Resource

Evidence-based gene predictions in plant genomes

Chengzhi Liang,[1,2,4] Long Mao,[2] Doreen Ware,[1,3] and Lincoln Stein[1]

[1]*Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA;* [2]*Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing 100081, China;* [3]*USDA Robert Holley Center, NAA Plant, Soil, and Nutrition Laboratory, Ithaca, New York 14850, USA*

Automated evidence-based gene building is a rapid and cost-effective way to provide reliable gene annotations on newly sequenced genomes. One of the limitations of evidence-based gene builders, however, is their requirement for transcriptional evidence—known proteins, full-length cDNAs, or expressed sequence tags (ESTs)—in the species of interest. This limitation is of particular concern for plant genomes, where the rate of genome sequencing is greatly outpacing the rate of EST- and cDNA-sequencing projects. To overcome this limitation, we have developed an evidence-based gene build system (the Gramene pipeline) that can use transcriptional evidence across related species. The Gramene pipeline uses the Ensembl computing infrastructure with a novel data processing scheme. Using the previously annotated plant genomes, the dicot *Arabidopsis thaliana* and the monocot *Oryza sativa*, we show that the cross-species ESTs from within monocot or dicot class are a valuable source of evidence for gene predictions. We also find that, using only EST and cross-species evidence, the Gramene pipeline can generate a plant gene set that is comparable in quality to the human genes based on known proteins and full-length cDNAs. We compare the Gramene pipeline to several widely used ab initio gene prediction programs in rice; this comparison shows the pipeline performs favorably at both the gene and exon levels with cross-species gene products only. We discuss the results of testing the pipeline on a 22-Mb region of the newly sequenced maize genome and discuss potential application of the pipeline to other genomes.

[Supplemental material is available online at http://www.genome.org. The Gramene pipeline software packages, all gene product data sets, and the full-length complementary DNA (FLcDNA)-based standard genes in *A. thaliana* and *O. sativa* are available at ftp://ftp.gramene.org/pub/gramene/genebuild.]

The prediction of protein-coding genes is one of the most critical steps in genome annotations. As shown in the EGASP assessment of gene prediction algorithms in humans, while only a small portion of the human genes are missed by computational predictions, the best gene prediction systems are able to predict entirely correct gene structures only 50% of the time (Guigo et al. 2006). Therefore, increasing the accuracy of predicted protein-coding genes remains a key goal. Another finding from the EGASP study is that the most accurate gene prediction systems are those that use transcriptional evidence—such as sequenced proteins, expressed sequence tags (ESTs), and full-length complementary DNAs (FLcDNAs)—to identify genes and deduce their splicing patterns. This technique has become the mainstay of gene structure annotation predictions for organisms that have transcriptional data available.

In plants, two sequenced model organisms (*Arabidopsis thaliana* and *Oryza sativa*) have been annotated using a mixture of evidence-based gene models and ab initio predictions (Ouyang et al. 2007; Zhu and Buell 2007; Swarbreck et al. 2008). However, while it is clear that evidence-based gene builds perform better than ab initio systems on average, there has been no systematic study of the accuracy of evidence-based gene prediction when given different balances of FLcDNAs, ESTs, and proteins. It is also unknown how effective it is to use expression data from one species to derive genes in a closely related species. In plant genomics, this is a particularly important question, because in most species the existing FLcDNA sets are small due to the high expense of sequencing FLcDNAs; most expression data come from smaller EST sequencing projects or from cross-species expression sets.

The Ensembl gene build pipeline (Curwen et al. 2004) is an accurate evidence-based gene prediction protocol that has been validated in multiple animal species. The process uses the Ensembl computing infrastructure, which contains automated job management for efficient data processing in conjunction with a software application programming interface (API) for easy data management and visualization. This pipeline begins by aligning known proteins to predict gene structures in coding regions and proceeds to use FLcDNAs and ESTs to add untranslated regions (UTRs) and FLcDNA-based genes in empty regions. Ensembl also provides an independent EST-based gene build (Eyras et al. 2004), but they were mainly used to determine possible alternative splicing of predicted genes.

Gramene (http://www.gramene.org) is a database that supports comparative genome mapping among multiple plant species (Liang et al. 2008). To provide a suitable platform for this endeavor, we must generate consistent gene sets for each plant genome using a standardized gene prediction system. (Throughout the remainder of this article, we refer to protein-coding genes simply as "genes" for the sake of brevity.) Our gene build pipeline is based on Ensembl, but we introduce a new data processing scheme to make it more suitable for plant genomes. In this article, we evaluate the accuracy of the Gramene gene build pipeline with various combinations of plant same-species and cross-species expression sets. We also provide information on our application of this pipeline to the new maize genome sequence. The analyses will provide a practical guideline for gene annotations using incomplete or cross-species gene products (as low-confidence evidences) in genomes lacking species-specific FLcDNAs and known proteins.

[4]Corresponding author.
E-mail liang@cshl.edu; fax (516) 367-6851.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.088997.108.

19:1912–1923; ISSN 1088-9051/09; www.genome.org

Results

Data processing pipeline workflow

The Gramene gene build is based on the Ensembl pipeline, but with several important modifications. First, while Ensembl uses GeneWise (Birney et al. 2004) for protein-to-genome alignments and Exonerate (Slater and Birney 2005) for cDNA/EST alignments, we use Exonerate for both tasks, due to the flexibility of its multiple alignment models and its ability to associate each alignment with a sequence identity score (see below). Although we have observed that GeneWise is in some cases slightly better than Exonerate for low-identity cross-species proteins (data not shown), in practice choosing Exonerate for protein alignment has little effect on plant genomes because of the small number of independently derived plant protein data sets.

A second difference between the Gramene and Ensembl pipelines is that the former makes heavy use of cross-species cDNAs (including ESTs), which are the major source of gene evidence for less studied organisms, such as sorghum. This is a significant departure from the original Ensembl pipeline.

Finally, the Gramene pipeline keeps the Exonerate alignment score of each raw transcript, thereby allowing us to rank each alternative splice form according to the quality of the evidence supporting it. The design principle of the Gramene pipeline is to select high-confidence transcripts first and then use low-confidence transcripts to improve those of higher confidence if necessary. The Gramene pipeline retains all high-confidence alternatively spliced transcripts based on the alignment score at the transcript level. In the original Ensembl pipeline, the number of predicted alternative splicing forms can become quite large, and it is difficult to distinguish well-supported from poorly supported forms. In contrast, the Gramene pipeline can dynamically adjust the threshold for predicting a splice form and can apply different thresholds to different classes of genes. When reviewing gene predictions, users can easily determine the quality of a gene based on its supporting evidence and aligned sequence identity.

The major steps of the Gramene pipeline data processing are shown in Figure 1. The pipeline begins with the mapping of gene products (FLcDNAs, ESTs, protein sequences) to the genome using Exonerate to create a raw transcript set for each evidence type (for details, see Methods). Exonerate uses various alignment models depending on data type. It aligns species-specific FLcDNAs and ESTs using DNA-to-DNA alignment; it aligns cross-species FLcDNAs and EST using translated DNA-to-translated DNA alignment; and it aligns proteins using protein-to-translated DNA alignment. All models contain a built-in intron model to account for the spliced introns in the alignments. We routinely repeat-mask the genome, but this step is optional. Gene products can be grouped arbitrarily by the application of different processing filters. For example, we can separate species-specific proteins from cross-species proteins to adjust the alignment threshold applied to the two sets. After mapping, we process each raw transcript set using the following strategies.

First, we filter each set to remove all transcripts that have poor alignment scores. We generally use a sequence identity threshold of 90% for same-species alignment and of 30% (protein sequence similarity) for cross-species alignments. We use a higher threshold (e.g., 99%) for single-exon alignments of same-species ESTs to reduce genomic DNA contaminants. We also attempt to detect and correct incorrectly assigned strands, as EST data often contain a mixture of sense and antisense gene products. In the case of multi-exon genes, we use the splice site consensus sequences to detect and correct strand mapping errors. For example, if a predicted gene has multiple "CT|AC" splice sites, which are the reverse complement of the canonical "GT|AG" sites, we automatically change the entire transcript to the opposite strand. This strategy is not feasible for single-exon transcripts, however. In such cases, we either keep the strand Exonerate assigns, or we flip the strand if there is a preponderance of transcript evidence supporting a gene model on the opposite strand.

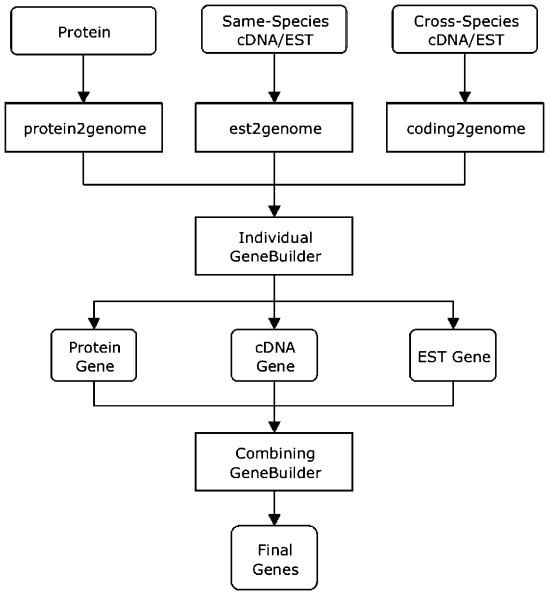

Figure 1. Overview of gene build data processing in the Gramene pipeline.

For transcripts with <99.5% alignment identity to the underlying EST, FLcDNA, or protein after initial filtering, we perform intron correction. We use overlapping "GT|AG" introns to correct the introns with noncanonical splicing sites if their boundaries are within a short distance (up to 25 nucleotides [nt], based on alignment score). Very long introns of low-confidence transcripts are cut if there exists another transcript that has exons in the region covered by the long intron—this is often due to mapping errors in tandemly duplicated genes. Short introns with noncanonical splicing sites are removed if the intron is covered by an exon in other transcripts and the open reading frame (ORF) can be maintained. After these steps, transcripts with alignment identity <99.5% that have too many introns with noncanonical splicing sites are removed; these transcripts generally come from paralogous or cross-species gene products. This usually limits the number of predictions with noncanonical splicing sites to <2%–4%, close to the ratio among FLcDNA-confirmed introns in *A. thaliana* and rice (Sparks and Brendel 2005).

Next we process the predicted transcripts to merge partial gene models and to remove those that are redundant. This is a critical step, as gene expression evidence is highly redundant by nature. For example, many ESTs are fragments originating from the same gene. We first remove transcripts that are completely covered by others. We then merge overlapping transcript models that have no incompatible introns (i.e., overlapping introns with different boundaries). If two transcripts share at least one exon, and all overlapping introns are identical, they are merged into a single

transcript. Alternatively spliced transcripts are preserved if their sequence identity level is above a predefined threshold (typically 99.5%). We always set this threshold higher than the typical rice FLcDNA mapping identity (95%) used in other studies (e.g., Satoh et al. 2007; Zhu and Buell 2007), to limit the overall number of alternative splicing forms.

The low-confidence transcripts are utilized in two ways. First, if a low-confidence transcript extends the ends of a partial high-confidence transcript, we will use it to extend the ends. Second, if a low-confidence transcript does not overlap a high-confidence gene model at all, we will retain it in the final gene set. Low-confidence transcripts that overlap with high-confidence models with incompatible introns between them are removed to limit the number of artefactual alternative splicing predictions that result from the gain and lost of splice sites during evolution.

After processing, we obtain a set of predicted transcripts for each type of evidence. We mix these sets and reprocess them as described above. The processed nonredundant transcripts are classified by their translational potential arbitrarily. We typically require a minimum ORF length of 50 amino acids (aa) as used in the method of Ouyang et al. (2007) for a transcript to be labeled as protein coding. We use less-stringent criteria for known protein-supported genes (e.g., 25 aa) and more-stringent criteria for single-exon transcripts with scant EST support (e.g., 100 aa). The remainders of the predicted transcripts that have a maximum ORF less than these values are labeled as either untranslated (e.g., <25 aa) or short proteins (which are usually single-exon transcripts derived from ESTs). All of these parameters are configurable.

We then subgroup the protein-coding transcripts into two categories by labeling transcripts with a coding sequence (CDS) that includes both start and stop codons as full-length transcripts; conversely, we label transcripts that lack either a start or stop codon as partial transcripts. A partial transcript that lacks a start codon must have its ORF begin within the first three nucleotides.

The last step of the pipeline is to group the protein-coding transcripts into genes, as in the original Ensembl pipeline: A gene is defined as a set of transcripts that together share at least one exon. A full-length gene is defined as a transcript set whose longest CDS is full-length. A partial gene is defined as a transcript set whose longest CDS is partial.

Evaluating the quality of predicted genes

To evaluate the quality of the genes predicted by the Gramene pipeline, we use the metrics developed by Ensembl (Curwen et al. 2004; Eyras et al. 2004), EGASP (Guigo et al. 2006), and the earlier GASP (Reese et al. 2000) gene prediction assessment tests to measure the sensitivity (Sn) and specificity (Sp) of the predicted genes. In addition to the metrics defined in those articles, we define the following terms: Two genes or exons *touch* if they share at least one nucleotide on the same strand. Two exons are *identical* if they have the same start and end coordinates. If two exons overlap, but the region of nonoverlapping involves a putative splice site, then they are called *different exons*. If two exons overlap, and the non-overlapping parts are beyond the sequence end of the transcript (or CDS) that the smaller exon is in, then the smaller is a *partial exon* and the longer is an *extended exon*. An *extended CDS* covers all the exons (either extended or identical with at least one extended) of another and does not have additional or different internal exons. A *partial CDS* is the counterpart of an extended CDS. If two CDSs (or transcripts) contain at least one different exon or a missed internal exon, then they are called *different* (or *incompatible*).

At the gene level, we use two different metrics. The first is the locus sensitivity measure used by Ensembl; it considers only the gene locus but not the gene structure. A gene is considered to be found if it overlaps a standard gene by at least one nucleotide. This will measure the number of missed genes (for the sensitivity test) or extra genes (for the specificity test). The second metric is derived from EGASP; it considers only genes with identical CDSs: Two genes are considered to be the same if and only if they contain at least one identical CDS. In our assessment, we did not directly compare CDSs, but instead we compared their translations. We call two CDSs identical if their protein translations are exactly the same. This is a very stringent measure of correctness. Predicted genes that are not identical can be extended, partial, or different depending on their CDS comparisons.

We also have made comparisons at the transcript level, which takes alternative splicing patterns into account. Such metrics, however, are less informative in this study than are CDS- or translation-level metrics, since none of the standard gene sets available to us have significant information on alternative splicing. We believe that the cutoff we use to select alternative transcripts (99%–99.5%) is more stringent than that used by other genome annotation projects to predict alternative transcripts. Thus, we will not discuss transcript-level results in this article. Interested users can find the results in the Supplemental material.

At the exon and base-pair level, we calculate the sensitivity and specificity of gene prediction using transcript pairs among overlapping genes only, following the method used by Ensembl with slight modifications. If two genes overlap each other by at least one nucleotide (on the same strand), we pair up all their transcripts and calculate the sensitivity and specificity of the identical exons within the matching transcripts. If one transcript is paired up with two or more nonoverlapping transcripts, we treat the exons from these nonoverlapping transcripts as from one transcript, largely to reduce the bias introduced by the split gene predictions that are originated from ESTs. A standard gene is considered split if it overlaps more than one predicted gene. Our exon-level assessment compares unique exons at the transcript level among overlapping genes, which is different than that used in EGASP, which compares only unique exons at the gene level and among all genes. However, our method gives more information about the exon structure of the predicted transcripts than does the EGASP method and reduces the bias introduced by the different degree of alternative splicing between the predicted gene sets. The base-pair-level comparison is based on exons only.

Gene builds using cross-species ESTs

Because there are abundant cross-species EST data from related species among the crop monocots, one of our major goals is to leverage this type of data in the Gramene pipeline. We applied the Gramene pipeline using cross-species plant ESTs on the repeat-masked *A. thaliana* (Swarbreck et al. 2008) and *O. sativa* ssp. *japonica* (rice) (Ouyang et al. 2007) genomes. We evaluated the sensitivity of the predicted gene sets against the FLcDNA-supported standard genes assembled as described in the Methods (11,378 and 11,785 genes for *A. thaliana* and *O. sativa*, respectively).

We first grouped the ESTs according to the taxonomy of their source organisms to generate a gene set for each group on both genomes. By comparing the sensitivity of the CDS predictions (see Supplemental Table S1), it is clear that the evolutionary distance between the EST source and the target genome plays an important role in the quality of the gene predictions. Most notably, ESTs from

dicot species are much more effective for predicting dicot genes than monocot ESTs are, and vice versa (also see below). Within the dicot and monocot classes, it seems that ESTs coming from very closely related species are more effective for accurate predictions than those from distant species. For example, the ESTs from *Brassica* and *Raphanus*, which are in the same taxonomic family as *A. thaliana* (all in tribe *Brassiceae*), give more correct genes in *A. thaliana* than similar numbers of ESTs from other dicot tribes. Nevertheless, adding ESTs from distant dicot taxonomic families to the ESTs from *Brassica* and *Raphanus* can still improve the predicted gene quality significantly. We expect that the gene prediction quality will also be influenced by the quality of the ESTs. However, the EST data quality is not readily available in many of these data sets, so we do not consider this criterion further.

Based on these observations, we grouped the ESTs into dicot and monocot categories. We used in total 5.13 million monocot ESTs and 7.88 million dicot ESTs. We evaluated the sensitivity of the predicted gene sets given increasing numbers of cross-species ESTs against the FLcDNA-supported standard genes. The results, shown in Figure 2, A (*A. thaliana*) and B (*O. sativa*), show that the predicted gene sets' coverage of confirmed genes ("di-touch" or "mo-touch") and CDS sensitivity ("di-same" or "mo-same") increase as a function of the number of cross-species ESTs made available to the pipeline, and begin to plateau as the number of mapped ESTs exceeds 3 million sequences. In *A. thaliana*, >94% of the genes in the confirmed set are touched by a predicted gene, with ~68% of the confirmed genes' structures predicted correctly across their entire coding region. In rice both the gene coverage (>89%) and CDS sensitivity (>50%) are reduced due to the smaller number of monocot ESTs available (or due to lack of within-tribe ESTs or both).

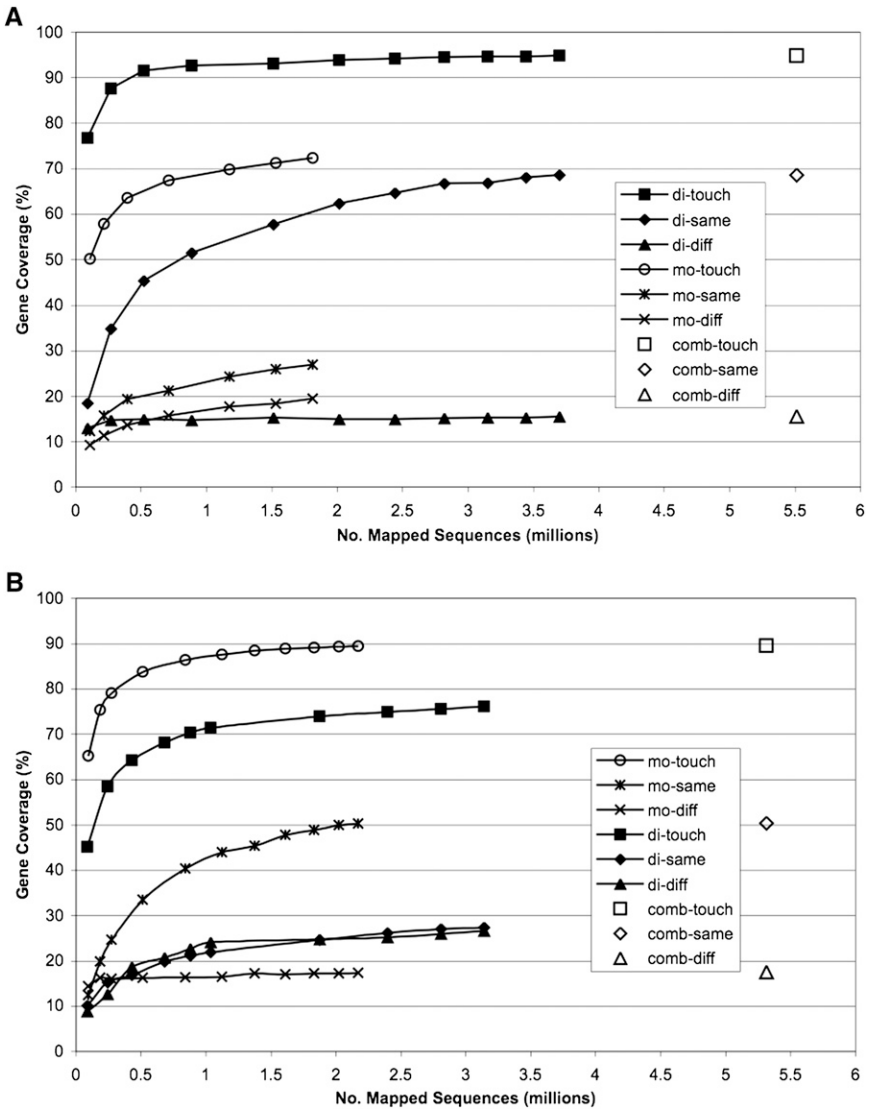

Figure 2. Gene build results using cross-species ESTs on *A. thaliana* (*A*) and *O. sativa* ssp. *japonica* (*B*). Mo- indicates using monocot ESTs; di-, using dicot ESTs; comb-, using combination of all monocot and dicot ESTs; touch, a predicted gene overlapping a standard gene on the same strand by at least one nucleotide (gene locus sensitivity); same, identical CDS between the predicted gene and the standard gene (gene CDS sensitivity); and diff, the predicted gene's CDS is different than that of the standard gene.

The within-class ESTs are much more effective in gene prediction than cross-class ESTs. When 2 million dicot ESTs are applied to the *A. thaliana* pipeline, >90% of the confirmed genes are touched, but <75% of the genes are touched when the same number of monocot ESTs are used. Similarly, in *O. sativa* 2 million monocot ESTs produce gene models that touch ~89% of confirmed genes, but <75% of the confirmed genes are touched when we attempt to use dicot ESTs.

When both within-class and cross-class ESTs are combined, we see coverage and accuracy that is similar to using the within-class ESTs alone (see the rightmost data points on Fig. 2A,B). For *A. thaliana,* where the number of within-class ESTs is saturating, the effect of supplementing dicot ESTs with monocot ESTs is negligible. However, for *O. sativa,* where the number of monocot ESTs have not yet saturated, there is a small but still observable increase in coverage and accuracy when all the dicot ESTs are added.

In addition to the correctly identified genes, each predicted gene set also includes partial genes and incorrect genes. When the EST data size is small, most of the predicted genes are partial (data not shown). The change of number of the correct genes and incorrect genes in Figure 2 reflects the internal properties of the gene-building process: As more ESTs are added, the genome coverage by their alignments increases, leading to an increase in correct gene models and a slower increase in incorrect models, presumably due to the gene structural difference between the species. The number of incorrect gene models increases faster when using cross-class ESTs than within-class ESTs. For within-class ESTs, the number of incorrect gene models plateaus at 15.5% in *A. thaliana* and 17.3%

in *O. sativa*; for cross-class ESTs, the number of incorrect gene models increases to 19.3% in *A. thaliana* and 26.6% in *O. sativa*. It is important to note, however, that the number of incorrect gene models increases only slightly (<0.3%) in either species when all the cross-class data are combined with the in-class ESTs.

Gene builds using FLcDNA, EST, and protein evidence

In addition to ESTs, independently sequenced proteins and FLcDNA data resources are available for some species. How do various combinations of these different evidence types affect gene coverage and quality?

There are substantial numbers of full-length and partial cDNA sequences (Seki et al. 2002; Kikuchi et al. 2003) available for *Arabidopsis* and rice (*O. sativa*) in addition to ESTs. We will refer to the combination of FLcDNA/ESTs as mRNAs. There are also small numbers of independently sequenced proteins available. To create comprehensive evidence-based gene sets, we used a combination of both species-specific and cross-species proteins, FLcDNAs, and ESTs in both *A. thaliana* and rice (for the data set details, see Methods). The cross-species proteins include all SWISS-PROT proteins and plant proteins from TrEMBL. We treated all gene products from the same genus as species-specific evidence, after observing no significant differences in alignment scores between within-genus and within-species proteins (data not shown).

Using all available transcriptional and translational evidence, we predicted 25,298 genes (22,502 full-length and 2796 partial) in *A. thaliana* and 33,836 genes (27,190 full-length and 6646 partial) in rice after repeat-masking the genomes. There were 7598 (30%) and 9919 (29.3%) genes displaying alternative splicing with an average 2.53 and 2.77 transcripts among them in *A. thaliana* and rice, respectively. Table 1 shows the distribution of genes based on the type of evidence used to predict them. More than 86.5% of predicted *A. thaliana* genes and more than 79% of the predicted rice genes are supported by two or more sources of evidence. Cross-species expression data support a large portion of the genes (92.18% in *A. thaliana* and 82.6% in rice). The gene coverage by cross-species data is close to the numbers (94% and 89%, respectively) described in the previous section. Proteins provide a minor contribution to the rice gene set due to the limited number of proteins available, whereas ESTs (both same-species or cross-species) are the major contributory source for both *A. thaliana* and rice.

To evaluate the quality of the genes predicted from evidence and the performance of the Gramene pipeline on different evidence types, we performed separate gene builds using different combinations of evidence types. Since one of our goals in this study is to provide guidance on the best strategy for annotating the genomes that lack of known proteins and FLcDNAs, we focus on predictions made without the benefit of same-species FLcDNAs and proteins. These gene sets were evaluated with the standard gene sets described in Methods, and the results are discussed in the following sections.

Evaluation of evidence-based genes in *A. thaliana*

Using the metrics described above, we tested each predicted gene set by comparing it to the set of 11,378 FLcDNA-based standard genes to evaluate sensitivity, and to the TAIR7 (Swarbreck et al. 2008) annotated genes (27,029 genes) for specificity. The latter is a set of predicted and confirmed *Arabidopsis* genes that have been hand-curated over a period of years, while the former is essentially a highly reliable subset of the latter. The gene-level assessment metrics are shown in Table 2 . For our purpose mentioned in the previous section, among the gene sets listed in Table 2, the most interesting categories are those based on species-specific ESTs (Arabi-EST), cross-species dicot ESTs (Dicot-EST), cross-species proteins (OProtein), and combinations of them (AllEST for combinations of same-species and cross-species ESTs; AllEST-OPro for combinations of AllEST and OProtein).

We find that the gene locus sensitivity and specificity of these gene sets—with the exception of the OProtein set—are uniformly high (Sn, 96.6%–98.4%; Sp, 93.6%–96.8%). As described later, the accuracy of the OProtein set is limited by the small number of independently derived protein sequences for *Arabidopsis*. At the more stringent CDS level, however, there are important differences between the evidence sets. When using species-specific ESTs to predict *Arabidopsis* genes, the major artifact is the generation of partial CDSs. The gene-level CDS sensitivity and specificity of the Arabi-EST genes are 62% and 41.7%, respectively. Among the partial proteins, the dominant error is "split genes," in which a gene is split into two or more predicted genes due to the lack of transcript evidence joining them. A total of 18.8% of the genes are incorrectly split into two or more predicted genes in the Arabi-EST set.

Compared with gene models built using same-species ESTs, those built from cross-species dicot ESTs give a higher CDS sensitivity and specificity (67.5% and 51%, respectively), most likely due to the larger number of cross-species ESTs available to the gene build pipeline since the similar number (1.48 million) of random selected cross-species ESTs give a lower CDS sensitivity (<60% for 1.5 million mapped dicot ESTs in Fig. 2A). The number of split genes (4.3%) and partial proteins (11.5%) is correspondingly lower than those generated from same-species ESTs.

Table 1. Percentage of gene set supported by different evidence types

Evidence type	Only source of evidence	Contributory source of evidence
A. Total *A. thaliana* evidence genes: 25,298 (36,960 transcripts)		
Arabi cDNA	0.7	56.76
Arabi EST	3.96	80.58
Arabi protein	0.8	21.14
Non-Arabi protein	0.76	64.27
Non-Arabi EST	7.25	92.18
B. Total rice evidence genes: 33,836 (51,369 transcripts)		
Rice cDNA	1.72	72.3
Rice EST	5.31	82.07
Rice protein	0.07	18.71
Non-rice protein	0.23	6.56
Non-rice mRNA	12.85	82.6

Arabi cDNA indicates FLcDNAs originated from *Arabidopsis*; Arabi EST, ESTs from *Arabidopsis*; Arabi protein, SWISS-PROT proteins from *Arabidopsis*; non-Arabi protein, proteins of non-*Arabidopsis* source in SWISS-PROT, TrEMBL, and GenBank; non-Arabi EST, ESTs from non-*Arabidopsis* dicot species; rice cDNA, rice FLcDNA; rice EST, rice ESTs; rice protein, rice proteins in SWISS-PROT; non-rice protein, non-rice proteins in SWISS-PROT, TrEMBL, and GenBank; and non-rice mRNA, monocot non-rice FLcDNAs and ESTs.

Table 2. Gene-level assessments in *A. thaliana*

	Locus Sn	Missed	CDS Sn	Locus Sp	Extra	CDS Sp	Split
Arabi-EST	98.4	0.8	62	96.8	1.5	41.7	18.8
Dicot-EST	96.6	2.5	67.5	95.4	3.6	51	4.3
AllEST	98.3	0.2	82.1	93.8	4.6	59	3.7
OProtein	88.4	11.4	25.8	97.7	1.6	23.4	1.4
AllEST-OPro	98.1	0.2	82.2	93.6	4.7	60.6	3.1
Arabi-protein	29.7	70.3	27.2	99.6	0.3	85.3	0.4
Arabi-cDNA	99.8	0	99.6	97.5	2	85.1	0.7[a]
All-evidence	98.6	0	94.5	93.2	6.0	70.4	3.4[a]

Sensitivity (Sn) is measured on the FLcDNA-based standard genes, and specificity (Sp) measured on the TAIR7 set. The CDS Sn and Sp values are calculated using identical proteins only. Two overlapping genes on opposite strands were classified as "missed" genes to each other if both have multiple exons. Arabi indicates *Arabidopsis*; AllEST, combination of *Arabidopsis* ESTs and other dicot ESTs; and OProtein, all cross-species proteins. All values are percentages.
[a]All split genes are measured on FLcDNA-standard genes, except the Arabi-cDNA and the All-evidence sets, which are measured using TAIR7 set.

On the other hand, cross-species ESTs introduce more incorrect gene structures (15.8%) than do same-species ESTs (5.1%) (see Supplemental Table S2). However, the combination of all species-specific and cross-species dicot ESTs increases the predicted gene accuracy significantly compared with either source type only (CDS Sn, 82.1%; Sp, 59% at the gene level).

We also tested the predictive power of cross-species proteins. Although we used all plant non-*Arabidopsis* proteins from SWISS-PROT and TrEMBL, the OProtein gene quality (CDS Sn, 25.8%; Sp, 23.4% at gene level) is much lower than that of cross-species ESTs, largely due to the limited number of suitable proteins available. Among the three evidence types, the OProtein gene set has the largest proportion of missed genes and the smallest number of false-positive genes. The addition of cross-species proteins only slightly improves the gene quality relative to that of all dicot ESTs (AllEST-OPro CDS Sn, 82.2%; Sp, 60.6% at the gene level).

The assessment results at the exon and base-pair levels are provided as Supplemental material. Here we provide only a brief summary of the predicted transcript properties. Among the overlapping transcripts, the base-level specificity for any gene set is at least 96.2%. The base-level sensitivity is at least 92.2%, except for EST-only genes (80.5%). The exon-level sensitivity and specificity of AllEST-OPro genes (Sn, 92%; Sp, 88.3%) are slightly higher than that of AllEST genes (Sn, 91.8%; Sp, 88.1%). The exon-level sensitivity and specificity for either Arabi-EST (Sn, 76.7%; Sp, 81.7%) or cross-species Dicot-EST (Sn, 86.9%; Sp, 87.3%) are lower, largely due to partial exons (13.4% and 12.7%, respectively) and missed exons (10.5% and 6.3%, respectively). These partial exons and missed exons can explain the large number of partially predicted genes found in EST-based gene predictions.

Evaluation of evidence-based genes in rice

The rice genome is the second well-annotated plant genome available. Rice is an economically important monocot species. Relative to *A. thaliana*, rice has a larger genome and more genes (Ouyang et al. 2007). For reference purposes, we compare the genes predicted by the Gramene pipeline to three gene sets generated by ab initio methods: an Fgenesh (Solovyev et al. 2006) set (56,453 predicted genes) available in the Gramene database, a Twinscan (Korf et al. 2001) set (50,975 predicted genes) graciously provided by C. Zhang and B. Barbazuk (The Donald Danforth Plant Science Center, St. Louis, MO), and an ExonHunter (Brejova et al. 2005) set (29,970 predicted genes) graciously provided by B. Brejova and T. Vinar (Cornell University, Ithaca, NY). We evaluated the predicted rice genes as we did for *A. thaliana* using 11,785 confirmed genes supported by FLcDNAs (see Methods) for sensitivity testing and 41,042 TIGR5 gene predictions (Ouyang et al. 2007) for specificity testing. The former set can be viewed as a highly reliable subset of the latter. There exists another annotated rice gene set from the Rice Annotation Project (RAP) (Tanaka et al. 2008), which contains 30,192 protein-coding genes that are supported by species-specific expression evidence. We did not compare our gene sets with the RAP set since the RAP set uses a different genome assembly than TIGR5.

The gene-level sensitivity and specificity for all rice gene categories are shown in Table 3 . When all evidence was used, the Gramene pipeline predicted 33,836 rice genes (All-evidence, which is also mentioned in Table 1B), compared with 41,042 genes not related to transposable elements (TEs) in TIGR5. The All-evidence set contains 3201 (9.5%) genes that are not present in the TIGR5 reference set (with an additional 7.3% on the opposite strand—both treated as false-positives in this test). On the other hand, there are 13,830 TIGR5 gene predictions not present in the All-evidence set for which there was no supporting EST, FLcDNA, or protein evidence, or which were filtered out due to low-confidence evidence during the gene build process. Comparison of the TIGR and Fgenesh sets reveals that 95.1% of these missing genes overlap Fgenesh genes and 12,505 (86.3%) of them are identical to Fgenesh gene predictions. This reflects a fundamental difference between the evidence-based gene build methods and ab initio methods: The evidence-based methods predict a gene only if there is evidence for its transcription; they can therefore potentially miss real genes that ab initio methods can catch based on the latter's statistical models on gene structure. However, as the abundance of the available gene products increases, the missing genes from evidence-based methods will decline commensurately. On the other hand, ab initio predictions usually suffer from relatively low accuracy compared to evidence-based gene predictions (Guigo et al. 2006). Therefore, one should treat the gene specificity in Table 3 with care as it is inevitably inflated for the Fgenesh set but may underestimate the

Table 3. Gene-level assessments in rice

	Locus Sn	Missed	CDS Sn	Locus Sp	Extra	CDS Sp	Split
Fgenesh	98.9	1	32.9	97.3	2.3	66.8	2.7
Twinscan	95.4	4.1	34.1	76.2	20.7	22.9	2.3
ExonHunter	96	3.9	35.8	90.4	8.4	26.8	4.8
Rice-EST	94.2	3.2	42.8	85.4	7.7	26.5	21.2
OmRNA	92.5	5.4	49.9	89.2	4.2	36.1	5.2
EST-OmRNA	98.1	0.3	67.8	82.8	9.0	40.7	7.5
OProtein	27.5	72.3	5	86.5	9.5	14.3	2.5
EST-OmRNA-OPro	98.1	0.3	67.8	82.8	9.1	40.8	7.4
Rice-cDNA	99.5	0.1	95.7	85.6	7.9	64.8	1.2[a]
Rice-protein	9.1	90.9	6.4	89.8	9.4	61.4	0.7
All-evidence	99.6	0	89.3[b]	83.1	9.5	54.7[b]	3.9[a]

Sn values are measured on FLcDNA-based standard genes, and Sp values measured on the TIGR5 set. See Table 2 for further explanation. OmRNA indicates monocot non-rice FLcDNAs and ESTs.
[a]Measured on the TIGR5 set; all other split predictions are measured on FLcDNA-based standard genes.
[b]The low values of CDS Sn/Sp are largely due to extended proteins not included for Sn calculation.

specificity for the other gene sets. Please see Supplemental Table S7 for a comparison between ab initio predictions and FLcDNA-based standard genes.

For our purposes, the gene sets of greatest practical interest are based on rice ESTs (rice-EST), cross-species monocot mRNAs (OmRNA), or cross-species proteins (OProtein) and on combinations of them (EST-OmRNA and EST-OmRNA-OPro). As expected from our results on *Arabidopsis*, it seems that the depth of the raw EST coverage is more important than the evolutionary distance of the ESTs, provided that they remain within the same dicot/monocot class. Predicted genes derived from 1.2 million same-species ESTs give lower gene-level protein sensitivity (42.8%) than the sensitivity obtained using 3.5 million cross-species monocot mRNAs (49.9%). Combining both types of evidence (EST-OmRNA) improves the gene quality markedly (gene-level protein sensitivity, 67.8%). The rice-EST gene predictions suffer from a large number of split genes (21.2%) and partial genes (38.7%). When cross-species mRNAs are added, the number of split genes and partial genes reduces to 7.5% and 13%, respectively. As with the previously described results in *A. thaliana*, the addition of non-rice proteins does not improve predicted gene set quality appreciably over the combination of species-specific ESTs and cross-species monocot mRNAs.

To evaluate how well the EST and cross-species data perform in practical gene predictions, we compared the predicted genes against ab initio predictions. The gene locus sensitivity of EST-OmRNA-OPro genes (92.5%–98.1%) is comparable to that of the three reference ab initio gene predictors (95.4%–98.9%), but their gene-level CDS sensitivity is substantially better than that achieved by the ab initio predictors (67.8% vs. 32.9%–35.8%). The gene locus specificity of the Gramene pipeline (82.8%–89.2%) is comparable to that achieved by the two ab initio predictors, Twinscan and ExonHunter (76.1%–90.4%); the Fgenesh set was not included in the specificity comparison due to the large number of the Fgenesh predictions in the TIGR5 set.

We calculate the exon-level sensitivity and specificity of the non-FLcDNA-based gene sets as described earlier using the subset of predicted genes that overlap the confirmed genes (Table 4). We observe several interesting differences between the evidence-based predicted genes and ab initio predicted genes. The three ab initio methods missed only a small number of coding exons (1.2%–5.5% false-negatives) but added many extra, apparently incorrect, exons (17.7%–24% false-positives). In comparison, the EST-OmRNA-predicted genes missed 4.1% of the coding exons but added only 5.2% extra coding exons. The ab initio methods also introduce slightly more incorrect internal exons than do the evidence genes (ab initio, 4.8%–8.4%; evidence, 1.5%–3.5%). The base-pair-level assessments are provided as Supplemental material.

Among the genes built from a combination of all evidence types, there are 1516 genes (6% in Table 2) not found in TAIR7 and 3201 (9.5% in Table 3) genes absent from TIGR5. We treat these extra genes as false-positives in our tests. However, inspection of their supporting evidence type shows that 41.6% of the *A. thaliana* genes and 50% of the rice genes are supported by multiple sources of evidence. This suggests that in fact many of these extra genes could be real. While preparing this manuscript, a new TAIR annotation set, TAIR8 (http://www.arabidopsis.org), was released. Comparing this annotation set with TAIR7, we find that 136 (9%) of the extra All-evidence genes in *A. thaliana* appeared in TAIR8. Further work is needed to test how many of the other genes are real. For rice genes, a potential real gene absent from TIGR5 is shown in Figure 3. This gene is supported by many species-specific and cross-species ESTs but is not represented by any FLcDNAs. We also observed that some of these extra genes had been included in a new TIGR gene set (now called MSU gene set; see http://rice.plantbiology.msu.edu; data not shown).

How many genes are in the rice genome?

We now attempt to estimate how many genes exist in the current rice genome and thus how many of them are still missing in the rice gene set predicted by the Gramene pipeline. We divide the predicted non-TE-related genes by Fgenesh and Twinscan into two groups each—genes supported by rice cDNA/EST ("supported," 26,289 genes and 26,404 genes, respectively) and not supported ("unsupported," 12,686 genes and 15,470 genes, respectively)—and map these genes to the repeat-masked sorghum genome using TBLASTN (Altschul et al. 1997; see Methods). Sorghum (*Sorghum bicolor*) is a monocot species that is evolutionarily close to rice. The two species have been diverged for 50 Myr (Wolfe et al. 1989). Sorghum genome has been recently sequenced (Paterson et al. 2009), showing a close gene content to rice genome. The percentage of rice genes mapped to the sorghum genome (the mapping ratio) plotted against their TBLASTN alignment *P*-value is shown in Figure 4. We note that the unsupported genes exhibit an appreciably lower mean sequence identity to sorghum genome matches than do supported genes, as expected.

Table 4. Exon-level per transcript-pair comparison to standard genes in rice

	Sn						Sp					
	Same	Ext	Part	Diff	Missed	All[a]	Same	Ext	Part	Diff	Extra	All[a]
Fgenesh	81.7	2.4	2	8.4	5.5	91.7	71.4	2.0	1.8	7.1	17.7	83.9
Twinscan	85.1	3.1	1.5	7.1	3.2	94.6	66.4	2.3	1.2	6	24	77.9
ExonHunter	87	6.2	0.9	4.8	1.2	98.5	70.7	5	1.1	3.9	19.3	83.6
Rice-EST	73.1	1.1	10.9	1.2	13.6	89.4	77.8	1.2	12.3	5.2	3.5	96.9
OmRNA	83	1.5	4.9	3.5	7.0	92.9	85.3	1.4	5	3.2	5.1	95.4
EST-OmRNA	89.3	1.5	3.4	1.5	4.2	97.9	84.7	1.4	4.1	4.9	5.2	95.4
OProtein	64.3	0.8	13.8	4.5	16.5	86.5	70.7	0.7	16.8	6.5	5.5	94.3
EST-OmRNA-OPro	89.4	1.6	3.3	1.5	4.2	97.9	84.2	1.4	4.3	4.9	5.2	95
Rice-protein	91.8	1.2	3.1	1.1	2.7	95.4	89.9	1.1	3.1	1	4.9	95.4

Only exons that are in predicted genes overlapping a FLcDNA-based standard gene are compared. The transcripts in overlapping genes are paired up with their best matching transcripts. The exons in each pair are compared and summed up for all transcript pairs for percentage calculation. Ext indicates extended; Part, partial; and Diff, different.

[a]These columns include all CDS exons and UTR exons touching a standard exon. All other columns are for CDS exons only.

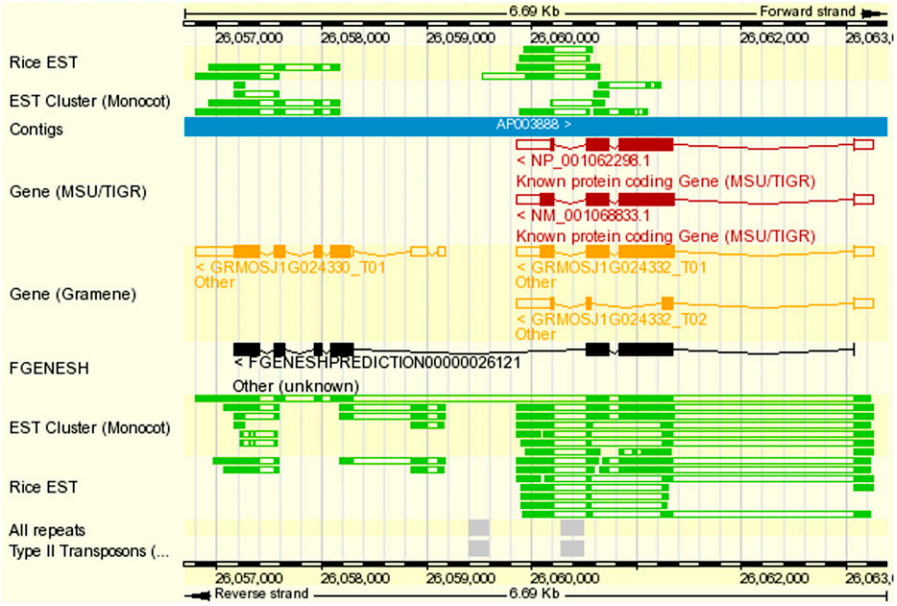

Figure 3. A rice gene on chromosome 8 that is predicted using EST only. The gene on the *right* side has FLcDNA support, which is correctly identified in both the TIGR5 set and the Gramene evidence set. The gene on the *left* side does not have FLcDNA support, which is missed from the TIGR5 set. Fgenesh incorrectly joins the two genes together. The *left* gene is supported by many same-species and cross-species ESTs.

We choose a stringent *P*-value cutoff of 1×10^{-5} to be the maximum *P*-value for a real rice gene mapped to sorghum genome based on the suggestions by Bennetzen et al. (2004). Above this cutoff, protein-to-DNA alignment is generally considered as background noise rather than due to true protein homology. We treat all supported genes as real genes. At *P*-values less than 1×10^{-5}, the mapping ratio for supported Fgenesh genes is 89.87% and that for supported Twinscan genes is 85.06%. The mapping ratios are close to the estimated rice gene coverage by the mixture of non-rice monocot ESTs described earlier (89%). If we assume the real genes among the unsupported category are mapped to the sorghum genome with the same mapping ratio as the supported genes, and none of the false-positive genes are mapped to the sorghum genome, we can estimate the number of the real genes in the unsupported group using the following formula (see Supplemental data for the derivation of this):

$$\text{Number of real genes unsupported} = \text{all genes unsupported} \times \text{mapping ratio unsupported}/\text{mapping ratio supported}.$$

Using this formula and the mapping ratio of the unsupported genes with a *P*-value less than 1×10^{-5} (35.57% for Fgenesh genes and 25.91% for Twinscan genes), we estimate the number of real genes in the unsupported group to be 5021 for Fgenesh and 4712 for Twinscan. By adding this to the supported genes, we estimate the total number of real genes in the Fgenesh set to be 31,310 and 31,116 in the Twinscan set. To estimate the total number of rice genes, we need another ratio: the gene locus sensitivity of the Fgenesh and Twinscan sets. If we assume that all the 33,836 rice evidence genes are real, the gene coverage of this set is 89.6% for Fgenesh and 88.9% for Twinscan. Dividing the number of real Fgenesh and Twinscan genes by their gene coverage, we estimate the number of rice genes to be 34,944 and 35,001, respectively.

There are a few factors that affect this estimate. Highly expressed genes are both more likely to have supporting evidence and are more likely to be strongly conserved. Therefore the mapping ratio of unsupported real genes might in fact be lower than that of supported genes, in which case the total number of real genes will be higher. Alternatively, there might be false predictions that map to the sorghum genome. In this case, the total number of real genes will be lower. Other factors to affect this estimation are the rates of split genes and joined genes in each gene set; however, the error rate due to split and joined genes is <3.9% for Gramene evidence genes, setting an upper bound on the estimation error due to this factor.

If we use the mean value of the two estimated gene numbers at the *P*-value cutoff of 1×10^{-5} (34,973) and ignore issues arising from split and joined genes, we find that the number of Gramene evidence-based genes is 3.3% lower. Based on this analysis, we confirm that the rice gene number is below the upper bound (40,000) estimated by Bennetzen et al. (2004); this is also well below the current TIGR5 annotation (41,042 genes), which are thought to contain many TE-related genes or pseudogenes (e.g., see Paterson et al. 2009). The missing genes from the Gramene evidence set will possibly be identified by adding new expression evidence and/or ab initio predictions.

An evidence-based gene build in a 22-Mb maize genome region

To test the effectiveness of the pipeline in a newly sequenced genome, we apply the Gramene pipeline to a 22-Mb testing region of maize (*Zea mays*) (The Maize Sequencing Consortium, unpubl.). The maize genome is known for its large transposon and repeat content: 76.4% of the maize region is masked as repetitive (see Methods). Using all the gene products used for rice, we obtained 1005 protein-coding genes (1266 transcripts), 642 apparently full-length, and 363 partial. There are 148 (14.7%) gene predictions that contain alternative splicing with an average of 2.97 transcripts among them. As expected from the order of magnitude difference in genome sizes, the maize gene density (~45 genes/Mb) is much lower than that in rice (~89 genes/Mb); however, relative to the maize genome as a whole (data not shown), this still represents a gene-rich region.

There were around 11,700 newly sequenced FLcDNAs from the maize full-length cDNA project (http://www.maizecdna.org) in GenBank when we did this analysis. To evaluate the quality of our gene build, we did not include these FLcDNAs among the transcriptional evidence, but instead used them to generate a separate set of well-supported genes (see Methods). We compared the 1005 non-FLcDNA evidence-based genes (non-cDNA) with the 148 FLcDNA-based genes (157 transcripts). For reference purposes, we also compared two intermediate gene sets based on maize EST (maize-EST) and cross-species mRNAs (OmRNA), and Fgenesh-predicted (Solovyev et al. 2006) ab initio genes to the FLcDNA-based genes. Using these FLcDNA-based genes as a standard set, we are essentially measuring sensitivity, though the

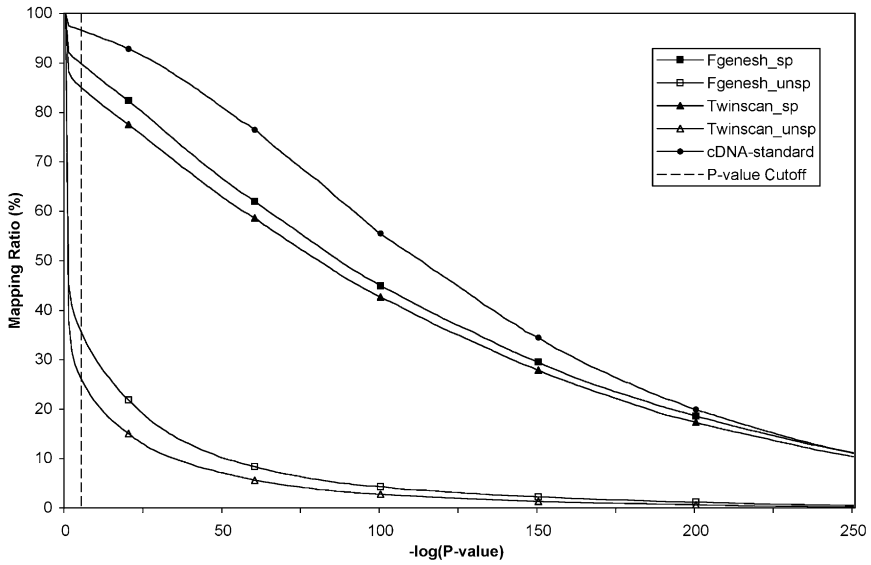

Figure 4. Mapping rate of rice genes on sorghum genome using TBLASTN. Rice genes predicted by Fgenesh and Twinscan are each divided into two groups: (1) supported by FLcDNA/EST (Fgenesh_sp and Twinscan_sp) and (2) not supported by FLcDNA/EST (Fgenesh_unsp and Twinscan_unsp). The genes are already filtered using MIPS TE library to remove TE-related genes. The *x*-axes −log(*P*-value) greater than 250 (*P*-value $< 1 \times 10^{-250}$) is taken as 250. The *P*-value cutoff 1×10^{-5} is labeled as a *vertical* dashed line. The confirmed genes with FLcDNA-support are included for reference purposes.

assessment might be biased due to the small number of FLcDNA-based genes. We are unable to evaluate the predictions' specificity due to the lack of a comprehensive gene set in this region.

The results are shown in Table 5 . The non-cDNA genes touch 144 (97.3%, as gene locus sensitivity) of the FLcDNA-based genes. At the gene level, identical CDSs cover 100 (67.6%, as gene CDS sensitivity) of them, with an additional 29 (19.6%) being covered by extended CDSs in the non-cDNA set. Many extended proteins are caused by extension at the 5′ end by ESTs, which could be due to alternative transcription start sites or, more likely, the truncation of the FLcDNAs at their 5′ ends. This suggests that the current available gene products, excluding maize FLcDNAs, enable us to identify >87.1% of the genes (in full CDS) supported by these FLcDNAs.

In comparison, the maize-EST and OmRNA evidence sets yield 64.9% identical (and 12.8% extended) and 54.7% identical (and 16.2% extended) CDSs, respectively. Fgenesh predictions touch all the FLcDNA-based genes except one (which was on the opposite strand). However, ~46% of the Fgenesh genes give different (incorrect) CDSs. We obtained slightly more identical CDSs in maize than in rice based on cross-species mRNAs as expected since there were many more rice FLcDNAs used for maize gene predictions than maize FLcDNAs used for rice gene predictions. The Fgenesh assessment results are comparable between maize and rice. In addition to the FLcDNA-based genes, we also compare the non-cDNA gene predictions to Fgenesh predictions. The number of non-cDNA genes that Fgenesh genes touch is 704 (70%), with an additional 73 (7.3%) genes on the opposite strand. Among the overlapping genes, even though around 60% of them had different protein translations, we find good agreement at the nucleotide and exon levels: 87% of bases in CDS regions among the evidence-based genes were shared by Fgenesh predictions and 65% of the CDS exons predicted by the Gramene pipeline are identical to Fgenesh exons.

As discussed in earlier sections, split gene predictions can be a problem when there is insufficient evidence to cover the whole CDS region. We find four of the non-cDNA genes are partial gene pairs that overlap two FLcDNA-based genes, for a split gene rate of only 1.4%. To check other potential split predictions, we use EST pairs (one 5′-EST and one 3′-EST) that originated from the same FLcDNA clones. Using a match cutoff of 90% sequence identity, we are able to map 2047 pairs to the region. The pipeline built 2006 of these pairs (98%) into single genes. The remaining 2% either exist as two partial predictions, or the pipeline failed to accept one or both of them. This suggests that up to 2% of real genes are split in the non-cDNA predicted set.

Discussion

We report here the design and implementation of an evidence-based gene prediction system based on multiple sources of gene expression information, both within and between species. This system has several advantages over nonevidence methods. The major improvement over other gene prediction methods is the Gramene pipeline's ability to prioritize the gene prediction process according to the confidence of underlying supporting evidence, thus increasing the specificity of the predicted gene sets. When using any of the evidence types available or combination of them, we achieved high specificity at the gene locus level (>93.2% in *A. thaliana* and >82.8% in rice), which compares favorably to the specificity of the unmodified Ensembl pipeline in humans using known proteins and FLcDNAs (72%–77%) (Curwen et al. 2004). Most importantly, although we are not trying to provide a generalized score for each gene predicted by the Gramene pipeline, the Ensembl system can store and display the supporting evidence for each gene, allowing researchers to determine for themselves how well they trust the prediction. Another advantage of evidence-based predictions is that they can identify alternatively spliced transcripts. The major limitation is that the accuracy of this method will vary depending on the nature and quantity of the available expression data.

One of our major goals was to study how well the Gramene pipeline performs on genomes with only incomplete or cross-species gene products. We have found that cross-species mRNA data within a dicot- or monocot-class are a valuable source of evidence

Table 5. Comparing maize-predicted genes with FLcDNA-based genes

				Protein (CDS)			
	Overlapping	Opposite	Missed	Same	Ext	Part	Diff
Non-cDNA	97.3	1.4	1.4	67.6	19.6	2.0	8.1
Fgenesh	99.3	0.7	0	38.5	12.2	2.7	46
Maize-EST	98.6	1.4	0	64.9	12.8	14.2	6.8
OmRNA	93.9	3.4	2.7	54.7	16.2	7.4	16.2

All numbers are percentages relative to 148 maize FLcDNA-based genes. Non-cDNA indicates using all evidence except maize FLcDNAs; OmRNA, all monocot non-maize FLcDNAs and ESTs.

for gene prediction. Further, we showed that species-specific EST and cross-species mRNAs are highly effective in combination, yielding predicted gene locus sensitivity in excess of 98% in both *Arabidopsis* and rice. The predicted gene CDS sensitivity is 82.1% and 67.8% in the two species, close to or better than the best human gene prediction programs described in EGASP (71.6%). The accuracy of the Gramene pipeline is also significantly superior to ab initio gene prediction programs that we evaluated.

The Gramene pipeline uses a rule-based method to select relatively high-confidence transcripts first, which are then improved if necessary using low-confidence transcripts. The high-confidence genes can be improved by low-confidence genes in two ways: (1) identify and fix some of the incorrect introns with non-canonical splicing sites and (2) connect a partial gene pair or extend the partial translation frame. Although the resulting genes do not necessarily all have correct gene structure, the presence of high-confidence exons in the model usually guarantees at least partially correct translations, which make them useful in genome annotations. In much the same way that the Gramene pipeline uses low-confidence transcripts to incrementally improve those of high confidence, we have also used ab initio genes to improve evidence-based genes (e.g., to connect partial gene pairs using ab initio exons or extend the ORF of the partial evidence-based genes). This is a different approach from the more traditional gene annotation approach in which evidence is used to improve or support ab initio genes (see Zhu and Buell 2007), but it more closely mirrors the human curation process where high-confidence genes are selected first and low-confidence genes are used only if necessary. Our experience in rice shows that roughly 50% of partial-evidence gene pairs can be incorporated into a single gene using ab initio-predicted exons from the three programs used in this study (data not shown). As this manuscript was being written, the whole maize genome was sequenced (http://www.maizesequence.org) and the Gramene pipeline (by combining ab initio predictions) was being used to generate a gene set for further annotations.

Application of the Gramene pipeline to new genomes

The strategy of using cross-species transcriptional evidence should be applicable to gene prediction in other closely related species. For example, in sorghum there are almost no FLcDNAs available besides a small number of ESTs (~230,000 ESTs) in GenBank. Due to the short evolutionary distance between sorghum and maize—diverged ~11.9 Mya (Swigonova et al. 2004)—the maize cDNA/ESTs are ideal cross-species gene products. Based on our testing in rice, one could apply the several million cross-species monocot mRNAs (including ~40,000 newly deposited maize FLcDNAs in GenBank not used in this study) to make a predicted sorghum gene set with an expected sensitivity (at the gene CDS level) of 50%–70%. In dicots, *A. lyrata* (http://www.jgi.doe.gov/sequencing/why/3066.html) is estimated to have diverged from *A. thaliana* only ~5 Mya (Koch et al. 2000). We can use all FLcDNAs and ESTs from *A. thaliana*, *Brassica*, *Raphnus*, and 1–2 million other dicot ESTs to produce a predicted gene set with higher sensitivity (>70%–80% at gene CDS level); the within-genus mRNAs can be used as within-species mRNAs. We anticipate various but significant success rates for the *Caenorhabditis* nematodes (http://www.wormbase.org) as well as the *Drosophila* insects (Clark et al. 2007), for which the research community has aggressive genomic sequencing programs, but few plans for new FLcDNA or EST sequencing.

For other closely related animal species, such as mammals, we expect that a similar gene prediction strategy might yield good results. A potential confounding factor, however, is the large intron size in mammals. The accuracy of the Gramene pipeline largely depends on Exonerate, which is able to correctly call introns up to 20 kb in rice (and introns up to 110 kb in humans in our preliminary tests). In humans, fewer than 10% of introns are >11 kb in length (Sakharkar et al. 2004). Therefore, for species-specific ESTs, we think the large intron size should not be a major problem; however, more study is needed to test whether cross-species mRNAs will be as effective in mammals and other long-intron organisms as they are in higher plants.

Using the Gramene pipeline with data from new sequencing technologies

Same-species EST- and FLcDNA-sequencing data are always to be preferred to cross-species data, (e.g., for identifying species-specific genes and alternative splicing), but the expense of acquisition is high using traditional sequencing technologies. However, new short-read sequencing technologies (for a review, see Mardis 2008) such as Illumina and ABI SOLiD can quickly and inexpensively generate high-coverage EST sequencing information; the drawback is that the reads are short, generally on the order of 30–50 bp. The 454 Life Sciences (Roche) sequencing technology costs more per base but produces longer reads—on the order of 200–300 bp. While it is hard to estimate how many ESTs are required to generate a high-quality gene set for a new sequenced genome, we are expecting a much higher sequencing depth to achieve the similar gene sensitivity from short sequence reads than that for current existing ESTs. For example, 1.4 million EST sequences with a mean length of 265 bp in *A. thaliana* and 1.2 million EST sequences with a mean length of 480 bp in rice are equivalent to 10.3 million and 16 million short reads of length 36 bp for the same sequencing depth, respectively.

Exonerate is good at aligning short reads (e.g., 20 bp) in our tests (data not shown). However, for evidence-based gene builds, a major challenge is to correctly identify the exon–intron boundaries. For 454 data, many alignments will span exon–intron boundaries, so that they can be used directly. For Illumina and ABI SOLiD data, on the other hand, we expect that a high coverage in depth and mate pairs or paired-end reads will be required to identify introns accurately. The required large data size will increase the number of the sequence alignments dramatically, thus increasing the data storage and decreasing the running speed of the pipeline. A solution is to use short read assemblies to alleviate the problem and potentially for more confident alignments. The constant improvements on the short-read sequencing technologies to increase the read length will also help us to incorporate this source of transcription data into the Gramene pipeline.

Methods

Genomes, gene products, and mapping

The genome assemblies of *A. thaliana* (TAIR7, Swarbreck et al. 2008), *O. sativa* ssp. *japonica* (TIGR5, Ouyang et al. 2007), and *S. bicolor* (Paterson et al. 2009) are stored in the Gramene database (http://www.gramene.org) using the Ensembl system. Their chloroplast and mitochondria genomes were not included for simplicity. The maize 22-Mb genome region was obtained from the Maize Sequencing Consortium (http://www.maizesequence.org). The genomes were repeat-masked with RepeatMasker (http://www.repeatmasker.org) and the MIPS plant repeat library RE-dat (http://mips.gsf.de/proj/plant/webapp/recat). The unmasked (nonrepeat) sequences are ~84.8%, 61.4%, 49.6%, and 23.6% in *A. thaliana*, rice, sorghum, and maize, respectively.

The gene products include FLcDNAs, ESTs, and proteins. All nucleotide sequences were downloaded from GenBank (http://www.ncbi.nlm.nih.gov). The FLcDNAs were downloaded from the core nucleotide database, with entries from genome annotation projects being removed. All protein sequences were downloaded from SWISS-PROT and TrEMBL (http://expasy.org/sprot; only plant proteins were downloaded from TrEMBL). Specifically, the data sets include the following: cDNAs from *Arabidopsis* (57,918), *Oryza* (72,919 including many duplicated entries from RefSeq), *Zea* (18,124), and other monocot species (~14,000); ESTs from *Arabidopsis* (1,478,777), *Brassica* (839,215), *Raphanus* (287,482), *Asterales* (875,854), *Fabales* (1,228,828), *Malpighiales* (640,931), *Malvales* (380,596), *Rosales* (432,206), *Sapindales* (473,991), *Solanales* (868,521), *Vitales* (380,597), *Oryza* (1,217,859), *Zea* (1,443,805), *Triticum* (1,066,552), *Hordeum* (499,423), and other monocots (902,666). For protein, there are 2222 *Oryza* proteins, 6668 *Arabidopsis* proteins, and ~350,000 other proteins from SWISS-PROT and 836,692 non-*Arabidopsis* plant proteins from TrEMBL. Additional newly deposited ~11,700 full or partial maize cDNAs were used in maize gene builds.

All gene products were mapped to each genome using Exonerate (Slater and Birney 2005; the program is available at http://www.ebi.ac.uk/~guy/exonerate) with suitable alignment models. All alignments were done using Exonerate version 1.0, since all the newer versions lack an optimization option, which can increase the alignment accuracy significantly. The optimization step decreases the speed by five to 10 times. Due to the large quantity of cross-species ESTs, their alignment represents the most time-consuming step of the process. On a 1000-CPU computer cluster, it takes up to 3–5 d to map 6–7 million cross-species ESTs to *A. thaliana* or rice genome. The computation speed also depends on the disk usage and database server usage. For *A. thaliana*, computation can be completed within 1 d with the full usage of both the cluster and database server. The running of the whole Gramene pipeline on a computer cluster requires job management and many other required Perl modules from the Ensembl pipeline (Potter et al. 2004; http://www.ensembl.org).

We computed the Exonerate mapping rate. For species-specific mapping at 90% sequence identity, the mapping rate of FLcDNA is 98.9% for *Arabidopsis* and 93.7% for rice; the mapping rate of EST is 90.9% for *Arabidopsis* and 84.3% for rice. For cross-species ESTs at 40% sequence identity, the respective mapping rates are as follows: dicots to *Arabidopsis*, 46.3%–83.4%; monocots to *Arabidopsis*, 25.1%–45.1%; monocots to rice, 40.2%–65.9%; and dicots to rice, 17.4%–54.4%. The protein mapping rate for same-species is 98.8% at 90% alignment identity and 35%–40% at 30% alignment identity for cross-species. Multiple mappings from each sequence on the genome were all used for gene builds.

The Fgenesh gene sets in rice and maize were generated using the default parameters for monocot ab initio gene predictions. We filtered the rice genes (proteins) predicted by Fgenesh and Twinscan by removing TE-related genes based on the MIPS and TIGR (Ouyang and Buell 2004) repeat library with TBLASTN (WU-BLAST, http://blast.wustl.edu/) using a *P*-value cutoff of 1×10^{-5}. The resulting non-TE genes were compared with cDNA/EST-supported evidence-based genes and were classified as supported genes and unsupported genes. These genes (proteins) and the rice FLcDNA-supported standard genes were aligned to the whole sorghum genome using TBLASTN.

Generation of standard gene sets

To evaluate the quality of predicted genes, we constructed a gene set using pure FLcDNAs in both *A. thaliana* (Seki et al. 2002) and rice (Kikuchi et al. 2003).

For *A. thaliana*, we constructed the standard gene set as follows: (1) map full-length *A. thaliana* cDNAs to repeat-masked TAIR7 genome; (2) select alignments with coverage of at least 95% and identity of at least 99.5%, and do a gene build; (3) filter out genes with non-GT|AG introns, with excessively long UTRs (to remove incomplete splicing forms or pseudogenes), and with incomplete CDS or protein length less than 50 aas; and (4) compare with the TAIR7 gene set, and select only those genes that have the same protein (UTR length not necessarily the same).

For rice, the first three steps were performed similarly on the TIGR5 genome, and an extra step was added to remove TE-related genes based on the MIPS and TIGR repeat library with a TBLASTN *P*-value cutoff 1×10^{-5}. The resulting genes were compared with the TIGR5 set; we found that <3% of them were different or absent from the TIGR5 set. These different or absent genes were inspected to ensure their similarity to other genes in translation length and exon number.

Finally, 11,378 genes (12,011 transcripts) were selected as the standard genes in *A. thaliana*. The rice standard gene set consists of 11,785 FLcDNA-based genes (12,324 transcripts). The maize genes based on pure FLcDNAs were generated using the first three steps listed above.

Acknowledgments

We thank the Maize Sequencing Consortium for the maize sequences and thank Sandra Clifton for her critical review of the manuscript. We also thank the Ensembl team, especially Guy Slater and Laura Clarke, for their help in running Exonerate and the Ensembl pipeline. Lastly, we thank the reviewers for their critical comments on the first version of the manuscript. This work was supported by the National Science Foundation (0321685 and 0703908) and a U.S. Department of Agriculture-Agricultural Research Service specific cooperative agreement (58-1907-0-041). The rice Twinscan gene set was kindly provided by C. Zhang and B. Barbazuk (The Donald Danforth Plant Science Center, St. Louis, MO), and the rice ExonHunter gene set was kindly provided by B. Brejova and T. Vinar (Cornell University, Ithaca, NY).

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389–3402.

Bennetzen JL, Coleman C, Liu R, Ma J, Ramakrishna W. 2004. Consistent over-estimation of gene number in complex plant genomes. *Curr Opin Plant Biol* **7:** 732–736.

Birney E, Clamp M, Durbin R. 2004. GeneWise and Genomewise. *Genome Res* **14:** 988–995.

Brejova B, Brown DG, Li M, Vinar T. 2005. ExonHunter: A comprehensive approach to gene finding. *Bioinformatics* **21 (Suppl. 1):** i57–i65.

Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman TC, Kellis M, Gelbart W, Iyer VN, et al. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450:** 203–218.

Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, Searle SM, Clamp M. 2004. The Ensembl automatic gene annotation system. *Genome Res* **14:** 942–950.

Eyras E, Caccamo M, Curwen V, Clamp M. 2004. ESTGenes: Alternative splicing from ESTs in Ensembl. *Genome Res* **14:** 976–987.

Guigo R, Flicek P, Abril JF, Reymond A, Lagarde J, Denoeud F, Antonarakis S, Ashburner M, Bajic VB, Birney E, et al. 2006. EGASP: The human ENCODE Genome Annotation Assessment Project. *Genome Biol* **7:** S2. doi: 10.1186/gb-2006-7-s1-s2.

Kikuchi S, Satoh K, Nagata T, Kawagashira N, Doi K, Kishimoto N, Yazaki J, Ishikawa M, Yamada H, Ooka H, et al. 2003. Collection, mapping, and annotation of over 28,000 cDNA clones from *japonica* rice. *Science* **301:** 376–379.

Koch MA, Haubold B, Mitchell-Olds T. 2000. Comparative evolutionary analysis of chalcone synthase and alcohol dehydrogenase loci in

Arabidopsis, Arabis, and related genera (Brassicaceae). *Mol Biol Evol* **17:** 1483–1498.
Korf I, Flicek P, Duan D, Brent MR. 2001. Integrating genomic homology into gene structure prediction. *Bioinformatics* **17 (Suppl. 1):** S140–S148.
Liang C, Jaiswal P, Hebbard C, Avraham S, Buckler ES, Casstevens T, Hurwitz B, McCouch S, Ni J, Pujar A, et al. 2008. Gramene: A growing plant comparative genomics resource. *Nucleic Acids Res* **36:** D947–D953.
Mardis ER. 2008. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* **9:** 387–402.
Ouyang S, Buell CR. 2004. The TIGR Plant Repeat Databases: A collective resource for the identification of repetitive sequences in plants. *Nucleic Acids Res* **32:** D360–D363.
Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, Thibaud-Nissen F, Malek RL, Lee Y, Zheng L, et al. 2007. The TIGR Rice Genome Annotation Resource: Improvements and new features. *Nucleic Acids Res* **35:** D883–D887.
Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, et al. 2009. The sorghum bicolor genome and the diversification of grasses. *Nature* **457:** 551–556.
Potter SC, Clarke L, Curwen V, Keenan S, Mongin E, Searle SM, Stabenau A, Storey R, Clamp M. 2004. The Ensembl analysis pipeline. *Genome Res* **14:** 934–941.
Reese MG, Hartzell G, Harris NL, Ohler U, Abril JF, Lewis SE. 2000. Genome annotation assessment in *Drosophila melanogaster*. *Genome Res* **10:** 483–501.
Sakharkar MK, Chow VT, Kangueane P. 2004. Distributions of exons and introns in the human genome. *In Silico Biol* **4:** 387–393.
Satoh K, Doi K, Nagata T, Kishimoto N, Suzuki K, Otomo Y, Kawai J, Nakamura M, Hirozane-Kishikawa T, Kanagawa S, et al. 2007. Gene organization in rice revealed by FLcDNA mapping and gene expression analysis through microarray. *PLoS One* **2:** e1235. doi: 10.1371/journal.pone.0001235.
Seki M, Narusaka M, Kamiya A, Ishida J, Satou M, Sakurai T, Nakajima M, Enju A, Akiyama K, Oono Y, et al. 2002. Functional annotation of a full-length *Arabidopsis* cDNA collection. *Science* **296:** 141–145.
Slater GS, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6:** 31. doi: 10.1186/1471-2105-6-31.
Solovyev V, Kosarev P, Seledsov I, Vorobyev D. 2006. Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biol* **7:** S10. doi: 10.1186/gb-2006-7-s1-s10.
Sparks ME, Brendel V. 2005. Incorporation of splice site probability models for non-canonical introns improves gene structure prediction in plants. *Bioinformatics* **21 (Suppl. 3):** iii20–iii30.
Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, Foerster H, Li D, Meyer T, Muller R, Ploetz L, et al. 2008. The *Arabidopsis* Information Resource (TAIR): Gene structure and function annotation. *Nucleic Acids Res* **36:** D1009–D1014.
Swigonova Z, Lai J, Ma J, Ramakrishna W, Llaca V, Bennetzen JL, Messing J. 2004. Close split of sorghum and maize genome progenitors. *Genome Res* **14:** 1916–1923.
Tanaka T, Antonio BA, Kikuchi S, Matsumoto T, Nagamura Y, Numa H, Sakai H, Wu J, Itoh T, Sasaki T, et al. 2008. The Rice Annotation Project Database (RAP-DB): 2008 update. *Nucleic Acids Res* **36:** D1028–D1033.
Wolfe KH, Gouy M, Yang YW, Sharp PM, Li WH. 1989. Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. *Proc Natl Acad Sci* **86:** 6201–6205.
Zhu W, Buell CR. 2007. Improvement of whole-genome annotation of cereals through comparative analyses. *Genome Res* **17:** 299–310.

Received November 6, 2008; accepted in revised form June 8, 2009.

Evidence-based gene predictions in plant genomes

Chengzhi Liang, Long Mao, Doreen Ware, et al.

Genome Res. 2009 19: 1912-1923 originally published online June 18, 2009
Access the most recent version at doi:10.1101/gr.088997.108