the split between jawless vertebrates (i.e. lampreys) and jawed vertebrates[23]. The position of *P. marinus* on the evolutionary tree should allow researchers to address the question of whether duplication occurred before or after lamprey divergence by using both the genome sequence and its annotations. While the fragmented nature of the genome makes this type of analysis somewhat difficult (as synteny is not always apparent), the MAKER2 produced annotation database will provide a valuable resource to help address this and other questions (for example, preliminary results indicate the presence of two HOX clusters). Further analysis of this genome is ongoing by the lamprey research community.

Maximizing annotation sensitivity in *Schmidtea mediterranea*

Schmidtea mediterranea is a model system for the study of stem cells, regeneration, and tissue homeostasis[24, 25]. It is a protostome and a free-living species of Platyhelminthes (flatworms), an understudied yet evolutionarily important phylum. While flatworms are among the simplest bilaterally symmetric animals, they exhibit many features that make them important for research. They

are acoelomates yet they possess derivatives of all three germ layers organized into complex organ systems. They have one of the most primitive centralized nervous systems, and many species can regenerate complete individuals from only small fragments of tissue[26, 27]. Research of *S. mediterranea* benefits from the fact that these organisms are flat and nearly translucent, making them particularly easy to study in gene expression, gene knock-out, and gene knock-in type experiments (since effects in all organs and tissues can be directly visualized under a microscope without the need for dissection). Annotation of the *S. mediterranea* genome is providing researchers with the knowledgebase necessary to design these types of gene manipulation experiments, so we can better understand the mechanisms underlying regeneration and stem cell development in this organism.

MAKER2 was used to annotate the ~800 megabase *S. mediterranea* genome. The genome assembly for this organism was particularly fragmented, consisting of 43,295 contigs, the longest of which is just over 670,000 base pairs in length. The high level of fragmentation makes *S. mediterranea*, a particularly difficult substrate for genome annotation, and overall gene counts are expected to be inflated with many genes being split across contigs.

Because of the difficult nature of this genome, I had to explore new strategies to maximize MAKER2's ability to identify partial and fragmented gene models. This was done primarily by integrated short mRNA-seq data into the MAKER2 analysis. The use of mRNA-seq data should increase the sensitivity of the MAKER2 pipeline as any assembly fragment containing even a piece of a gene should have significant overlap of mRNA-seq reads. Unfortunately this strategy also has the potential to decrease gene prediction specificity (i.e., increase gene over-prediction and false

positives), especially since repetitive sequence can create regions of false homology. Repeat identification is therefore given a great degree of priority in my analysis.

An *S. mediterranea* specific repeat library was prepared by using the RepBase[28] repeat library together with the RepeatRunner[29] transposable element protein library to identify known repetitive elements. The program PILER[30] was then used to identify novel repeats specific to *S. mediterranea*. I classified the dataset of novel species-specific repeats based on comparison to conserved known repeats in RepBase (see Methods section below). This species-specific dataset was then combined with the RepBase and RepeatRunner libraries to mask out all identified repetitive elements in the genome.

Analysis of the repeats showed that the *S. mediterranea* genome is A/T rich, with low-complexity regions comprising 22% of all sequence, and high-complexity dispersed repeats comprising another 5.38%. The high-complexity repeats were largely composed of Helitrons, LINE, and viral sequences.

Following preparation of a repeat library, short-read mRNA-seq alignments were processed using TopHat[31] and Cufflinks[32] and then passed to MAKER2, resulting in 38,924 gene models. Examination of the position of these gene models indicates that over 27,000 of these occur at the edge of contigs suggesting that as much as 70% of all annotations may be partial or split. When I looked at the InterPro domain content of these genes, a total of 53% contained known domains. The InterPro domain enrichment is comparable to what is seen in other eukaryotic genomes (Figures 3.1 and 3.2), which is surprising given that many genes appear to be partial and fragmented (partial models make it more difficult to capture enough sequence for domain identification). The domain enrichment thus provides a degree of confidence that the high gene count is not just the result of an abundance of false

positive gene predictions. The 53% InterPro domain enrichment seen in *S. mediterranea* is also comparable to the 54% seen in the similarly fragmented *P. marinus* genome. Overall domain content in both these organisms would be expected to improve with more genome sequence and assembly, making them excellent targets for future re-annotation (another analysis that is also facilitated by MAKER2).

Of great interest to researchers studying *S. mediterranea* is its phylogenetic relationship to other organisms, including humans. Where exactly flatworms fit into the tree of life is still somewhat controversial. While Platyhelminthes are classified as protostomes, it is still debatable as to whether they belong to the Ecdysozoa (in general protostomes that molt their exoskeleton) or the Lophotrochozoa (most have a feeding appendage bearing hollow tentacles). Using the whole genome annotation set, I performed phylogenetic analysis in comparison to other eukaryotic genomes. The resulting phylogenetic tree firmly places *S. mediterranea* in the Lophotrochozoa with bootstrap support of 98/100 (Figure 3.4).

Additionally, a comparison of total genes shared with other organisms (Figure 3.5) reveals that *S. mediterranea* shares more genes in common with human (5,390 genes) and mouse (5,467 genes) than it does with fellow protostomes *C. elegans* (4,411 genes) and *D. melanogaster* (4,617 genes) – both are Ecdysozoans. This data suggest that *S. mediterranea* in many ways is more like the ancestral Metazoan than certain other protostomes. These results are largely consistent with previous findings that gene loss has been one of the primary architects of metazoan evolution (with Ecdysozoans experiencing the greatest gene loss in comparison to deuterostomes and Lophotrochozoans)[33]. Because of the greater gene overlap with deuterostomes, *S. mediterranea* may in fact be a better model organism for studying

certain human biological processes than classic Ecdysozoan systems like *C. elegans* or *D. melanogaster*. All genome annotations, for *S. mediterranea* are available online at the SmedGD[34] online database.

Conclusions

The successful annotation of eight phylogenetically diverse emerging model organism genomes of varying genetic structure and assembly quality demonstrates MAKER2's utility for producing database ready genome annotations for virtually any eukaryotic genome. The easy-to-use design of MAKER2 means it is especially suited for projects that may have limited resources and lack bioinformatics experience (e.g., the *F. circunatum* genome, a project to develop bioinformatics expertise in South Africa). The performance of the algorithm on the evolutionarily isolated Oomycete, *Pythium ultimum*, demonstrates MAKER2's capability to annotate exotic/divergent genomes. The *S. mediterranea* and *Petromyzon marinus* projects highlight MAKER2's ability to work on highly fragmented genomes, which are generally difficult substrates for annotation. Additionally, genome annotation of the Loblolly pine, *Pinus taeda*, where only ~1 megabase of sequence was initially available shows how MAKER2 can still produce usable, biologically relevant results even when available training data for *ab initio* gene-prediction is extremely limited or non-existent. The high-throughput parallelization capability of MAKER2 may also provide the only reasonable solution for future annotation of mega-genomes genomes like pine, which is greater than 20 gigabases in length. Finally, the ability to directly integrate MAKER2's output into GMOD tools for downstream analysis has been shown time and time again to be extremely valuable. This is especially true of the GMOD community annotation system (CAS), which allowed our collaborators

on the *Pogonomyrmex barbatus*, *L. humile*, and *A. cephalotes* genome projects to directly access and edit MAKER2 annotations over the internet; thus better integrating the research community and producing a major feat of human parallelization to accelerate downstream analyses. In fact, the efficiency derived from integrating MAKER2's output into this system allowed the ant genome projects to advance from annotation to manuscript submission in only 3 months, a monumental feat considering that these collaborative projects involved many researchers spread over several continents.

Methods

Pythium ultimum genome annotation

MAKER2 was configured to use both spliced EST alignments as well as single-exon ESTs greater than 250 bp in length as evidence for producing evidence-based gene predictions. MAKER2 was also set to filter out gene models for short and partial gene predictions that produce proteins with fewer than 28 amino acids. This value was selected because the smallest predicted protein in the related Oomycete *Phytophthora infestans* was 30 amino acids in length. The MAKER2 pipeline was set to produce *ab initio* gene predictions from both the repeat-masked and unmasked genomic sequence using SNAP[35] (trained using CEGMA[36]), FGENESH[37] (using existing *Phytophthora* parameter file), and GeneMark[38] (self-trained).

The EST sequences used in the annotation process were derived from Sanger and 454 sequenced *Pythium ultimum* BR144 ESTs[39] considered together with ESTs from dbEST[40] for *Aphanomyces cochlioides*, *Phytophthora brassicae*, *Phytophthora capsici*, *Phytophthora infestans*, *Phytophthora parasitica*, *Phytophthora sojae*, and *Pythium oligandrum*. Protein evidence was derived from

the UniProt/Swiss-Prot[41, 42] protein database and from predicted proteins for *Phytophthora infestans*, *Phytophthora ramorum*, and *Phytophthora sojae*. MAKER2 was also provided with a *Pythium ultimum* species-specific repeat library prepared for this work (created using PILER with no downstream classification).

Both annotations and *ab initio* gene predictions not overlapping a MAKER2 annotation were scanned for Pfam[43] and InterPro[44] protein domains using InterProScan[45]. All nonoverlapping predictions containing a domain were added to the annotation set.

The MAKER2 produced gene annotation set was then submitted to the *Pythium ultimum* community for manual curation using the annotation editing tool, Apollo[11]. Manual annotations from Apollo were saved in GFF3[46] format and passed back through MAKER2 via the internal GFF3-passthrough option to both standardize the annotations and to calculate quality metrics for each gene model.

Finally, putative functions were assigned to each annotated *Pythium ultimum* protein using BLASTP[47] to identify the best homologs from the UniProt/Swiss-Prot protein database. The functions of each best BLAST hit were then mapped to the corresponding *Pythium ultimum* protein. Individual researchers also assigned some putative protein functions during manual curation.

Pinus taeda genome annotation

The EST/cDNA sequences used by MAKER2 were derived from *P. taeda* as part of the sequencing project[8] and were combined with EST/cDNA sequences from all other Pinaceae species found in dbEST. The UniProt/Swiss-Prot protein database and all proteins from *Arabidopsis thaliana*[48] were used as the protein homology dataset for the MAKER2 run. Repeat elements were identified using MAKER2

default settings and pre-computed repeats from the program CENSOR[49] (passed to MAKER2 via the algorithms GFF3-passthrough option).

MAKER2 was first run using the *ab initio* gene prediction algorithms SNAP, Augustus[50], and GeneMark (all trained for *Arabidopsis thaliana*) together with FGENESH (trained with generic dicot parameters). The second run of MAKER2 was performed using SNAP and GeneMark (both trained for *Oryza sativa* – rice) in conjunction with Augustus (trained for *Zea mays* – corn). Summary statistic for intron/exon structure and transcript lengths were calculated using the program Eval[51].

Fusarium circunatum genome annotation

MAKER2 was configured with an EST set prepared for the genome project together with all *F. circunatum* ESTs available from dbEST. The protein homology set consisted of all proteins from the following genomes: *Aspergillus niger*, *Batrachochytrium dendrobatidis*, *Candida albicans*, *Coccidioides immitis*, *Cryptococcus neoformans*, *Fusarium graminearum*, *Fusarium oxysporum*, *Fusarium verticillioides*, *Laccaria bicolor*, *Magnaporthe grisea*, *Monosiga brevicollis*, *Nectria haematococca*, *Neurospora crassa*, *Phycomyces blakesleeanus*, *Pichia stipitis*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Sclerotinia sclerotiorum*, *Sporobolomyces roseus*, *Stagonospora nodorum*, *Trichoderma atroviride*, *Trichoderma reesei*, *Trichoderma virens*, and *Ustilago maydis*. These proteins were combined with the UniProt/Swiss-Prot database, and all *Fusarium* proteins in GenBank[52]. Repeat masking options were set to default values in MAKER2. Annotations and nonoverlapping gene predictions were scanned for InterPro and

Pfam domains using InterProScan, and all nonoverlapping predictions with a domain were added to the final annotation set.

Resulting annotations were loaded into a Chado[10] database and then accessed using Apollo. Some contigs were saved back to GFF3 format, in which case they were merged back into the formal annotation dataset by running MAKER2 with the `model_gff` options set to the location of the manually curated GFF3 file.

Pogonomyrmex barbatus genome annotation

MAKER2 was provided with a species-specific repeat library consisting of novel repeats identified by RepeatModeler[53] and PILER[30] in both the *P. barbatus* and *L. humile* genomes. For the protein evidence, proteins from the UniProt/Swiss-Prot database, *D. melanogaster*, *A. mellifera*, *N. vitripennis*, and insect chemosensory proteins from GenBank were combined into a single dataset. The *P. barbatus* EST dataset consisted of ESTs sequenced as part of the genome project together with Hymenopteran and *L. humile* ESTs from dbEST.

Prior to running MAKER2, we independently trained the *ab initio* predictors SNAP (trained with CEGMA), Augustus (trained with packaged training script `autoAug.pl`), and GeneMark (self-training).

Following automatic annotation, we ran InterProScan over the set of MAKER2 annotations as well as *ab initio* gene predictions that did not overlap a MAKER2 annotation. Nonoverlapping gene predictions that contained a domain were then added to the final annotation set. The MAKER2-generated annotations were also subjected to further human review and curation (see ‘Community Annotation System’ below).

Linepithema humile genome annotation

The *L. humile* EST dataset consisted of ESTs sequenced as part of the genome project together with all Hymenopteran and *L. humile* ESTs from dbEST. For the protein evidence, the UniProt/Swiss-Prot database, *D. melanogaster*, *A. mellifera*, *N. vitripennis*, and insect chemosensory proteins from GenBank were combined.

Prior to running MAKER2, we independently trained the *ab initio* predictors SNAP (trained using CEGMA), Augustus (trained with the packaged autoAug.pl script), and GeneMark (self-trained). MAKER2 was also provided with a species-specific repeat library consisting of novel repeats identified by RepeatModeler and PILER in both the *P. barbatus* and *L. humile* genomes.

Following automatic annotation, we ran InterProScan over the set of MAKER2 annotations as well as the set of non-overlapping *ab initio gene* predictions (do not overlap an annotation). Nonoverlapping gene predictions that contained a domain were added to the final annotation set. The MAKER2-generated annotations were then subjected to further human review and curation (see ‘Community Annotation System’ below).

Atta cephalotes genome annotation

The EST dataset for *A. cephalotes* consisted of an EST dataset sequenced specifically for that genome together with Hymenopteran and *A. cephalotes* ESTs from dbEST. The protein dataset was identical to that used for *P. barbatus* and *L. humile* annotation, as was the repeat library. Prior to running MAKER2, we trained the *ab initio* predictors SNAP (used CEGMA training), Augustus (trained with the packaged autoAug.pl script), and GeneMark (self-trained).

Following automatic annotation, we ran InterProScan over the set of MAKER2 annotations as well as the set of non-overlapping *ab initio* gene predictions. Nonoverlapping gene predictions that contained a domain were added to the final annotation set. The MAKER2-generated annotations were then subjected to further human review and curation (see ‘Community Annotation System’ below).

Petromyzon marinus genome annotation

Inputs for MAKER2 included the *P. marinus* genome assembly, *P. marinus* ESTs sequenced for this project, and a protein databases containing all annotated proteins for human, mouse, chicken, *D. melanogaster*, *C. elegans*, *Xenopus tropicalis*, *Strongylocentrotus purpuratus* (sea urchin), *Branchiostoma floridae* (lancelet), *Lottia gigantea* (sea snail), *Ciona intestinalis* (sea squirt), *Trichoplax adhaerens*, *Nematostella vectensis* (sea anemone), *Danio rerio* (zebra fish), and *Takifugu rubripes* (puffer fish) combined with the UniProt/Swiss-Prot protein database and all sequences for Chondrichthyes (cartilaginous fish) and Myxinidae (hagfishes) in the NCBI protein database. *Ab initio* gene predictions were produced inside of MAKER2 by the programs SNAP and Augustus. MAKER2 was also passed *P. marinus* mRNA-seq data processed by the programs TopHat[31] and Cufflinks[54]. MAKER2 was run in a bootstrap-fashion, with the output gene models of one run acting as inputs for retraining *ab initio* gene-predictors, thus better informing mRNA-seq alignment junctions for TopHat and Cufflinks. A total of three iterative runs of MAKER2 were performed.

Following genome annotation, gene models were analyzed using the program InterProScan to identify putative Pfam and InterPro protein domains. *Ab initio* gene

predictions that did not overlap a MAKER2 annotation but contained a domain were added to the annotation set.

The GMOD Community Annotation System (CAS)

All gene annotations and supporting evidence alignments produced by MAKER2 (as well as protein domain information derived from InterProScan) were loaded into a Chado database to facilitate community access. The annotation curation tool Apollo was then used to allow researchers to view and manually edit the genome annotations in the database. Apollo allows users to connect remotely to the Chado database over the Internet, thus providing a way for researchers to curate the genome annotations from distinct locations. Apollo was configured as a Java Web Start application, which can be pre-configured and pushed onto a computer via a web-browser. This kept configuration of the program under the control of a central server and ensured consistency in the way data could be viewed and accessed.

Schmidtea mediterranea genome annotation

The annotation pipeline MAKER2 was used to annotate the *S. mediterranea* genome. The algorithm identifies repetitive elements, aligns ESTs and protein homology to the genome, integrates pre-processed mRNA-seq alignments, produces *ab initio* gene predictions, and integrates these data to synthesize downstream genome annotations.

The EST dataset used consisted of ESTs sequenced for the *S. mediterranea* genome project[34, 55]. The protein homology set consisted of all proteins in the UniProt/Swiss-Prot protein database, all *Schistosoma mansoni* proteins from Sanger, and all GenBank proteins for *Nematostella vectensis*, *H. sapiens*, *C.*

elegans, and *S. mediterranea*. Unpublished short read mRNA-seq transcriptome datasets prepared for the genome project were processed with the programs TopHat and Cufflinks. The scripts `tophat2gff3` and `cufflinks2gff3` (both included with MAKER2) were used to process the results into GFF3 format. The resulting GFF3 files were provided to the `est_gff` option in MAKER2. A species-specific repeat library prepared for *S. mediterranea* was also supplied to MAKER2.

Prior to running MAKER2, the predictor SNAP was trained using CEGMA to generate an initial training set of conserved universal eukaryotic genes. After running MAKER2, SNAP was re-trained using the resulting first round MAKER2 genome annotations identified as high quality via the script `maker2zff` (which is included with the MAKER2 distribution). Augustus was also trained using this same first round MAKER2 training set. MAKER2 was then run one last time using the updated training files to produce final annotations.

InterProScan was used to identify InterPro and Pfam domains for all resulting *S. mediterranea* annotations as well as nonoverlapping *ab initio* gene predictions. Any nonoverlapping prediction containing a known domain was added to the final annotation set.

Schmidtea mediterranea repeat library preparation

We used RepeatMasker together with the RepBase library of repetitive elements to identify repeats in the *S. mediterranea* genome. The program PILER was used to identify novel repeats not identified by RepeatMasker.

To classify these novel repeats, we used BLAST to locate (via BLAST score) the five best hits for each repeat element within the unmasked *S. mediterranea* genome. Each of the five best hits was then extended 500 bp upstream and

downstream. The resulting extended repeat sequences served as queries for searching the RepBase library via BLASTN. This was done because many novel repeats may just be divergent extensions of repeats already found in RepBase. RepBase also served as a query against itself in BLASTN to identify homology within the database between repeats.

All BLAST hits found in these analyses were clustered and compared via single-linkage clustering. Linkage was determined by two hits sharing an e-value above a given threshold. Consensus classification was then made by looking for statistically significant overrepresentation of RepBase repeat-families as identified by chi-squared analysis with a p value of 0.001. The linkage threshold was gradually relaxed to create the maximally large cluster with statistically significant RepBase association. The threshold initially started at 1×10^{-304} and was gradually relaxed to 1×10^{-4} . Once repeats were clustered, each cluster was classified or named based on the most common repeat type within that cluster.

Schmidtea mediterranea orthology comparison

To determine how orthology of *S. mediterranea* genes compares to what is seen in other model systems, we performed an all-by-all BLASTP analysis of *H. sapiens*, *C. elegans*, *D. melanogaster*, *M. musculus*, *N. vectensis*, *C. intestinalis*, *S. purpuratus*, and *S. mediterranea*. We then estimated the number of orthologs between each genome using reciprocal-best-hits analysis (wherein, if the best hit of gene A is gene B and the best hit of gene B is also gene A, then gene A and B are considered orthologs).

Schmidtea mediterranea phylogenetic tree construction

To perform phylogenetic analysis of the *S. mediterranea* genome, we first identified a subset of conserved genes that could be used to produce a concatenated multi-gene alignment. This was done by performing an all-by-all BLASTP analysis using the following genomes: *X. tropicalis*, *T. rubripes*, *H. sapiens*, *S. mansoni*, *Schistosoma japonicum*, *D. melanogaster*, *C. elegans*, *N. vectensis*, *Ciona intestinalis*, *L. gigantea*, *Capitella teleta*, *Tribolium castaneum*, *Tricoplax adherens*, *S. purpuratus*, *S. cerevisiae*, and *Monosiga brevicollis*. WUBLAST[56] BLASTP was configured with the following parameters: E=0.001, matrix=BLOSUM45, Q=14, R=2, gapK=0.032, gapL=0.195, gapH=0.10, gspmax=10, hspmax=10, and wordmask=seg. Reciprocal-best-hits analysis was then used to identify orthologs across species. Only genes that contained orthologs for all 16 species were kept (i.e., conserved in all species). The sequence for each gene was concatenated for every species (the transcript with the best alignment to human was used in cases where multiple transcripts existed). The concatenated sequences were then aligned using ClustalW[57] and processed by GBlocks[58] to identify conserved blocks of homology. The concatenated alignment was then processed into 100 bootstrap datasets using the Phylip[59] program SEQBOOT, and tree topology was produced for each dataset using Phylobayes[60] (CAT model). Phylobayes was run to 500 generations with the first 100 trees of each run being discarded as 'burn in' (in accordance with Phylobayes documentation). Phylobayes generated a consensus tree for each dataset using the remaining 400 generations. The Phylip program Consense was then used to obtain the bootstrap consensus tree from the 100 datasets, and the Phylip program PROML was used to calculate branch distances.

References

1. Forster H, Coffey MD, Elwood H, Sogin ML: **Sequence analysis of the small subunit ribosomal RNAs of three zoosporic fungi and implications for fungal evolution.** *Mycologia* 1990, **82**:306-312.
2. Tyler BM, Tripathy S, Zhang X, Dehal P, Jiang RHY, Aerts A, Arredondo FD, Baxter L, Bensasson D, Beynon JL, et al: **Phytophthora genome sequences uncover evolutionary origins and mechanisms of pathogenesis.** *Science* 2006, **313**:1261-1266.
3. Kamoun S: **A catalogue of the effector secretome of plant pathogenic oomycetes.** *Annual Review of Phytopathology* 2006, **44**:41-60.
4. Levesque CA, Brouwer H, Cano L, Hamilton J, Holt C, Huitema E, Raffaele S, Robideau G, Thines M, Win J, et al: **Genome sequence of the necrotrophic plant pathogen *Pythium ultimum* reveals original pathogenicity mechanisms and effector repertoire.** *Genome biology* 2010, **11**:R73.
5. Jeffares DC, Mourier T, Penny D: **The biology of intron gain and loss.** *Trends in Genetics* 2006, **22**:16-22.
6. Dyrlov Bendtsen J, Nielsen H, von Heijne G, Brunak S: **Improved prediction of signal peptides: SignalP 3.0.** *Journal of Molecular Biology* 2004, **340**:783-795.
7. Judelson HS, Randall TA: **Families of repeated DNA in the oomycete *Phytophthora infestans* and their distribution within the genus.** *Genome* 1998, **41**:605-615.
8. Kovach A, Wegrzyn J, Parra G, Holt C, Bruening G, Loopstra C, Hartigan J, Yandell M, Langley C, Korf I, Neale D: **The *Pinus taeda* genome is characterized by diverse and highly diverged repetitive sequences.** *BMC Genomics* 2010, **11**:420.
9. Khoshoo TN: **Polyploidy in gymnosperms.** *Evolution* 1959, **13**:24-39.
10. Mungall C, Emmert D: **A Chado case study: an ontology-based modular schema for representing genome-associated biological information.** *Bioinformatics (Oxford, England)* 2007, **23**:i337 - 346.
11. Lewis SE, Searle SMJ, Harris N, Gibson M, Iyer V, Richter J, Wiel C, Bayraktaroglu L, Birney E, Crosby MA, et al: **Apollo: a sequence annotation editor.** *Genome Biology* 2002, **3**:research0082.0081 - 0082.0014.
12. **Insights into social insects from the genome of the honeybee *Apis mellifera*.** *Nature* 2006, **443**:931-949.

13. Werren JH, Richards S, Desjardins CA, Niehuis O, Gadau Jr, Colbourne JK, Group TNGW: **Functional and evolutionary insights from the genomes of three parasitoid nasonia species.** *Science*, **327**:343-348.
14. Munoz-Torres MC, Reese JT, Childers CP, Bennett AK, Sundaram JP, Childs KL, Anzola JM, Milshina N, Elsiek CG: **Hymenoptera Genome Database: integrated community resources for insect species of the order Hymenoptera.** *Nucleic acids research*, **39**:D658-D662.
15. Bonasio R, Zhang G, Ye C, Mutti NS, Fang X, Qin N, Donahue G, Yang P, Li Q, Li C, et al: **Genomic comparison of the ants *Camponotus floridanus* and *Harpegnathos saltator*.** *Science*, **329**:1068-1071.
16. Bell E: **Lampreys diversify differently.** *Nat Rev Immunol* 2004, **4**:580-580.
17. Yin HS, Selzer ME: **Axonal regeneration in lamprey spinal cord.** *J Neurosci* 1983, **3**:1135-1144.
18. Selzer ME: **Mechanisms of functional recovery and regeneration after spinal cord transection in larval sea lamprey.** *J Physiol* 1978, **277**:395-408.
19. Ohno S: *Evolution by gene duplication.* London: George Alien & Unwin Ltd. Berlin, Heidelberg and New York: Springer-Verlag.; 1970.
20. Dehal P, Boore JL: **Two rounds of whole genome duplication in the ancestral vertebrate.** *PLoS Biol* 2005, **3**:e314.
21. Donoghue PCJ, Purnell MA: **Genome duplication, extinction and vertebrate evolution.** *Trends in ecology & evolution (Personal edition)* 2005, **20**:312-319.
22. Kuraku S, Meyer A, Kuratani S: **Timing of genome duplications relative to the origin of the vertebrates: did cyclostomes diverge before or after?** *Molecular Biology and Evolution* 2009, **26**:47-59.
23. Putnam NH, Butts T, Ferrier DEK, Furlong RF, Hellsten U, Kawashima T, Robinson-Rechavi M, Shoguchi E, Terry A, Yu J-K, et al: **The amphioxus genome and the evolution of the chordate karyotype.** *Nature* 2008, **453**:1064-1071.
24. Zayas RM, Hernandez A, Habermann B, Wang Y, Stary JM, Newmark PA: **The planarian *Schmidtea mediterranea* as a model for epigenetic germ cell specification: analysis of ESTs from the hermaphroditic strain.** *Proceedings of the National Academy of Sciences* 2005, **102**:18491-18496.

25. Alvarado AS, Newmark PA, Robb SMC, Juste R: **The Schmidtea mediterranea database as a molecular resource for studying platyhelminthes, stem cells and regeneration.** *Development* 2002, **129**:5659-5665.
26. Randolph H: **Observations and experiments on regeneration in planarians.** *Arch Entw Mech Org* 1897, **7**:352-372.
27. Morgan TH: **Experimental studies of the regeneration of Planaria maculata.** *Arch Entw Mech Org* 1898, **7**:364-397.
28. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: **Rebase Update, a database of eukaryotic repetitive elements.** *Cytogenetic and Genome Research* 2005, **110**:462-467.
29. Smith CD, Edgar RC, Yandell MD, Smith DR, Celniker SE, Myers EW, Karpen GH: **Improved repeat identification and masking in dipterans.** *Gene* 2007, **389**:1-9.
30. Edgar RC, Myers EW: **PILER: identification and classification of genomic repeats.** *Bioinformatics* 2005, **21**:i152-158.
31. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-seq.** *Bioinformatics* 2009, **25**:1105-1111.
32. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation.** *Nat Biotech* 2010, **28**:511-515.
33. Zmasek C, Godzik A: **Strong functional patterns in the evolution of eukaryotic genomes revealed by the reconstruction of ancestral protein domain repertoires.** *Genome biology* 2011, **12**:R4.
34. Robb S, Ross E, Alvarado A: **SmedGD: the Schmidtea mediterranea genome database.** *Nucleic Acids Res* 2007:D599 - 606.
35. Korf I: **Gene finding in novel genomes.** *BMC Bioinformatics* 2004, **5**:59.
36. Parra G, Bradnam K, Korf I: **CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes.** *Bioinformatics* 2007, **23**:1061-1067.
37. Salamov AA, Solovyev VV: **Ab initio gene finding in drosophila genomic DNA.** *Genome Res* 2000, **10**:516-522.
38. Lomsadze A, Ter-Hovhannisyanyan V, Chernoff YO, Borodovsky M: **Gene identification in novel eukaryotic genomes by self-training algorithm.** *Nucl Acids Res* 2005, **33**:6494-6506.

39. Cheung F, Win J, Lang J, Hamilton J, Vuong H, Leach J, Kamoun S, Andre Levesque C, Tisserat N, Buell CR: **Analysis of the *Pythium ultimum* transcriptome using Sanger and pyrosequencing approaches.** *BMC Genomics* 2008, **9**:542.
40. Boguski MS, Lowe TMJ, Tolstoshev CM: **dbEST - database for expressed sequence tags.** *Nat Genet* 1993, **4**:332-333.
41. UniProt C: **The Universal Protein Resource (UniProt).** *Nucleic Acids Res* 2007:D193 - 197.
42. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000.** *Nucl Acids Res* 2000, **28**:45-48.
43. Finn RD, Mistry J, Schuster-Bockler B, Griffiths-Jones S, Hollich V, Lassmann T, Moxon S, Marshall M, Khanna A, Durbin R, et al: **Pfam: clans, web tools and services.** *Nucl Acids Res* 2006, **34**:D247-251.
44. The InterPro C, Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Biswas M, Bradley P, Bork P, et al: **InterPro: An integrated documentation resource for protein families, domains and functional sites.** *Brief Bioinform* 2002, **3**:225-235.
45. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R: **InterProScan: protein domains identifier.** *Nucl Acids Res* 2005, **33**:W116-120.
46. **GFF3** [<http://www.sequenceontology.org/gff3.shtml>]
47. Altschul SF, Gish W, Miller W, Meyers EW, Lipman DJ: **Basic Local Alignment Search Tool.** *Journal of Molecular Biology* 1990, **215**:403-410.
48. The Arabidopsis Genome Initiative: **Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*.** *Nature* 2000, **408**:796-815.
49. Jurka J, Klonowski P, Dagman V, Pelton P: **Censor--a program for identification and elimination of repetitive elements from DNA sequences.** *Computers & Chemistry* 1996, **20**:119-121.
50. Stanke M, Waack S: **Gene prediction with a hidden Markov model and a new intron submodel.** *Bioinformatics* 2003, **19**:ii215-225.
51. Keibler E, Brent M: **Eval: A software package for analysis of genome annotations.** *BMC Bioinformatics* 2003, **4**:50.
52. Benson D, Karsch-Mizrachi I, Lipman D, Ostell J, Wheeler D: **GenBank.** *Nucleic acids research* 2007:D21 - 25.
53. **RepeatModeler** [<http://repeatmasker.org>]

54. Roberts A, Trapnell C, Donaghey J, Rinn J, Pachter L: **Improving RNA-seq expression estimates by correcting for fragment bias.** *Genome biology*, **12**:R22.
55. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, Holt C, Sanchez Alvarado A, Yandell M: **MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes.** *Genome Res* 2008, **18**:188-196.
56. **WUBLAST** [<http://blast.wustl.edu>]
57. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD: **Multiple sequence alignment with the Clustal series of programs.** *Nucl Acids Res* 2003, **31**:3497-3500.
58. Castresana J: **Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis.** *Mol Biol Evol* 2000, **17**:540-552.
59. **PHYLIP** [<http://evolution.genetics.washington.edu/phylip.html>]
60. Lartillot N, Philippe H: **A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process.** *Mol Biol Evol* 2004, **21**:1095-1109.

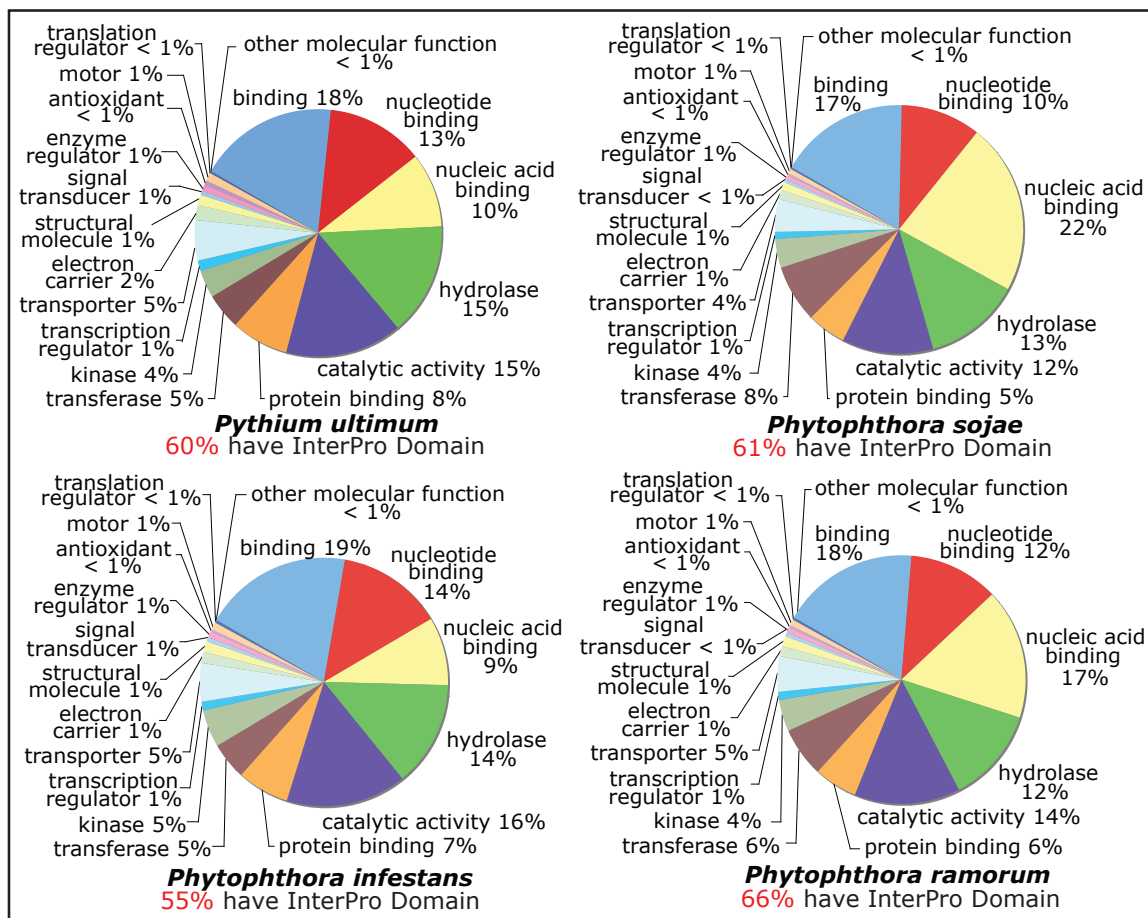


Figure 3.1 Comparison of domain content in Oomycetes.

The percentage of genes containing a known InterPro protein domain is persistent across all species of Oomycetes. These similarities extend to the functional makeup of these genomes as well (as seen using GO terms associations for molecular function). The consistency of patterns in *Pythium ultimum* in comparison to *Phytophthora* species suggest there is no obvious systematic functional bias inherent from the gene annotation process, and that the quality of the MAKER2 gene annotations is comparable to the quality seen in other published Oomycete genomes.

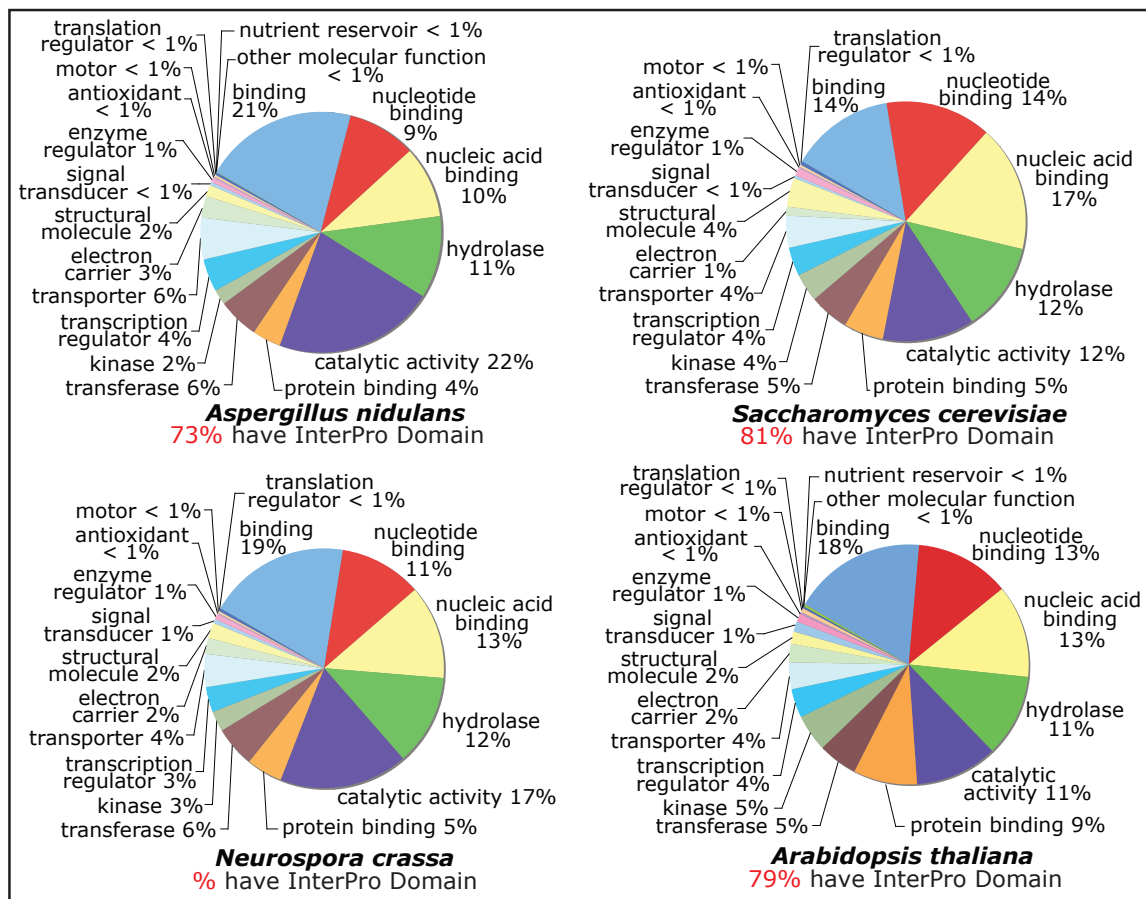


Figure 3.2 Domain content in reference eukaryotes.

The overall similarity in the functional makeup of eukaryotic genomes is clear from a visual comparison of the GO function pie charts. This relatively invariant pattern of domain content and GO term distribution makes it simple to derive expected patterns of genome functional makeup for virtually any eukaryotic organism. Deviation from this pattern would indicate problems with genome annotation. These genome have slightly higher InterPro domain content than Oomycetes do, which is less indicative of differences in biology for these organism, and more related to the fact that Oomycetes are evolutionarily distant from most sequenced eukaryotes (some Oomycete genes may contain new or divergent protein domains that would not be present in the existing InterPro database).

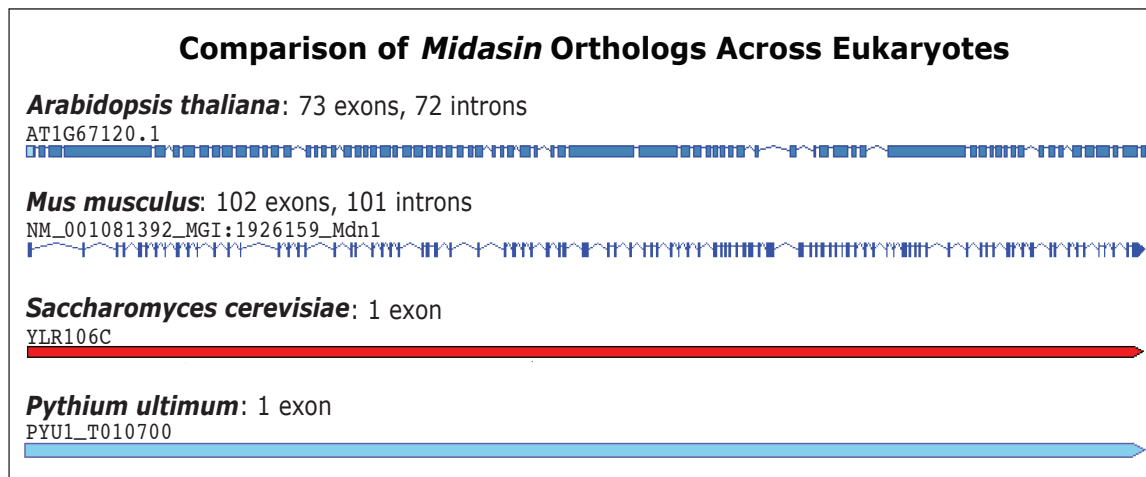


Figure 3.3 Comparison of intron/exon structure across eukaryotes.

Many genes in *Pythium ultimum* are large single exon genes. However, orthologs of these single exon genes in other eukaryotic organisms tend to be intron rich, suggesting a possible evolutionary trend in *P. ultimum* for intron loss. The example in the figure shows orthologs of the gene midasin, which encodes a highly conserved ~600 kDa nuclear chaperone protein. Orthologs in both *Mus musculus* and *Arabidopsis thaliana* are intron rich; however, orthologs in *Saccharomyces cerevisiae* and *P. ultimum* are both encoded by a large single exon. The similarity in gene structure between *S. cerevisiae* and *P. ultimum* is surprising given the evolutionary distance between these organisms (Oomycetes are equidistant from plants, animals, and fungi) and suggests that intron loss may be a convergent adaptation related to a similar lifestyle.

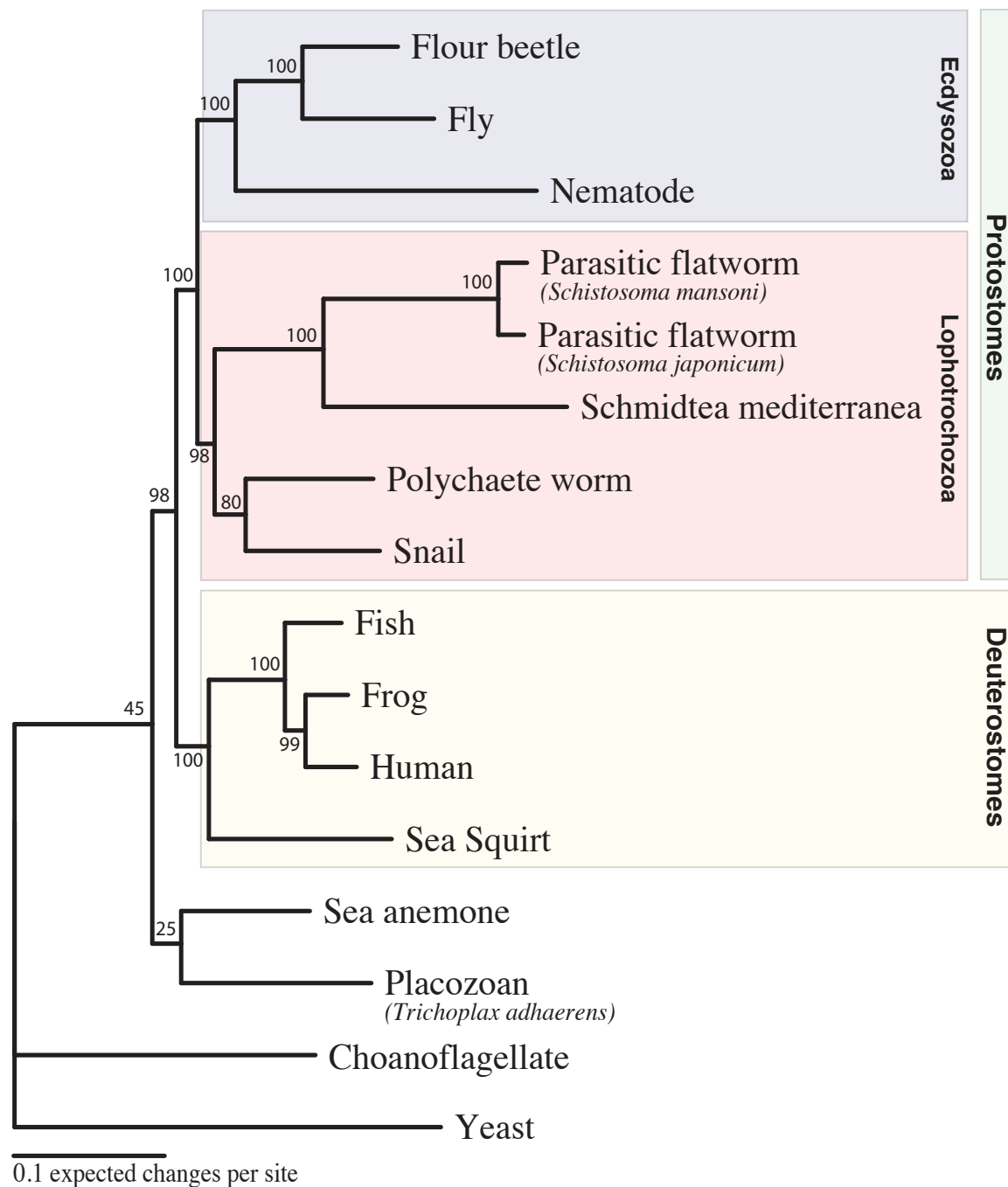


Figure 3.4 Phylogenetic analysis of *Schmidtea mediterranea*.

Whole genome phylogenetic analysis of sixteen eukaryotes using Bayesian methods firmly places *Schmidtea mediterranea* in the Lophotrochozoa, with a high bootstrap score of 98/100. Two related species of parasitic flatworms fall into in the same grouping (further supporting this classification). The relatively long branch lengths for the three Platyhelminthes species suggest that these organisms may have experienced greater rates of evolution than other metazoans.

Deuterostomes

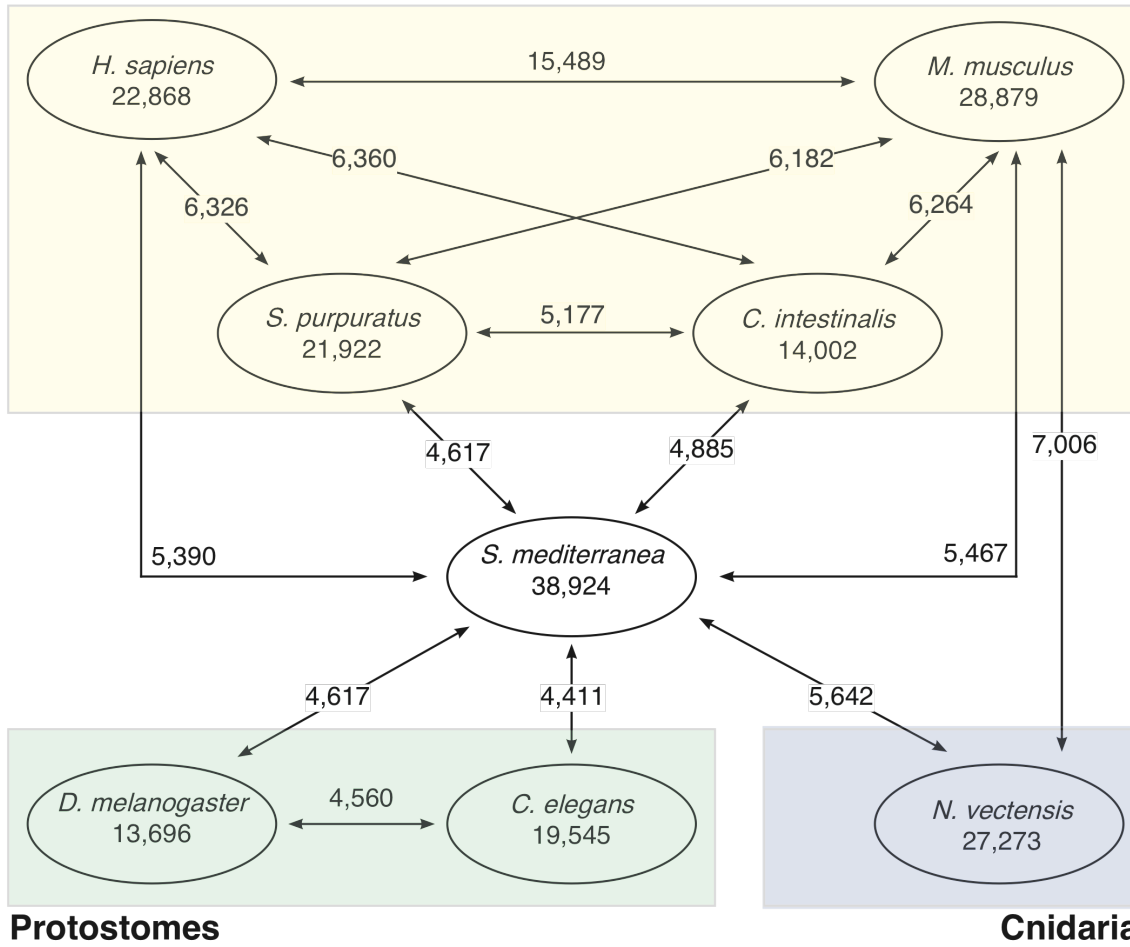


Figure 3.5. Genes shared by *Schmidtea mediterranea* with other metazoans. The level of orthology for *Schmidtea mediterranea* is similar to what is seen between other metazoans, which reflects well on the quality of the gene models (considering the high level of assembly fragmentation). We see that on average *S. mediterranea* shares more genes with deuterostomes and cnidarians than it does with more closely related protostomes (*Drosophila melanogaster* and *Caenorhabditis elegans*). The *S. mediterranea* genome may therefore be basal in nature (i.e. having more in common with the ancestral metazoan than with other more closely related organisms).

CHAPTER 4

EDUCATIONAL OUTREACH

While second-generation technologies are making genome and transcriptome sequencing routine even for small laboratories, the large volumes of data they produce are overwhelming many research groups with issues of data management and downstream analysis. This is especially true for smaller groups with limited bioinformatics expertise. While we have developed MAKER2, a tool for automated genome annotation and database management, overhead related to the installation of prerequisite software as well as the fact that MAKER2 is a command line tool may intimidate many researchers (especially if they have no experience navigating a Linux environment). To help overcome these issues, we have made a concerted effort through educational outreach to train the scientific community in strategies for genome annotation as well as in the use of Generic Model Organism Database project (GMOD)[1] tools like MAKER2.

As part of our effort to educate the scientific community in strategies for genome annotation, MAKER2 classes and workshops have been taught at several international meetings and conferences. MAKER2 was the opening workshop for the GMOD Summer School of the Americas in 2009 and 2010; and at the 2009 GMOD summer school, MAKER2 received the highest marks of any workshop. MAKER2 was also presented as a course at the 2011 GMOD spring training session at

NESCent. At the Plant and Animal Genome Conference, MAKER2 workshops were presented in 2010 and 2011, and a MAKER2 workshop was taught at the 2010 Arthropod Genomics Symposium.

There have also been a number of independent workshops on using MAKER2 presented as part of bioinformatics courses: the University of Maryland (Brandi Cantarel), Texas A&M University (Jim Hu and Rodolfo Aramayo), and the University of Utah (Karen Eilbeck). Additionally, Christopher Smith at San Francisco State University used MAKER2 and Apollo[2] as hybrid teaching/research tools to finish up genome annotations for the *Pogonomyrmex barbatus*[3], *Linepithema humile*[4], and *Atta cephalotes*[5] genome projects. Similarly, researchers at the 2010 *Fusarium Circunatum* genome annotation jamboree held at the University of Pretoria in South Africa were trained in using MAKER2 and Apollo as part of an effort to develop bioinformatics expertise in that nation. Through my work on MAKER2, I have also become an advisor for DNA Subway[6], a tool produced by Cold Spring Harbor Laboratory's Dolan DNA Learning Center to introduce high school and undergraduate students to genome annotation.

While efforts in educational outreach have made major inroads in the scientific community, the development and distribution of web-based annotation and analysis tools could even-further increase the availability of MAKER2 as well as provide educational opportunities to a wider array of scientists. In this effort, I have developed the MAKER Web Annotation Service, which makes genome annotation and analysis as easy as typing in a URL.

References

1. **GMOD** [<http://www.gmod.org>]
2. Lewis SE, Searle SMJ, Harris N, Gibson M, Iyer V, Richter J, Wiel C, Bayraktaroglu L, Birney E, Crosby MA, et al: **Apollo: a sequence annotation editor**. *Genome Biology* 2002, **3**:research0082.0081 - 0082.0014.
3. Smith CR, Smith CD, Robertson HM, Helmkamp M, Zimin A, Yandell M, Holt C, Hu H, Abouheif E, Benton R, et al: **Draft genome of the red harvester ant *Pogonomyrmex barbatus***. *Proceedings of the National Academy of Sciences* 2011, **108**:5667-5672.
4. Smith CD, Zimin A, Holt C, Abouheif E, Benton R, Cash E, Croset V, Currie CR, Elhaik E, Elsik CG, et al: **Draft genome of the globally widespread and invasive Argentine ant (*Linepithema humile*)**. *Proceedings of the National Academy of Sciences* 2011, **108**:5673-5678
5. Suen G, Teiling C, Li L, Holt C, Abouheif E, Bornberg-Bauer E, Bouffard P, Caldera EJ, Cash E, Cavanaugh A, et al: **The genome sequence of the leaf-cutter ant *Atta cephalotes* reveals insights into its obligate symbiotic lifestyle**. *PLoS Genet* 2011, **7**:e1002007.
6. **DNA Subway** [<http://dnasubway.iplantcollaborative.org/>]

CHAPTER 5

MAKER WEB ANNOTATION SERVICE: AN ONLINE PORTAL FOR GENOME ANNOTATION AND ANALYSIS

Abstract

I have developed the MAKER Web Annotation Service (MWAS), a web-based, distributed multiuser environment that permits automated genome annotation and downstream analysis via a simple web-browser. MWAS serves as a portal to annotate genomic datasets for users lacking sufficient bioinformatics expertise (or computer resources) to install, configure, and run the command-line program MAKER2. MWAS is also an online educational tool where users can explore strategies for genome annotation and analysis using applications from the Generic Model Organism Database project (GMOD)[1], the European Bioinformatics Institute (EBI), and the Sequence Ontology project[2].

Rationale

The challenges of modern genomics extend beyond merely annotating genomes, as annotations must also be subjected to diverse and complex downstream analyses. To this end, I have developed the MAKER Web Annotation Service (MWAS), in which I have integrated the MAKER2 genome annotation pipeline

together with other downstream applications and GMOD tools. MWAS acts as a portal for biological research by providing annotations and downstream analyses for submitted genomic sequences. MWAS is also an educational tool, allowing researchers to explore strategies for genome annotation as well as methods for integrating annotations into external applications (all just by typing a URL into a web-browser). Through the web-based environment, users can annotate genome sequences, identify protein domains, annotate putative gene functions, and add associated Gene Ontology (GO)[3] terms to annotations. All results are automatically incorporated into GMOD and Sequence Ontology tools. Thus, MWAS provides a unified web-based resource for easy genome annotation, distribution of results, and global analyses of annotation features. MWAS also permits concurrent viewing of annotations via GBrowse[4] and JBrowse[5] (online annotation visualization tools), and remote manual editing of annotations using Apollo[6] (an annotation curation tool). This seamless integration with external tools greatly increases the efficiency of small project collaboration by enabling researchers in different locations to edit and analyze a shared genome dataset remotely.

Algorithm/website design

Development of a web-based annotation and analysis tool imposes several design constraints. Because of the large datasets involved in genomics as well as the potential confidential nature of those datasets, researchers need to reuse the same files without having to upload them multiple times, and at the same time, they want to restrict access to their experimental data and results. Both of these requirements are met by having individual accounts with a password-protected login. However, because casual users and those using MWAS solely for educational purposes may be

intimidated by a registration and login requirement, these are optional. Depending on the size of datasets submitted, annotation can take anywhere from a few minutes to several hours. Therefore, users need a way to submit an analysis to MWAS and return at a later time once it has completed. While this requirement is easily met for datasets linked to a password-protected account, providing return access to anonymous users must be met by a different mechanism and is achieved via links that can be copied and bookmarked by the user. These links will automatically return the web-browser to the correct datasets and results.

While genome annotations provide a starting point for genomics-based research, they are not an end in themselves. After annotating a genome, researchers also need to carry out downstream analyses; these analyses include finding protein domains, adding GO terms to annotations, identifying orthologs in related species, and elucidating potential gene functions. Additionally, researchers need global statistics that can be compared between organisms, such as gene count, average intron length, etc. But most importantly, researchers want to visualize the data so they can better interpret the results. While there are numerous software tools that can perform these tasks, requiring users to install and configure these tools before using MWAS would produce a barrier inconsistent with the easy-to-use aspects of the service. To meet these downstream analysis needs as well as to provide proof-of-principle examples for educational users, I have integrated software tools from GMOD, the EBI, and the Sequence Ontology project into MWAS.

Entering MWAS

When a user first enters the MWAS website, they are presented with a login page that lets them access an existing account, register as a new user, or access the

site anonymously as a guest (Figure 5.1). Once inside of MWAS, the user is presented with the main page for the website, which shows the results of finished analyses as well as the status of unfinished jobs that were submitted by the user. At the top of the main page, there is a series of tabs for navigation of the website. These tabs allow the user to submit new datasets for annotation, manage/upload analysis files, or seek help via e-mail and written documentation.

Submitting datasets and configuring MWAS for annotation

To submit datasets for analysis, users must first upload genome sequence and any desired analysis support files to MWAS. This is done by selecting the 'Manage Files' tab at the top of the main page. From there, users can submit genomic sequence for annotation as well as EST and protein homology datasets for alignment against the genome. If users have trained *ab initio* gene-predictors such as SNAP[7] or GeneMark[8], they can also submit species parameter files for those programs. Additionally, if users have datasets of pre-existing *ab initio* gene predictions or legacy annotations, and they wish to use them for re-annotation of a genome, these files can be submitted in GFF3[9] format.

When users are ready to annotate an uploaded genome, they must select the 'New Job' tab at the top of the screen. From there, users customize the MWAS annotation engine (MAKER2) via a series of dropdown menus. Using these menus the user selects the genome file to be annotated as well as a set of ESTs and a protein homology evidence to align against the genome (Figure 5.2). Advanced options for repeat masking and the *ab initio* gene prediction are configured here as well. Alternatively, users who are using MWAS for educational purposes can select from a list of preconfigured example annotation jobs. When finished selecting all

desired parameters, users submit their job to the annotation queue where MAKER2 will process it.

Viewing and downloading annotation results

After submitting a new job, the status of that job is displayed on the MWAS main page. Each submitted job has an associated 'JobID' that can be used to review the previously supplied MAKER2 configuration parameters. There is also a 'Log' that displays any problems encountered by MAKER2 during its analysis. Upon completion of a submitted annotation job, a new icon appears that allows users to access resulting gene annotations. From there, users can simply download results to their own computer for further analysis, or they can view the annotations using integrated GMOD tools. If the user selects 'View in GBrowse' or 'View in JBrowse', the annotations are immediately displayed in a separate window using those programs. By using viewing programs like these, annotation structure can be verified and evaluated in relation to aligned evidence from ESTs and protein homology, and users can decide whether to accept the annotations or make changes to MAKER2's configuration parameters for job resubmission. By clicking on 'View in Apollo' in the results menu (Figure 5.3), users have the option to manually edit gene structure using a preconfigured Java Web Start version of Apollo that will automatically be pushed over the Internet and installed on their machine. Any changes made to the annotations using Apollo can then be saved locally as GFF3 files.

Postprocessing of gene annotations

While MAKER2-produced structural annotations comprise a valuable resource, researchers also have needs for other types of downstream analyses. MWAS provide users with access to external tools that can perform those tasks. To have access to summary statistics for genome annotations, the user can select ‘SOBA Statistics’ from the results menu (Figure 5.4). This will upload MAKER2-produced annotations to SOBA[10], a summary statistics tool from the Sequence Ontology. From there, users can explore statistics on gene numbers, exon structure, coding sequence and transcript length, intron density, and many other useful summary values. Users can also view the Sequence Ontology relationships between different annotated features in the genome and produce reports that can be downloaded back to a local computer.

If a user is satisfied with MAKER2-produced annotations based on manual inspection of gene models in GBrowse, JBrowse, and Apollo, he/she can then submit annotations to downstream functional analyses by selecting ‘Do postprocessing of annotations’ from the results menu. From there, users have the option to add putative gene functions via comparison to the UniProt/Swiss-Prot[11, 12] protein database or to identify InterPro[13] domains and associated GO terms using the program InterProScan[14]. Users can also rename genes to comply with NCBI suggested naming formats using a registered genome project prefix. Additionally, if the user does not yet have a trained *ab initio* parameter file for the organism being analyzed then MWAS can attempt to train the gene-predictor SNAP for them.

After selecting and submitting desired postannotation analyses, a new job will appear on the MWAS main screen. Just as with the earlier MAKER2 analysis, each submitted job has an associated ‘JobID’ that can be used to review the

configuration parameters and a 'Log' that displays any problems encountered during postprocessing.

Upon completion, users can access the results via a new icon that appears next to the job. The postannotation analysis results menu is nearly identical to the MAKER2 results menu, and users can still view datasets in GBrowse, JBrowse, and Apollo or download the data locally for further experimentation. However, when results are loaded into GBrowse, you will see that the annotations have changed relative to the earlier MAKER2-produced results. With all postanalyses options selected, gene names will now be updated to use the user-supplied prefix followed by a unique identifier, and putative gene functions will be displayed below each gene model (Figure 5.5). Additionally, in the GBrowse track options, 'InterPro Protein Domains' will appear as an option, and when selected, it will display protein domains as physical features aligned against the genome. When users click on one of the gene models shown in GBrowse, a new page will be displayed containing all information for that gene including integrated protein domains, associated GO terms, putative gene functions, and any MAKER2-produced quality control statistics (Figure 5.6). By providing and incorporating these data automatically into the annotations, MWAS allows users to more quickly and efficiently transition to downstream comparative analyses and experimentation using the genome annotations as substrates.

Annotation of *Pinus taeda* BAC clones

As a proof-of-principle example of the performance of MWAS, I used it to annotate a set of 111 contigs assembled from the sequencing of *Pinus taeda* BAC clones (~12 megabases of total sequence). The entire ~20-30 gigabase *P. taeda*

genome has yet to be fully sequenced, and these contigs exemplify the type of smaller dataset that might be submitted to MWAS. In general these are datasets that are not large enough to justify the installation of the entire MAKER2 annotation pipeline, but they are still rich enough to provide a valuable resource for further study of an organism. Results from the *P. taeda* BAC clone contigs can potentially serve as a test dataset for future annotation of the entire genome; and many groups are using them as preliminary results for justification of genome sequencing proposals. I also performed postprocessing analyses of the annotations to identify protein domains, GO term associations, putative gene functions, and to provide a trained SNAP parameter file for *P. taeda*.

Using this set of 111 contigs assembled from the sequencing of *Pinus taeda* BAC clones (~12 megabases of total sequence), MWAS identified 220 genes, of which 64% contained known InterPro domains. In comparison *Arabidopsis thaliana*, a very well annotated reference genome, has a domain enrichment of 79%. The similarity in domain content between these organisms reflects well on the quality of the MWAS produced annotations, especially given the extremely limited nature of the *P. taeda* dataset.

When I further analyze the annotations via SOBA (a summary statistic tool), I see that the average gene contained 2.6 exons, the average coding length of each gene was 572 bp, and coding regions make up < 1% of the total DNA sequence. These kinds of summary statistics provide the foundation researchers need for comparing genes from a newly sequenced genome to those of other related organisms. The summary statistics and annotation functional data also provide a basis for analyzing the quality of the gene annotations. But what is amazing about these data is the fact that they were conveniently and automatically generated via a

simple web-browser. There was no complex data manipulation required from the user, instead pointing and clicking was sufficient.

While our simple exercise with using *P. taeda* BAC sequence demonstrates the ease of genome annotation and analysis using MWAS, I was also able to extract much useful information that could be applied to any future pine genome project. For example, the basic statistics for patterns of gene structure (produced by SOBA) and the SNAP training file produced by MWAS, all provide the kind of high-quality reference data that can help jumpstart a genome project, primarily by eliminating the need to pre-train *ab initio* gene predictors.

Conclusions

MWAS serves as a powerful tool that can provide detailed structural and functional annotations from genomic datasets. The easy-to-use interface makes genome annotation as simple as entering a URL and provides a convenient mechanism for visualizing annotations and distributing them to collaborators. By further leveraging the power of tools from GMOD, the EBI, and the Sequence Ontology, MWAS provides an efficient pathway to downstream analyses. The *P. taeda* genome sequence serves as a proof-of-principle example of how groups can use this tool in their own research. Also, the simple web-based interface and supplied example datasets make it easy for anyone to use MWAS as an educational tool for learning how genome annotation works and to explore different annotation strategies.

Availability and requirements

MWAS can be accessed through the Yandell Lab website at <http://www.yandell-lab.org>. It requires nothing more than a web-browser with JavaScript enabled. MWAS can also be installed locally on machines running Linux, Mac OS X, or other Unix-like operating systems, and it is bundled into the standard MAKER2 software package (versions 2.11 and greater). For local installation of MWAS, the Apache web server is required as well as GBrowse (2.10 or greater), JBrowse (1.2 or greater), Apollo (1.11.6 or greater), and InterProScan (4.X) in addition to all standard MAKER2 prerequisites.

Methods


P. taeda BAC sequence was obtained from GenBank[15] and uploaded to MWAS together with an EST dataset consisting of all *P. taeda* and Pinaceae ESTs found in dbEST[16]. The set of all Viridiplantae proteins from GenBank served as the input protein homology dataset.

An annotation job for *P. taeda* sequence was first submitted using the `est2genome` prediction option, which produces gene models directly from EST alignments. This was done because no pine specific *ab initio* prediction parameters were yet available. The `est2genome` set of annotations was used as the basis for SNAP training during later postprocessing in MWAS. Once SNAP was trained, the annotation job was resubmitted using the newly trained parameter file. The resulting MWAS-SNAP annotation set was then submitted to another round of post-processing to identify protein domains and putative gene functions. Gene names were changed to use the prefix PINE. Final summary statistics were produced via SOBA.

References

1. **GMOD** [<http://www.gmod.org>]
2. Eilbeck K, Lewis S, Mungall C, Yandell M, Stein L, Durbin R, Ashburner M: **The Sequence Ontology: a tool for the unification of genome annotations.** *Genome biology* 2005, **6**:R44.
3. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, et al: **Gene Ontology: tool for the unification of biology.** *Nat Genet* 2000, **25**:25-29.
4. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW, Arva A, Lewis S: **The Generic Genome Browser: a building block for a model organism system database.** *Genome Res* 2002, **12**:1599-1610.
5. Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH: **JBrowse: a next-generation genome browser.** *Genome research* 2009, **19**:1630-1638.
6. Lewis SE, Searle SMJ, Harris N, Gibson M, Iyer V, Richter J, Wiel C, Bayraktaroglu L, Birney E, Crosby MA, et al: **Apollo: a sequence annotation editor.** *Genome Biology* 2002, **3**:research0082.0081 - 0082.0014.
7. Korf I: **Gene finding in novel genomes.** *BMC Bioinformatics* 2004, **5**:59.
8. Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M: **Gene identification in novel eukaryotic genomes by self-training algorithm.** *Nucl Acids Res* 2005, **33**:6494-6506.
9. **GFF3** [<http://www.sequenceontology.org/gff3.shtml>]
10. Moore B, Fan G, Eilbeck K: **SOBA: sequence ontology bioinformatics analysis.** *Nucl Acids Res*, **38**:W161-164.
11. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000.** *Nucl Acids Res* 2000, **28**:45-48.
12. UniProt C: **The Universal Protein Resource (UniProt).** *Nucleic Acids Res* 2007:D193 - 197.
13. The InterPro C, Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Biswas M, Bradley P, Bork P, et al: **InterPro: an integrated documentation resource for protein families, domains and functional sites.** *Brief Bioinform* 2002, **3**:225-235.




14. Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R: **InterProScan: protein domains identifier**. *Nucl Acids Res* 2005, **33**:W116-120.
15. Benson D, Karsch-Mizrachi I, Lipman D, Ostell J, Wheeler D: **GenBank**. *Nucleic acids research* 2007:D21 - 25.
16. Boguski MS, Lowe TMJ, Tolstoshev CM: **dbEST - database for expressed sequence tags**. *Nat Genet* 1993, **4**:332-333.

 **Maker Web Annotation Service**

User Name

Password

Remember User Name

 [New user registration](#)  [Forgot login?](#)  [Help](#)

New Guest Account **User Sign In**

Figure 5.1 MWAS login screen.

MWAS uses a password-protected login to secure users' experimental data as well as resulting gene annotations and analyses (as shown in this screen shot).




	MAKER Job Details	Assigned id: 3313
Denovo Annotation		
<p>Begin by selecting a sequence file to annotate. You also need to select the type of organism the sequence was derived from. This is important because annotation strategies and gene prediction algorithms differ for Eukaryotic and Prokaryotic organisms.</p>		
<p>Choose a genome fasta file: Submissions must be 500,000 base pairs or less.</p> <p>Organism Type: <input checked="" type="radio"/> Eukaryotic <input type="radio"/> Prokaryotic</p>	<p>Select a file below</p> <p>No file selected... Upload File </p> <p>No file selected...</p> <p>Server Provided Fasta Files</p> <ul style="list-style-type: none"> D. melanogaster : example contig De novo Annotation : example contig E. coli : example contig Legacy Annotation : example contig Pass-through : example contig <p>User Provided Fasta Files</p> <ul style="list-style-type: none"> Chromosome 17 Human cDNAs Human proteins 	<p>save</p>
EST Evidence		
<p>ESTs aligned against the genome are used to infer correct intron/exon structure of a gene as well as the location of 5' and 3' UTR. Two different types of ESTs can be used: (1) ESTs from the same source as your genomic sequence; and (2) (optionally) ESTs from a closely related organism, for example if your genomic sequence is human, this second set of ESTs might be from mouse.</p>		
<p>Upload a multi-fasta file of ESTs to be aligned from the same source as your genomic sequence.</p>	<p>Select a file below</p> <p>No file selected... Upload File </p> <p style="text-align: center;">View File Contents</p>	<p>save</p>

Figure 5.2 The MWAS job submission screen.

MWAS allows users to configure MAKER2 just as they would on the command line by using dropdown menus to select files uploaded by the user (as shown in this screen shot).

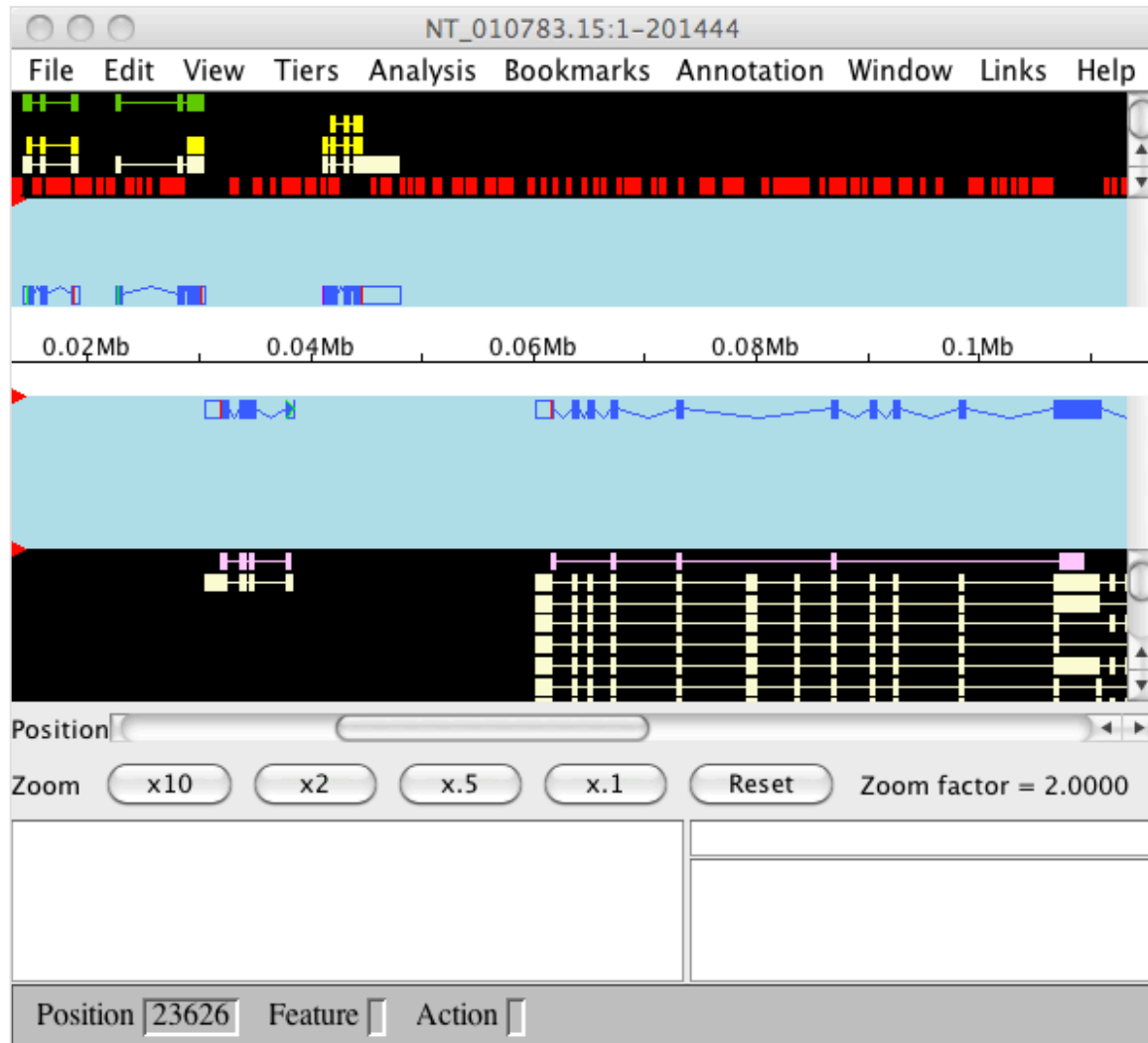


Figure 5.3 View of MWAS annotations in Apollo.

Once an annotation job is complete, the user has many viewing and analysis options that can be accessed directly from the website. In this screen shot, we see how Apollo has been launched by MWAS with the correct dataset already loaded and configured for the user. DNA sequence runs from left to right, gene annotations are shown in the light blue panels, and supporting evidence is displayed in the dark panels.

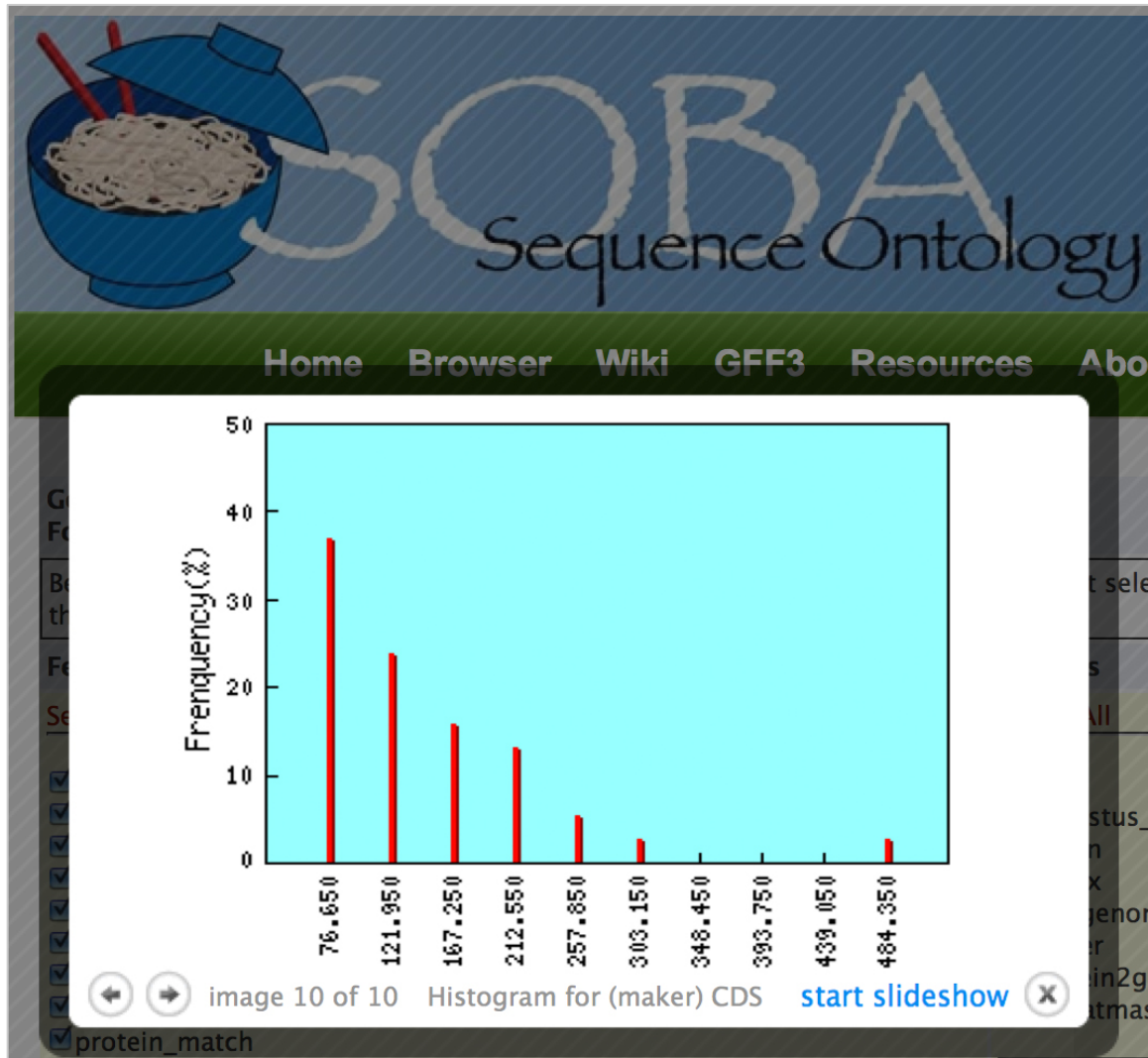


Figure 5.4 Annotation summary statistics.

MWAS can provide users with basic statistics for the genome annotations by using the tool SOBA from the Sequence Ontology project. In this screen shot, we are looking at the length distribution of coding sequence (CDS features) in the genome. SOBA can also produce written reports of annotation statistics that can be downloaded locally for further analysis. MWAS lets the user launch SOBA from the annotation results menu.

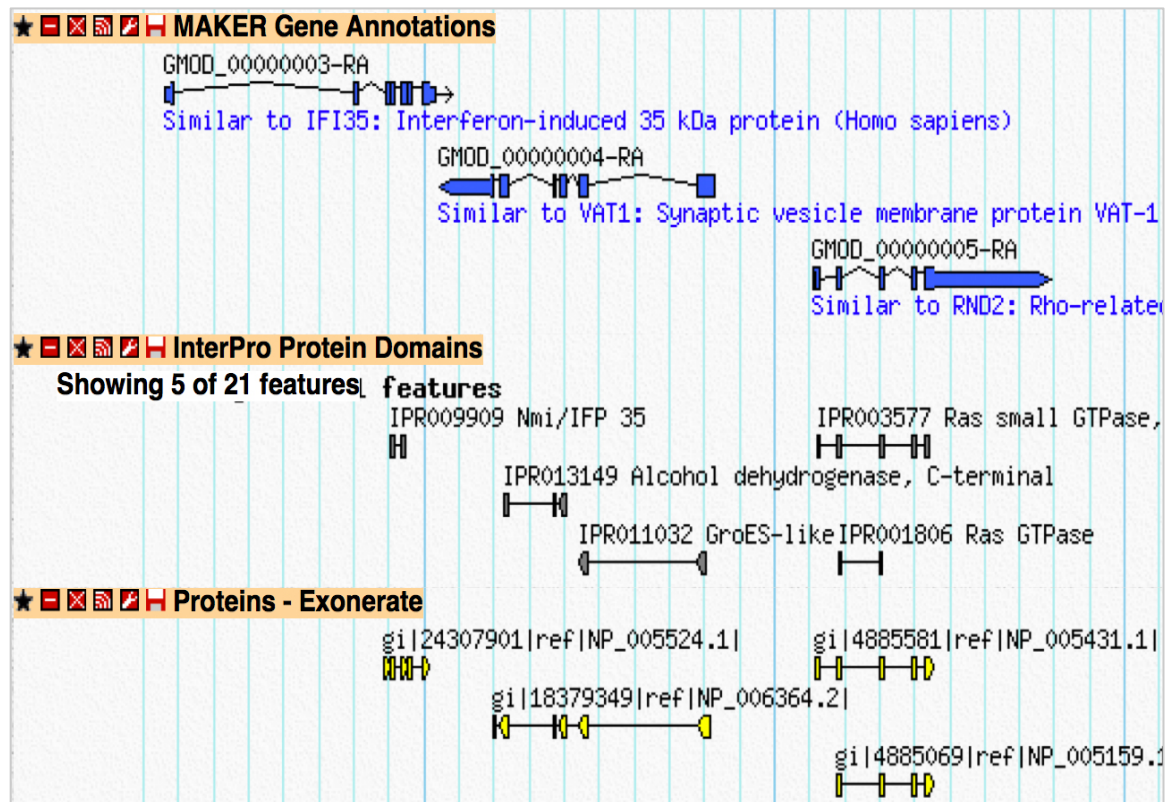


Figure 5.5 View of MWAS functional annotation integration.

In addition to MAKER2 provided structural annotations, users can perform downstream analyses in MWAS to identify protein domains, putative gene functions, and gene associated GO terms. Users can also reformat gene names to correspond to registered genome project prefixes. In this screen shot, we see gene models and downstream functional annotations identified and loaded into GBrowse by MWAS. Note that gene models (in blue) have been named with the prefix 'GMOD', and the putative gene function is displayed below each gene annotation. MWAS also identified InterPro domains in the genes and displays them as physical features mapped against the genome.

GMOD_00000004-RA Details	
Name:	GMOD_00000004-RA
Type:	mRNA
Description:	Similar to VAT1: Synaptic vesicle membrane protein VAT-1 homolog (Homo sapiens)
Source:	maker
Position:	NT_010783.15:30454..38291 (- strand)
Length:	7838
Alias:	augustus_masked-NT_010783.15-abinit-gene-0.8-mRNA-1
Dbxref:	Gene3D:G3DSA:3.40.50.720
	InterPro:IPR002085
	InterPro:IPR002364
	InterPro:IPR011032
	InterPro:IPR013149
	InterPro:IPR013154
	InterPro:IPR016040
	InterPro:IPR020843
	PANTHER:PTHR11695
	PANTHER:PTHR11695:SF29
	Pfam:PF00107
	Pfam:PF08240
	Prosite:PS01162
	SMART:SM00829
	superfamily:SSF50129
	superfamily:SSF51735
Note:	Similar to VAT1: Synaptic vesicle membrane protein VAT-1 homolog (Homo sapiens)
Ontology_term:	GO:0003824
	GO:0005488

Figure 5.6 Association of results and analyses with gene annotations.

MWAS is able to associate a wide body of information to each gene as part of its analyses. By clicking on a gene model in GBrowse, the user is taken to a summary page (shown in this screen shot) that displays all information available for that gene including: protein domain content, associated GO terms, putative gene functions, all MAKER2 produced quality control statistics, and many other useful types of information. The ability of MWAS to rapidly collect these data and make them available to the user demonstrates the power of MWAS. This same data is also integrated into the final GFF3 file that the user can download locally.

CHAPTER 6

ALGORITHM DESIGN

MAKER2 is a genome annotation pipeline designed for research groups with little bioinformatics experience. The program identifies and collects evidence for building gene annotations using EST and protein alignments, third-party gene prediction programs, and other evidence as supplied by the user. Using these input datasets, MAKER2 produces a final set of database-ready gene annotations in GFF3[1] format.

The basic strategy behind MAKER2 is to automatically take a sequence from a genome assembly and guide that sequence through a series of evidence collection steps. The genome sequence is passed from one external executable to another, and results are collected and interpreted by MAKER2. At the end of this process, MAKER2 uses the collected evidence to supply hints about the locations of introns, exons, and protein coding regions to *ab initio* gene-prediction algorithms. The hints provided by MAKER2 increase the accuracy of the final gene models. Gene models are also tagged with codes that identify the types and quantity of evidence overlapping them. These tagged gene models are then saved together with their associated evidence trails as final database-ready gene annotations.

Supported system architectures

MAKER2 is designed to support laptop and desktop machines running Unix-like operating systems (i.e., Linux or Mac OS X). These are machines likely to be found in the average research laboratory. However, MAKER2 is also compatible with the distributed system architectures found in advanced computer clusters. MAKER2 can therefore be run on systems as simple as a laptop or as complex as a computer cluster with thousands of processors.

Input to MAKER2

MAKER2 accepts input files in FASTA and GFF3 formats. At the very least, a user is expected to supply MAKER2 with a genome sequence file, an EST dataset, and a protein homology dataset. Datasets in FASTA format will be aligned and processed by MAKER2, while datasets in GFF3 format (which contain pre-aligned features) will bypass many of these computational steps.

Because of the large number of potential annotation parameters and input files, it would be inconvenient to configure MAKER2 using only command line options. MAKER2 therefore uses a set of three default configuration files to setup options for genome annotation, these files are: `maker_opts.ctl` (contains parameters most likely to be configured by the user, i.e., the location of the all input FASTA and GFF3 files), `maker_bopts.ctl` (contains filtering statistics for the programs BLAST[2] and Exonerate[3] which MAKER2 uses to process evidence alignments), and `maker_exe.ctl` (contains the path information to prerequisite executables that MAKER2 uses to analyze input datasets). The annotation configuration files are run-specific meaning they must be generated and configured anew for each genome to be annotated.

MAKER2's output

The output format for MAKER2 produced genome annotations is GFF3; additionally, protein and transcript sequence from final gene models is produced in FASTA format (which is the *de facto* standard for sequence data used throughout biology). GFF3 is the file format commonly used by the Generic Model Organism Database project (GMOD)[4]. By using GFF3, MAKER2's output becomes automatically compatible with the extensive library of GMOD tools already available for easy downstream analysis of genome annotations.

Step-by-step overview of MAKER2

As a software pipeline, MAKER2 leverages the capabilities of many external tools, and integrates their output to produce final results. Because of the large number of steps and interactions involved I have created a flowchart (Figure 6.1) to help assist in my explanation of how exactly the MAKER2 algorithm is designed to work.

Dataset initialization in MAKER2

When MAKER2 starts, it searches for and loads the three default annotation control files I described earlier (`maker_opts.ctl`, `maker_bopts.ctl`, and `maker_exe.ctl`). From these files MAKER2 extracts the location of any input datasets for the pipeline. FASTA format files that are supplied to MAKER2 are parsed for errors and then divided into smaller datasets (each file is split into 10 smaller files). Splitting the files this way allows for faster indexing and rapid access to entries within the file. Having large files split into smaller datasets also permits MAKER2 to improve performance as each small dataset can be distributed to a separate CPU and

processed in parallel. For GFF3 format files, MAKER2 checks the entries for errors and loads them into an SQLite database. This is done because these datasets contain pre-aligned features, so they do not need to be processed and aligned the way FASTA files do. MAKER2 only requires rapid indexed access to the entries in the GFF3 file, and this is achieved via the database.

Parallelization of MAKER2

MAKER2 supports parallelization via Message Passing Interface (MPI), a distributed communication protocol. As opposed to other parallelization method MPI has the advantage of splitting threads between systems, meaning the same program can run on multiple computers (treating them all as if they were a single powerful machine). While MPI compatible programs can be installed and run on laptop and desktop machines, there are systems that are specially optimized for this mechanism of parallelization. These systems are referred to as computer clusters, and they can contain thousands of CPUs on hundreds of interconnect computers working in tandem.

While parallelization via MPI does greatly expand the number of CPUs available for computation, it also introduces many design constraints. Because two threads of the same program might not be running on the same computer, it cannot be assumed that they will have access to each other's memory or even each other's hard drives. Every effort must therefore be made to divide computation into small jobs that can be fully self-contained (i.e., each thread will not need to know the results of another thread). These small jobs can then run from start to finish while minimizing the need for communication and coordination between threads.

Once initialized, MAKER2 begins to divide datasets among CPUs depending on the number of processors indicated by the user. MAKER2 uses a host/client hierarchy for distributing data. Under this model one thread is dedicated to receiving and answering requests (the host) while all other threads take instructions and ask for datasets to work with from the host (the clients). MAKER2 at first attempts to divide contigs from the input genomic sequence among all CPUs; in this way, each CPU can analyze a single contig from start to finish (see Figure 6.1, blue parallelization arrows). This allows MAKER2 to avoid, for the most part, any issues with synchronizing annotation steps between CPUs. However, when the host thread runs out of contigs to distribute, client threads are given the option to break contigs currently in their possession into smaller chunks (using genomic sequence). Client threads then distribute those sequence chunks to other threads (see Figure 6.1, green parallelization arrows).

These chunks are usually 100,000 base pairs in length (this can be configured in the MAKER2 control files). Normally each client thread processes contigs one chunk at a time, running all annotation analyses on that sequence region to completion and then moving on to the next chunk (this is like walking across the contigs in 100,000 base pair steps, and it helps minimize the memory footprint of MAKER2). When a client thread has been told by the host that it is ok to distribute chunks to other clients, then that thread will walk down the contig distributing chunks until there is no more sequence or there are no more clients. Because all annotation steps on each chunk are run to completion on the same thread, overhead related to synchronizing analyses is largely avoided just as with the earlier contig distribution step.

As a MAKER2 annotation job advances to completion, the host and client nodes eventually run out of both contigs and sequence chunks for distribution. During these final stages of analysis, MAKER2 will attempt to finish annotation by moving to a third level of parallelization. Because most of the compute time in MAKER2 is spent aligning experimental evidence from ESTs and protein sequence via BLAST[2], MAKER2 can divide the BLAST databases among the threads, thus splitting up the work (see Figure 6.1, purple parallelization arrows). Distributing analyses in this way is facilitated by the fact that FASTA datasets were divided into smaller files during MAKER2 initialization. MAKER2 can then just distribute those datasets among the different threads. Unlike previous parallelization steps, though, MAKER2 now requires synchronization between threads, which will induce a performance penalty. However this third level of parallelization is only encountered toward the end of whole genome analysis so the resulting total overhead is minimal compared to total runtime. As a result MAKER2's performance scales almost linearly without noticeable loss in data throughput as the number of CPUs is increased (this can be seen in Chapter 2, Figure 2.6).

Repeat masking

The first step to MAKER2 is repeat masking; but why do we need to do this? Repetitive elements can make up a significant portion of a eukaryotic genome. Many of these elements are simple/low-complexity repeats of C's or G's or even dinucleotide repeats. Other repeats are more complex (i.e., transposable elements). These high-complexity repeats often encode real proteins like retrotranscriptase or even Gag, Pol, and Env viral proteins. Because repeats can encode real proteins, they can play havoc with *ab initio* gene-predictors. For example, a transposable

element that occurs next to or even within the intron of a ‘real’ protein-encoding gene might cause a gene-predictor to include extra exons as part of a gene model. These exons, however, really only belongs to the transposable element and not to the coding sequence of the gene. Other repeats may even be annotated as being individual genes. Because repeat families can exist in high copy numbers (numbering into the thousands), the final count of gene predictions can then be inflated by orders of magnitude, thus washing out the signal of the organisms true proteome in a flood of nearly identical bad gene calls.

In addition to issues of false gene prediction, repeats can also induce problems of false cross-species homology. Low-complexity repeat regions tend to align with certain kinds of experimental evidence such as structural proteins (which often have repetitive amino acid sequence to induce proper folding). While the length of these alignments can result in high statistical significance, closer analyses reveals that the alignment is more likely the result of random expansion of repetitive sequence. These types of repetitive sequence therefore create spurious protein and EST alignments throughout the genome, which undermines the accuracy of the annotation process.

To avoid these complications, it is convenient to identify and mask all repeat elements as the very first step of genome annotation. MAKER2 does this as a two-part process. First, a program called RepeatMasker[5] is used to identify low-complexity and high-complexity repeats that match entries in either the RepBase[6] repeat library or any species specific repeat library supplied by the user (this is an analysis in nucleotide space). Next, MAKER2 uses RepeatRunner[7] to identify transposable element and viral proteins from the RepeatRunner protein database. Because protein sequences diverge at a slower rate than nucleotide sequence, this

step helps pick up the problematic regions of divergent repeats that may have been missed by RepeatMasker analysis in nucleotide space.

Regions identified during repeat analysis are masked out so as not to complicate other downstream steps of the annotation process. High-complexity repeats are hard-masked, a technique in which nucleotide sequence is replaced with the letter N to prohibit any alignments to that region. Low-complexity regions are soft-masked, a technique in which nucleotides are made lower case so they can be treated as masked under certain situations without losing sequence information[8].

The idea of masking out sequence may appear, at first, as if a great deal of information is being lost. It is after all true that proteins exist which have integrated repeats into their true structure, thus repeat masking will affect one's ability to annotate these proteins. However, such proteins are rare, and the number of gene models and homology alignments improved by this step far exceed the few gene models that may be negatively impacted. However, if users are concerned about the effect that repeat masking will have on gene model sensitivity (i.e., false negatives), they have the option to run *ab initio* gene-predictors on both masked and unmasked sequences by setting the 'unmask' parameter in the maker_opts.ctl file to 1.

Ab initio gene prediction

Following repeat masking, MAKER2 runs *ab initio* gene-predictors specified by the user to produce preliminary gene models. *Ab initio gene-predictors* produce gene predictions based on underlying mathematical models describing patterns of intron/exon structure, codon usage, and consensus start signals. Because these aspects of gene structure differ from organism to organism, gene-predictors must be trained before they can be used on a newly sequence genome. Training requires pre-

existing gene models, which usually do not exist for newly sequenced organisms, but MAKER2 is capable of producing rough gene models using evidence alignments. These rough models can then be used for initial training of algorithms like SNAP[9] (trained according to its internal documentation). See the section in Chapter 2 on 'Gene prediction/annotation in second-generation genomes' for more explanation of issues related to training gene-finders.

MAKER2 currently supports the following gene-prediction programs: SNAP, Augustus[10], GeneMark[11], and FGENESH[12]. However MAKER2 can also accept gene-predictions produced by other programs if they are supplied in GFF3 format to the `pred_gff` options of the `maker_opts.ctl` file.

EST and protein evidence alignment

A simple way to indicate if a sequence region contains a gene is to identify (A) if the region is actively being transcribed or (B) if the region has homology to a known protein. This can be done by aligning Expressed Sequence Tags (ESTs) and proteins to the genome using alignment algorithms like BLAST.

ESTs are sequences derived from a cDNA library. Because of the difficulties associated with working with mRNA and depending on how the cDNA library was prepared, EST databases usually represent bits and pieces of transcribed mRNAs (with only a few full length transcripts). MAKER2 aligns these sequences to the genome using BLASTN. If ESTs from the organism being annotated are unavailable or sparse, then ESTs from a closely related organism can be used. However, ESTs from closely related organisms are unlikely to align using BLASTN since nucleotide sequences can diverge quite rapidly. For these divergent ESTs, MAKER2 uses TBLASTX to align them in protein space. Protein sequence generally diverges quite

slowly and over large evolutionary distances; as a result, proteins from even evolutionarily distant organisms can often be used to identify regions of homology. MAKER2 uses BLASTX to align protein datasets.

As stated previously evidence is being aligned against the repeat-masked genomic sequence (remember repeat masking was the first step of the analysis to avoid spurious false positive alignments). One of the effects of masking is that sequences will not be able to align against low-complexity regions. Unfortunately some real proteins do contain low-complexity regions (and they are not all that uncommon). It would be beneficial if there were a way to identify regions of true low-complexity homology without opening up the entire genome to spurious alignments. Fortunately the program BLAST provides just such a mechanism. Soft-masking is the use of lower case letters to identify a region as masked without losing the sequence information. BLAST can restrict any alignments from being seeded in these soft masked regions, thus avoiding spurious hits in the BLAST report. But at the same time, BLAST can allow alignments that have already reached statistical significance outside the masked region to extend through them. This means proteins with true homology can align using nonrepetitive conserved domains within their structure and then capture true repetitive regions through extension. Proteins without true homology, though, should not align because the only regions of similarity are the repetitive regions. There will be no conserved region to seed an initial alignment, and extension is thus impossible. Setting 'softmask' to 0 in the maker_opts.ctl file, however, can turn off this behavior.

Polishing evidence alignments

Because of oddities associated with how BLAST statistics work, BLAST alignments are not as informative as they could be. BLAST tries to align sequence anywhere it can, in every way it can (i.e., reversing the strand, changing the alignment order, aligning in fragments, or allowing large gaps). This can become especially confusing for genes with repeated domains (they will align to multiple places and multiple times), or genes with neighboring paralogous duplications (alignments will bridge duplications every way imaginable even splicing together unrelated exons from separate genes). To get more informative alignments, MAKER2 uses the program Exonerate[3] to polish BLAST hits. MAKER2 uses Exonerate to realign each sequence identified by BLAST around splice sites and forces the alignments to occur in order. The result is a high quality alignment that can be used to suggest near exact intron/exon positions. Polished alignments are produced using the `est2genome` and `protein2genome` options for the program Exonerate.

One of the benefits of polishing EST alignments is the ability to identify the proper strand an EST derives from. Because of amplification steps involved in building a cDNA library and limitations involved in some high-throughput sequencing technologies, there is no way of telling which strand a sequenced EST belongs to. However, if canonical splice sites are taken into account, the EST can only align to one strand correctly.

Integrating evidence to synthesize annotations

Once MAKER2 has gather evidence from *ab initio* gene predictions, EST alignments, and protein homology, the combined results can be integrated and

evaluated to produce even better gene models. MAKER2 does this by ‘talking’ to gene-prediction programs like SNAP, Augustus, and FGENESH and providing them ‘hints’ as to the most probable locations of introns, exons, splice sites, and coding regions. The evidence-based ‘hints’ are provided in parameter files that are accepted by these algorithms.

Selecting and revising the final gene model

MAKER2 next takes the entire pool of *ab initio* and evidence informed gene-predictions that correspond to a given locus, and updates them with features such as five and three prime UTRs based on EST evidence overlap. The pipeline then tries to determine alternative splice forms where EST data permits (i.e., using long ESTs with mutually exclusive intron/exon patterns). Finally MAKER2 chooses from among all the gene model possibilities the one that best matches the evidence. This is done using the modified sensitivity/specificity distance metric AED[13] from the Sequence Ontology (also see ‘Calculating Annotation Edit Distance’ in the Methods section of Chapter 2). The gene with the lowest AED score is the best match to the evidence.

Quality control statistics

AED is calculated using the standard sensitivity (SN) and specificity (SP) equations[14] with the only difference being that the reference is the aligned experimental evidence rather than a high quality gene model. When calculating AED, the base pair level sensitivity equation is $SN = |i \cap j| / |j|$; where $|i \cap j|$ represents the number of overlapping nucleotides between a gene prediction i and the aligned evidence j , and $|j|$ represents the total number of nucleotides in the

evidence. Specificity is calculated using the equation $SP = |i \cap j| / |j|$. The average of sensitivity and specificity is *congruency*, where $C = (SN+SP)/2$. AED represents the *incongruency*, or distance between i and j , using the equation $AED = 1-C$. An AED value of 0 indicates no distance between the evidence and an annotation, so the annotation is in perfect agreement with aligned evidence. In contrast a value of 1 means no evidence support.

In addition to AED, MAKER2 also calculates other metrics that summarize how different evidence types support gene annotations. These are MAKER2's Quality Indices (QI) and are reported for each gene as part of the GFF3 output. Statistics calculated for the Quality Indices are shown in Figure 6.2. These metrics are invaluable resource and assist in future downstream management and curation of gene models.

Algorithm stability and error handling

Because MAKER2 is expected to operate on large, genome-wide datasets, computation times can range from several hours to several days. With such long run times, issues unrelated to the algorithm itself can arise (i.e., power failures, user error, etc.). Stability of the program is therefore a critical issue. MAKER2 uses log files to automatically register the progress of each analysis MAKER2 performs. In this way, the program can restart where it left off without unnecessary reprocessing of existing data. Also, when there is a failure processing any of the contigs from the genome input file, MAKER2 skips to the next contig and continues processing. Failed contigs are then retried at the end of the whole-genome analysis, and any contigs that fail multiple times are separated into external files for user review. The robustness of this mechanism allows a researcher to start MAKER2 and walk away

with little time spent monitoring the status of the compute. The log files produced by MAKER2 as part of its normal operation also serve as indicators of what analyses should be re-run or maintained if a user decides to restart annotation using different parameters. MAKER2 is, thus, highly efficient and avoids redoing any analyses unnecessarily.

References

1. **GFF3** [<http://www.sequenceontology.org/gff3.shtml>]
2. Altschul SF, Gish W, Miller W, Meyers EW, Lipman DJ: **Basic Local Alignment Search Tool**. *Journal of Molecular Biology* 1990, **215**:403-410.
3. Slater G, Birney E: **Automated generation of heuristics for biological sequence comparison**. *BMC Bioinformatics* 2005, **6**:31.
4. **GMOD** [<http://www.gmod.org>]
5. **RepeatMasker** [<http://repeatmasker.org>]
6. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: **Rebase Update, a database of eukaryotic repetitive elements**. *Cytogenetic and Genome Research* 2005, **110**:462-467.
7. Smith CD, Edgar RC, Yandell MD, Smith DR, Celniker SE, Myers EW, Karpen GH: **Improved repeat identification and masking in dipterans**. *Gene* 2007, **389**:1-9.
8. Frith M, Hamada M, Horton P: **Parameters for accurate genome alignment**. *BMC Bioinformatics*, **11**:80.
9. Korf I: **Gene finding in novel genomes**. *BMC Bioinformatics* 2004, **5**:59.
10. Stanke M, Waack S: **Gene prediction with a hidden Markov model and a new intron submodel**. *Bioinformatics* 2003, **19**:ii215-225.
11. Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M: **Gene identification in novel eukaryotic genomes by self-training algorithm**. *Nucl Acids Res* 2005, **33**:6494-6506.
12. Salamov AA, Solovyev VV: **Ab initio gene finding in drosophila genomic DNA**. *Genome Res* 2000, **10**:516-522.

13. Eilbeck K, Moore B, Holt C, Yandell M: **Quantitative measures for the management and comparison of annotated genomes.** *BMC Bioinformatics* 2009, **10**:67.
14. Guigo R, Dermitzakis ET, Agarwal P, Ponting CP, Parra G, Reymond A, Abril JF, Keibler E, Lyle R, Ucla C, et al: **Comparison of mouse and human genomes followed by experimental verification yields an estimated 1,019 additional genes.** *Proceedings of the National Academy of Sciences* 2003, **100**:1140-1145.

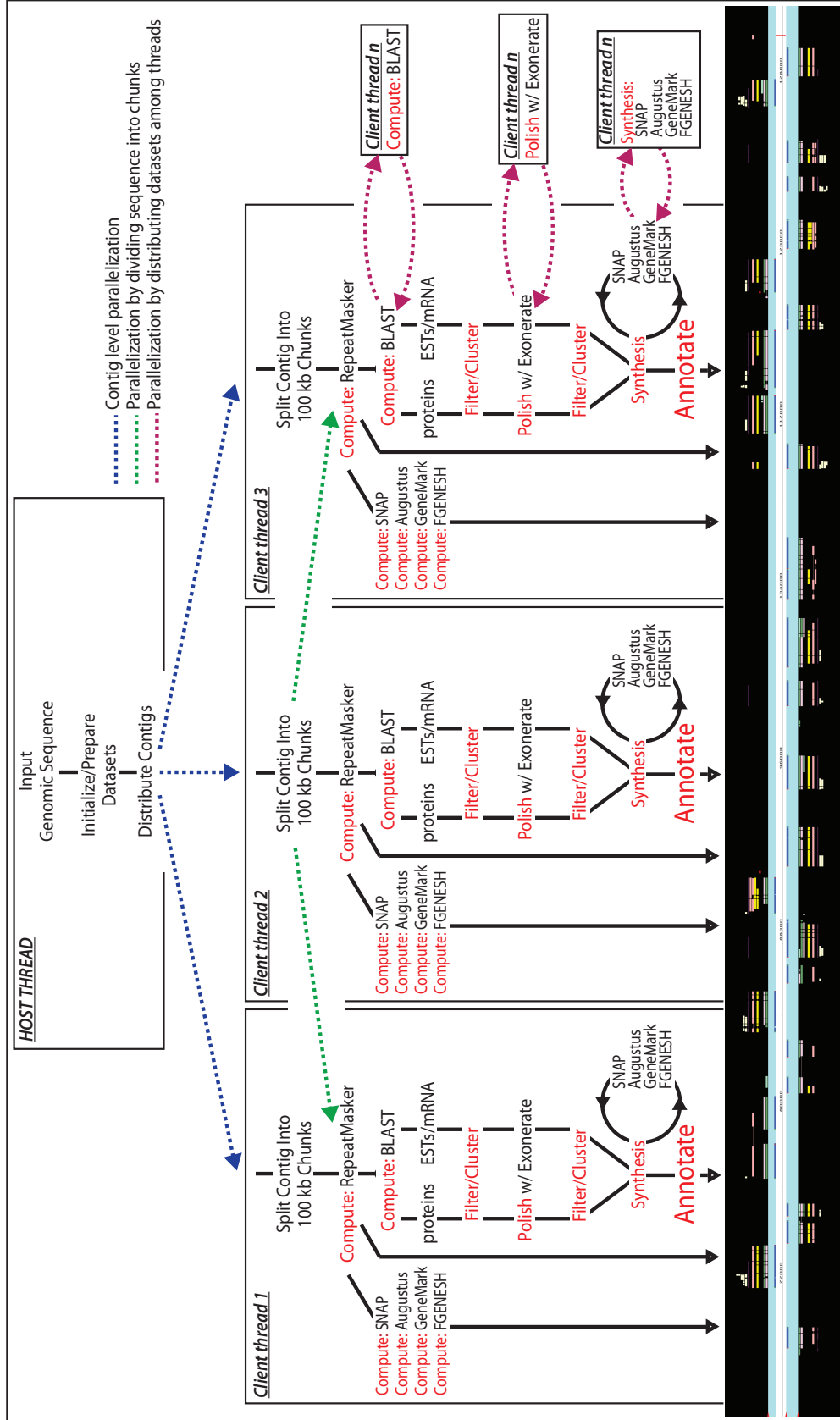


Figure 6.1 Flowchart of MAKER2's design and operation. MAKER2 is an automated software pipeline that passes data from one external program to another (each basic step of annotation is shown in red). MAKER2 parallelizes annotation by dividing contigs among threads (blue arrows), dividing sequence into smaller chunks (green arrows), and dividing individual analyses among threads (purple arrows).

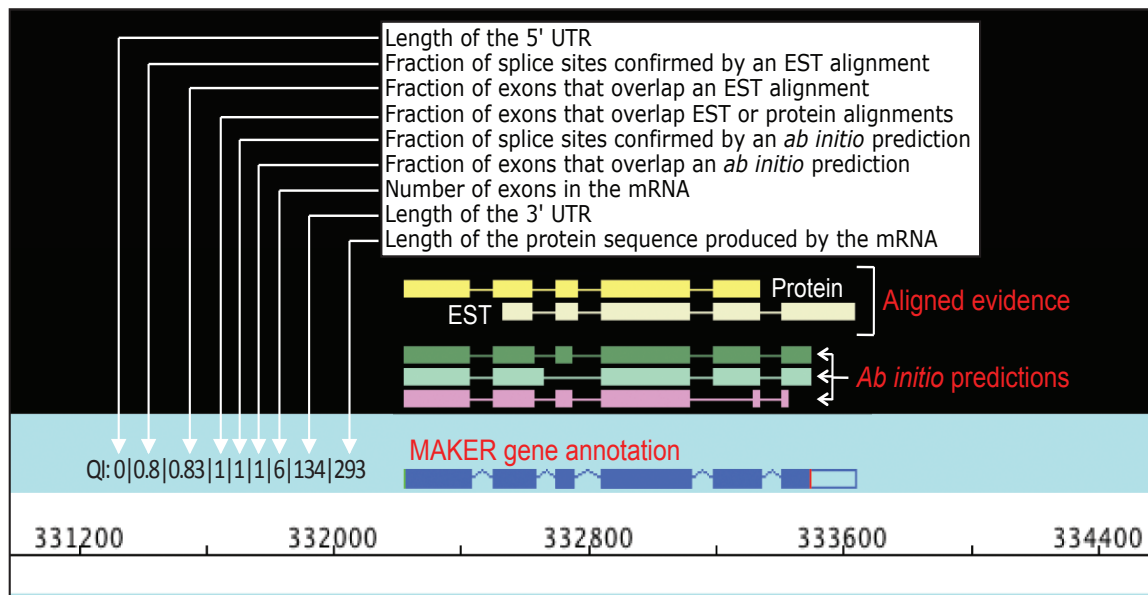


Figure 6.2 Summary of MAKER2's Quality Indices.

MAKER2's Quality Indices provide basic summary statistics for how well gene annotations are supported by different types of experimental and computationally derived evidence.

CHAPTER 7

SUMMARY AND CONCLUSIONS

MAKER2 is seeing widespread adoption by the scientific community. There are over 700 research projects around the world actively using MAKER2 (registered on Yandell-lab.org), and already over a dozen published genomics manuscripts have listed MAKER2 as an integral tool in their analyses[1-13]. There will undoubtedly continue to be a need for MAKER2 as sequencing costs continue to fall, thus making it possible for even individual labs to sequence and annotate entire genomes.

In the course of my thesis work, I have had the opportunity to develop new techniques and methods in order to solve the annotation challenges presented by the different genomes I have worked with and annotated. This has produced the robust and powerful pipeline that MAKER2 has become today. My many annotation collaborations have contributed critically to the biological understanding of multiple new model organisms, and the annotation databases I have helped to create are currently informing downstream experimental work in hundreds of laboratories around the world.

References

1. Suen G, Teiling C, Li L, Holt C, Abouheif E, Bornberg-Bauer E, Bouffard P, Caldera EJ, Cash E, Cavanaugh A, et al: **The genome sequence of the leaf-cutter ant *Atta cephalotes* reveals insights into its obligate symbiotic lifestyle.** *PLoS Genet* 2011, 7:e1002007.

2. Smith CR, Smith CD, Robertson HM, Helmkampf M, Zimin A, Yandell M, Holt C, Hu H, Abouheif E, Benton R, et al: **Draft genome of the red harvester ant *Pogonomyrmex barbatus***. *Proceedings of the National Academy of Sciences* 2011, **108**:5667-5672.
3. Smith CD, Zimin A, Holt C, Abouheif E, Benton R, Cash E, Croset V, Currie CR, Elhaik E, Elsik CG, et al: **Draft genome of the globally widespread and invasive Argentine ant (*Linepithema humile*)**. *Proceedings of the National Academy of Sciences* 2011, **108**:5673-5678
4. Levesque CA, Brouwer H, Cano L, Hamilton J, Holt C, Huitema E, Raffaele S, Robideau G, Thines M, Win J, et al: **Genome sequence of the necrotrophic plant pathogen *Pythium ultimum* reveals original pathogenicity mechanisms and effector repertoire**. *Genome biology* 2010, **11**:R73.
5. Baxter SW, Nadeau NJ, Maroja LS, Wilkinson P, Counterman BA, Dawson A, Beltran M, Perez-Espona S, Chamberlain N, Ferguson L, et al: **Genomic hotspots for adaptation: the population genetics of Mullerian mimicry in the *Heliconius melpomene* clade**. *PLoS Genet* 2010, **6**:e1000794.
6. Ferguson L, Lee SF, Chamberlain N, Nadeau N, Joron M, Baxter S, Wilkinson P, Papanicolaou A, Kumar S, Kee T-J, et al: **Characterization of a hotspot for mimicry: assembly of a butterfly wing transcriptome to genomic sequence at the HmYb/Sb locus**. *Molecular Ecology* 2010, **19**:240-254.
7. Kovach A, Wegrzyn J, Parra G, Holt C, Bruening G, Loopstra C, Hartigan J, Yandell M, Langley C, Korf I, Neale D: **The *Pinus taeda* genome is characterized by diverse and highly diverged repetitive sequences**. *BMC Genomics* 2010, **11**:420.
8. MacDonald J, Doering M, Canam T, Gong Y, Guttman DS, Campbell MM, Master ER: **Transcriptomic responses of the softwood-degrading white-rot fungus *Phanerochaete carnosae* during growth on coniferous and deciduous wood**. *Appl Environ Microbiol* 2011:AEM.02490-02410.
9. Legeai F, Shigenobu S, Gauthier JP, Colbourne J, Rispé C, Collin O, Richards S, Wilson ACC, Murphy T, Tagu D: **AphidBase: a centralized bioinformatic resource for annotation of the pea aphid genome**. *Insect Molecular Biology* 2010, **19**:5-12.
10. Martin J, Abubucker S, Wylie T, Yin Y, Wang Z, Mitreva M: **Nematode.net update 2008: improvements enabling more efficient data mining and comparative nematode genomics**. *Nucleic acids research* 2009, **37**:D571-D578.

11. Robb S, Ross E, Alvarado A: **SmedGD: the Schmidtea mediterranea genome database.** *Nucleic Acids Res* 2007;D599 - 606.
12. Wurm Y, Wang J, Riba-Grognuz O, Corona M, Nygaard S, Hunt BG, Ingram KK, Falquet L, Nipitwattanaphon M, Gotzek D, et al: **The genome of the fire ant *Solenopsis invicta*.** *Proceedings of the National Academy of Sciences* 2011, **108**:5679-5684.
13. Hauser PM, Burdet FX, Cisse OH, Keller L, Taffe P, Sanglard D, Pagni M: **Comparative genomics suggests that the fungal pathogen *Pneumocystis* is an obligate parasite scavenging amino acids from its host's lungs.** *PLoS ONE* 2010, **5**:e15152.