Published online 17 November 2015 *Nucleic Acids Research, 2016, Vol. 44, Database issue* ***D1161–D1166***
doi: 10.1093/nar/gkv1215

GREENC: a Wiki-based database of plant lncRNAs

Andreu Paytuví Gallart[1,†], Antonio Hermoso Pulido[2,3,†], Irantzu Anzar Martínez de Lagrán[1], Walter Sanseverino[1,*] and Riccardo Aiese Cigliano[1,*]

[1]Sequentia Biotech SL, Calle Comte D'Urgell 240, Barcelona, Spain, [2]CRG Bioinformatics Facility, Centre for Genomic Regulation (CRG), Dr Aiguader 88, 08003 Barcelona, Spain and [3]Universitat Pompeu Fabra (UPF), Dr Aiguader 88, 08003 Barcelona, Spain

Received July 27, 2015; Revised October 27, 2015; Accepted October 28, 2015

ABSTRACT

Long non-coding RNAs (lncRNAs) are functional non-translated molecules greater than 200 nt. Their roles are diverse and they are usually involved in transcriptional regulation. LncRNAs still remain largely uninvestigated in plants with few exceptions. Experimentally validated plant lncRNAs have been shown to regulate important agronomic traits such as phosphate starvation response, flowering time and interaction with symbiotic organisms, making them of great interest in plant biology and in breeding. There is still a lack of lncRNAs in most sequenced plant species, and in those where they have been annotated, different methods have been used, so making the lncRNAs less useful in comparisons within and between species. We developed a pipeline to annotate lncRNAs and applied it to 37 plant species and six algae, resulting in the annotation of more than 120 000 lncRNAs. To facilitate the study of lncRNAs for the plant research community, the information gathered is organised in the *Green Non-Coding Database* (GreeNC, http://greenc.sciencedesigners.com/).

INTRODUCTION

The Encyclopaedia of DNA Elements (ENCODE) was launched by the US National Human Genome Research Institute (NHGRI) in September 2003. The aim was to uncover the role of the non-coding regions of the human genome, concluding that 80.4% of the human genome participated in at least one biochemical RNA or chromatin associated event (1). Non-coding RNAs (ncRNAs) are arbitrarily grouped into short (<200 nt), and long ncRNA (lncRNAs, >200 nt). The mechanisms and the role played in gene expression regulation by short ncRNAs, such as miRNA, siRNA and piRNA, established in several species (2–4), have been linked to chromatin modifications, transcriptional regulation, and conformational changes in proteins (5).

Research in lncRNAs is far more advanced in humans and mice than in plants, although there are a few well-known exceptions. In *Arabidopsis thaliana*, *IPS1* is a lncRNA expressed upon phosphate starvation and it is thought to counteract the activity of miR399 on *PHO2*, which in turn regulates the expression of phosphate transporter genes (6). It has been shown that the lncRNA *COLDAIR* recruits the histone methylase PRC2 to interact with the PRC2 complex, so maintaining a stable silenced state of *FLC* to repress flowering during vernalization (7). *COOLAIR* is another Arabidopsis lncRNA that represses *FLC* expression by interfering with the binding of PolII (8). In rice, the lncRNA *LDMAR* has been found to control photo-sensitive male sterility by regulating DNA methylation levels in the promoter region of *LDMAR* (9). Finally, in *Medicago truncatula*, the lncRNA *Enod40* has been shown to participate in establishing symbiotic interactions with soil–bacteria by affecting nodule formation (10). These findings highlight the potential interest of lncRNAs in plant biology and in regulating important agronomic traits.

To further our knowledge of lncRNAs in plant biology, their comprehensive annotation is very important. Genome-wide studies have been performed in several plant species (11–16), however, different pipelines for lncRNAs annotation were used and neither their sequences or other information organized in a database.

Many lncRNA databases exist but most of them are focused on human and vertebrate lncRNAs. Those databases with entries from plants include:

- The NONCODE (17) includes the annotation of different classes of ncRNAs from different species. The database has about 3,800 entries from *A. thaliana*, the only plant species represented, gathered from the literature, specialized DB and GenBank;
- The PNRD database (18) includes the annotation of different classes of ncRNAs. About 5,000 lncRNAs are an-

[*]To whom correspondence should be addressed. Tel: +34930107368; Fax: +34934034510; Email: raiesecigliano@sequentiabiotech.com
Correspondence may also be addressed to Walter Sanseverino. Tel: +34930107368; Fax: +34934034510; Email: wsanseverino@sequentiabiotech.com
[†]These authors contributed equally to the paper as first authors.

notated, from *A. thaliana*, *O. sativa*, *P. trichocarpa* and *Z. mays*. The entries are from the integration of data from other databases and publications;

- The PLncDB database (19) is an *A. thaliana*-specific database, with more than 13 000 ncRNAs obtained from various data resources;
- The PlncRNADB database (http://bis.zju.edu.cn/PlncRNADB/index.php) includes about 5100 lncRNAs from *A. thaliana*, *A. lyrata*, *P. trichocarpa* and *Z. mays*. The sequences were obtained either from the literature or by annotation based on reference-guided transcriptome assembly;
- The PLNlncRbase database (20) is a manually curated database of experimentally validated lncRNAs from several plant species and includes around 1000 sequences;
- The lncRNAdb database (21) is a manually curated database of experimentally validated lncRNAs from several species and contains two entries from *A. thaliana*.

In this work, we developed a pipeline to annotate lncRNAs from official genome annotations. We applied our pipeline to 37 plant genome annotations and to six algae, and organized the results in the *Green Non-Coding Database* (GreeNC). This database provides information on the sequence, genome position, coding potential and folding energies of >120 000 lncRNAs. The aim of GreeNC is as a meeting point for the plant lncRNA research and is freely available at http://greenc.sciencedesigners.com.

METHODS

Genomes and annotations

The FASTA sequences of the transcripts of the analyzed species were downloaded from Phytozome v10.3 (22). The assembly version of each genome is given in Supplementary Table S1. Only the genomes available for genomic studies according to the restriction of data usage were used (23–63).

Identification of lncRNAs

Two bash scripts were written to identify lncRNAs among the downloaded transcript sequences. The first script followed the approach developed at the McGinnis lab to identify lncRNAs in transcriptomes (Supplementary Figure S1), and is based on identifying the coding potential of each transcript and on similarity with known proteins (11). The script retains transcripts longer than 200 nt and with an ORF shorter than 120 aa by using Ugene (1.13) (http://ugene.unipro.ru/). Sequences were then blasted (blastx, 2.2.28+) (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST) against SwissProt (2013/11) (64). CPC (0.9-r2) (http://cpc.cbi.pku.edu.cn/) (65) was also used, with the FrameFinder parameter -r set to ‘True’ or ‘False’ and the BLASTX parameter -S set to ‘3’ or ‘1’, depending on the group of transcripts being analyzed. The second script was written to discriminate other non-coding transcripts from lncRNAs and to identify possible miRNA precursors (Supplementary Figure S1). Transcripts were analyzed by cmscan (Infernal 1.1rc4) against the RFAM database (release 11). In addition, BLASTn (2.2.28+) was used against a database of mature plant miRNA sequences from miRBase (release 20) (66) and the results validated with MIReNA (v2.0) (http://www.lcqb.upmc.fr/mirena/index.html). Finally, MIReNA was called again, using the parameters –valid, –x, –mfei -0.69, –amfe -32, –ratiomin 0.83, and –ratiomax 1.17.

The final set of lncRNAs was divided into high-confidence and low-confidence. The transcripts without hits in BLASTX described as non-coding by CPC, and considered non-precursors of miRNA, were classified as high-confidence lncRNAs. Transcripts without hits in the BLASTx step and described as coding by CPC, and transcripts with hits in the BLASTx step but described as non-coding by CPC, were considered low-confidence lncRNAs, as well as the transcripts identified as putative precursors of miRNAs. Transcripts having predicted repetitive regions by RepeatMasker (http://www.repeatmasker.org/) were also classified as low-confidence in order to exclude putative transposons. The first script for the annotation of lncRNAs was tested with 480 lncRNAs and 1268 coding genes annotated in *Arabidopsis thaliana* (TAIR10) resulting in a sensitivity of 92% and a specificity of 94.95%. The second script was tested with 480 lncRNAs annotated in *Arabidopsis thaliana* (TAIR10) resulting in a sensitivity of 93% and a specificity of 97.6%.

Annotation of repetitive elements

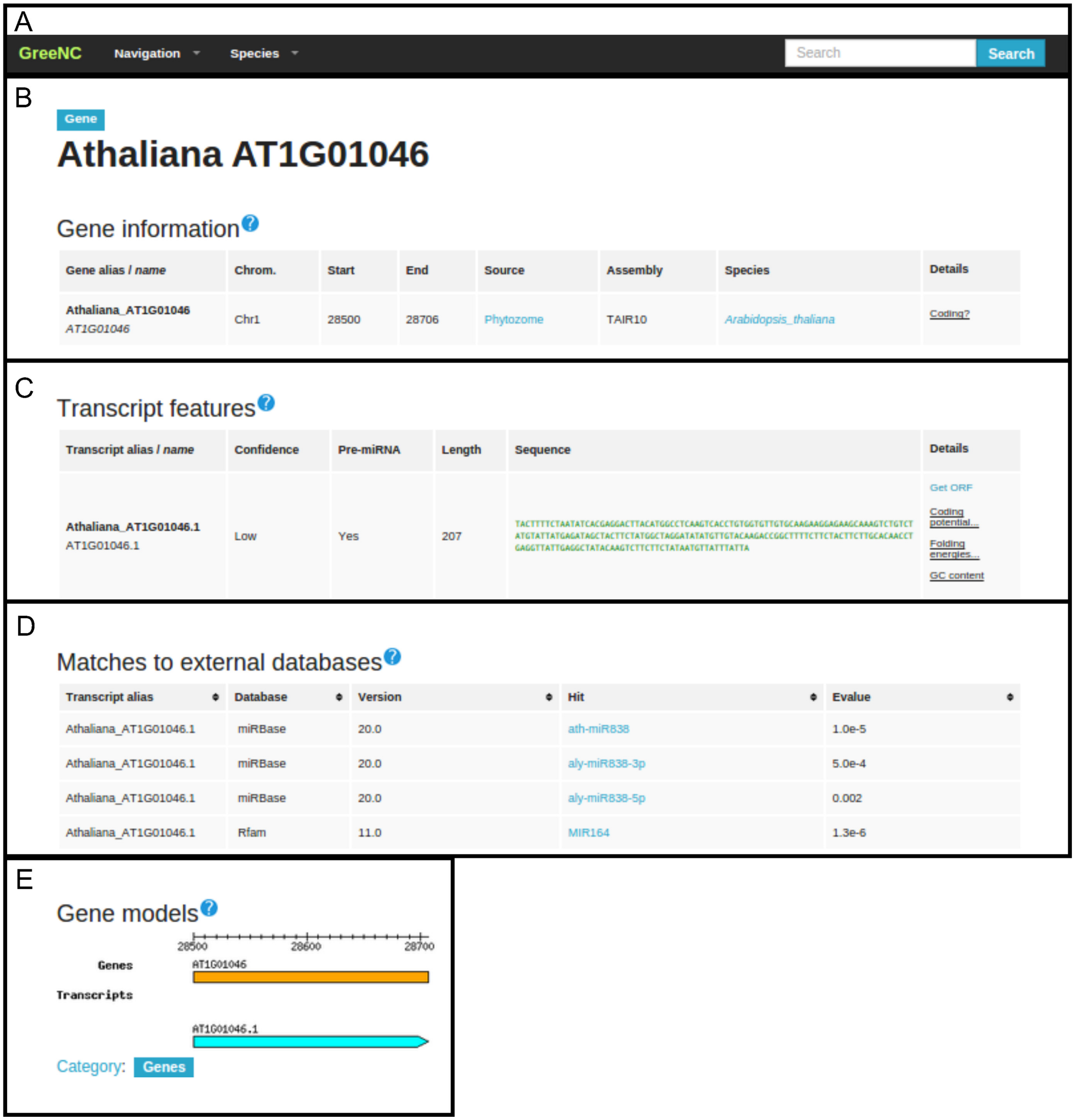
RepeatMasker (open-4.0.5) (http://www.repeatmasker.org/) was used for repetitive element identification with the parameters: -species Viridiplantae, -no_—-is, -gff, and -nolow. The search engine used was RMBLAST (2.2.23+) against the RepBase database (released: 31 January 2014) (67).

Relational database

Data was imported into a MySQL (5.5) based relational database stored in an Ubuntu server (14.04). This database was then integrated into a MediaWiki (1.23) by mapping relational data fields against predefined templates via Semantic MediaWiki. Transcript sequences in a FASTA file were formatted using makeblastdb. Sequence retrieval is based on blastdbcmd. An Express Node.js API web service was created to expose both sequence retrieval and BLAST searches via client JavaScript from the MediaWiki interface.

AIMS OF THE DATABASE

GreeNC is a repository of lncRNAs annotated in 37 plant species and six algae. By using the same pipeline to annotate lncRNAs we make it possible to compare lncRNA sequences and distribution from different species. By organizing the sequences in a central database we aim to provide a tool for the scientific community that can boost research on this class of transcripts. The GreeNC database provides information on sequence, genome coordinates, coding potential and folding energy of the lncRNAs. In future updates we will add more species, expression information and conservation. The GreeNC is also integrated with other databases, i.e. NONCODE (17), Swissprot (64) and RFAM (68), LNCRNADB, Phytozome (22), and miRBase (66) so users can easily obtain information from different sources.



Gene

Athaliana AT1G01046

Gene information

Gene alias / *name*	Chrom.	Start	End	Source	Assembly	Species	Details
Athaliana_AT1G01046 *AT1G01046*	Chr1	28500	28706	Phytozome	TAIR10	*Arabidopsis_thaliana*	Coding?

Transcript features

Transcript alias / *name*	Confidence	Pre-miRNA	Length	Sequence	Details
Athaliana_AT1G01046.1 AT1G01046.1	Low	Yes	207	TACTTTTCTAATATCACGAGGACTTACATGGCCTCAAGTCACCTGTGGTGTTGTGCAAGAAGGAGAAGCAAAGTCTGTCT ATGTATTATGAGATAGCTACTTCTATGGCTAGGATATATGTTGTACAAGACCGGCTTTTCTTCTACTTCTTGCACAACCT GAGGTTATTGAGGCTATACAAGTCTTCTTCTATAATGTTATTTATTA	Get ORF Coding potential... Folding energies... GC content

Matches to external databases

Transcript alias	Database	Version	Hit	Evalue
Athaliana_AT1G01046.1	miRBase	20.0	ath-miR838	1.0e-5
Athaliana_AT1G01046.1	miRBase	20.0	aly-miR838-3p	5.0e-4
Athaliana_AT1G01046.1	miRBase	20.0	aly-miR838-5p	0.002
Athaliana_AT1G01046.1	Rfam	11.0	MIR164	1.3e-6

Gene models

Category: Genes

Figure 1. A snapshot of an *Arabidopsis thaliana* entry from the GreeNC database. (**A**) Header, to navigate through the website and access to the tools and the pages of the species; (**B**) table of gene information reporting genomic coordinates, genome version, the source of the genome assembly and if the gene encodes at least one coding transcript; (**C**) table of transcript features reporting the kind of lncRNA (low-/high-confidence), if it is a precursor of miRNAs, length, sequence and links to get the Open Reading Frame (ORF), the Coding Potential, the folding energy and the GC content; (**D**) an optional table that provides links to other databases, when applicable, and giving information about the version of the database and the e-value of the match; (**E**) a schematic representation of the gene and transcript models.

DATABASE STRUCTURE

The GreeNC database is a MySQL relational database and it is freely available at: http://greenc.sciencedesigners.com/. Data was integrated into a MediaWiki by mapping relational data fields against wiki predefined templates via Semantic MediaWiki. Using templates makes it easy to print information and style it for different page types (e.g. genes and species). The template approach exposes the fields which may be queried, enhancing the search possibilities of the site. All transcript sequences were kept in a FASTA file with the same IDs as in the MySQL, and then formatted using NCBI makeblastdb. In this way, sequences can be retrieved using their ID with blastdbcmd and, at the same time, other BLAST programs can be run against the resulting BLAST database. Taking advantage of this, an Express Node.js API web service was created to expose both sequence retrieval and BLAST searches via client JavaScript from the MediaWiki interface (Supplementary Figure S2).

GREENC CONTENT

LncRNAs were annotated using the criteria defined by Boerner and McGinnis for the prediction of maize lncRNAs (11). In addition, we scanned the lncRNAs sequences against miRbase (66) and RepBase (67) to discriminate between proper lncRNAs from precursors of smallRNAs and transposable elements.

GreeNC includes ~200 000 pages with information on >190 000 transcripts from 37 plants and six algae. More than 120 000 transcripts were annotated as high confidence lncRNAs, 30% of them from the *T. aestivum* (17.8%) and *Z. mays* (8.2%). The lowest number of lncRNAs was annotated in the three algae *C. rehinardtii* (0.1%), *M. pusilla* (0.15%) and *O. lucimarinus* (0.16%). More than 25 000 and 8000 transcripts were annotated as repetitive elements and miRNA precursors, respectively.

For each species it is possible to browse and search for lncRNAs at the gene or transcript level, and both link to the main locus pages. These pages include information on the version of the genome assembly and the chromosome position of the loci. In addition, the 'Transcript Features' table contains the list of transcripts encoded by each locus showing the sequence and several annotations, such as the type of lncRNA, the length, coding potential, folding energies and the GC content, and a link to the NCBI ORF Finder tool to further investigate the coding potential of the transcripts. Finally, there is an additional table, 'Matches to external databases', where one of the transcripts has a match in Swissprot (64) or RepBase (67), with links and information on the matched sequences (Figure 1).

The Search button at the top of each page lets the user search the database by using keywords. The Advanced Search gives the possibility of looking for lncRNAs in all the species, using criteria such as confidence level and being a precursor of miRNAs. Finally the BLAST page, gives the possibility of querying the database with a user-supplied sequence and searching for it in the whole database or in a specific species.

FUTURE DIRECTIONS

GreeNC will be updated annually in order to add new sequences from other species and to update existing genome annotations. New information will also be made available, such as expression levels obtained from publicly available RNA-seq data, conservation across different species and phylogeny.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Riccardo Aversano, Clara Conicella, M. Federica Consiglio and Ortrun Mittelsten Scheid for valuable suggestions and critical reading of the article, Ermanno Battista and Simone Cossu for IT support.

FUNDING

Funding for open access charge: Self funding.
Conflict of interest statement. None declared.

REFERENCES

1. ENCODE Project Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
2. Weick,E.-M. and Miska,E.A. (2014) piRNAs: from biogenesis to function. *Development*, **141**, 3458–3471.
3. Chen,X. (2012) Small RNAs in development - insights from plants. *Curr. Opin. Genet. Dev.*, **22**, 361–367.
4. Dogini,D.B., Pascoal,V.D.B., Avansini,S.H., Vieira,A.S., Pereira,T.C. and Lopes-Cendes,I. (2014) The new world of RNAs. *Genet. Mol. Biol.*, **37**, 285–293.
5. Au,P.C.K., Zhu,Q.-H., Dennis,E.S. and Wang,M.-B. (2011) Long non-coding RNA-mediated mechanisms independent of the RNAi pathway in animals and plants. *RNA Biol*, **8**, 404–414.
6. Franco-Zorrilla,J.M., Valli,A., Todesco,M., Mateos,I., Puga,M.I., Rubio-Somoza,I., Leyva,A., Weigel,D., García,J.A. and Paz-Ares,J. (2007) Target mimicry provides a new mechanism for regulation of microRNA activity. *Nat. Genet.*, **39**, 1033–1037.
7. Heo,J.B. and Sung,S. (2011) Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. *Science*, **331**, 76–79.
8. Swiezewski,S., Liu,F., Magusin,A. and Dean,C. (2009) Cold-induced silencing by long antisense transcripts of an Arabidopsis Polycomb target. *Nature*, **462**, 799–802.
9. Ding,J., Lu,Q., Ouyang,Y., Mao,H., Zhang,P., Yao,J., Xu,C., Li,X., Xiao,J. and Zhang,Q. (2012) A long noncoding RNA regulates photoperiod-sensitive male sterility, an essential component of hybrid rice. *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 2654–2659.
10. Campalans,A., Kondorosi,A. and Crespi,M. (2004) Enod40, a short open reading frame-containing mRNA, induces cytoplasmic localization of a nuclear RNA binding protein in Medicago truncatula. *Plant Cell*, **16**, 1047–1059.
11. Boerner,S. and McGinnis,K.M. (2012) Computational identification and functional predictions of long noncoding RNA in Zea mays. *PLoS ONE*, **7**, e43047.
12. Li,L., Eichten,S.R., Shimizu,R., Petsch,K., Yeh,C.-T., Wu,W., Chettoor,A.M., Givan,S.A., Cole,R.A., Fowler,J.E. *et al.* (2014) Genome-wide discovery and characterization of maize long non-coding RNAs. *Genome Biol.*, **15**, R40.
13. Lu,T., Zhu,C., Lu,G., Guo,Y., Zhou,Y., Zhang,Z., Zhao,Y., Li,W., Lu,Y., Tang,W. *et al.* (2012) Strand-specific RNA-seq reveals widespread occurrence of novel cis-natural antisense transcripts in rice. *BMC Genomics*, **13**, 721.
14. Shuai,P., Liang,D., Tang,S., Zhang,Z., Ye,C.-Y., Su,Y., Xia,X. and Yin,W. (2014) Genome-wide identification and functional prediction

of novel and drought-responsive lincRNAs in Populus trichocarpa. *J. Exp. Bot.*, **65**, 4975–4983.
15. Wen,J., Parker,B.J. and Weiller,G.F. (2007) In Silico identification and characterization of mRNA-like noncoding transcripts in Medicago truncatula. *In Silico Biol. (Gedrukt)*, **7**, 485–505.
16. Xin,M., Wang,Y., Yao,Y., Song,N., Hu,Z., Qin,D., Xie,C., Peng,H., Ni,Z. and Sun,Q. (2011) Identification and characterization of wheat long non-protein coding RNAs responsive to powdery mildew infection and heat stress by using microarray analysis and SBS sequencing. *BMC Plant Biol.*, **11**, 61.
17. Xie,C., Yuan,J., Li,H., Li,M., Zhao,G., Bu,D., Zhu,W., Wu,W., Chen,R. and Zhao,Y. (2014) NONCODEv4: exploring the world of long non-coding RNA genes. *Nucleic Acids Res.*, **42**, D98–D103.
18. Yi,X., Zhang,Z., Ling,Y., Xu,W. and Su,Z. (2015) PNRD: a plant non-coding RNA database. *Nucleic Acids Res.*, **43**, D982–D989.
19. Jin,J., Liu,J., Wang,H., Wong,L. and Chua,N.-H. (2013) PLncDB: plant long non-coding RNA database. *Bioinformatics*, **29**, 1068–1071.
20. Xuan,H., Zhang,L., Liu,X., Han,G., Li,J., Li,X., Liu,A., Liao,M. and Zhang,S. (2015) PLNlncRbase: A resource for experimentally identified lncRNAs in plants. *Gene*, **573**, 328–332.
21. Quek,X.C., Thomson,D.W., Maag,J.L.V., Bartonicek,N., Signal,B., Clark,M.B., Gloss,B.S. and Dinger,M.E. (2015) lncRNAdb v2.0: expanding the reference database for functional long noncoding RNAs. *Nucleic Acids Res.*, **43**, D168–D173.
22. Goodstein,D.M., Shu,S., Howson,R., Neupane,R., Hayes,R.D., Fazo,J., Mitros,T., Dirks,W., Hellsten,U., Putnam,N. *et al.* (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res.*, **40**, D1178–D1186.
23. Amborella Genome Project (2013) The Amborella genome and the evolution of flowering plants. *Science*, **342**, 1241089–1241089.
24. Hu,T.T., Pattyn,P., Bakker,E.G., Cao,J., Cheng,J.-F., Clark,R.M., Fahlgren,N., Fawcett,J.A., Grimwood,J., Gundlach,H. *et al.* (2011) The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. *Nat. Genet.*, **43**, 476–481.
25. Lamesch,P., Berardini,T.Z., Li,D., Swarbreck,D., Wilks,C., Sasidharan,R., Muller,R., Dreher,K., Alexander,D.L., Garcia-Hernandez,M. *et al.* (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.*, **40**, D1202–D1210.
26. International Brachypodium Initiative. (2010) Genome sequencing and analysis of the model grass Brachypodium distachyon. *Nature*, **463**, 763–768.
27. Slotte,T., Hazzouri,K.M., Ågren,J.A., Koenig,D., Maumus,F., Guo,Y.-L., Steige,K., Platts,A.E., Escobar,J.S., Newman,L.K. *et al.* (2013) The Capsella rubella genome and the genomic consequences of rapid mating system evolution. *Nat. Genet.*, **45**, 831–835.
28. Ming,R., Hou,S., Feng,Y., Yu,Q., Dionne-Laporte,A., Saw,J.H., Senin,P., Wang,W., Ly,B.V., Lewis,K.L.T. *et al.* (2008) The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). *Nature*, **452**, 991–996.
29. Merchant,S.S., Prochnik,S.E., Vallon,O., Harris,E.H., Karpowicz,S.J., Witman,G.B., Terry,A., Salamov,A., Fritz-Laylin,L.K., Maréchal-Drouard,L. *et al.* (2007) The Chlamydomonas genome reveals the evolution of key animal and plant functions. *Science*, **318**, 245–250.
30. Wu,G.A., Prochnik,S., Jenkins,J., Salse,J., Hellsten,U., Murat,F., Perrier,X., Ruiz,M., Scalabrin,S., Terol,J. *et al.* (2014) Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. *Nat. Biotechnol.*, **32**, 656–662.
31. Blanc,G., Agarkova,I., Grimwood,J., Kuo,A., Brueggeman,A., Dunigan,D.D., Gurnon,J., Ladunga,I., Lindquist,E., Lucas,S. *et al.* (2012) The genome of the polar eukaryotic microalga Coccomyxa subellipsoidea reveals traits of cold adaptation. *Genome Biol.*, **13**, R39.
32. Bartholomé,J., Mandrou,E., Mabiala,A., Jenkins,J., Nabihoudine,I., Klopp,C., Schmutz,J., Plomion,C. and Gion,J.-M. (2015) High-resolution genetic maps of Eucalyptus improve Eucalyptus grandis genome assembly. *New Phytol.*, **206**, 1283–1296.
33. Yang,R., Jarvis,D.E., Chen,H., Beilstein,M.A., Grimwood,J., Jenkins,J., Shu,S., Prochnik,S., Xin,M., Ma,C. *et al.* (2013) The Reference Genome of the Halophytic Plant Eutrema salsugineum. *Front Plant Sci*, **4**, 46.
34. Shulaev,V., Sargent,D.J., Crowhurst,R.N., Mockler,T.C., Folkerts,O., Delcher,A.L., Jaiswal,P., Mockaitis,K., Liston,A., Mane,S.P. *et al.* (2011) The genome of woodland strawberry (Fragaria vesca). *Nat. Genet.*, **43**, 109–116.
35. Schmutz,J., McClean,P.E., Mamidi,S., Wu,G.A., Cannon,S.B., Grimwood,J., Jenkins,J., Shu,S., Song,Q., Chavarro,C. *et al.* (2014) A reference genome for common bean and genome-wide analysis of dual domestications. *Nat. Genet.*, **46**, 707–713.
36. Schmutz,J., Cannon,S.B., Schlueter,J., Ma,J., Mitros,T., Nelson,W., Hyten,D.L., Song,Q., Thelen,J.J., Cheng,J. *et al.* (2010) Genome sequence of the palaeopolyploid soybean. *Nature*, **463**, 178–183.
37. Wang,Z., Hobson,N., Galindo,L., Zhu,S., Shi,D., McDill,J., Yang,L., Hawkins,S., Neutelings,G., Datla,R. *et al.* (2012) The genome of flax (Linum usitatissimum) assembled de novo from short shotgun sequence reads. *Plant J.*, **72**, 461–473.
38. Velasco,R., Zharkikh,A., Affourtit,J., Dhingra,A., Cestaro,A., Kalyanaraman,A., Fontana,P., Bhatnagar,S.K., Troggio,M., Pruss,D. *et al.* (2010) The genome of the domesticated apple (Malus × domestica Borkh.). *Nat. Genet.*, **42**, 833–839.
39. Prochnik,S., Marri,P.R., Desany,B., Rabinowicz,P.D., Kodira,C., Mohiuddin,M., Rodriguez,F., Fauquet,C., Tohme,J., Harkins,T. *et al.* (2012) The Cassava Genome: Current Progress, Future Directions. *Trop Plant Biol*, **5**, 88–94.
40. Young,N.D., Debellé,F., Oldroyd,G.E.D., Geurts,R., Cannon,S.B., Udvardi,M.K., Benedito,V.A., Mayer,K.F.X., Gouzy,J., Schoof,H. *et al.* (2011) The Medicago genome provides insight into the evolution of rhizobial symbioses. *Nature*, **480**, 520–524.
41. Worden,A.Z., Lee,J.-H., Mock,T., Rouzé,P., Simmons,M.P., Aerts,A.L., Allen,A.E., Cuvelier,M.L., Derelle,E., Everett,M.V. *et al.* (2009) Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes Micromonas. *Science*, **324**, 268–272.
42. Droc,G., Larivière,D., Guignon,V., Yahiaoui,N., This,D., Garsmeur,O., Dereeper,A., Hamelin,C., Argout,X., Dufayard,J.-F. *et al.* (2013) The banana genome hub. *Database (Oxford)*, bat035–bat035.
43. Ouyang,S., Zhu,W., Hamilton,J., Lin,H., Campbell,M., Childs,K., Thibaud-Nissen,F., Malek,R.L., Lee,Y., Zheng,L. *et al.* (2007) The TIGR Rice Genome Annotation Resource: improvements and new features. *Nucleic Acids Res.*, **35**, D883–D887.
44. Palenik,B., Grimwood,J., Aerts,A., Rouzé,P., Salamov,A., Putnam,N., Dupont,C., Jorgensen,R., Derelle,E., Rombauts,S. *et al.* (2007) The tiny eukaryote Ostreococcus provides genomic insights into the paradox of plankton speciation. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 7705–7710.
45. Tuskan,G.A., Difazio,S., Jansson,S., Bohlmann,J., Grigoriev,I., Hellsten,U., Putnam,N., Ralph,S., Rombauts,S., Salamov,A. *et al.* (2006) The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science*, **313**, 1596–1604.
46. Paterson,A.H., Wendel,J.F., Gundlach,H., Guo,H., Jenkins,J., Jin,D., Llewellyn,D., Showmaker,K.C., Shu,S., Udall,J. *et al.* (2012) Repeated polyploidization of Gossypium genomes and the evolution of spinnable cotton fibres. *Nature*, **492**, 423–427.
47. International Peach Genome Initiative, Verde,I., Abbott,A.G., Scalabrin,S., Jung,S., Shu,S., Marroni,F., Zhebentyayeva,T., Dettori,M.T., Grimwood,J. *et al.* (2013) The high-quality draft genome of peach (Prunus persica) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nat. Genet.*, **45**, 487–494.
48. Chan,A.P., Crabtree,J., Zhao,Q., Lorenzi,H., Orvis,J., Puiu,D., Melake-Berhan,A., Jones,K.M., Redman,J., Chen,G. *et al.* (2010) Draft genome sequence of the oilseed species Ricinus communis. *Nat. Biotechnol.*, **28**, 951–956.
49. Banks,J.A., Nishiyama,T., Hasebe,M., Bowman,J.L., Gribskov,M., dePamphilis,C., Albert,V.A., Aono,N., Aoyama,T., Ambrose,B.A. *et al.* (2011) The Selaginella genome identifies genetic changes associated with the evolution of vascular plants. *Science*, **332**, 960–963.
50. Bennetzen,J.L., Schmutz,J., Wang,H., Percifield,R., Hawkins,J., Pontaroli,A.C., Estep,M., Feng,L., Vaughn,J.N., Grimwood,J. *et al.* (2012) Reference genome sequence of the model plant Setaria. *Nat. Biotechnol.*, **30**, 555–561.
51. Tomato Genome Consortium. (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, **485**, 635–641.

52. Potato Genome Sequencing Consortium, Xu,X., Pan,S., Cheng,S., Zhang,B., Mu,D., Ni,P., Zhang,G., Yang,S., Li,R. *et al.* (2011) Genome sequence and analysis of the tuber crop potato. *Nature*, **475**, 189–195.
53. Paterson,A.H., Bowers,J.E., 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.
54. Wang,W., Haberer,G., Gundlach,H., Gläßer,C., Nussbaumer,T., Luo,M.C., Lomsadze,A., Borodovsky,M., Kerstetter,R.A., Shanklin,J. *et al.* (2014) The Spirodela polyrhiza genome reveals insights into its neotenous reduction fast growth and aquatic lifestyle. *Nat Commun*, **5**, 3311.
55. Motamayor,J.C., Mockaitis,K., Schmutz,J., Haiminen,N., Livingstone,D., Cornejo,O., Findley,S.D., Zheng,P., Utro,F., Royaert,S. *et al.* (2013) The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. *Genome Biol.*, **14**, r53.
56. Jaillon,O., Aury,J.-M., Noel,B., Policriti,A., Clepet,C., Casagrande,A., Choisne,N., Aubourg,S., Vitulo,N., Jubin,C. *et al.* (2007) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, **449**, 463–467.
57. Prochnik,S.E., Umen,J., Nedelcu,A.M., Hallmann,A., Miller,S.M., Nishii,I., Ferris,P., Kuo,A., Mitros,T., Fritz-Laylin,L.K. *et al.* (2010) Genomic analysis of organismal complexity in the multicellular green alga Volvox carteri. *Science*, **329**, 223–226.
58. Schnable,P.S., Ware,D., Fulton,R.S., Stein,J.C., Wei,F., Pasternak,S., Liang,C., Zhang,J., Fulton,L., Graves,T.A. *et al.* (2009) The B73 maize genome: complexity, diversity, and dynamics. *Science*, **326**, 1112–1115.
59. Huang,S., Li,R., Zhang,Z., Li,L., Gu,X., Fan,W., Lucas,W.J., Wang,X., Xie,B., Ni,P. *et al.* (2009) The genome of the cucumber, Cucumis sativus L. *Nat. Genet.*, **41**, 1275–1281.
60. International Wheat Genome Sequencing Consortium (IWGSC). (2014) A chromosome-based draft sequence of the hexaploid bread wheat (Triticum aestivum) genome. *Science*, **345**, 1251788–1251788.
61. Hellsten,U., Wright,K.M., Jenkins,J., Shu,S., Yuan,Y., Wessler,S.R., Schmutz,J., Willis,J.H. and Rokhsar,D.S. (2013) Fine-scale variation in meiotic recombination in Mimulus inferred from population shotgun sequencing. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 19478–19482.
62. Rensing,S.A., Lang,D., Zimmer,A.D., Terry,A., Salamov,A., Shapiro,H., Nishiyama,T., Perroud,P.-F., Lindquist,E.A., Kamisugi,Y. *et al.* (2008) The Physcomitrella genome reveals evolutionary insights into the conquest of land by plants. *Science*, **319**, 64–69.
63. Zimmer,A.D., Lang,D., Buchta,K., Rombauts,S., Nishiyama,T., Hasebe,M., Van de Peer,Y., Rensing,S.A. and Reski,R. (2013) Reannotation and extended community resources for the genome of the non-seed plant Physcomitrella patens provide insights into the evolution of plant gene structures and functions. *BMC Genomics*, **14**, 498.
64. UniProt Consortium. (2015) UniProt: a hub for protein information. *Nucleic Acids Res.*, **43**, D204–D212.
65. Kong,L., Zhang,Y., Ye,Z.-Q., Liu,X.-Q., Zhao,S.-Q., Wei,L. and Gao,G. (2007) CPC: assess the protein-coding potential of transcripts using sequence features and support vector machine. *Nucleic Acids Res.*, **35**, W345–W349.
66. Griffiths-Jones,S., Grocock,R.J., van Dongen,S., Bateman,A. and Enright,A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, **34**, D140–D144.
67. Bao,W., Kojima,K.K. and Kohany,O. (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA*, **6**, 11.
68. Nawrocki,E.P., Burge,S.W., Bateman,A., Daub,J., Eberhardt,R.Y., Eddy,S.R., Floden,E.W., Gardner,P.P., Jones,T.A., Tate,J. *et al.* (2015) Rfam 12.0: updates to the RNA families database. *Nucleic Acids Res.*, **43**, D130–D137.