Crossing genome and transcriptome: deciphering links between structure and function in *Arabidopsis thaliana* genes

Véronique Brunaud[1], Virginie Bernard[1], David Armisén[1], Jean-Philippe Tamby[1], Séverine Gagnot[1], Sandra Derozier[1], Franck Samson[1], Cécile Guichard[1], Marie-Laure Martin-Magniette[1,2], Alain Lecharny[1] and Sébastien Aubourg[1]

[1] Unité de Recherche en Génomique Végétale (URGV) - UMR INRA 1165-CNRS 8114-UEVE, 2 Rue Gaston Crémieux, 91057 Evry Cedex, France.

[2] Unité de Mathématiques et Informatique Appliquées (MIA) - UMR 518 AgroParisTech-INRA, 16 Rue Claude Bernard, 75231 Paris Cedex, France.

(brunaud,bernard,armisen,tamby,gagnot,derozier,samson,guichard,martin, lecharny,aubourg)@evry.inra.fr

Abstract: *One of the challenges of bioinformatics today is to cross and analyse data differing in type, origin and quality in order to increase our knowledge on genomes. We have developed an information system focused on the model plant Arabidopsis and allowing the improvement of the functional annotation of genes by combining transcriptome and structural data.*

Keywords: transcription, annotation, evolution, regulation, plant, database, integration

1 Introduction

Since the sequencing of the *Arabidopsis thaliana* whole genome 8 years ago, gene annotation has been improved through several releases taking advantage of resources as like as expert curation, new genome availability, transcript sequencing projects and gene prediction software improvement. In parallel, several transcriptomic platforms based on DNA chips have been developed and they now give access to several thousands of transcriptomes [1,2]. Nevertheless, only 14% of the Arabidopsis genes have a biological function that was experimentally assessed while around 20% of genes remain without any functional information. In the context of functional genomic projects for plants, we have developed two databases for the management of genomic (FLAGdb^{++}, [3]) and transcriptomic data (CATdb, [4]). This information system allows us to integrate, through holistic approaches, gene models and expression profiles in order to improve the genome annotation and to decipher relationships between the organization, evolution and function of Arabidopsis genes.

2 Results

A combined approach of genome annotation and transcript analysis was firstly performed to identify new genes in the Arabidopsis genome. Probes on the CATMA microarrays were based on the gene models predicted by the EuGène software [5] and 677 probes were located within regions that were considered as intergenic by the official TAIR annotation. The statistical analysis of the results for more than 500 hybridized samples distributed among 12 organs provided an experimental validation for 465 novel genes [6]. These novel genes were characterized by their small size (encoding proteins with an average size of 137 aa) and very specific expression patterns.

Another illustration of the advantage to combine genomic and transcriptomic data is the characterization of a particular class of genes in plants: the unique genes [7]. Despite the major role of duplications in genome evolution, all characterized genomes include unique (single-copy) genes, i.e. genes without apparent paralog. Mining the FLAGdb^{++} database, we identified the unique genes within both Arabidopsis and rice genomes and classified them according to the number of homologs in the alternative species. Unique gene sets share structural features. In particular, the conserved unique gene pairs are characterized by a relatively small protein size, a high intron

density, a rare occurrence of TATA-box and a high occurrence of TELO-box. These structural features predict these genes as preferentially house-keeping genes with a slow evolution. Even if no shared transcription factor binding site (TFBS) can be detected in their promoter, the orthology relationship in Arabidopsis-rice gene pairs was strongly supported by a high conservation of their transcription levels. Furthermore, many unique genes have been conserved in single-copy throughout evolution from Prasinophytes to angiosperms, indicating that the uniqueness is under a strong selective pressure. A high proportion of conserved unique genes was also observed in other life phylums and we showed a link between protein targeting towards plastids and homology with bacterial proteins [7].

The expression of mature transcripts is controlled by the intron-exon structures and by the TFBS content of promoter regions. Also we have initiated a genomic study of the links between the core promoter architecture and gene function in Arabidopsis. Firstly, we identified different motifs with topological features that are strongly similar to canonical TATA-box features suggesting that they are functional motifs. Based on theses sequences, we established a novel classification of promoters and described links between promoter gene classes and the Gene Ontology categories. Secondly, we focused on the house-keeping genes, i.e. the genes expressed in almost all the conditions and organs, and found that TATA-box is under-represented in these genes. This atypical class of genes is also characterized by a compact structure (shorter introns and coding sequences).

3 Conclusion

The CATMA transcriptomes available in the CATdb database are fully independent of other public transcriptome resources. For instance, more than 4000 gene probes (including miRNA genes) are only present on CATMA chips. Using CATMA, it is therefore possible (i) to cross-validate results inferred from other resources [8] and (ii) to improve our Arabidopsis gene knowledge through the ‘guilt by association’ strategy.

References

[1] Graham NS, Broadley MR, Hammond JP, White PJ, May ST. Optimising the analysis of transcript data using high density oligonucleotide arrays and genomic DNA-based probe selection. BMC Genomics. 8:344, 2007.

[2] Allemeersch J, Durinck S, Vanderhaeghen R, Alard P, Maes R, Seeuws K, Bogaert T, Coddens K, Deschouwer K, Van Hummelen P, Vuylsteke M, Moreau Y, Kwekkeboom J, Wijfjes AH, May S, Beynon J, Hilson P, Kuiper MT. Benchmarking the CATMA microarray. A novel tool for Arabidopsis transcriptome analysis. Plant Physiol. 137:588-601, 2005.

[3] Samson F, Brunaud V, Duchêne S, De Oliveira Y, Caboche M, Lecharny A, Aubourg S. FLAGdb++: a database for the functional analysis of the Arabidopsis genome. Nucleic Acids Res. (Database issue) 32:D347-350, 2004. **http://urgv.evry.inra.fr/FLAGdb**

[4] Gagnot S, Tamby JP, Martin-Magniette ML, Bitton F, Taconnat L, Balzergue S, Aubourg S, Renou JP, Lecharny A, Brunaud V. CATdb: a public access to Arabidopsis transcriptome data from the URGV-CATMA platform. Nucleic Acids Res. (Database issue) 36:D986-990, 2008. **http://urgv.evry.inra.fr/CATdb**

[5] Schiex T, Moisan A, Rouzé P. Eugene, an eukaryotic gene finder that combines several sources of evidence. Lect. Notes Computational Sciences 2066:111-125, 2001.

[6] Aubourg S, Martin-Magniette ML, Brunaud V, Taconnat L, Bitton F, Balzergue S, Jullien PE, Ingouff M, Thareau V, Schiex T, Lecharny A, Renou JP. Analysis of CATMA transcriptome data identifies hundreds of novel functional genes and improves gene models in the Arabidopsis genome. BMC Genomics. 8:401, 2008.

[7] Armisén D, Lecharny A, Aubourg S. Unique genes in plants: specificities and conserved features throughout evolution. BMC Evol. Biol. 8:280, 2008.

[8] Hughes TR. 'Validation' in genome -scale research. Journal of Biology, 8:3, 2009.