BMC
Bioinformatics

Open Access

PlantTFcat: an online plant transcription factor and transcriptional regulator categorization and analysis tool

Xinbin Dai[†], Senjuti Sinharoy[†], Michael Udvardi and Patrick Xuechun Zhao[*]

Abstract

Background: Plants regulate intrinsic gene expression through transcription factors (TFs), transcriptional regulators (TRs), chromatin regulators (CRs), and the basal transcription machinery. An understanding of plant gene regulatory mechanisms at a systems level requires the identification of these regulatory elements on a genomic scale.

Results: Here, we present PlantTFcat, a high-performance web-based analysis tool that is designed to identify and categorize plant TF/TR/CR genes from genome-scale protein and nucleic acid sequences by systematically analyzing InterProScan domain patterns in protein sequences. The comprehensive prediction logics that are included in PlantTFcat are based on relationships between gene families and conserved domains from 108 published plant TF/TR/CR families. These prediction logics effectively distinguish TF/TR/CR families with common conserved domains. Our systematic performance evaluations indicate that PlantTFcat annotates known TF/TR/CR families with high coverage and sensitivity.

Conclusions: PlantTFcat provides an analysis tool to identify and categorize plant TF/TR/CR genes on a genomic scale. PlantTFcat is freely available to the public at http://plantgrn.noble.org/PlantTFcat/.

Keywords: Transcription factor, Transcriptional regulator, Chromatin regulator, High-throughput prediction, Plant genome

Background

Plants regulate intrinsic gene expression through transcription factors (TFs), transcriptional regulators (TRs), chromatin regulators (CRs) and the basal transcription machinery to control development and respond to environmental changes. To understand the fundamental mechanisms of plant development and environmental responses, we need a systematic way to identify and categorize TFs, TRs, and CRs on a genomic scale.

TFs are proteins that bind to specific DNA sequences, usually to a motif in the target gene promoter, to control the transcription of the target gene. TFs may perform this function with other proteins, *e.g.*, TRs. A classic example of the control of gene expression via TFs and TRs is the control that the auxin signal has over all major aspects of plant development. Auxin enhances the proteolysis of auxin/IAA (Aux/IAA) transcriptional repressors, which heterodimerize with the auxin response factor (ARF) TFs to block the transcriptional activation of auxin-responsive genes [1]. This example demonstrates that TFs and TRs play equally vital roles in plants.

CRs are also essential components in the gene expression regulation machinery. CRs are divided into two categories: 1) chromatin remodelers that reposition and restructure nucleosomes and 2) chromatin modifiers that add or remove covalent marks from histones. Chromatin remodeling complexes are widely conserved among plants, animals, and fungi. The maintenance of the chromatin structure conserves epigenetic marks on the histones, which are required for both proper plant development and the maintenance of perception to environmental cues such as light and temperature [2]. In *Arabidopsis thaliana*, mutations of histone mono-ubiquitination enzymes lead to premature flowering as well as defects in plant size and leaf development. These changes indicate that histone mono-ubiquitination enzymes are master regulators of plant development [3].

* Correspondence: pzhao@noble.org
[†]Equal contributors
Plant Biology Division, The Samuel Roberts Noble Foundation, 2510 Sam Noble Parkway 73401 Ardmore, OK, USA

Over the past several years, next-generation sequencing (NGS) technologies have enabled life scientists to collect genome sequences both rapidly and economically. However, although plant gene transcription regulatory elements, including TFs, TRs, and CRs, have been broadly studied for several years, only a few tools have been developed to systematically categorize and analyze published TF/TR/CR gene families on a genomic scale. One popular database, DBD (http://www.transcriptionfactor.org), features PFAM [4] and superfamily [5] domain information and is organized by families and genomes [6]. Furthermore, DBD provides a supplemental function that allows users to submit single protein sequence to search for DNA binding domains within the protein. This database, however, focuses on TFs only.

Another popular database, PlnTFDB (http://plntfdb.bio.uni-potsdam.de/), hosts sequences and PFAM domain models of 84 TF/TR families in plants [7]. However, PlnTFDB lacks an analytic function and serves as a reference database only. ITAK is an unpublished analysis tool (http://bioinfo.bti.cornell.edu/cgi-bin/itak/index.cgi) that has adopted the PFAM domain rules described in PlnTFDB to predict TF and TR genes. However, although the web version of ITAK is user-friendly, this tool can analyze a maximum of 50 sequences for each submission, which makes it impractical for analyzing sequences on a genomic scale. The standalone version of ITAK may be capable of large-scale data analyses if the Linux command line program is optimized and deployed on a parallel computing cluster. However, this implementation poses a challenge for most biologists.

The above analysis tools were developed on the basis of domain patterns from PFAM and superfamily databases, which are not sufficient for the systematic identification and categorization of TF/TR domains because some TF/TR domains are not covered by either the PFAM or the superfamily databases. For example, the BTB/POZ-like domain is the featured domain of the BTB-POZ, BTB-POZ-MATH, and ABTB families, but the domain model is available only in the PROSITE profile database [8] (*i.e.* PS50097) or the SMART [9] database (*i.e.* SM00225). As another example, the Chromo and shadow domains of the CHROMO-DOMAIN family are also found only in the PROSITE and SMART domain pattern databases.

InterProScan [10] is a comprehensive program that integrates 14 domain signature search programs and corresponding domain pattern libraries and has been used for TF/TR gene prediction. Kakar *et al* (2008) scanned all *Medicago truncatula* genome sequences using InterProScan and identified TF/TR gene families from a curated mapping table between InterProScan domains and TF/TR families [11]. Wang *et al* (2010) also developed a TF/TR prediction protocol using InterProScan and the mapping table previously described in

their soybean database, SoyDB [12]. However, to date, neither of these efforts has led to a web server or software that is available for public use.

We have developed a web-based analysis server, Plant TFcat (http://plantgrn.noble.org/PlantTFcat/), which utilizes InterProScan to systematically search proteins for TF/TR/CR-related domain signatures. Furthermore, we have developed comprehensive prediction logics based on relationships between gene families and conserved domains to effectively distinguish families with common conserved domains, which are often overlooked in traditional BLAST-based searches. PlantTFcat therefore systematically predicts and categorizes plant TF, TR, and CR genes with high coverage and sensitivity.

To provide high-throughput genome-scale analysis capability, we customized InterProScan to include only relevant domain information. We further accelerated the back-end prediction module by deploying PlantTFcat on our in-house BioGrid parallel computing platform, which is equipped with approximately 300 AMD processor cores at 2.5GHz clock frequency. With these optimizations, PlantTFcat is able to analyze the *Arabidopsis thaliana* genome release TAIR10 (http://www.arabidopsis.org/) in less than 10 minutes.

Implementation
Comprehensive prediction logics
The InterProScan domain patterns of 108 TF/TR/CR families were compiled in a back-end database by mining the published datasets [7,11] and reviewing the existing literature. For each family, the prediction logic is composed of a *"must-have domain"* pattern and a *"prohibited domain"* pattern, respectively. The domain signatures of a protein in InterProScan that meet the former pattern may satisfy the criterion for such a family. However, any signature that meets the latter pattern will lead to a negative prediction.

The *"must-have domain"* pattern was designed to handle logical relationships such as "AND" and "OR". For example, the ARF TF family is expected to include either IPR010525 or IPR011525 ("OR") as well as IPR003340 ("AND"). We therefore use a logical expression (IPR010525|IPR011525)&IPR003340 to represent the *"must-have domain"* pattern for the ARF TF family.

The *"prohibited domain"* pattern was designed to divide superfamilies into more specific families based on the existing literature. For example, experimental evidence implicates the BTB/POZ-like domain, which is a protein-protein interaction module, in the regulation of gene expression through the local control of chromatin conformation. In some cases, the BTB-POZ domain has been observed with MATH (meprin and TRAF-C homology), which is known to be involved in the regulation of protein processing and ubiquitination. However, the BTB-POZ domain has also been associated with the

ankyrin repeat, which is another protein-protein interaction domain [13,14]. Therefore, we categorized BTB-POZ proteins into the BTB-POZ-MATH and ABTB families based on the differences in their companion domain using the "*prohibited domain*" pattern. As another example, TFs with the DNA-binding domain MADS box (IPR002100) can be subdivided into two families: MADS type1 and MADS-MICK. The MADS type1 family includes only the MADS box. However, the MADS-MICK family includes both the MADS box and a keratin-like box (K-box) (IPR002487) that promotes protein dimerization. To discriminate between these two families, the K-box domain can be included in the "*prohibited domain*" pattern of the MADS type1 family.

The comprehensive prediction logics, including the "*must-have domain*" and "*prohibited domain*" patterns, are available in Additional file 1 for all 108 families. These prediction logics are also listed by family on the PlantTFcat website (http://plantgrn.noble.org/PlantTF-cat/familylist.do).

Infrastructure and implementation

The PlantTFcat consists of an intuitive web interface that allows users to submit large numbers of sequences and retrieve analysis results as well as a sophisticated back-end high-performance prediction module that uses InterProScan to search potential domains from the user-submitted sequences. This back-end module screens potential TFs/TRs/CRs by referring to the featured conserved domain patterns of each family. The PlantTFcat web server runs on an Oracle Java Virtual Machine and Resin J2EE server. The high-performance prediction module is deployed on our in-house BioGrid parallel computing platform, which runs on a CentOS 6 Linux Operating System. The *PlantTFcat* web interface was written in the Groovy language and the back-end pipeline was written in both Java and Groovy.

High-throughput capability

InterProScan is a computationally intensive program that completes the search for the domains of a typical protein sequence in approximately 5 minutes. To analyze genome-scale sequences, we reduced the domain pattern libraries in InterProScan to include only TF/TR/CR-related domains, which accelerated the search by 2 to 3 orders of magnitude. This customized InterProScan installation coupled with our in-house BioGrid parallel computing deployment allows PlantTFcat to analyze the entire *Arabidopsis thaliana* genome (~30,000 genes) within 10 minutes.

Session management

User-submitted jobs are scheduled into the back-end pipeline by a queue management system on a first-come first-serve basis. Only two sessions can run simultaneously in the back-end pipeline. All other sessions wait in the queue until the current jobs have been completed. An exception is made for small jobs (<20 kb), which are submitted and run directly without waiting in the queue.

Results

Interface

Users can submit either nucleic acid or protein sequences under the "*Analysis*" tab (Figure 1) of PlantTFcat web server. The back-end pipeline detects the sequence type and translates nucleic acid sequences into protein sequences according to their six open reading frames (ORFs). Most submissions can be completed within several minutes. Large datasets, *e.g.*, approximately 100 MB of nucleic acid sequences, may require up to an hour to complete. PlantTFcat returns a web-based table that allows users to query, sort, and filter the results when the analysis session is complete. The prediction results are also available for batch download (Figure 2).
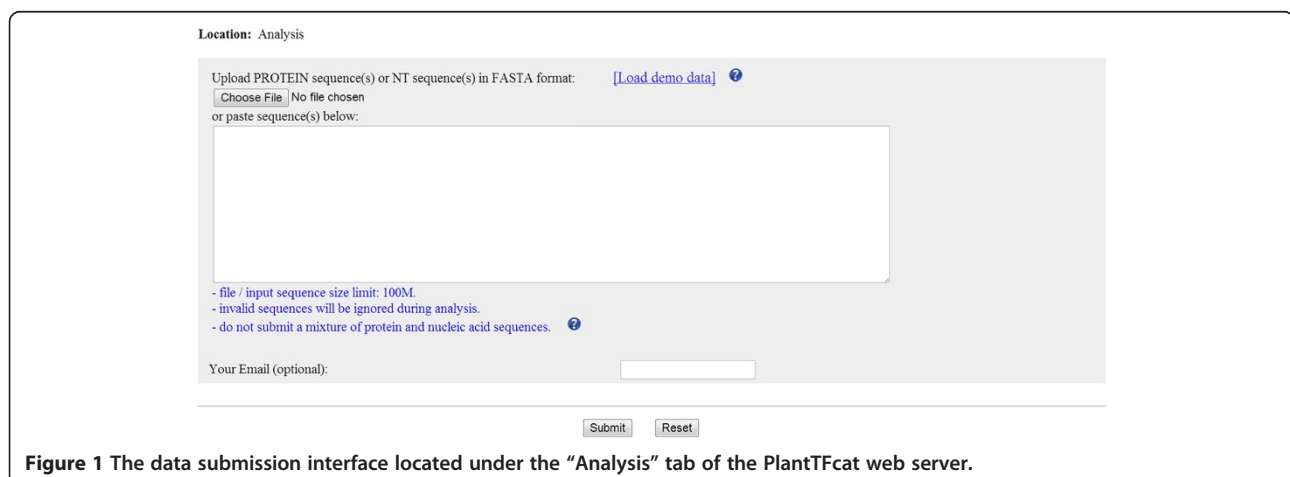
Figure 1 The data submission interface located under the "Analysis" tab of the PlantTFcat web server.
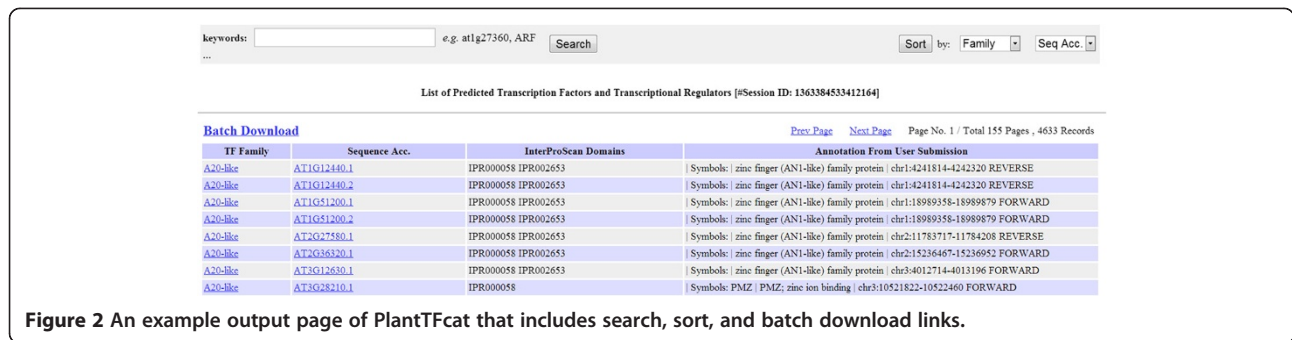
Figure 2 An example output page of PlantTFcat that includes search, sort, and batch download links.

Performance evaluation

We compared the predictions from PlantTFcat with the benchmark data in PlnTFDB [7] using the *Arabidopsis thaliana* genome release V8 (TAIR8) and the *Zea mays* genome release 3b.50 (http://www.maizesequence.org/). PlantTFcat missed 199 genes out of 2,757 TF/TR gene models from the *Arabidopsis thaliana* data in PlnTFDB. Moreover, PlantTF-cat reported 1,744 additional gene models as TF/TR/CR candidates. Of these identified gene models, only 95 were not confirmed, because they were annotated as either unknown or non-TF/TR genes per the TAIR8 annotation (see Additional file 2). In the monocot species, *Zea mays*, PlantTFcat achieves a similar prediction performance: 370 genes were missed and 6,050 genes were newly predicted as TF/TR/CR genes (see Additional file 3) (Table 1). Such results suggest that PlantTFcat can predict TF/TR/CR genes with high coverage and sensitivity.

We compared the false positive rate of PlantTFcat with traditional BLAST-based methods. The *Arabidopsis thaliana* TAIR8 was chosen as a test dataset. A BLAST search (e-value < =1e-04) was run against the TAIR8 TF/TR genes that were downloaded from PlnTFDB. The BLAST search correctly hit all of the 2,757 TF/TR reference genes because the test dataset contains these reference genes. In addition, the BLAST-based method reported 3,870 more homologous genes as TF/TR candidates that

had been excluded by PlnTFDB (Figure 3a). In contrast, PlantTFcat reported only 95 false positives, as described above. We also tested these methods against the *Populus trichocarpa* JGI gene models v1.1 and achieved similar results (Figure 3b). These results suggest that PlantTFcat is a better choice for TF/TR/CR gene annotation over a traditional BLAST search against a reference dataset due to a lower rate of false positives. The details for both comparisons are available in Additional file 4.

Discussion

The back-end pipeline of PlantTFcat was originally developed to annotate *Medicago truncatula* TFs and other TRs as a part of the International Medicago Genome Sequencing and Annotation Project. The latest genome analysis and gene models from this project were recently published [15]. The PlantTFcat web interface and server were made publicly available in late 2011.

The PlantTFcat web server provides the scientific community with a convenient high-throughput tool to annotate TF/TR/CR genes on a genomic scale. Compared with BLAST-based methods, PlantTFcat does not require biologists to install Linux command line programs, parse

Table 1 Benchmark evaluation of PlantTFcat using the PlnTFDB reference database

	Matches	Conflicts	Missed genes	Newly predicted genes
Arabidopsis thaliana				
(32,825 gene models)	2,225	333	199	1,744
Zea Mays				
(147,837 gene models)	4,094	782	370	6,050

The 'Matches' column denotes the number of genes that were categorized into the same/synonymous/super/sub families of PlnTFDB by PlantTFcat. The 'Conflicts' column denotes the number of genes that were identified by PlantTFcat as TF/TR/CR genes, but with different family annotations than PlnTFDB. The 'Missed genes' column represents the genes that were omitted by PlantTFcat but identified as TF/TR genes in the PlnTFDB database. The 'Newly predicted genes' column denotes the genes predicted as TF/TR/CR genes by PlantTFcat that were missed by PlnTFDB.
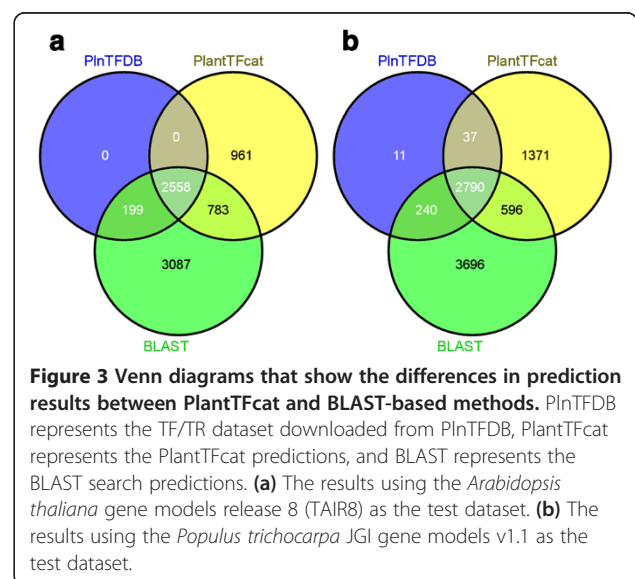
Figure 3 Venn diagrams that show the differences in prediction results between PlantTFcat and BLAST-based methods. PlnTFDB represents the TF/TR dataset downloaded from PlnTFDB, PlantTFcat represents the PlantTFcat predictions, and BLAST represents the BLAST search predictions. **(a)** The results using the *Arabidopsis thaliana* gene models release 8 (TAIR8) as the test dataset. **(b)** The results using the *Populus trichocarpa* JGI gene models v1.1 as the test dataset.

outputs, or deploy programs on complicated high performance clusters. With the support of our parallel computing platform and customized InterProScan domain pattern libraries, PlantTFcat is able to analyze large-scale datasets, such as the data from the next generation sequencing platform for the genomic sequencing projects.

PlantTFcat uses signature domain information to ascribe proteins to different families depending on the presence or absence of multiple domains in a single protein, which are based on published TF/TR/CR functional characterizations. Our performance evaluations indicated that PlantTFcat identifies TF/TR/CR genes with a lower false positive rate than traditional BLAST-based approaches without compromising the true positive rate.

PlantTFcat is also able to distinguish families with "*prohibited domain*" patterns. For example, the B3, AP2-EREBP, RAV, and ARF families have common domains, but exclude some domains in their sequences. The B3 and AP2-EREBP families each contain one DNA binding domain, the B3 DNA binding domain and the AP2 DNA binding domain, respectively. The RAV family contains both the B3 and AP2 DNA binding domains in a single protein, and the ARF family contains a B3 DNA binding domain as well as a protein-protein interaction domain. Such complicated families cannot be predicted easily with a single BLAST search. For example, on the NCBI website, At3g25730.1 is annotated as an AP2/ERF gene as well as a B3 family transcription factor, namely ARF14. However, At3g25730.1 is actually a member of the RAV family because it contains both B3 and AP2 domains.

Conclusions

In conclusion, PlantTFcat performs a systematic analysis of protein domain signatures in InterProScan to produce high coverage and sensitivity for TF/TR/CR annotations. PlantTFcat provides more accurate functional classifications than BLAST-based methods.

Availability and requirements

Project name: PlantTFcat
Project home page: http://plantgrn.noble.org/PlantTFcat/
Operating system: Platform independent (web-based application)
Programming language: Groovy, Java
Other requirements: A web browser
License: None for usage.
Any restrictions to use by non-academics: None.

Additional files

Additional file 1: List of comprehensive prediction logics for all TF/TR/CR families included in the PlantTFcat.

Additional file 2: Performance analysis by systematically comparing PlantTFcat predictions and the PlnTFDB database on *Arabidopsis thaliana* genome.

Additional file 3: Performance analysis by systematically comparing PlantTFcat predictions and the PlnTFDB database on Maize (*Zea mays*) genome.

Additional file 4: Performance comparisons between PlantTFcat and traditional BLAST method on *Arabidopsis thaliana* genome.

References
1. K-i H: **The interaction and integration of auxin signaling components.** *Plant Cell Physiol* 2012, **53**(6):965–975.
2. Cubas P, Jarillo JA, Martínez-Zapater JM, Piñeiro M: **Chromatin remodeling in plant development.** *The Int J Dev Biol* 2009, **1596**:1581–1596.
3. Liu Y, Koornneef M, Soppe WJJ: **The absence of histone H2B monoubiquitination in the Arabidopsis hub1 (rdo4) mutant reveals a role for chromatin remodeling in seed dormancy.** *The Plant Cell Online* 2007, **19**(2):433–444.
4. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, *et al*: **The Pfam protein families database.** *Nucleic Acids Res* 2012, **40**(D1):D290–D301.
5. Gough J, Karplus K, Hughey R, Chothia C: **Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure.** *J Mol Biol* 2001, **313**(4):903–919.
6. Kummerfeld SK, Teichmann SA: **DBD: a transcription factor prediction database.** *Nucleic Acids Res* 2005, **34**(suppl 1):D74–D81.
7. Pérez-Rodríguez P, Riaño-Pachón DM, Corrêa LGG, Rensing SA, Kersten B, Mueller-Roeber B: **PlnTFDB: updated content and new features of the plant transcription factor database.** *Nucleic Acids Res* 2010, **38**(suppl 1):D822–D827.
8. Sigrist CJA, Cerutti L, de Castro E, Langendijk-Genevaux PS, Bulliard V, Bairoch A, Hulo N: **PROSITE, a protein domain database for functional characterization and annotation.** *Nucleic Acids Res* 2010, **38**(suppl 1):D161–D166.
9. Letunic I, Doerks T, Bork P: **SMART 7: recent updates to the protein domain annotation resource.** *Nucleic Acids Res* 2012, **40**(D1):D302–D305.
10. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, *et al*: **InterPro: the integrative protein signature database.** *Nucleic Acids Res* 2009, **37**(suppl 1):D211–D215.
11. Kakar K, Wandrey M, Czechowski T, Gaertner T, Scheible W-R, Stitt M, Torres-Jerez I, Xiao Y, Redman J, Wu H, *et al*: **A community resource for high-throughput quantitative RT-PCR analysis of transcription factor gene expression in Medicago truncatula.** *Plant Methods* 2008, **4**(1):18.

12. Wang Z, Libault M, Joshi T, Valliyodan B, Nguyen H, Xu D, Stacey G, Cheng J: SoyDB: a knowledge database of soybean transcription factors. *BMC Plant Biol* 2010, **10**(1):14.
13. Albagli O, Dhordain P, Deweindt C, Lecocq G, Leprince D: **The BTB/POZ domain: a new protein-protein interaction motif common to DNA- and actin-binding proteins.** *Cell Growth Differ* 1995, **6**(9):1193–1198.
14. Weber H, Hellmann H: **Arabidopsis thaliana BTB/ POZ-MATH proteins interact with members of the ERF/AP2 transcription factor family.** *FEBS J* 2009, **276**(22):6624–6635.
15. Young N, Debellé F, Oldroyd G, Geurts R, Cannon S, Udvardi M, Benedito V, Mayer K, Gouzy J, Schoof H, *et al*: **The Medicago genome provides insight into the evolution of rhizobial symbioses.** *Nature* 2011, **480**(7378):520–524.