Supplemental file

1. Supplemental Table

Supplemental Table 1. Sources of FLcDNAs.

Species	Monocot / Dicot	No. of FLcDNAs	Source	Reference
O. sativa (*japonica*)	Monocot	37,139	DDBJ/EMBL/GenBank databases	[1]
O. sativa (*indica*)	Monocot	10,083	DDBJ/EMBL/GenBank databases	[2]
O. rufipogon	Monocot	2,044	DDBJ/EMBL/GenBank databases	[3]
H. vulgare	Monocot	5,006	DDBJ/EMBL/GenBank databases	[4]
T. aestivum	Monocot	6,162	Triticeae Full-Length CDS DataBase (http://trifldb.psc.riken.jp/)	[5,6]
Z. mays	Monocot	60,963	DDBJ/EMBL/GenBank databases	[7-9]
A. thaliana	Dicot	35,991	DDBJ/EMBL/GenBank databases	[10-12]
G. max	Dicot	4,712	DDBJ/EMBL/GenBank databases	[13]
P. trichocarpa	Dicot	4,664	DDBJ/EMBL/GenBank databases	[14]
S. lycopersicum	Dicot	13,227	Kazusa Full-length Tomato cDNA Database (http://www.pgb.kazusa.or.jp/kaftom/index.html)	[15]
Total		179,991		

Supplemental Table 2. Sources of genome sequences and annotations.

Species	Monocot / Dicot	Version or download date	Source
O. sativa (*japonica*)	Monocot	IRGSP build4	http://rgp.dna.affrc.go.jp/E/IRGSP/Build4/build4.html, http://rapdb.dna.affrc.go.jp/archive/build4.html
Z. mays	Monocot	ZmB73_AGPv1 Release 4a.53	http://www.maizesequence.org/
S. bicolor	Monocot	1/7/2009	http://www.phytozome.net/sorghum
B. distachyon	Monocot	JGI v1.0 8x assembly	http://www.phytozome.net/brachy.php
A. thaliana	Dicot	TAIR9 assembly	http://www.arabidopsis.org/
P. trichocarpa	Dicot	Assembly release v1	http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html
V. vinifera	Dicot	1/7/2009	http://www.genoscope.cns.fr/externe/GenomeBrowser/Vitis/
L. japonicus	Dicot	5/28/2009	http://www.kazusa.or.jp/lotus/index.html
C. papaya	Dicot	1/7/2009	DDBJ/EMBL/GenBank databases

Supplemental Table 3. Reference sets of gene structures.

Species	Number of gene structures: Common (Reference)	Intraspecies mapping (RAP method) [a]	Annotation data (Other resources)
O. sativa	12,975	19,617[b]	52,327[c]
Z. mays	12,664	54,743	51,623[d]
A. thaliana	12,881	36,103	31,280[e]

[a] CDSs were predicted by intraspecies mapping following the method described by Itoh et al. (2006) [16].

[b] RAP representative with FLcDNA evidence.

[c] MSU 6.1.

[d] B73 RefGen_v1 Filtered Gene Set.

[e] TAIR 9.0 representative CDS.

Supplemental Table 4. Specificity of introns in CDSs and UTRs.

	O. sativa	*Z. mays*	*A. thaliana*
CDS	92.6% (40,023/43,208)	92.2% (32,674/35,439)	93.9% (16,695/17,789)
UTR	28.9% (425/1,473)	32.1% (640/1,992)	28.1% (9/32)

Supplemental Table 5. Improvement of SP and SN for each step in our pipeline

Program	Process	Intron		All introns	
		SP	SN	SP	SN
This study	All transcripts				
	Est2genome	87.4	77.6	44.7	52.5
	Detection of tandemly duplicated genes	87.4	77.8	45.0	52.9
	Correction of exon-intron boundaries	89.5	77.6	48.5	52.9
	Removal of short introns	91.3	77.6	52.2	55.0
	Representative transcripts	92.6	74.2	58.5	49.5
GeneSeqer[a]	All transcripts	61.9	79.9	13.9	45.2
	Representative transcripts	83.7	75.5	42.3	38.9
Sim4cc[b]	All predictions	87.9	73.0	40.5	47.2

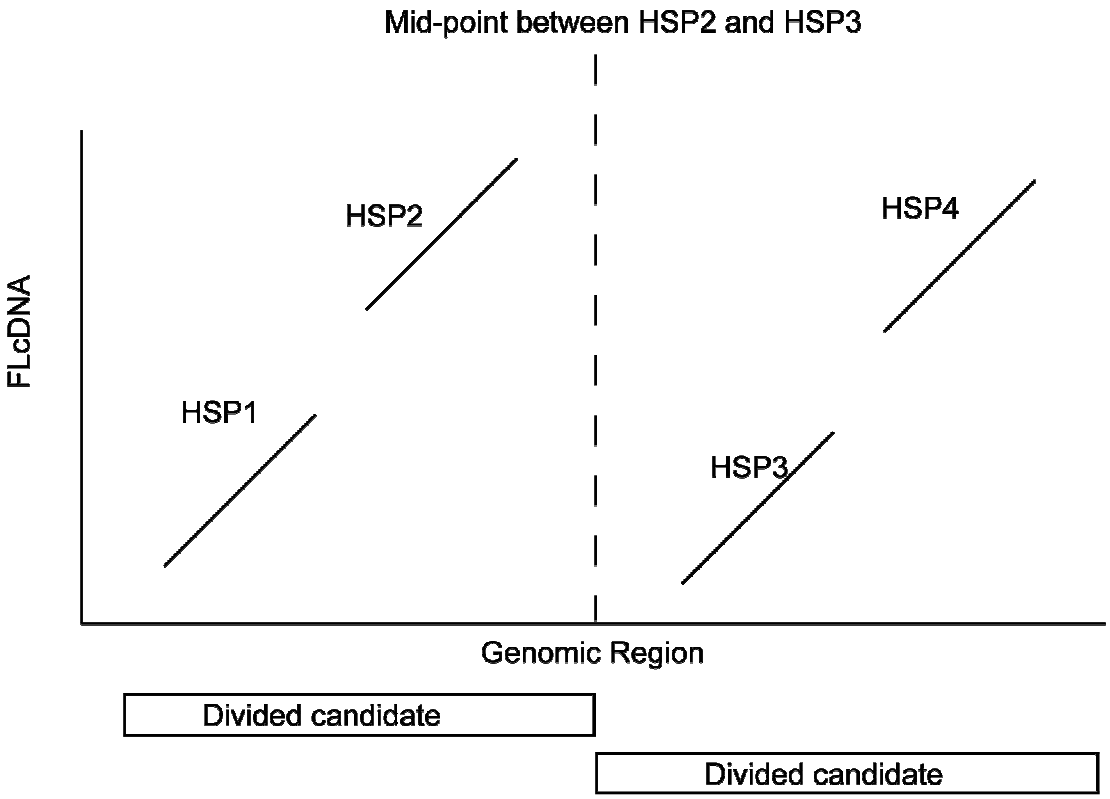
[a]Only CDS predictions were used.

[b]Sim4cc does not report CDSs, so that all results were used.

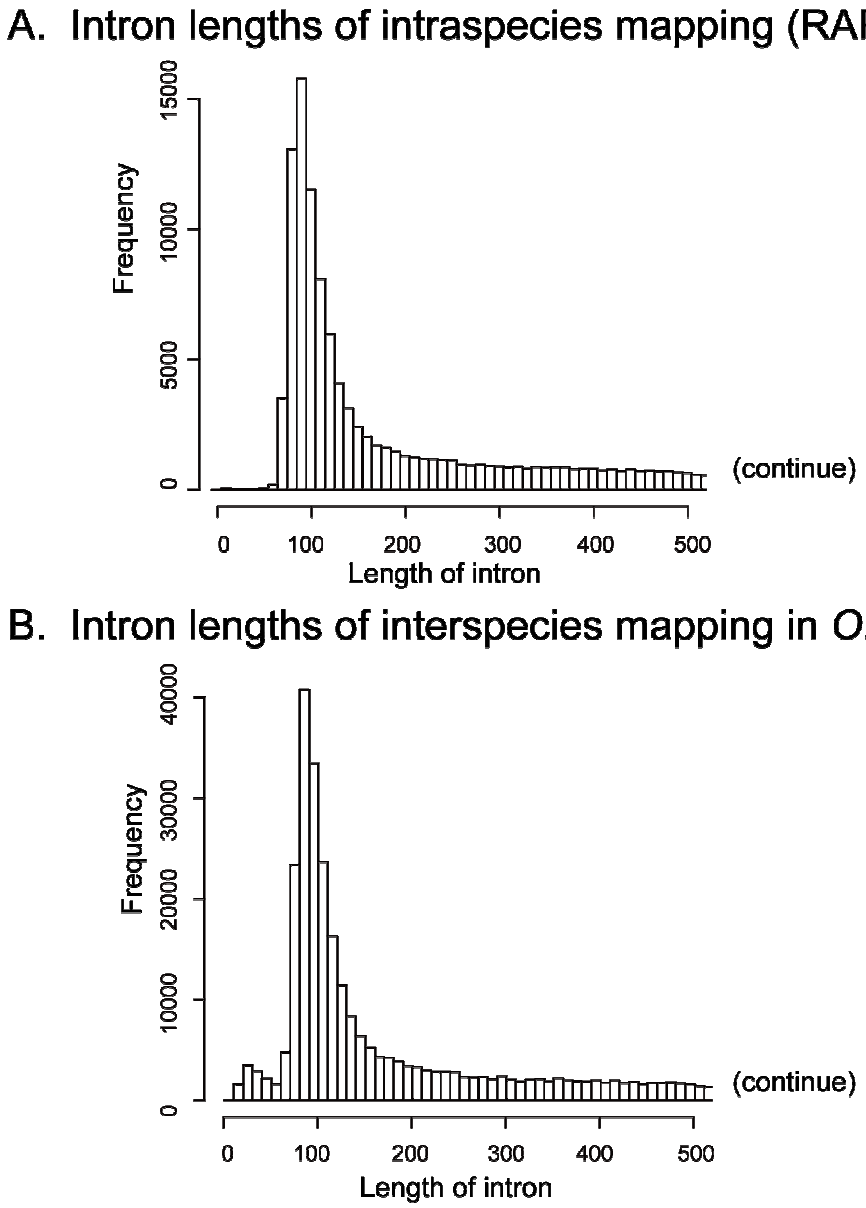
Supplemental Table 6. Mapping results for ten genomes.

	Intraspecies mapping		Interspecies mapping		Combined result	
Species	Loci	Mapped FLcDNAs	Loci	Mapped FLcDNAs	Loci	Mapped FLcDNAs
O. sativa	25,993	45,417	22,500	54,004	31,387	99,421
Z. mays	26,634	54,743	29,375	106,351	38,093	161,094
S. bicolor	-	-	29,842	96,142	29,842	96142
B. distachyon	-	-	25,120	87,174	25,120	87,174
A. thaliana	16,220	35,863	5,576	8,464	17,133	44,327
P. trichocarpa	3,887	4,655	15,377	26,502	16,943	31,157
G. max	4,715	4,594	18,124	22,738	19,891	27,332
C. papaya	-	-	8,294	26,717	8,294	26,717
V. vinifera	-	-	11,119	27,031	11,119	27,031
L. japonicus	-	-	12,729	22,729	12,729	22,729
Total	77,449	145,272	178,056	477,852	210,551	623,124

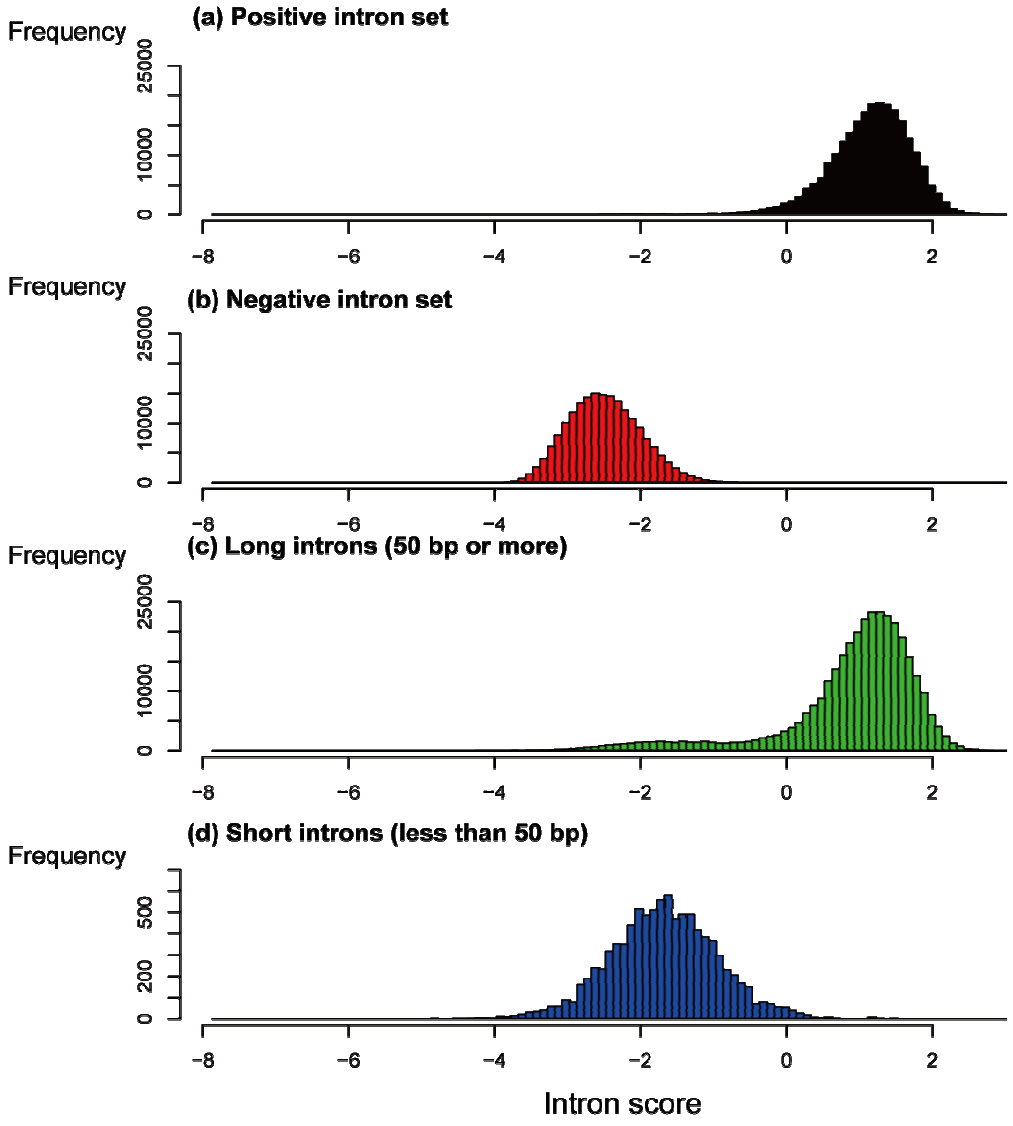
2. Supplemental Figures



Supplemental Figure 1. Dot plot-like representation of a candidate region of tandemly duplicated genes. The sequences of an FLcDNA and a genomic region were aligned. The horizontal axis indicates the genome, and the vertical axis indicates the FLcDNA. Diagonal lines designate high-scoring segment pairs (HSPs) identified by blastn. At the mid-point between the 3'-end of HSP2 and the 5'-end of HSP3, the candidate is divided in two, such that each portion contains a single candidate region.



Supplemental Figure 2. Distribution of lengths of introns. The ratio of short introns (<50 bp) of interspecies mapping and intraspecies mapping were 2.9% (8,817/345,180) and 0.2% (217/133,968), respectively.



Supplemental Figure 3. Distribution of intron scores. The horizontal axis shows the intron score, and the vertical axis shows the occurrence frequency of introns. **(a)** Positive data set of the interspecies predictions that were identical to the RAP introns. **(b)** Negative data set created from GT-AG sites within exons. **(c)** Predicted introns with lengths of 50 bp or more. **(d)** Predicted short introns with lengths of less than 50 bp.

3. Supplemental methods

Interspecies cDNA mapping and CDS identification

1. Algorithm of interspecies mapping

In the interspecies mapping, possible alignment errors were observed mainly for the following three reasons (Figure 2). First, tandemly duplicated genes tended to be accidentally combined into one gene. Second, some alignment gaps between FLcDNAs and genome sequences were regarded as short introns, but the possibility that they were large insertions or deletions (indels) could not be excluded. In fact, intraspecies mapping of *O. sativa* indicated that such short introns were quite rare (Supplemental Figure 2). Third, splice sites were particularly prone to misalignments. Our mapping algorithm solved these problems as follows.

To map FLcDNAs to a genome, a blastn comparison between the FLcDNAs and a given genome was used to identify exon candidates, which were defined as high-scoring segment pairs (HSPs) with *E*-values $\leq 10^{-10}$. HSPs were concatenated if their distances were ≤20 Kb. We added 5 Kb margins to both ends of the concatenated candidate regions. To cope with the first problem of tandem duplications, FLcDNA regions that appeared repeatedly with an *E*-value of ≤0.01 were identified as tandemly duplicated genes (Supplemental Figure 1). The FLcDNAs were then mapped to their candidate regions using the est2genome program [17]. Only the forward strands of the FLcDNAs were examined. Mapping assignments that exhibited coverage against an FLcDNA of less than 40% were excluded from further analyses. To deal with the second problem, short introns of less than 50 bp were discarded.

Finally, to precisely align the splice sites, an "intron score", which was defined as a linear combination of nucleotide identity and two splice site signals, was introduced as follows. Using an alignment between an FLcDNA and a genome, nucleotide identities were

calculated 20-bp upstream of the 5'-splice site of an intron and 20-bp downstream of the 3'-splice site of the same intron, and the two identities were averaged. Signals of the 5'- and 3'-ends of an intron were evaluated with a positional weight matrix (PWM) as described in the next section. The intron score was defined as an integrated score of the averaged nucleotide identity and the splice site signals; weights for the identity and signals were determined by linear discriminant analysis (LDA) implemented in the lda program of the MASS library of the R statistical package[18,19]. To determine the weights of LDA, a positive training set was constructed from introns that were identical between intraspecies and interspecies mappings, and all of the other introns that were not identical between intraspecies and interspecies mappings were regarded as a negative data set. To assess the validity of our intron score, we compared the positive and negative data sets, as well as the long and short introns (Supplemental Figure 3). If two or more different exon-intron boundaries were predicted within a 20 bp distance in a genome, the intron with the highest intron score among the overlapping introns was selected.

2. PWM and scores

To generate PWMs around splice sites of known genes, results from intraspecies mapping were used. We collected sequences from 3 bp upstream to 10 bp downstream of the 5' exon-intron boundaries, and from 15 bp upstream to 3 bp downstream of the 3' exon-intron boundaries. These ranges included significantly biased sites [20], and are predicted to be sufficient to detect splice signals. In fact, we examined larger regions (e.g., 50 bp), but did not observe significant improvement for the SP and SN (data not shown). The frequency of each nucleotide was calculated at each site. For a given alignment between an FLcDNA and a genome, an average of the summed frequencies of each 5' and 3' splice site was defined as the 5' or 3' PWM score. PWMs for *O. sativa*, *A. thaliana*, *P. trichocarpa,* and *G.max* were

created separately. For other species, the *O. sativa* PWMs were applied to the monocot genomes and the *A. thaliana* PWMs to the dicot genomes.

3. CDS identification

CDSs in predicted transcripts were determined on the bases of homology searches using blastx [21] against the Uniprot [22] and RefSeq [23] databases with a threshold of 50% or more identity. If more than one CDS was predicted within a single transcript, we selected the region that was positioned to the most 5'-upstream end of the transcript. If no homologs were found, GeneMark [24] was used to determine the reading frame of a CDS. If no CDS was predicted, the longest ORF with 70 a.a. or more was adopted.

The predicted ORFs were extended to both 5' and 3' regions on a genome until stop codons were found. The methionine (Met) that made up the longest ORF was used as the predicted start codon. If an appropriate Met was not found, the amino acid next to the stop codon at the 5'-end was used as the tentative start codon.

4. Definition of a locus

Mapped transcripts with at least 1 bp overlap were clustered. If multiple transcripts were mapped to a single locus, the transcript with the longest ORF was selected as a representative. When two or more transcripts possessed ORFs of the same size, the transcript having the largest number of exons was adopted as a representative for that locus.

Reference

1. Kikuchi, S., Satoh, K., Nagata, T., et al. 2003, Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice. *Science*, **301**, 376-379.

2. Liu, X., Lu, T., Yu, S., et al. 2007, A collection of 10,096 indica rice full-length cDNAs reveals highly expressed sequence divergence between Oryza sativa indica and japonica subspecies. *Plant Mol Biol*, **65**, 403-415.
3. Lu, T., Yu, S., Fan, D., et al. 2008, Collection and comparative analysis of 1888 full-length cDNAs from wild rice Oryza rufipogon Griff. W1943. *DNA Res*, **15**, 285-295.
4. Sato, K., Shin, I. T., Seki, M., et al. 2009, Development of 5006 full-length CDNAs in barley: a tool for accessing cereal genomics resources. *DNA Res*, **16**, 81-89.
5. Mochida, K., Yoshida, T., Sakurai, T., Ogihara, Y. and Shinozaki, K. 2009, TriFLDB: A Database of Clustered Full-Length Coding Sequences from Triticeae with Applications to Comparative Grass Genomics. *Plant Physiol*, **150**, 1135-1146.
6. Triticeae Full-Length CDS DataBase Triticeae Full-Length CDS DataBase.
7. Alexandrov, N. N., Brover, V. V., Freidin, S., et al. 2009, Insights into corn genes derived from large-scale cDNA sequencing. *Plant Mol Biol*, **69**, 179-194.
8. Jia, J., Fu, J., Zheng, J., et al. 2006, Annotation and expression profile analysis of 2073 full-length cDNAs from stress-induced maize (Zea mays L.) seedlings. *Plant J*, **48**, 710-727.
9. Soderlund, C., Descour, A., Kudrna, D., et al. 2009, Sequencing, Mapping, and Analysis of 27,455 Maize Full-Length cDNAs. *PLoS Genet*, **5**, e1000740.
10. Alexandrov, N. N., Troukhan, M. E., Brover, V. V., Tatarinova, T., Flavell, R. B. and Feldmann, K. A. 2006, Features of Arabidopsis genes and genome discovered using full-length cDNAs. *Plant Mol Biol*, **60**, 69-85.
11. Seki, M., Narusaka, M., Kamiya, A., et al. 2002, Functional annotation of a full-length Arabidopsis cDNA collection. *Science*, **296**, 141-145.
12. Seki, M., Satou, M., Sakurai, T., et al. 2004, RIKEN Arabidopsis full-length (RAFL) cDNA and its applications for expression profiling under abiotic stress conditions. *J Exp Bot*, **55**, 213-223.
13. Umezawa, T., Sakurai, T., Totoki, Y., et al. 2008, Sequencing and analysis of approximately 40,000 soybean cDNA clones from a full-length-enriched cDNA library. *DNA Res*, **15**, 333-346.
14. Ralph, S. G., Chun, H. J., Cooper, D., et al. 2008, Analysis of 4,664 high-quality sequence-finished poplar full-length cDNA clones and their utility for the discovery of genes responding to insect feeding. *BMC Genomics*, **9**, 57.
15. Aoki, K., Yano, K., Suzuki, A., et al. 2010, Large-scale analysis of full-length cDNAs from the tomato (Solanum lycopersicum) cultivar Micro-Tom, a reference system for the Solanaceae genomics. *BMC Genomics*, **11**, 210.
16. Itoh, T., Tanaka, T., Barrero, R. A., et al. 2007, Curated genome annotation of Oryza sativa ssp. japonica and comparative genome analysis with Arabidopsis thaliana. *Genome Res*, **17**, 175-183.

17. Mott, R. 1997, EST_GENOME: a program to align spliced DNA sequences to unspliced genomic DNA. *Comput Appl Biosci*, **13**, 477-478.
18. Ripley, B. D. 1996, *Pattern REcognition and Neural Networks*. Cambridge University Press.
19. Venables, W. N. and Ripley, B. D. 2002, *Modern Applied Statistics with S*. Springer.
20. Schwartz, S. H., Silva, J., Burstein, D., Pupko, T., Eyras, E. and Ast, G. 2008, Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. *Genome Res*, **18**, 88-103.
21. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J. 1990, Basic local alignment search tool. *J Mol Biol*, **215**, 403-410.
22. Uniprot Consortium 2009, The Universal Protein Resource (UniProt) 2009. *Nucleic Acids Res*, **37**, D169-D174.
23. Pruitt, K. D., Tatusova, T., Klimke, W. and Maglott, D. R. 2009, NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res*, **37**, D32-D36.
24. Borodovsky, M. and McIninch, J. 1993, Recognition of genes in DNA sequence with ambiguities. *Biosystems*, **30**, 161-171.