CURRENT OPINION
www.current-opinion.com

Comparison of rice and *Arabidopsis* annotation

Heiko Schoof* and Wojciech M Karlowski†

Several versions of the rice genome were published in 2002, providing a first overview of the genome content of this model monocot. At the same time, the genome of the model dicot, *Arabidopsis thaliana*, reached a new level of annotation as thousands of full-length cDNA sequences were integrated with the genome sequence.

Addresses
*Technical University of Munich, Genome Oriented Bioinformatics, Wissenschaftszentrum Weihenstephan, 85350 Freising, Germany
e-mail: h.schoof@wzw.tum.de
†GSF Forschungszentrum für Umwelt und Gesundheit, IBI/MIPS, Ingolstaedter Landstrasse 1, 85764 Neuherberg, Germany
e-mail: w.karlowski@gsf.de

Current Opinion in Plant Biology 2003, **6**:106–112

This review comes from a themed issue on
Genome studies and molecular genetics
Edited by Takuji Sasaki and Ronald R Sederoff

1369-5266/03/$ – see front matter

DOI 10.1016/S1369-5266(03)00003-7

Abbreviations

AGI	Arizona Genomics Institute
BAC	bacterial artificial chromosome
BGI	Beijing Genomic Institute
CUGI	Clemson University Genomic Institute
EST	expressed sequence tag
GO	gene ontology
IRGSP	International Rice Genome Sequencing Project
MIPS	Munich Information Center for Protein Sequences
NCBI	National Center for Biotechnology Information
NCGR	National Center for Gene Research
RGP	Rice Genome Research Program
TIGR	The Institute for Genomic Research

Introduction

The first complete plant genome, that of the model dicotyledonous plant *Arabidopsis thaliana*, was published in 2000 [1]. In 2002, several rice genome projects have published their data [2,3]. With this recent explosion of rice genomic data (both in quantity and diversity), the accurate and precise characterization, description and classification of genetic elements encoded by rice DNA have become extremely important tasks. Meanwhile, large-scale cDNA sequencing projects for *Arabidopsis* have allowed the confirmation or correction of the gene structures of thousands of genes [4••]. This progress allows critical review of the initial annotation of the *Arabidopsis* genome, and expands the annotation of this genome to include data on untranslated regions (UTRs) and splicing anomalies. Therefore, it seems that now is the right time to compare the annotation results from the *Arabidopsis* and rice genomes, and to take a critical view of the strategies and quality standards used for genome annotation.

Rice and *Arabidopsis*, which are widely accepted models for monocotyledonous and eudicotyledonous plants, respectively, diverged from a common ancestor about 200 million years ago [5]. Fundamental differences between the *Arabidopsis* and rice genomes include their size and gene content, the rice genome being four times larger and containing up to twice as many genes as that of *Arabidopsis*. (*Arabidopsis*: 125 Mbp and 26 422 genes [1], rice: 420–466 Mbp and a maximum gene number of 50 000–55 615 [2,3].) The two genomes show very limited synteny: the largest reported syntenic region maps to *Arabidopsis* chromosome 5 and rice chromosome 4 and covers 119 *Arabidopsis* proteins, which show at least 70% identity over a minimum of 30 contiguous amino acids [2].

Arabidopsis is exclusively of scientific interest, whereas rice is a major food source. The establishment of several rice genome projects (both public and proprietary) with the primary goal of sequencing the whole genome has manifested the scientific and economical interest in rice. Whereas a multinational public effort produced about 115 Mbp of *Arabidopsis* sequence over a period of four years, three projects have independently sequenced the genomes of two different rice subspecies. They have already produced more then ten times the sequence data present in *Arabidopsis* databases. For a status report on the rice genome projects, please see the review by Michel Delseny in this issue. The centers involved in the International Rice Genome Sequencing Project (IRGSP) and the commercial efforts are summarized in Table 1, as they will be referenced frequently within this review.

The independent sequencing and annotation of three rice genome datasets has resulted in a plethora of available data. Interpretation of these data requires reliable annotation tools and the integration of the data with extrinsic information. The first genome-wide annotation sets for rice are just emerging, but *Arabidopsis* annotation has received a boost through the integration of full-length cDNA data [4••]. In this review, we attempt to summarize the approaches utilized to annotate the *Arabidopsis* and rice genomes and the information gained.

Annotation

Annotation can be defined as the attachment of information to a sequence. For example, features such as genes or

Table 1

Rice genome centers and resources.

Name	Website	Sequence and resources
Beijing Genomic Institute (BGI)	http://btn.genomics.org.cn/rice	Whole-genome shotgun sequence of *Oryza sativa* ssp. *indica* cv. 93-11 [3]
Syngenta/Torrey Mesa	http://www.tmri.org/index.html	Whole-genome shotgun sequence of *Oryza sativa* ssp. *japonica* cv. 'Nipponbare' [2]
Monsanto	http://www.monsanto.com	Working draft of the *O. sativa* ssp. *japonica* cv. 'Nipponbare' genome contributed to the IRGSP [33]
Rice Genome Research Program (RGP)*	http://rgp.dna.affrc.go.jp	*O. sativa* ssp. *japonica* cv. 'Nipponbare' Chromosomes 1, 2, 6, 7, 8 and 9.
Arizona Genomics Institute (AGI)*/	http://www.genome.arizona.edu	*O. sativa* ssp. *japonica* cv. 'Nipponbare' Chromosomes 3 and 10.
Clemson University Genomic Institute (CUGI)*	http://www.genome.clemson.edu	
Cold Spring Harbor Laboratory (CSHL)*	http://nucleus.cshl.org/riceweb	*O. sativa* ssp. *japonica* cv. 'Nipponbare' Chromosomes 3 and 10.
The Institute for Genomic Research (TIGR)*	http://www.tigr.org/tdb/e2k1/osa1	*O. sativa* ssp. *japonica* cv. 'Nipponbare' Chromosomes 3, 10 and 11.
Korea Rice Genome Research Program (KRGRP)*	http://biogen.niast.go.kr	*O. sativa* ssp. *japonica* cv. 'Nipponbare' Chromosomes 1 and 9.
National Center for Gene Research (NCGR)	http://www.ncgr.ac.cn/index.html	Chromosome 4 of *Oryza sativa* spp. *japonica* cv. 'Nipponbare'† and *indica* cv. 93-11.
Academia Sinica Plant Genome Center (ASPGC)*	http://genome.sinica.edu.tw	*O. sativa* ssp. *japonica* cv. 'Nipponbare' Chromosomes 3 and 5.
Genoscope*	http://www.genoscope.cns.fr/externe/English/Projets/Projet_CC/CC.html	*O. sativa* ssp. *japonica* cv. 'Nipponbare' Chromosome 12.
The Plant Genome Initiative at Rutgers (PGIR)*	http://pgir.rutgers.edu	*O. sativa* ssp. *japonica* cv. 'Nipponbare' Chromosomes 10 and 11.
Indian Initiative for Rice Genome Sequencing (IIRGS)*	http://www.nrcpb.org/rgp.html	*O. sativa* ssp. *japonica* cv. 'Nipponbare' Chromosome 11.
National Center for Genetic Engineering and Biotechnology (BIOTEC)*	http://www.cs.ait.ac.th/nstda/biotec/biotec.html	*O. sativa* ssp. *japonica* cv. 'Nipponbare' Chromosome 9.
John Innes Centre*	http://www.jic.bbsrc.ac.uk	*O. sativa* ssp. *japonica* cv. 'Nipponbare' Chromosome 2.
Wisconsin Rice Genome Project* (GCOW)	http://www.gcow.wisc.edu/Rice/index.htm	*O. sativa* ssp. *japonica* cv. 'Nipponbare' Chromosome 11.

*IRGSP member. †See note added in proof.

repeat elements are marked and metadata are added. The metadata might include information such as the sequencing center or experimental source used for functional assignment. Genome projects generally take a two-layered approach to annotation. On one level, the coordinates of elements such as genes, repeats, clones or expressed sequence tag (EST) matches are detected or marked. On the second level, additional information is attached to these elements; this may include a name, description, classification, source or confidence value.

This approach does not intrinsically differentiate between manual and automated annotation. However, the need to label clearly the source and/or reliability of the annotation becomes obvious. We have learned to accept that the annotations for the majority of genes in genome databases rely only on prediction methods and are probably not completely correct, but the quality of annotation can vary greatly. For anyone not involved in the annotation process, judging the reliability of a given dataset is difficult. Although estimations of the performance of gene prediction methods are generally available, the reliability of genome annotation datasets is not widely known.

To provide an overview of the rice and *Arabidopsis* genome annotation, and to aid readers in appreciating its quality, we discuss the approaches applied by different contributors. Details are summarized in Table 2.

Gene prediction

The coding portion of the sequence is a primary focus in genome analysis. The detection of protein-coding genes is an important first step that forms the basis for further functional analyses. Accurate predictions of the complete structures of protein-coding genes, including their 5′- and 3′-untranslated regions, are crucial for full interpretation of genome sequence [6••].

In general, gene prediction programs can be divided into three classes: alignment-based, *ab initio* and hybrid algorithms [7•,8]. The last of these three classes combines *ab initio* gene predictions with alignments between related sequences. A good introduction on how to build your own

Table 2

Comparison of rice annotation pipelines*.

Genomic/annotation center	EST/cDNA mapping database	*In silico* predictions	Manual curation	Functional annotation†
BGI	No	FGENESH (monocots)	No	*Arabidopsis* homologs
Syngenta	Various plant and fungal sequences	FGENESH (monocots), GeneMark.hmm (*Arabidopsis*, rice), Genscan (*Arabidopsis*)	–	Protein databases, Pfam, Prosite
RGP (RiceGAAS)	National Center for Biotechnology Information (NCBI) nr, internal rice cDNA database	Genscan (maize, *Arabidopsis*), RiceHMM, MZEF, SplicePredictor, tRNAscan-SE	Yes/No	NCBI nr, Pfam, Prosite; (MIPS functional categories)
TIGR	NCBI nr, TIGR Plant Gene Index	FGENESH, GeneMark.hmm (rice), Genscan (maize), Genscan+ (*Arabidopsis*), GlimmerR, GeneSplicer, tRNAscan-SE	Yes (only for BACs sequenced at TIGR)	No
ASPGC	NCBI nr, TIGR Rice Gene Index	Genscan, GlimmerR	–	–
NCGR	NCBI nr, EST database at NCGR	FGENESH, Genscan, GeneMark.hmm, tRNAscan-SE	–	–
PGIR	Various ESTs	Genscan, FGENESH++, tRNAscan-SE	–	–
CSHL	–	–	Yes	–
CUGI/AGI	–	Genscan, Genscan+, GeneMark.hmm, Xgrail, NetGene2, tRNAscan-SE	–	–
IIRGS	GenBank (full cDNA sequences)	Genscan, GeneMark, Gene Finder	–	NCBI nr, Swiss-Prot
Gramene	–	Ensembl [25]	Yes	Ensembl

– no data available. The organism(s) on which the gene finders were trained are given in parentheses. *This list is incomplete and subject to change as the methods are rapidly updated. †Sequence databases listed here were used for homology-based function prediction.

gene finder is given by Pertea and Salzberg [9]. These authors also discuss the accuracy of several algorithms when applied to rice or *Arabidopsis*. They show clearly that gene predictors that are trained on sequences from dicotyledonous plants will not work well on monocot sequences. Another important conclusion is that combining the output of several algorithms can detect protein-coding regions more accurately than the use of a single algorithm. Yuan and colleagues published a discussion on the design and training of their gene finder, GlimmerR [10•]. The commercially available FGENESH, a hybrid algorithm, is currently accepted as the most useful prediction tool for rice sequences [3].

The initial annotation of the *Arabidopsis* genome relied on the manual synthesis of the output from several prediction programs, including Genscan and GenemarkHMM, with extrinsic data such as protein similarities and EST matches [1]. Recently, the *Arabidopsis* gene models were corrected on the basis of full-length cDNA sequence data available from several large-scale projects [4••,11]. About one-third of *Arabidopsis* gene models are now supported by full-length cDNA alignments (H Schoof, unpublished data). 35% of these cDNA sequences led to the adjustment of the respective gene structures, and 5% identified new genes [4••]. The frequency of these necessary corrections suggest that the initial annotation provided about 60% exact gene models (i.e. genes in which a start, stop and all exon–intron boundaries are exactly correct), suggesting that the performance of the gene finders was better than had been expected [9,12]. This may be due to manual curation or the integration of extrinsic data such as EST and protein alignments.

For rice, the gene-prediction strategies generally include both homology-based methods and initial predictions. One of the most important handicaps when attempting to predict rice genes is the lack of high-quality EST sequences and full-length cDNAs. The training set of full-length cDNAs for rice gene finders is pitifully small at present. It may be possible to use homology-based methods, such as Doublescan, to overcome this problem [13]. A useful approach to assessing the quality of the current predictions is to classify the predicted rice genes on the basis of the homology of their predicted proteins to known proteins. This approach was utilized by the Beijing Genomic Institute (BGI) to create a trustworthy set of rice genes. However, this work revealed a feature that greatly impairs the performance of current gene finders: rice genes show a gradient of GC content, with a high proportion of GC nucleotides at the 5′ end [3,14] (Table 3).

Prediction of function

Once genes have been identified, the protein sequence that they encode forms the basis for a second layer of annotation. Although the number of characterized plant

Table 3

Gene prediction software used to annotate rice.

Name	Web server	Reference(s)
Genscan	http://genes.mit.edu/GENSCAN.html	[34]
FGENESH	http://www.softberry.com/berry.phtml?topic=gfind	[35]
GeneMark.hmm	http://opal.biology.gatech.edu/GeneMark/eukhmm.cgi	(a)
GlimmerR	http://www.tigr.org/tdb/glimmerm/glmr_form.html	[36]
RiceHMM	http://rgp.dna.affrc.go.jp/RiceHMM/index.html	[37]
tRNAscan-SE	http://www.genetics.wustl.edu/eddy/tRNAscan-SE	[38]
SplicePredictor	http://bioinformatics.iastate.edu/cgi-bin/sp.cgi	[39]
GeneSplicer	http://www.tigr.org/tdb/GeneSplicer/gene_spl.html	[40]
Gene Finder (MZEF)	http://argon.cshl.org/genefinder	[41]
NetGene2	http://www.cbs.dtu.dk/services/NetGene2	[42,43]

(a) M Borodovsky, A Lukashin, unpublished data.

genes with known function is small, a large percentage of genes can be assigned a putative function on the basis of the homology of their predicted proteins to either proteins of known function or domain profiles. For *Arabidopsis* annotation, manual inspection of a major part of the genome could be used to control the quality of the assigned functions. Whereas just 10% of *Arabidopsis* genes have a characterized function, more than 70% could be functionally classified [1]. Classification systems such as Gene Ontology [15••] and the Munich Information Center for Protein Sequences (MIPS) functional catalog [16] have been used to this end. The Interpro domain database has proven especially useful for genome-wide comparisons with other species [1].

It is possible to predict the function of rice genes on the basis of homology searches against different protein (nr, Swiss-Prot; see Table 2) and protein-domain databases (Pfam, Prosite). A major part of the functional assignment is based on genes whose function has been predicted in another organism, mostly *Arabidopsis*. So the quality of *Arabidopsis* annotation directly influences the functional annotation of rice genes.

Non-coding features

The advantage of whole-genome sequencing, as opposed to sequencing cDNA or gene-rich islands, is the possibility to study non-coding genetic elements. The initial *Arabidopsis* annotation included RNA genes (tRNA, rRNA, snoRNA, spliceosomal RNA), an incomplete and inhomogeneous (i.e. performed by different groups, each using their own definitions and standards) analysis of transposable elements and pseudogenes, and some characterized repeats. Since the publication of this annotation, more detailed analyses — such as those of microsatellites [17], mobile elements [18], and small RNAs [19••,20] — have been published but only partially integrated into the genome databases.

More than 50% of the rice genome consists of repetitive DNA. A number of different approaches, including specialized repeat databases or detection algorithms, have been adopted to identify those elements. The Institute for Genomic Research (TIGR) has compiled a rice repeat database that includes transposons, retroelements and miniature inverted-repeat transposable elements (MITEs) [10•]. On the other hand, the Rice Genome Program (RGP) identifies transposable elements by similarity to *gag* and *pol* genes, which encode polyproteins that are associated with retroelement function. Matches are then sandwiched with long terminal repeats (LTRs) [21]. Syngenta detects rRNA units by searching for consensus sequence patterns [2].

Annotation pipelines

The rate at which genome-sequencing projects generate data makes some form of high-throughput annotation pipeline necessary; annotation usually becomes the rate-limiting step once the sequencing machines are in full swing. This is especially true for whole-genome shotgun projects in which annotation can usually begin only once the complete dataset is assembled.

The initial annotation of the *Arabidopsis* project involved much manual work. However, the need for automated procedures for the long-term curation of *Arabidopsis* annotation soon became evident. Both TIGR [4••] and MIPS have implemented automatic updating of gene models on the basis of cDNA alignments, with manual intervention in problematic cases. The aim is to allow manual annotation to focus on tasks that cannot be solved by automated methods, thereby gaining the maximum benefit from human knowledge.

Some early rice genome fragments were annotated by hand [22]. But automated procedures, augmented by manual work, have been used for the majority of available data. RiceGAAS [23] is a fully automatic system that is designed specifically to provide continually current and consistent annotation of the rice genome. TIGR utilizes an automated pipeline for preliminary analysis before manual curation [10•].

The Gramene grass comparative genomics resource [24], like RiceGAAS, downloads all available sequence data

regularly, but this system includes a rice-specific adaptation of the Ensembl [25] system that provides consistent baseline annotation.

Databases

The two original bioinformatics data management centers — the *Arabidopsis* Genome Initiative Centers TIGR and MIPS — maintain annotated *Arabidopsis* genome databases that are continually enhanced with new functionality and data [4••,26]. The *Arabidopsis* Information Resource (TAIR), traditionally a 'one-stop-shop' for *Arabidopsis* data, has also integrated the genome annotation [27].

One of the richest resources for rice-related data is provided by TIGR. A database containing information from manually supervised annotation of TIGR-sequenced bacterial artificial chromosome (BAC) clones is complemented by a database containing automatically generated annotation of all of the sequences contributed by IRGSP members (see Table 2). Moreover, TIGR also provides automatic annotation for their unfinished rice clones [10•].

The INE (Integrated Rice Genome Explorer) database system was developed at RGP to store data from sequencing and annotation projects and to integrate these data with map-based information [21,28,29].

Gramene [24] provides full access to all publicly available rice sequences, including homology searches and an advanced retrieval system. Gramene is not restricted to rice, but includes information on other members of the *Gramineae* family. It concentrates especially on comparative genomics, providing comparative maps between rice and other grasses. These are based upon orthologous sequences and phenotype information as well as gene ontology (GO) functional classifications [30]. To complement the automated annotations, Gramene will be developed to provide the tools and expertise necessary to classify rice genes by GO. Gramene will add electronically generated SWISS-PROT and InterPro annotation, as well as manually augmented annotation, to GO mappings [24].

The MIPS *Oryza sativa* database (MOsDB) is another emerging resource for rice annotation [31]. In this database, automatically collected public data are fed into an annotation management system that has been adapted and developed to work alongside the MIPS *Arabidopsis thaliana* database (MAtDB) [26]. The PEDANT protein analysis tool [16] provides detailed functional and structural analysis of all of the predicted proteins stored at MOsDB. EST data from various plants, including grasses, is integrated with MOsDB through the SPUTNIK system [31]. This system is designed to provide a comprehensive resource that allows the transfer of knowledge between different plant species (W Karlowski, unpublished data).

It is worth noting that access to rice sequences from shotgun projects is very limited, despite their publication in a scientific journal [2,3]. Syngenta provides license-based access to genomic DNA and annotations. On the other hand, the BGI grants full access to DNA sequence data, that is to genomic sequences plus the EST sequences used for quality validation and annotation, but not to annotation data.

Conclusions

The annotation of the rice genome is still impaired by unreliable gene-prediction methods, but various efforts are being undertaken to address this problem. Meanwhile, *Arabidopsis* annotation has been corrected through the integration of full-length cDNA data. Previously unavailable information, such as data on alternative splicing, is now accessible. The initial annotation of the *Arabidopsis* genome was a manual process, during which the avoidance of duplication of effort was of prime importance. In contrast, automated methods with minimal manual supervision dominate rice annotation. This has allowed several groups to complete the annotation of the whole genome. Each of these groups has very different standards and quality requirements, but such diversity may actually help to improve quality by allowing comparisons of different versions and datasets. However, this requires the development and implementation of new standards that allow the integration of several annotation sources [32]. Interfaces between various annotation datasets also need to be implemented to enable comparative genomics and knowledge transfer between species.

The evolution from manual to fully automatic annotation has been paralleled by an increase in both the volume of available data and the accuracy of automatic methods. Although human inspection can still improve the quality of automatic predictions, this may be outweighed by the advantage of rapid re-annotation and thus constant improvement.

The massive use of *Arabidopsis* annotation for the analysis of the rice genomes emphasizes the need for reliable annotation of model genomes: erroneous annotation will propagate. On the other hand, the effort put into sequencing and annotating the model genomes will be well rewarded: annotation will be transferred to related sequences and used again and again.

Note added in proof

Recently, Sasaki *et al.* [44] and Feng *et al.* [45] reported the sequences of rice spp. *japonica* chromosomes 1 and 4, respectively. These publicly available, essentially complete, sequences are annotated and functionally categorized. Comparison with restricted regions of the draft whole-genome sequence (of *O. sativa* spp. *indica*) indicates that 43% of the predicted genes from the whole-

genome shotgun were incomplete, which hints at the importance of the sequence quality used for gene prediction and annotation.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. The *Arabidopsis* Genome Initiative: **Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana***. *Nature* 2000, **408**:796-815.

2. Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, Glazebrook J, Sessions A, Oeller P, Varma H *et al.*: **A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*)**. *Science* 2002, **296**:92-100.

3. Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, Deng Y, Dai L, Zhou Y, Zhang X *et al.*: **A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*)**. *Science* 2002, **296**:79-92.

4. •• Haas BJ, Volfovsky N, Town CD, Troukhan M, Alexandrov N, Feldmann KA, Flavell RB, White O, Salzberg SL: **Full-length messenger RNA sequences greatly improve genome annotation**. *Genome Biol* 2002, **3**:RESEARCH0029.

This paper is a must for anyone working on the structures of plant genes. Beside a good technical overview of problems in aligning cDNA and genomic sequences, the authors provide a discussion of the anomalies that have been observed in the cDNA data. For example, alternative transcription start or polyadenylation sites may play a larger role in plants than previously assumed.

5. Wolfe KH, Gouy M, Yang YW, Sharp PM, Li WH: **Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data**. *Proc Natl Acad Sci USA* 1989, **86**:6201-6205.

6. •• Brent MR: **Predicting full-length transcripts**. *Trends Biotechnol* 2002, **20**:273-275.

A short update on finding eukaryotic genes that tackles an old source of inaccuracy in gene prediction, the consideration of only translated exons, and brings the attention they deserve to first exons. Although Brent is optimistic about computational approaches, a reply by MQ Zhang calls for large-scale experimental data.

7. • Stormo GD: **Gene-finding approaches for eukaryotes**. *Genome Res* 2000, **10**:394-397.

A clearly written summary of the art of gene prediction.

8. Mathé C, Sagot M-F, Schiex T, Rouzé P: **Current methods of gene prediction, their strengths and weaknesses**. *Nucleic Acids Res* 2002, **30**:4103-4117.

9. Pertea M, Salzberg SL: **Computational gene finding in plants**. *Plant Mol Biol* 2002, **48**:39-48.

10. • Yuan Q, Quackenbush J, Sultana R, Pertea M, Salzberg SL, Buell CR: **Rice bioinformatics. Analysis of rice sequence data and leveraging the data to other plant species**. *Plant Physiol* 2001, **125**:1166-1174.

An introduction to the annotation and analysis tools employed by TIGR for the rice genome. The authors provide an overview that includes background information on the IRGSP, resources, a description of GlimmerR (the TIGR gene finder) and ideas for knowledge transfer between species, such as Tentative Orthologous Groups.

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

12. Brendel V, Zhu W: **Computational modeling of gene structure in *Arabidopsis thaliana***. *Plant Mol Biol* 2002, **48**:49-58.

13. Meyer IM, Durbin R: **Comparative *ab initio* prediction of gene structures using pair HMMs**. *Bioinformatics* 2002, **18**:1309-1318.

14. Wong GK, Wang J, Tao L, Tan J, Zhang J, Passey DA, Yu J: **Compositional gradients in Gramineae genes**. *Genome Res* 2002, **12**:851-856.

15. •• The Gene Ontology Consortium: **Creating the gene ontology resource: design and implementation**. *Genome Res* 2001, **11**:1425-1433.

Gene Ontology is a tool for functional annotation that has the potential to integrate annotations from different projects and genomes. If functional annotation on a large scale is to remain tractable, standardization is necessary. Although the complexity of GO may be overwhelming, the basic concepts are clearly laid down in this paper.

16. Frishman D, Albermann K, Hani J, Heumann K, Metanomski A, Zollner A, Mewes HW: **Functional and structural genomics using PEDANT**. *Bioinformatics* 2001, **17**:44-57.

17. Morgante M, Hanafey M, Powell W: **Microsatellites are preferentially associated with non-repetitive DNA in plant genomes**. *Nat Genet* 2002, **30**:194-200.

18. Mette MF, Van Der Winden J, Matzke M, Matzke AJ: **Short RNAs can identify new candidate transposable element families in *Arabidopsis***. *Plant Physiol* 2002, **130**:6-9.

19. •• Llave C, Kasschau KD, Rector MA, Carrington JC: **Endogenous and silencing-associated small RNAs in plants**. *Plant Cell* 2002, **14**:1605-1619.

The world of non-coding RNAs is relatively unexplored. In this systematic study, the authors uncover a family of small RNAs, which they called miRNAs, from *Arabidopsis*. These RNAs are difficult to detect and may be much more abundant than suspected to date. The authors gather hints as to their role, opening up glimpses of a powerful regulatory system.

20. MacIntosh GC, Wilkerson C, Green PJ: **Identification and analysis of *Arabidopsis* expressed sequence tags characteristic of non-coding RNAs**. *Plant Physiol* 2001, **127**:765-776.

21. Sasaki T: **The progress in rice genomics**. *Euphytica* 2001, **118**:103-111.

22. Mayer K, Murphy G, Tarchini R, Wambutt R, Volckaert G, Pohl T, Dusterhoft A, Stiekema W, Entian KD, Terryn N *et al.*: **Conservation of microstructure between a sequenced region of the genome of rice and multiple segments of the genome of *Arabidopsis thaliana***. *Genome Res* 2001, **11**:1167-1174.

23. Sakata K, Nagamura Y, Numa H, Antonio BA, Nagasaki H, Idonuma A, Watanabe W, Shimizu Y, Horiuchi I, Matsumoto T *et al.*: **RiceGAAS: an automated annotation system and database for rice genome sequence**. *Nucleic Acids Res* 2002, **30**:98-102.

24. Ware D, Jaiswal P, Ni J, Pan X, Chang K, Clark K, Teytelman L, Schmidt S, Zhao W, Cartinhour S *et al.*: **Gramene: a resource for comparative grass genomics**. *Nucleic Acids Res* 2002, **30**:103-105.

25. Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, Cox T, Cuff J, Curwen V, Down T *et al.*: **The Ensembl genome database project**. *Nucleic Acids Res* 2002, **30**:38-41.

26. Schoof H, Zaccaria P, Gundlach H, Lemcke K, Rudd S, Kolesov G, Arnold R, Mewes HW, Mayer KF: **MIPS *Arabidopsis thaliana* Database (MAtDB): an integrated biological knowledge resource based on the first complete plant genome**. *Nucleic Acids Res* 2002, **30**:91-93.

27. Huala E, Dickerman AW, Garcia-Hernandez M, Weems D, Reiser L, LaFond F, Hanley D, Kiphart D, Zhuang M, Huang W *et al.*: **The *Arabidopsis* Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant**. *Nucleic Acids Res* 2001, **29**:102-105.

28. Sakata K, Antonio BA, Mukai Y, Nagasaki H, Sakai Y, Makino K, Sasaki T: **INE: a rice genome database with an integrated map view**. *Nucleic Acids Res* 2000, **28**:97-101.

29. Antonio BA, Sakata K, Sasaki T: **Rice at the forefront of plant genome informatics**. *Genome Informatics* 2000, **11**:3-11.

30. Jaiswal P, Ware D, Ni J, Chang K, Zhao W, Schmidt S, Pan X, Clark K, Teytelman L, Cartinhour S *et al.*: **Gramene: development and integration of trait and gene ontologies for rice**. *Comp Funct Genom* 2002, **3**:132-136.

31. Karlowski WM, Schoof H, Janakiraman V, Stuempflen V, Mayer KFX: **MOsDB: an integrated information resource for rice genomics**. *Nucleic Acids Res* 2003, **31**:in press.

32. Dowell RD, Jokerst RM, Day A, Eddy SR, Stein L: **The distributed annotation system**. *BMC Bioinformatics* 2001, **2**:7.

33. Barry GF: **The use of the Monsanto draft rice genome sequence in research**. *Plant Physiol* 2001, **125**:1164-1165.

34. Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA**. *J Mol Biol* 1997, **268**:78-94.

35. Solovyev VV: **Finding genes by computer: probabilistic and discriminative approaches**. In: *Current Topics in Computational Biology*. Edited by Jiang T, Smith T, Xu Y, Zhang M. Cambridge, MA, USA: MIT Press; 2002:365-401.

36. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL: **Improved microbial gene identification with GLIMMER**. *Nucleic Acids Res* 1999, **27**:4636-4641.

37. Sakata K, Nagasaki H, Idonuma A, Waki K, Kise M, Sasaki T: **A computer program for prediction of gene domain on rice genome sequence [abstract]**. In *The 2nd Georgia Tech International Conference on Bioinformatics*: 1999 November 11–14. Atlanta; 1999:78. http://opal.biology.gatech.edu/GeneMark/conference99/

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

39. Usuka J, Brendel V: **Optimal spliced alignment of homologous proteins to a genomic DNA template**. *J Mol Biol* 2000, **297**:1075-1085.

40. Pertea M, Lin X, Salzberg SL: **GeneSplicer: a new computational method for splice site prediction**. *Nucleic Acids Res* 2001, **29**:1185-1190.

41. Zhang MQ: **Identification of protein coding regions in the human genome based on quadratic discriminant analysis**. *Proc Natl Acad Sci USA* 1997, **94**:565-568.

42. Brunak S, Engelbrecht J, Knudsen S: **Prediction of human mRNA donor and acceptor sites from the DNA sequence**. *J Mol Biol* 1991, **220**:49-65.

43. Hebsgaard SM, Korning PG, Tolstrup N, Engelbrecht J, Rouze P, Brunak S: **Splice site prediction in *Arabidopsis thaliana* DNA by combining local and global sequence information**. *Nucleic Acids Res* 1996, **24**:3439-3452.

44. Sasaki T, Matsumoto T, Yamamoto K, Sakata K, Baba T, Katayose Y, Wu J, Niimura Y, Cheng Z, Nagamura Y *et al.*: **The genome sequence and structure of rice chromosome 1**. *Nature* 2002, **420**:312-316.

45. Feng Q, Zhang Y, Hao P, Wang S, Fu G, Huang Y, Li Y, Zhu J, Liu Y, Hu X *et al.*: **Sequence and analysis of rice chromosome 4**. *Nature* 2002, **420**:316-320.