Bioinformatics approaches in the study of circular RNAs in plants

mgr Katarzyna Nowis

ABSTRACT

Circular RNAs (circRNA) were discovered in the 70s of the last century. In the case of plants, it was viroid RNA (Sanger et al. 1976) and in the case of human, the transcript found in the cytoplasm of the HeLa cell line (Hsu et al. 1979). However, little attention was paid to circRNAs for a long time, because it was widely believed that they are products of aberrant splicing. Only the development of next-generation sequencing techniques in the 21st century made it possible to carry out more comprehensive studies of transcriptomes, including circRNA. These studies have revealed that circRNAs are evolutionarily conserved and wild spread in living organisms. These observations suggested that circRNAs have important biological functions. As a result, the number of publications on circRNA began to increase rapidly, but still little is known about the biogenesis and function of these molecules, particularly in plants.

The analysis of articles that have been already published on circRNA in plants revealed a number of problems faced by researchers. The first was the lack of well-validated laboratory and bioinformatics protocols dedicated to circRNA analysis. Most authors identified circRNAs, but did not perform their quantitative comparative analyses in organs or ecotypes. Another problem was the lack of a uniform method for depositing information on plant circRNAs in publicly available databases. As a consequence, it was not possible to directly compare the already collected data on circRNAs occurrence and their accumulation in different plant organs or lines. So far, three databases containing information on *Arabidopsis thaliana* circRNA have been established. These repositories provide information on circRNA identified by using different protocols and completely ignore the issues related to the reliability of the qualitative and quantitative analyzes performed.

In order to solve these problems, I have developed bioinformatic protocols that enable qualitative and quantitative analyses of circRNAs in various plant tissues or organs. Then I used these protocols to characterize circRNAs in the wild-type *A. thaliana* flowers, leaves, roots, and seedlings (ecotype Columbia, abbreviated as Col-0) and to identify splicing proteins that are involved in the biogenesis of circRNAs. The obtained results indicate that in *A. thaliana* most of the circRNAs are products of stochastic processes that are regulated independently of the transcription of the linear counterparts. In the next step, we attempted to identify proteins involved in the biogenesis of circRNA. Considering that they are splicing products, we have analyzed circRNA accumulation in *A. thaliana* mutants lacking individual splicing protein genes. Analysis of wild-type circRNA and 18 knock-out mutants revealed that silencing of three splicing-related genes had a significant impact on the production and accumulation of circRNAs (mutants: cbp80, flk, c2h2). It suggests that the lack of proteins encoded by these genes disrupts the transcript processing and contributes to an increase in circRNA production. Moreover, for two variants i.e. cbp80, c2h2, a large number of unique circRNAs not found in wild-type *A. thaliana* was identified. Additionally, I used previously developed circRNA dedicated protocols to analyze publicly available RNA-seq data for *A. thaliana*. All the obtained results were deposited in a specially created public database (http://plantcircrna.ibch.poznan.pl/). Thus, the possibility of qualitative and quantitative comparative analyses of all circRNAs identified in *A. thaliana* was created.