Opinion TRENDS in Plant Science Vol.10 No.1 January 2005

Community-based gene structure annotation

Shannon D. Schlueter[1], Matthew D. Wilkerson[1], Eva Huala[2], Seung Y. Rhee[2] and Volker Brendel[1,3]

[1]Department of Genetics, Development and Cell Biology, Iowa State University, 2112 Molecular Biology Building, Ames, IA 50011, USA
[2]Carnegie Institution, Department of Plant Biology, 260 Panama Street, Stanford, CA 94305, USA
[3]Department of Statistics, Iowa State University, Ames, IA 50011, USA

Uncertainty and inconsistency of gene structure annotation remain limitations on research in the genome era, frustrating both biologists and bioinformaticians, who have to sort out annotation errors for their genes of interest or to generate trustworthy datasets for algorithmic development. It is unrealistic to hope for better software solutions in the near future that would solve all the problems. The issue is all the more urgent with more species being sequenced and analyzed by comparative genomics – erroneous annotations could easily propagate, whereas correct annotations in one species will greatly facilitate annotation of novel genomes. We propose a dynamic, economically feasible solution to the annotation predicament: broad-based, web-technology-enabled community annotation, a prototype of which is now in use for *Arabidopsis*.

When is a genome finished?

For all plant and animal species, presentation of the 'finished genome' is considered to be a major milestone in the study of its genetics. However, ambiguous claims of this highly prized accomplishment beg the question of the meaning and worth of such announcements. Competitive and controversial claims concerning the completion of the human genome have been widely discussed [1]. In the area of plant genetics, the completed *Arabidopsis* genome was reported at the end of 2000 [2]. At that time, the genomic assembly comprised 115 409 949 base pairs covering the five chromosomes and leaving only an estimated 10 Mb of centromeric and ribosomal DNA (rDNA) repeat regions not sequenced. The total length of the assembled genome has increased by about 1 Mb per year (http://www.plantgdb.org/AtGDB/resource.php). A more demanding definition of a 'finished genome' requires extensive annotation of the assembled chromosome sequences in addition to the mere sequence report. In particular, researchers using the genome as a model system require annotation of the protein coding genes as the basis for assessing the transcriptome and proteome of the species. At the time of the *Arabidopsis* genome release, 25 498 protein-coding genes were annotated on the genome sequence. Since that time, this annotation challenge has continued to receive serious consideration for *Arabidopsis*, as evidenced by a ~10% increase in the number of annotated gene structures during the past three years [3] and continuing correction of erroneous initial annotations [4]. Perhaps the most ambitious and accurate definition of a 'finished genome' should include functional characterization of all the genes, a goal of the *Arabidopsis* 2010 project [5]. It is clear that each, successively more comprehensive, definition requires completion of the less ambitious tasks. The complexities of providing comprehensive annotation, whether that annotation is structural or functional, depend on an accurately defined gene structure. Because our collective understanding of genes and genome function continually advances, and users of the genome annotation naturally expect it to remain up to date with recent discoveries, the definition of a finished genome is necessarily a bit of a moving target.

Currently, a considerable time lag between completion of sequencing and completion of annotation appears to be unavoidable. This is because, even though sequencing is largely automated and robotic, and sequence assembly is largely routine (at least for genome regions that are not highly repetitive), accurate sequence annotation entirely by gene-finding software has remained elusive [6]. Current efforts towards more accurate and comprehensive gene structure annotation have focused on expressed sequence tag (EST) and full-length cDNA mapping onto the *Arabidopsis* genome [7–9] and combinations of computational and experimental approaches [10,11]. These studies have underscored the utility of spliced alignment to identify non-coding exons and to correct inaccurate computational gene predictions that formed the basis of the initial genome annotation. In particular, the results of cDNA mapping point to inherent limitations of high-throughput computational gene prediction, including difficulties in predicting exact exon borders, problems with distinguishing intergenic regions from introns and lack of models capable of identifying untranslated mRNA regions. However, these recent efforts have also not been entirely immune to the problems of large-scale automated annotation. For example, novel algorithmic changes incorporated into the newest annotation release [12] have inadvertently resulted in the ambiguous assignment of ESTs to multiple adjacent genes, thereby falsely extending their gene

Corresponding author: Brendel, V. (vbrendel@iastate.edu).
Available online 15 December 2004

 1360-1385/$ - see front matter doi:10.1016/j.tplants.2004.11.002

structure annotations (e.g. Figure 2 in Ref. [13]). Inclusion of draft sequences of clones that are too repetitive to finish with existing technology, although useful as a way to improve genome coverage with the available fragments of sequence data, has had some undesirable consequences, such as the inclusion of pBlueScript vector sequences in the genome sequence (http://www.plantgdb.org/AtGDB/Annotation/vector.php). The scope and complexity of the genome annotation task would seem to imply that shortcomings and mistakes are simply unavoidable in the early to middle stages of finishing a genome. Hild *et al.* [14] have discussed similar challenges with respect to the *Drosophila* genome annotation.

Arabidopsis genome annotation

The *Arabidopsis* research community currently has several ways to access genome data. TAIR (The *Arabidopsis* Information Resource; http://www.arabidopsis.org/ [15]), TIGR (The Institute for Genome Research; http://www.tigr.org/tdb/), MATdb (MIPS *Arabidopsis thaliana* Databases; http://mips.gsf.de/proj/thal/db/ [16]), SIGnAL (Salk Institute Genomic Analysis Laboratory; http://signal.salk.edu/ [10]), and AtGDB (The *Arabidopsis thaliana* Genome Database at PlantGDB; http://www.plantgdb.org/AtGDB/ [9,13]) provide web-based genome browsers for *Arabidopsis* that display gene structure annotation and comparisons with spliced alignment of ESTs and cDNAs. In addition to its genome browser, TAIR provides a comprehensive access point for *Arabidopsis* data, including information about genes, sequences, proteins, microarrays, germplasms, polymorphisms, seed and DNA stocks, and the research community. TAIR's curation efforts include the functional annotation of genes, with an emphasis on capturing experimental data from the literature and using controlled vocabularies [17].

Since the first release of the genome sequence in 2000, TIGR has maintained and updated the *Arabidopsis* genome annotation, making the updates publicly available in periodic releases, ending with the TIGR 5.0 release in January 2004, visible also at both AtGDB and TAIR. Because TIGR's role in maintaining and improving the genome annotation has come to an end, other mechanisms must be put in place to ensure that the genome data remain as error-free and up to date as possible. In response to this need, TAIR is currently setting up its own automated pipeline for improving gene models using new EST and cDNA data and manual methods for updating gene structures in response to community input. Although TAIR will work to eliminate the previously reported problems associated with automated gene structure annotation, automated methods will never be as flexible as a human curator in handling unusual cases or making use of new kinds of data. However, manual curation efforts by trained curators are limited by the size of the curation team and the amount of time needed to resolve each problematic gene structure annotation.

Even with well-organized community resources to support the informatics needs of a genome project, genome annotation remains a difficult task because, ultimately, all gene models will have to be evaluated by human experts. We have argued previously [18,19] that the only promising solution to this quandary is involvement of the user community and the development of enabling technology that streamlines user input, curation of user contributions and dissemination of approved user contributions. The purpose of this article is to introduce web-based gene structure annotation tools that are directly linked into AtGDB and TAIR and that will, we believe, facilitate broad-based community participation in the genome annotation task.

To assist in evaluating the quality of specific gene structure annotation and to determine the overall quality of the current *Arabidopsis* annotation, we have developed a system at AtGDB that allows gene structure comparison in the genomic context (http://www.plantgdb.org/AtGDB/Annotation/). The system, called Genome Annotation EVALuation (GAEVAL, pronounced 'gavel'), highlights inconsistencies between current gene structure annotation and the cognate placement of spliced aligned ESTs and (full-length) cDNAs. The reference for current gene structure annotation is provided by the mRNA fields in the GenBank deposited chromosome sequence files (Accession no. NC_003070, Accession no. NC_003071, Accession no. NC_003074, Accession no. NC_003075, Accession no. NC_003076). The cognate spliced alignments were derived with the GeneSeqer program as described previously [9] and provide the ability to identify non-coding exons, to confirm splicing boundaries and to correct inaccurate *ab initio* gene predictions [4,6]. Additionally, owing to the nature of cognate mapping, these spliced alignments provide higher accuracy when evaluating genes from multigene families by explicitly using only sequences native to the specific locus for annotation.

Quality assessment of predicted gene structures

Alignments are first evaluated to determine their native locus and, if necessary, the specific transcript isoform derived from the locus. A scoring system for comparing the spliced alignment with overlapping gene annotations was devised to aid in this determination (http://www.plantgdb.org/AtGDB/Annotation/gaeval/). Once a transcript isoform has been identified from which the EST or cDNA originated, all corresponding spliced alignments are compared with the predicted gene structure. This comparison is used to judge the accuracy of the gene annotation and to assign a quality flag for immediate appraisal of annotation validity. Five levels of annotation quality were established (Figure 1). The first quality level corresponds to an unconfirmed gene annotation for which no EST or cDNA evidence is currently available. These gene structure annotations are generally based entirely on *ab initio* computational prediction. Further analysis using homologous ESTs and cDNAs can be used to provide estimates of the annotation accuracy [20,21]. Annotations of quality levels beyond the first level benefit from the spliced alignment of ESTs and cDNAs. Increasing quality levels (Figure 1) represent increasing confidence in the accuracy and completeness of an annotation. Ultimately, the fifth level of quality assignment is given to gene annotations completely tiled by cognate ESTs or cDNAs, with all splice site boundaries supported. These annotations

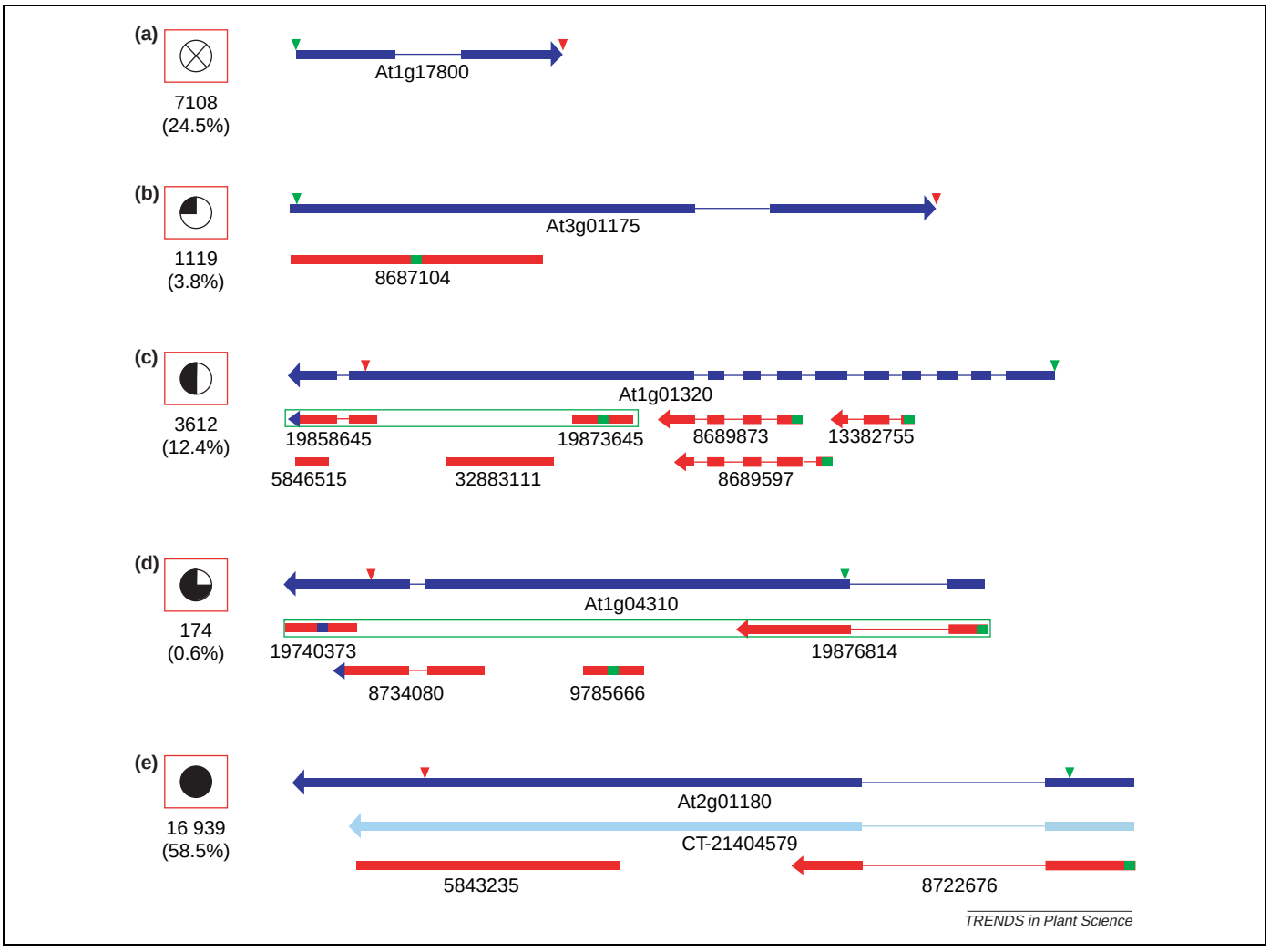

Figure 1. Levels of support for gene structure annotation. This figure uses existing gene annotations and their depiction (http://www.plantgdb.org/AtGDB/) to illustrate various levels of confirmation, evidenced by spliced alignment of cognate ESTs and cDNAs. As per AtGDB, the gene structure diagram consists of gene structure representations in which rectangular boxes are used to show exons, lines connecting these boxes depict introns and arrowheads imply forward and reverse strand transcription. GenBank-supplied gene structure predictions are shown as dark-blue arrow diagrams. Red arrow diagrams represent the spliced alignment of ESTs. Light-blue arrow diagrams are used for the spliced alignment of known full-length cDNAs. Each row also contains a flag surrounded by a red box to show the symbol used at AtGDB to indicate the degree of support for the gene structure annotation. **(a)** The least-confirmed level of gene structure annotation, in which there is no known EST or cDNA from this predicted gene. 7108 (24.5%) gene annotations within the current *Arabidopsis* pseudochromosome records fall into this category. **(b)** The next level of confirmation, in which an EST or cDNA sequence is shown for the region but no splice sites could be confirmed. This level of confirmation implies the existence of a gene yet tells us little about its gene structure. 1119 (3.8%) examples were noted. **(c)** A considerable improvement in confirmation of the given gene structure. These cases include annotations in which at least one splice site is confirmed by EST or cDNA spliced alignment. 3612 (12.4%) annotated genes fit this description. **(d,e)** Reserved for annotations in which all splice sites are confirmed. Level 4 annotations (d) differ from level 5 (e) only in their sequence coverage. As shown, level 4 annotations, 174 (0.6%) cases, include gaps in their sequence coverage (d), whereas level 5 annotations, 16939 (58.5%) cases, are completely covered from first exon to last (e).

represent well-supported gene structures that have the least anticipated need of future modification.

Inconsistencies found by this comparison are used to flag possible alternative splicing and gene structure deviations, as well as inaccurate prediction of introns and intergenic regions (Figure 2, http://www.plantgdb.org/AtGDB/Annotation/gaeval/gaeval_lists.php). Alternative splicing is suggested when the supported gene structure of a given locus is incongruent with that of the spliced alignment of one or more cognate expressed sequences. Validation of an alternative isoform can be based on criteria such as the number of ESTs and cDNAs supporting the alternative structure, the acceptability of the alternative splice junction relative to known models, and the surrounding context of the alternative isoform (e.g. proper open reading frame or presence of splicing enhancers). Consistent alignments can provide clues to incomplete or inaccurate annotation as well. For example, an expressed sequence alignment also matching to adjacent non-overlapping gene annotations is common evidence of a falsely predicted gene termination or intergenic region (e.g. center example in Figure 2). This mistake creates separate gene annotations representing fragments of a single gene. In addition, consideration of sequence vector properties, such as the source clone of an EST or the 5′ versus 3′ origination of the EST from the clone, can aid in determining the extent of a valid gene structure. Clone-pair ESTs (ESTs obtained from opposite ends of a cDNA clone) provide an often-overlooked indicator of fragmented gene structure annotation (Figure 2). Another less common mistake, whereby independent gene structures are incorrectly fused into a

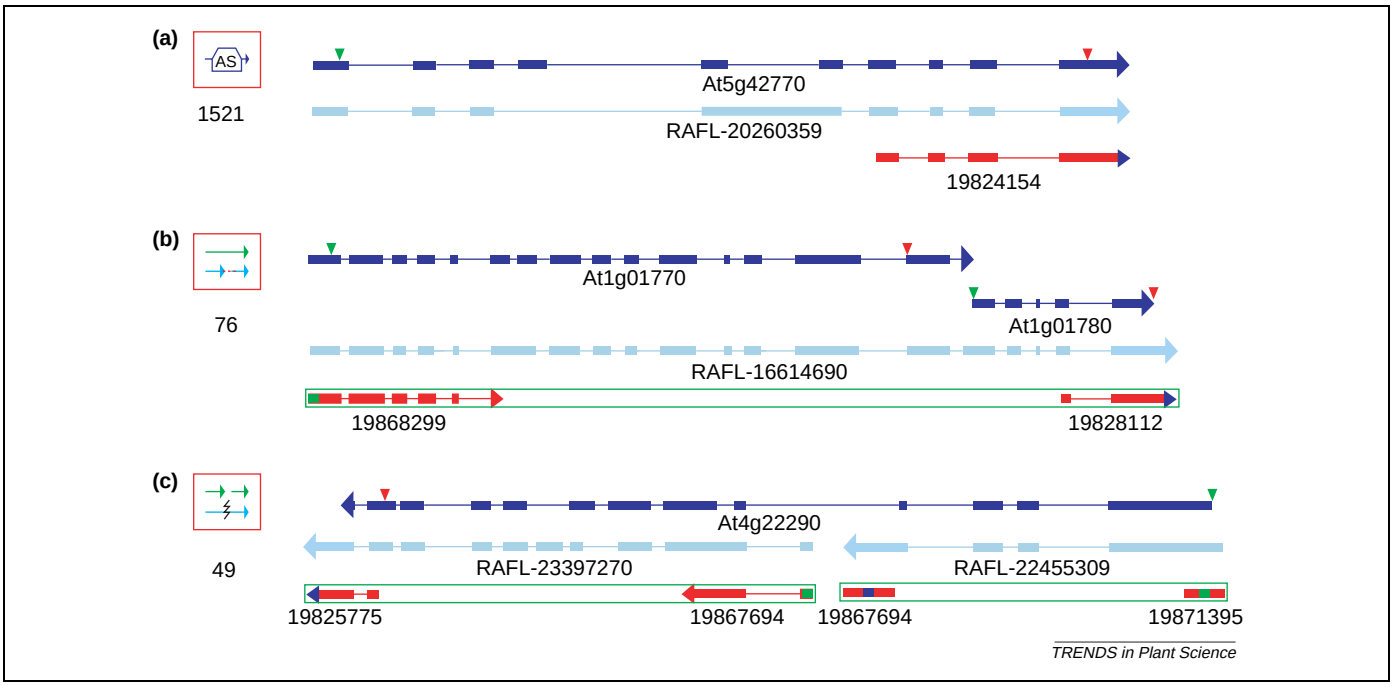

Figure 2. Automated selection of problematic annotations. Annotation flags are used to highlight annotations incongruent with spliced sequences. As in Figure 1, these flags are depicted with relevant examples and the number of automated finds for each event. GenBank-supplied gene structure predictions are shown as dark-blue arrow diagrams. Red arrow diagrams represent the spliced alignment of ESTs. Light-blue arrow diagrams are used for the spliced alignment of known full-length cDNAs. Each row also contains a flag surrounded by a red box to show the symbol used at AtGDB to indicate the degree of support for the gene structure annotation. **(a)** A case of alternative gene structure (or splicing) in which the annotated gene structure differs from that evidenced by the full-length cDNA (gi 20260359). 1521 such cases exist for the current *Arabidopsis* annotations. **(b)** Evidence of the false prediction of an intergenic region, necessitating the union of two adjacent gene structure predictions. 76 other such cases can be found. **(c)** The false prediction of an intron that causes the inaccurate union of two independent gene structures: 49 such cases exist.

single gene annotation, can be caught using clone-pair ESTs, groups of 3′ ESTs and 'full-length' cDNAs as evidence (http://www.plantgdb.org/AtGDB/Annotation/gaeval/cps.php). Although they require extremely robust algorithms to correct in an automated fashion, these anomalies are easily found and corrected by manual curation when presented appropriately.

Genomic context visualization, as provided at AtGDB [9,13], can be used to correct annotation mistakes for which automated correction is not feasible and to validate the behavior of novel automated annotation routines. For example, the inaccurate extension of some *Arabidopsis* gene models incorporated in the latest annotation release apparently resulted from changes in the automated annotation routines now in use [12]. Context visualization makes it possible to find these anomalies in a targeted and user-accessible manner (Figure 3). In addition, the genome context visualization and display of user comments allows for more complex annotation than can be captured with current GenBank feature tags. For example, some proportion of the gene structure pairs flagged as potentially needing to be merged by cDNA evidence correctly represent distinct translation products derived from dicistronic mRNAs (B. Haas, personal communication). Standards for annotating such cases have not been set, and therefore these cases are currently not represented in GenBank feature tags. It is also clear that such complex cases would be difficult to annotate automatically. The success of automated annotation pipelines relies on stringent criteria that capture the most reliable annotations [12]. In our view, the subsequent phase of completing the annotation can only be achieved by broad-based community input.

Community-based annotation

To ensure maximum participation by the community, the tools for updating annotations must be accessible and convenient for those viewing the data. Because TAIR is a heavily used resource, many users will first notice a structural annotation problem in TAIR. In addition, TAIR users already routinely submit comments and corrections on a range of types of data using a comment field on TAIR detail pages or by email directly to TAIR curators, suggesting that the TAIR user community is willing to contribute information. Therefore, we have added a link on TAIR data pages where gene structural annotation information is visible, allowing users wishing to correct a gene structure to connect automatically to AtGDB's GAEVAL system. By connecting TAIR and AtGDB through a centralized authentication service, we can enable a TAIR user's identity to be securely passed to AtGDB, ensuring proper attribution. The corrected structures are automatically sent back to TAIR, where they will be checked by a curator before being incorporated into the next version of the genome. The TAIR curator will examine the updated gene using the Apollo genome annotation tool [22] to confirm that existing cognate cDNAs and ESTs support the new gene model, to verify the translational start and stop for protein-coding genes, and to review and update any functional annotation attached to the gene. This review will insure that the new annotation conforms to TAIR's curation standards. If

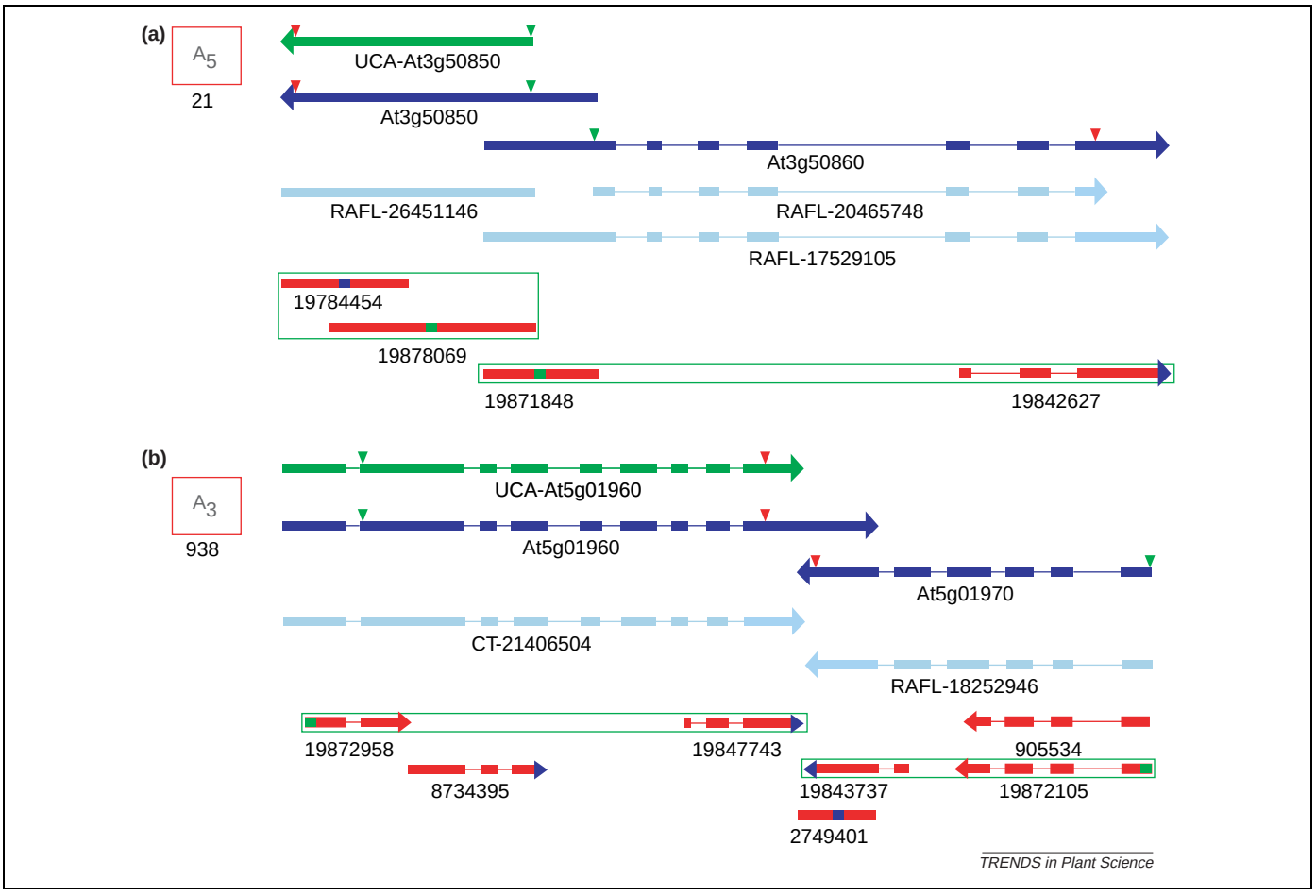

Figure 3. Erroneous assignment of 5′ and 3′ gene ends. GenBank-supplied gene structure predictions are shown as dark-blue arrow diagrams. Red arrow diagrams represent the spliced alignment of ESTs. Light-blue arrow diagrams are used for the spliced alignment of known full-length cDNAs. Each row also contains a flag surrounded by a red box to show the symbol used at AtGDB to indicate the degree of support for the gene structure annotation. Green user-contributed gene annotations represent the corrected gene structure supported by assignment of spliced aligned sequences. **(a)** The upstream extension of gene At3g50850 owing to the inclusion of EST gi-19871848. This EST clearly originates from gene At3g50860 as evidenced by its green bounding box with neighboring EST gi-19842627, representing the fact that these ESTs are the 5′ and 3′ respective ends of a single cDNA clone (clone-pair ESTs). **(b)** A similar likely wrong downstream extension of the gene model At5g01960 owing to EST gi-2749401.

TAIR curators detect a consistent pattern of error in user-submitted annotations, AtGDB will make use of this information to improve the interface to prevent the error, either by improving the tools available to the submitter or by alerting the submitter of the error at the time of submission.

Outreach

We believe that genome annotation could be an excellent vehicle for education, at both the high school and the undergraduate levels. To this end, we have developed a tutorial site at AtGDB (http://www.plantgdb.org/AtGDB/tutorial/) that guides users through the terminology and practice of gene structure annotation. This development was achieved in collaboration with local high schools in the Iowa State University area. In addition, talented high-school student interns have proven to be both eager and effective users of these gene annotation tools and have greatly contributed to improvements in tool design and scope (see http://www.plantgdb.org/AtGDB/Interns/ for project descriptions and results).

Conclusions

The community curation approach has the potential to solve the problem of how to maintain a high-quality genome annotation for the long term. Although some of the problems with existing automated annotation pipelines might eventually be corrected, manual curation remains the best method for producing high-quality genome annotation. The tools and resources presented here have made such community curation convenient and efficient while providing wide access to the resulting data. More than 300 such community annotations are currently accessible at AtGDB (see http://www.plantgdb.org/AtGDB/Annotation/UCAlist.php).

Acknowledgements

This work was supported by NSF Plant Genome Research Grant DBI-0321600 to V.B. and NSF grant number DBI-9978564 to S.Y.R. We thank Michael Lawler and Stephanie Haila, both school science teachers in Iowa, for working with us to develop the gene structure annotation tutorial. Their work was sponsored by an NSF RET grant to Iowa State University. We also thank our colleagues at The Institute for Genomic Research (TIGR) for critical reading and comments.

References

1 Roberts, L. (2001) Controversial from the start. *Science* 291, 1182–1188
2 The *Arabidopsis* Genome Initiative (2002) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408, 796–815

3 Wortman, J.R. *et al*. (2003) Annotation of the *Arabidopsis* genome. *Plant Physiol.* 132, 461–468
4 Brendel, V. and Zhu, W. (2002) Computational modeling of gene structure in *Arabidopsis thaliana*. *Plant Mol. Biol.* 48, 49–58
5 Ausubel, F.M. (2002) Summaries of National Science Foundation-sponsored *Arabidopsis* 2010 projects and National Science Foundation-sponsored plant genome projects that are generating *Arabidopsis* resources for the community. *Plant Physiol.* 129, 394–437
6 Pavy, N. *et al*. (1999) Evaluation of gene prediction software using a genomic data set: application to *Arabidopsis thaliana* sequences. *Bioinformatics* 15, 887–899
7 Seki, M. *et al*. (2002) Functional annotation of a full-length *Arabidopsis* cDNA collection. *Science* 296, 141–145
8 Haas, B.J. *et al.* (2002) Full-length messenger RNA sequences greatly improve genome annotation. *Genome Biol*. 3, research0029.1–research0029.12
9 Zhu, W. *et al*. (2003) Refined annotation of the *Arabidopsis thaliana* genome by complete EST mapping. *Plant Physiol.* 132, 469–484
10 Yamada, K. *et al*. (2003) Empirical analysis of transcriptional activity in the *Arabidopsis* genome. *Science* 302, 842–846
11 Castelli, V. *et al*. (2004) Whole genome sequence comparisons and 'full-length' cDNA sequences: a combined approach to evaluate and improve *Arabidopsis* genome annotation. *Genome Res.* 14, 406–413
12 Haas, B.J. *et al*. (2003) Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* 31, 5654–5666
13 Dong, Q. *et al*. (2004) PlantGDB, plant genome database and analysis tools. *Nucleic Acids Res.* 32, D354–D359
14 Hild, M. *et al*. (2003) An integrated gene annotation and transcriptional profiling approach towards the full gene content of the *Drosophila* genome. *Genome Biol.* 5, R3.1–R3.16
15 Rhee, S.Y. *et al*. (2003) The *Arabidopsis* Information Resource (TAIR): a model organism database providing a centralized, curated gateway to *Arabidopsis* biology, research materials and community. *Nucleic Acids Res.* 31, 224–228
16 Schoof, H. *et al*. (2004) MIPS *Arabidopsis thaliana* Database (MAtDB): an integrated biological knowledge resource for plant genomics. *Nucleic Acids Res.* 32, D373–D376
17 Berardini, T.Z. *et al*. (2004) Functional annotation of the *Arabidopsis* genome using controlled vocabularies. *Plant Physiol.* 135, 745–755
18 Rhee, S.Y. (2004) *Carpe diem*: retooling the 'publish or perish' model into the 'share and survive' model. *Plant Physiol.* 134, 543–547
19 Brendel, V. Novel tools for plant genome annotation and applications to *Arabidopsis* and rice. In *Genome Exploitation: Data Mining (Stadler Genetics Symposia Series, 23rd Symposium)* (Gustafson, J.P. *et al.*, eds), Kluwer Academic/Plenum (in press)
20 Schlueter, S.D. *et al*. (2003) GeneSeqer@PlantGDB: gene structure prediction in plant genomes. *Nucleic Acids Res.* 31, 3597–3600
21 Brendel, V. *et al*. (2004) Gene structure prediction from consensus spliced alignment of multiple ESTs matching the same genomic locus. *Bioinformatics* 20, 1157–1169
22 Lewis, S.E. *et al*. (2002) Apollo: a sequence annotation editor. *Genome Biol*. 3, research0082.1–research0082.14