Journal of
Data Mining in Genomics & Proteomics

Jiang et al., J Data Mining Genomics Proteomics 2012, 3:3
DOI: 10.4172/2153-0602.1000122

Research Article Open Access

Comparative Genomic Analysis of ADP-ribose-1''-monophosphatase in 19 *Arabidopsis thaliana* Ecotypes

Huifang Jiang#, Haichao Wei#, Chunyun Jiang, Wei Sun, Wen Dong, Nini Chen, Hui Zhang, Yanxiu Zhao and Zenglan Wang*

Key Lab of Plant Stress Research, Life Science College, Shandong Normal University, 88 Wenhua East Road, Jinan 250014, China
#Authors contributed equally

Abstract

ADP-ribose-1''-monophosphatase containing A1pp or MACRO domain is an important processing enzyme in cells, participating in splicing the t-RNA procedures and catalyzing ADP-ribose-1''- monophosphate into ADP-ribose. We identified two genes, *AT1G69340* and *AT2G40600*, in *Arabidopsis thaliana* and found that, although there were many differences in amino acid, the spatial structure of conserved region was similar. We also analyzed the difference in sequence of promoter, coding region and untranslated region using the data from the whole genome of 19 ecotypes and compared both genes' expression in different tissues in Col-0 and in seedlings in 19 ecotypes based on AtGenExpress database and the RNA-seq data, respectively. We found the same gene had different expression patterns in some ecotypes and the two genes had the similar expression patterns except in floral organs and seeds according to the data of Col-0. These results implied that the regulatory mechanisms of these genes' expressions had changed in these ecotypes for the diversities of transcription factors and transcription factor binding sites. Above all, our research will provide some information for description of the gene function and the ecotype candidates used to study the genes' function.

Keywords: Comparative genomic analysis; Appr-1''-pase; *Arabidopsis thaliana* ecotype; MACRO; 3D structure; RNA-seq

Introduction

ADP-ribose-1''-monophosphatase (Appr-1''-pase) which contains A1pp or MACRO domain is an important processing enzyme in cells, participating in pre-tRNA splicing and catalyzing ADP-ribose-1-phosphate (Appr-1''-p) pase into ADP-ribose [1,2]. In this pre-tRNA splicing, a specific phosphotransferase changes the spliced tRNA into ADP-ribose- 1'', 2''-cyclic phosphate (Appr>p). Appr>p will become ADP-ribose-1-phosphate ADP-ribose-1-phosphate (Appr-1''-p) under cyclic phosphodiesterase (CPDase) [3-6], and ADP-ribose-1-phosphate (Appr-1''-p) will be further hydrolyzed into ADP-ribose [6]. ADP-ribose pyrophosphatase (ADPRase) catalyzes the ADP-ribose into ribose-5-phosphate and AMP [7]. It has been suggested that Appr>p, or its hydrolysis product, may play some unknown regulatory functions in the cell.

A1pp or MACRO domain is composed of 130~190 amino acids and can bind ADP-ribose [8]. Early research of YBR022W in yeast finds that A1pp domain is associated with the activity of Appr-1''-p processing enzyme [6]. Kumaran et al. [2] find instead of the Appr-1''-pase generally exists as a dimer when they analyze the crystal structure of YMX7_YEAST protein in yeast. The monomer is consisted of two domains, with 200 (16–219) residues forming α/β domains and 60 (220–279) residues forming a helical bundle. The core of α/β domain forms a hydrophobic region which makes the two monomers into a dimer [2]. Appr-1''-pase together with ADP-ribose can form an Appr-1''-pase: ADP-ribose complex. The crystal structure of this complex shows that ADP-ribose can bind with β-sheet of A1pp (or MACRO) domain [2,9].

MACRO domains are ancient, highly evolutionarily conserved domains. They exist in many species, for example, homo (marcroH2A) [10], bacteria (POA8D6) [11], archaeal (O59182) [12]. But the MACRO domain family has few members in species. A1pp or MACRO domains can recognize the ADP-ribose or poly (ADP-ribose) and participate in ADP ribosylation which plays an important role in complex biological processes, such as DNA damage and repair [13,14], transcriptional activation and repression [15], microtubule formation [16], telomere biology [17,18], insulator activation [19], mitosis [16,20], cell proliferation and differentiation [21-24], apoptosis [25,26] and programmed cell death [27].

The MACRO domain family has high divergence among each member. Perhaps, these variances are correlated with their biological functions. We know each member of MACRO gene family can take part in different biological processes, however the functions of MACRO domain proteins in plants are known little. In our work, we used Appr-1-p as Keyword to search the related genes in Arabidopsis genome on TAIR10 and found that ADP-ribose-1''-monophosphatase family had only two members, *AT1G69340* and *AT2G40600*, which are two novel genes. To better study the function of both genes, we compared the promoters, coding sequence (CDS), amino acid sequence, untranslated region (UTR) of these two genes using the whole genomic data of 19 *Arabidopsis thaliana* ecotypes [28]. We compared the 3D structures of AT2G40600 in 19 *Arabidopsis thaliana* ecotypes using ESyPred3D Web Server 1.0. Moreover, we analyzed gene transcriptome data of these ecotypes and acquired their different expression patterns. According to our results, we hope to understand the relationship between diversity in gene conservation and expression patterns and to select the suitable

***Corresponding author:** Zenglan Wang, Key Lab of Plant Stress Research, Life Science College, Shandong Normal University, 88 Wenhua East Road, Jinan 250014, China, Tel: 86-531-86180764; Fax: 86-531-86180764; E-mail: wangzenglan666@yahoo.cn

Received November 05, 2012; **Accepted** December 26, 2012; **Published** January 03, 2013

Citation: Jiang H, Wei H, Jiang C, Sun W, Dong W, et al. (2012) Comparative Genomic Analysis of ADP-ribose-1''-monophosphatase in 19 *Arabidopsis thaliana* Ecotypes. J Data Mining Genomics Proteomics 3: 122. doi:10.4172/2153-0602.1000122

ecotypes to study the gene function.

Materials and Methods

Data collection

Appr-1″-pase contained A1pp or MACRO domain. We used Appr-1-p as keyword to search the related genes in Arabidopsis genome on TAIR10, and we found only two members, *AT1G69340* and *AT2G40600,* had the potential to be Appr-1″-p processing enzyme. Further, we used both genes as query to search other homologous genes on TAIR10.

Multiple sequence alignment and analysis

We extracted the A1pp or MACRO domain using simple modular architecture research tool (SMART, http://smart.embl-heidelberg.de/). Using DNAstar tool (http://www.dnastar.com/) with default parameters, we performed multiple sequence alignments on the obtained sequences of the A1pp or MACRO domains. We used weblogo 3.3 (http://weblogo.threeplusone.com/) to estimate the site-by-site variation and used ESyPred3D Web Server 1.0 (http://www.fundp.ac.be/sciences/biologie/urbm/bioinfo/esypred/) to predict the 3D structure of MACRO domain of YMX7_YEAST, AT1G69340 and AT2G40600, and the structures of AT2G40600 in 19 ecotypes. These 3D structures were shown with VEGA ZZ2.4.0 (http://www.vegazz.net/).

To gain the divergence and identity of *AT1G69340* and *AT2G40600* among 19 Arabidopsis ecotypes (Table 1), we chose the sequence of promoter (2 Kb regions upstream of the start codon), CDS, UTR, amino acid and performed multiple alignment analysis of each sequence using DNAstar and Clustalx 2.0 (http://www.clustal.org/). We used TESS (http://www.cbil.upenn.edu/cgi-bin/tess/tess) to predict transcription factors binding sites in DNA sequences using site or consensus strings and positional weight matrices from the TRANSFAC, JASPAR, IMD, and CBIL-Gibbs Mat database with the default parameters (http://www.cbil.upenn.edu/cgi-bin/tess/tess?RQ=WELCOME). All the genomic sequences are from the web site (http://mus.well.ox.ac.uk/19genomes).

Expression profile analysis

We used AtGenExpress database (http://jsp.weigelworld.org/expviz/expviz.jsp) to detect the expression patterns of *AT1G69340* and *AT2G40600* in different development stages and different organs in ecotype Col-0 [29]. We also analyzed the diversity of both gene expressions in seedling period in 19 ecotypes using RNA-seq data from the website (http://mus.well.ox.ac.uk/19genomes) (all data were normalized by RPKM (Reads Per Kilobase per Million mapped reads) and constructed expression profile with software R 2.12.0 (http://cran.r-project.org/bin/windows/base/old/2.12.0/).

Results

The identification of Appr-1″-pase gene

Appr-1″-pase has A1pp or MACRO domain which is associated with the activity of Appr-1″-P processing enzyme, so we used Appr-1-p as keyword to search related genes in Arabidopsis genome, and only found *AT1G69340* and *AT2g40600*, which encodes 562, 257 amino acid residues, respectively. To be sure if there were only these two genes, we used both genes as query to search in TAIR10. At last, we found there were no other genes having the processing enzyme activity except *AT1G69340* and *AT2G40600*.

We compared the amino acids of both proteins and found that the two proteins had few consensus sites (Figure 1). We further used SMART to predict the conservation domain of both proteins and found AT1G69340 had MACRO domain and SEC14 domain, whereas AT2G40600 had only MACRO domain (Figure 1).

Accession	Origin	AIMS Stock Centre
Bur-0	Ireland	CS6643
Can-0	Canary Isles	CS6660
Col-0	Columbia	CS6673
Ct-1	Italy	CS6674
Edi-0	Scotland	CS6688
Hi-0	Netherlands	CS6736
Kn-0	Lithuania	CS6762
Ler-0	Poland, formerly Germany	CS20
Mt-0	Libya	CS1380
No-0	Germany	CS6805
Oy-0	Norway	CS6824
Po-0	Germany	CS6839
Rsch-4	Russia	CS6850
Sf-2	Spain	CS6857
Tsu-0	Japan	CS6874
Wil-2	Russia	CS6889
Ws-0	Russia	CS6891
Wu-0	Germany	CS6897
Zu-0	Germany	CS6902

Table 1: The 19 different ecotypes of *Arabidopsis thaliana*.

The crystal structure of YMX7_YEAST protein in yeast has been reported. We want to know if the difference of the amino acid residues will change the spatial structure of MACRO or A1pp domain. Extracting the A1pp or MACRO domain of YMX7_YEAST, AT1G69340 and AT2G40600 with SMART database, we got the alignments of them and the statistic of site-by-site analysis (Figure 2). The result showed that the amino acid sequence of the A1pp or MACRO domain among the three proteins had high diversity. In spite of this, the predicted 3D structure of the MACRO domain of each protein demonstrated that the ADP-ribose binding regions (α/β domains) were the same (Figure 1S).

The analysis of *AT1G69340* and *AT2G40600* in 19 ecotypes

To estimate if there are changes in gene structures of *AT1G69340* and *AT2G40600* in 19 ecotypes (Table 1), we analyzed the sequences of each gene among the 19 ecotypes at the level of protein and nucleic acid, respectively.

The analysis of amino acid sequence of *AT1G69340* and *AT2G40600* in 19 ecotypes: We got the divergence and identity of amino acid sequence of AT1G69340 and AT2G40600 in 19 ecotypes using DNAstar tool (Table 1S). The result showed that in 19 ecotypes, the amino acid sequence had no variations in AT1G69340 while small difference in AT2G40600: the 139th site A was changed into E and the 235th site V was changed into I in zu and edi compared with other ecotypes (Supplement 1,2). Moreover, to explain whether the amino acid residue variations give rise to potential protein structural or functional changes, we further analyzed the 3D structure of AT2G40600 in 19 ecotypes and found that the spatial structures of AT2G40600 in all the ecotypes had no changes (Figure 2S).

The analysis of CDS of *AT1G69340* and *AT2G40600* in 19 ecotypes: We compared the CDS of *AT1G69340* and *AT2G40600* in 19 ecotypes and got the divergence and identity (Table 2). The result showed the 383th site A and 404th site A of *AT1G69340* in his were

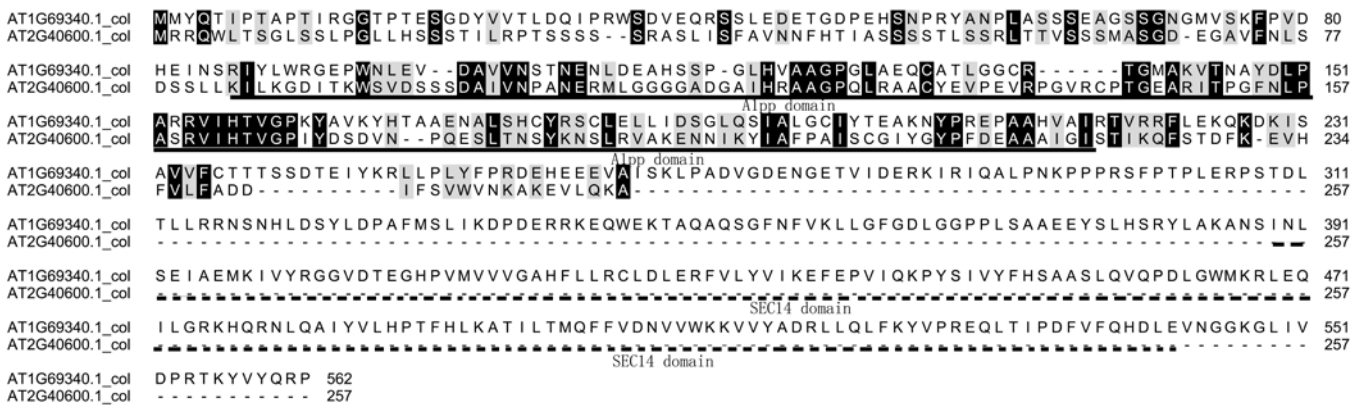

Figure 1: The amino acid alignment between AT1G69340 and AT2G40600 in Col-0. The solid line below the sequence was the conserved regions of A1pp or MACRO domain and the dashed line stood for SEC-14 domain in AT1G69340 and AT2G40600.

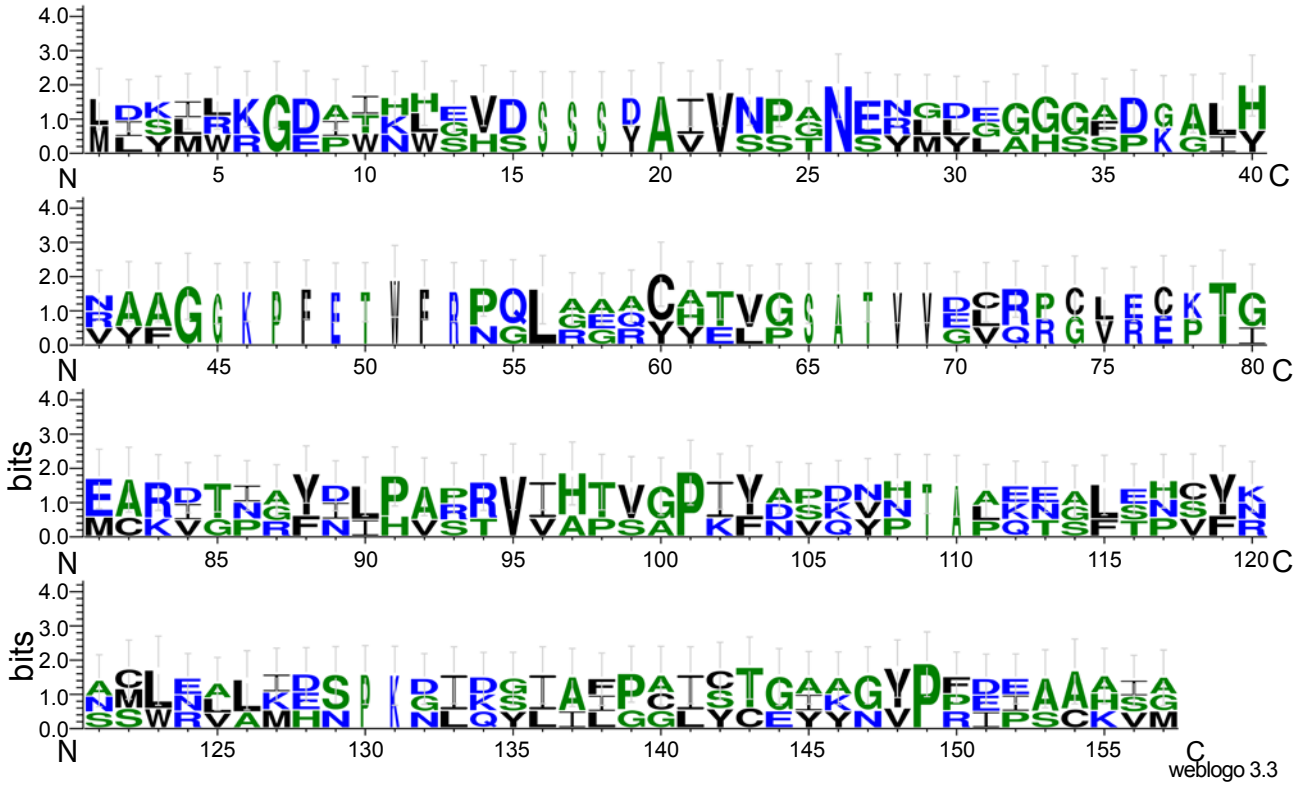

Figure 2: The site-by-site variation analysis of A1pp domains among AT1G69340, AT2G40600 and YMX7_YEAST. We extracted the conserved domain of AT1G69340, AT2G40600 and YMX7_YEAST by SMART. We took the alignment of the three proteins, and then gained the site-by-site statistic of the conserved domain using weblogo 3.3 (http://weblogo.threeplusone.com/). Error bars indicated an approximate 95% Bayesian confidence interval. The information content of a site (bits) is defined as the relative entropy of the monomers at the site to the background distribution.

turned into C and T compared with other ecotypes, respectively. The 175th site C of *AT1G69340* in ler, no, mt, ws, edi, wil was turned into T compared with other ecotypes. The 175th site A of *AT1G69340* in sf was turned into C. The 143th site T, 388th site C and 705th site G of *AT2G40600* in zu and edi were turned into G, A, A compared with other ecotypes, respectively (Supplement 3,4).

The analysis of 5'-UTR of *AT1G69340* and *AT2G40600* in 19 ecotypes: Through the alignments of 5'-UTR of *AT1G69340* and *AT2G40600* in 19 ecotypes, we found there were no variations in 5'-UTR of *AT2G40600*, while there was a variation in 142th site of *AT1G69340* in ler (G was turned into C) compared with other ecotypes (Supplement 5,6).

The analysis of promoter of *AT1G69340* and *AT2G40600* in 19 ecotypes: Taking the 2 Kb regions before ATG in genome as their promoters, we compared promoter sequence of *AT1G69340* and *AT2G40600* in Col-0 (Figure 3) and got the identity between

A: The identity and divergence of *AT1G69340* CDS among 19 ecotypes

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1		99.9	99.9	99.8	99.8	99.8	99.9	99.9	99.9	99.9	99.8	99.9	99.8	99.9	99.8	99.9	99.8	99.9	99.9
2	0.1		100.0	99.9	99.9	99.9	100.0	100.0	100.0	100.0	99.9	100.0	99.9	100.0	99.9	100.0	99.9	100.0	100.0
3	0.1	0.0		99.9	99.9	99.9	100.0	100.0	100.0	100.0	99.9	100.0	99.9	100.0	99.9	100.0	99.9	100.0	100.0
4	0.2	0.1	0.1		100.0	100.0	99.9	99.9	99.9	99.9	100.0	99.9	100.0	99.9	100.0	99.9	99.9	99.9	99.9
5	0.2	0.1	0.1	0.0		100.0	99.9	99.9	99.9	99.9	100.0	99.9	100.0	99.9	100.0	99.9	99.9	99.9	99.9
6	0.2	0.1	0.1	0.0	0.0		99.9	99.9	99.9	99.9	100.0	99.9	100.0	99.9	100.0	99.9	99.9	99.9	99.9
7	0.1	0.0	0.0	0.1	0.1	0.1		100.0	100.0	100.0	99.9	100.0	99.9	100.0	99.9	100.0	99.9	100.0	100.0
8	0.1	0.0	0.0	0.1	0.1	0.1	0.0		100.0	100.0	99.9	100.0	99.9	100.0	99.9	100.0	99.9	100.0	100.0
9	0.1	0.0	0.0	0.1	0.1	0.1	0.0	0.0		100.0	99.9	100.0	99.9	100.0	99.9	100.0	99.9	100.0	100.0
10	0.1	0.0	0.0	0.1	0.1	0.1	0.0	0.0	0.0		99.9	100.0	99.9	100.0	99.9	100.0	99.9	100.0	100.0
11	0.2	0.1	0.1	0.0	0.0	0.0	0.1	0.1	0.1	0.1		99.9	100.0	99.9	100.0	99.9	99.9	99.9	99.9
12	0.1	0.0	0.0	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.1		99.9	100.0	99.9	100.0	99.9	100.0	100.0
13	0.2	0.1	0.1	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.0	0.1		99.9	100.0	99.9	99.9	99.9	99.9
14	0.1	0.0	0.0	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.1	0.0	0.1		99.9	100.0	99.9	100.0	100.0
15	0.2	0.1	0.1	0.0	0.0	0.0	0.1	0.1	0.1	0.1	0.0	0.1	0.0	0.1		99.9	99.9	99.9	99.9
16	0.1	0.0	0.0	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.1		99.9	100.0	100.0
17	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1		99.9	99.9
18	0.1	0.0	0.0	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.1	0.0	0.1		100.0
19	0.1	0.0	0.0	0.1	0.1	0.1	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.1	0.0	0.1	0.0	

1-19 stood for hi, kn, rsch, ler, no, mt, col, can, bur, wu, ws, zu, edi, ct, wil, tsu, sf, po and oy, right upper region stood for the identity and the left bottom region stood for divergence

B: The identity and divergence of *AT2G40600* CDS among 19 ecotypes

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1		100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.6	99.6	100.0	100.0	100.0	100.0	100.0	100.0
2	0.0		100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.6	99.6	100.0	100.0	100.0	100.0	100.0	100.0
3	0.0	0.0		100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.6	99.6	100.0	100.0	100.0	100.0	100.0	100.0
4	0.0	0.0	0.0		100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.6	99.6	100.0	100.0	100.0	100.0	100.0	100.0
5	0.0	0.0	0.0	0.0		100.0	100.0	100.0	100.0	100.0	100.0	99.6	99.6	100.0	100.0	100.0	100.0	100.0	100.0
6	0.0	0.0	0.0	0.0	0.0		100.0	100.0	100.0	100.0	100.0	99.6	99.6	100.0	100.0	100.0	100.0	100.0	100.0
7	0.0	0.0	0.0	0.0	0.0	0.0		100.0	100.0	100.0	100.0	99.6	99.6	100.0	100.0	100.0	100.0	100.0	100.0
8	0.0	0.0	0.0	0.0	0.0	0.0	0.0		100.0	100.0	100.0	99.6	99.6	100.0	100.0	100.0	100.0	100.0	100.0
9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		100.0	100.0	99.6	99.6	100.0	100.0	100.0	100.0	100.0	100.0
10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		100.0	99.6	99.6	100.0	100.0	100.0	100.0	100.0	100.0
11	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0		99.6	99.6	100.0	100.0	100.0	100.0	100.0	100.0
12	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4		100.0	99.6	99.6	99.6	99.6	99.6	99.6
13	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.0		99.6	99.6	99.6	99.6	99.6	99.6
14	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.4		100.0	100.0	100.0	100.0	100.0
15	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.4	0.0		100.0	100.0	100.0	100.0
16	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.4	0.0	0.0		100.0	100.0	100.0
17	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.4	0.0	0.0	0.0		100.0	100.0
18	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.4	0.0	0.0	0.0	0.0		100.0
19	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.4	0.0	0.0	0.0	0.0	0.0	

1-19 stood for hi, kn, rsch, ler, no, mt, col, can, bur, wu, ws, zu, edi, ct, wil, tsu, sf, po and oy, right upper region stood for the identity and the left bottom region stood for divergence

Table 2: The identity and divergence of *AT1G69340* and *AT2G40600* CDS among 19 ecotypes.

them was 46%. Further, we performed the alignment of *AT1G69340* and *AT2G40600* promoters in 19 ecotypes, respectively and got the divergence and identity (Table 3). The result showed high conservation of promoter sequence in different ecotypes (Supplement 7,8) in *AT1G69340*, their identities were above 96.4%; in *AT2G40600*, their identities were above 99.7%. To gain the detail of promoter, we used TESS to predict the cis-elements in the promoter regions (Tables 2S, 3S). From these results, we found that among 19 ecotypes, the number of transcription factor binding sites of *AT1G69340* and *AT2G40600* changed in the range of 612 ± 9 and 662 ± 5, respectively (Figure 4), but the difference was not up to a great degree, which was consistent with the high identity of promoter sequence of each gene in 19 ecotypes. Compared with *AT1G69340*, *AT2G40600* promoters had more transcription factor binding sites (Figure 4). Although the number of transcription factors binding in the promoter of each gene also had no large difference among 19 ecotypes, the categories of the transcription factors were highly different. In Col-0, the number of specific transcription factors of *AT1G69340* and *AT2G40600* was 69 and 68, respectively. The difference in transcription factor types showed the same trend in other 18 ecotypes (Tables 2S and 3S).

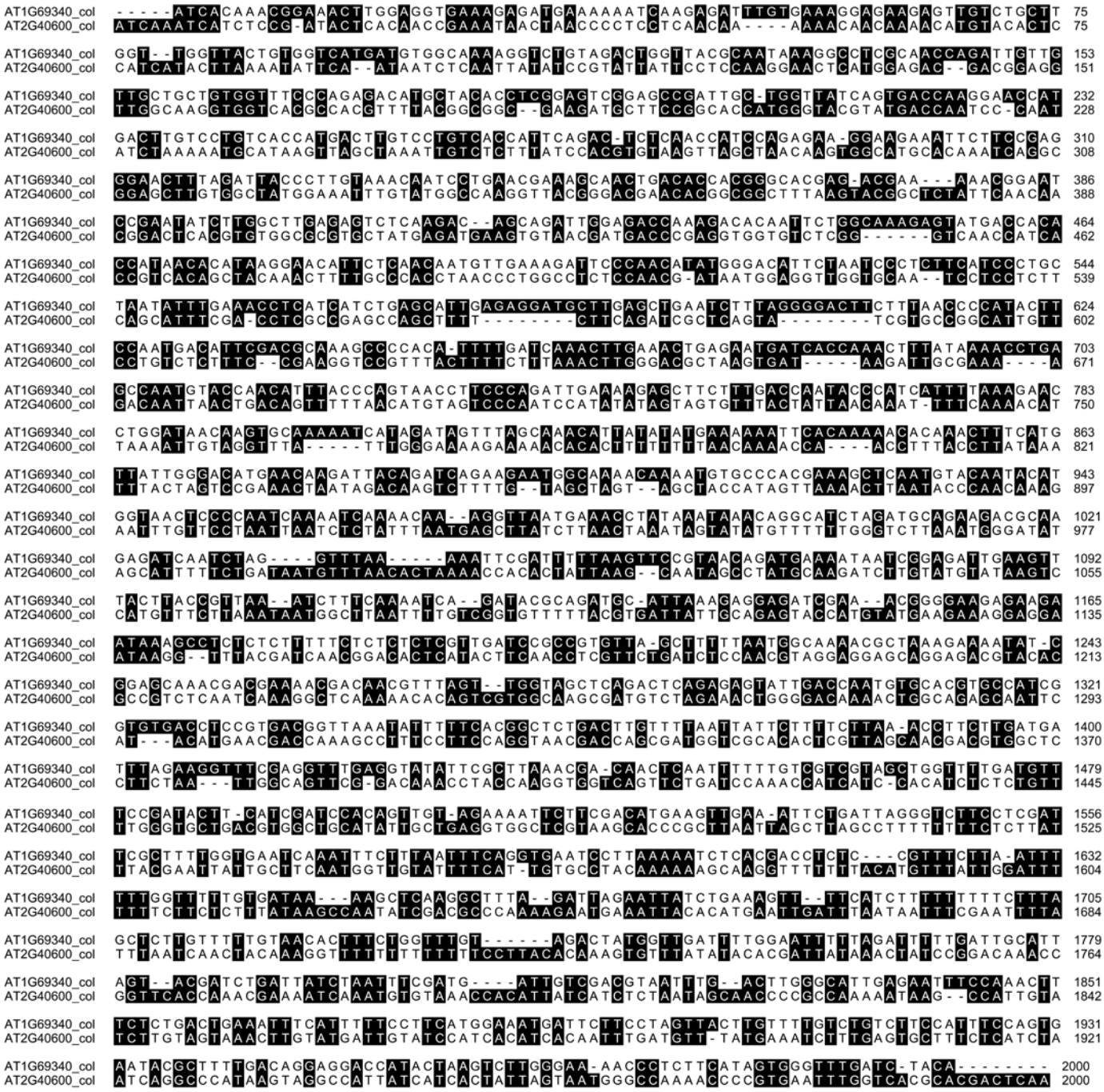

Figure 3: The alignment of promoter between *AT1G69340* and *AT2G40600* in Col-0.

The expression pattern of *AT1G69340* and *AT2G40600* in 19 ecotypes

Using AtGenExpress database, we analyzed the expression patterns of *AT1G69340* and *AT2G40600* in different tissues and development periods in ecotype Col-0 (Figure 5). The results showed both genes had the similar expression patterns except in floral organs and seeds. In floral organs and seeds, the expression of *AT2G40600* was less than *AT1G69340*. To further gain the expression information of these two genes in same tissue and same development stage of different ecotypes, we analyzed RNA-seq database (Figure 6) and found that the expression of the two genes in 19 ecotypes were difference: the expression of *AT2G40600* was higher than *AT1G69340* in different ecotypes except in zu; especially, in bur, ct and mt, the expression of *AT2G40600* was high as 2 times as *AT1G69340*. Moreover, we found the same gene had different expression in different ecotypes: the expression of *AT1G69340* in ws was 1.4 times as in rsch; the expression of *AT2G40600* in bur was 2.6 times as in zu (Figure 6).

A: The identity and divergence of *AT1G69340* promoter region among 19 ecotypes

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1		99.9	99.9	99.9	99.3	99.9	99.9	99.9	100.0	99.9	99.9	99.9	100.0	99.9	99.8	99.9	99.9	99.9	99.9
2	0.1		100.0	99.9	99.3	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.8	99.9	99.9	99.9	99.9
3	0.1	0.0		99.9	99.3	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.8	99.9	99.9	99.9	99.9
4	0.1	0.1	0.1		99.5	100.0	99.9	99.9	99.9	100.0	100.0	99.9	99.9	100.0	99.9	100.0	100.0	100.0	100.0
5	0.7	0.7	0.7	0.5		99.4	99.5	99.3	99.3	99.4	99.4	99.3	99.3	99.5	99.3	99.4	99.4	99.5	99.4
6	0.1	0.1	0.1	0.0	0.6		99.9	99.9	99.9	100.0	100.0	99.9	99.9	100.0	99.9	100.0	100.0	100.0	100.0
7	0.0	0.1	0.1	0.0	0.5	0.0		99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.8	99.9	99.9	99.9	99.9
8	0.1	0.1	0.1	0.1	0.7	0.1	0.1		99.9	99.9	99.9	99.9	99.9	99.9	99.8	99.9	99.9	99.9	99.9
9	0.0	0.1	0.1	0.1	0.7	0.1	0.0	0.1		99.9	99.9	99.9	100.0	99.9	99.8	99.9	99.9	99.9	99.9
10	0.0	0.1	0.1	0.0	0.6	0.0	0.0	0.1	0.1		100.0	99.9	99.9	100.0	99.9	100.0	100.0	100.0	100.0
11	0.1	0.1	0.1	0.0	0.6	0.0	0.0	0.1	0.1	0.0		99.9	99.9	100.0	99.9	100.0	100.0	100.0	100.0
12	0.1	0.1	0.1	0.1	0.7	0.1	0.1	0.1	0.1	0.1	0.1		99.9	99.9	99.9	99.9	99.9	99.9	99.9
13	0.0	0.1	0.1	0.1	0.7	0.1	0.0	0.1	0.0	0.1	0.1	0.1		99.9	99.8	99.9	99.9	99.9	99.9
14	0.1	0.1	0.1	0.0	0.5	0.0	0.0	0.1	0.1	0.0	0.0	0.1	0.1		99.9	100.0	100.0	100.0	100.0
15	0.2	0.2	0.2	0.1	0.7	0.1	0.1	0.2	0.2	0.1	0.1	0.1	0.2	0.1		99.9	99.9	99.9	99.9
16	0.1	0.1	0.1	0.0	0.6	0.0	0.0	0.1	0.1	0.0	0.0	0.1	0.1	0.0	0.1		100.0	100.0	100.0
17	0.1	0.1	0.1	0.0	0.6	0.0	0.0	0.1	0.1	0.0	0.0	0.1	0.1	0.0	0.1	0.0		100.0	100.0
18	0.1	0.1	0.1	0.0	0.5	0.0	0.0	0.1	0.1	0.0	0.0	0.1	0.1	0.0	0.1	0.0	0.0		100.0
19	0.1	0.1	0.1	0.0	0.6	0.0	0.0	0.1	0.1	0.0	0.0	0.1	0.1	0.0	0.1	0.0	0.0	0.0	

1-19 stood for col , bur, can, edi, hi , kn, po, ler, rsch, mt, sf , no, tsu, oy, wil, ws, wu, zu and ct, right upper region stood for the identity and the left bottom region stood for divergence

B: The identity and divergence of *AT2G40600* promoter region among 19 ecotypes

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
1		99.8	99.8	99.8	99.8	99.8	99.8	99.8	99.8	99.8	99.8	99.7	99.8	99.8	99.9	99.8	100.0	99.8	99.8
2	0.2		100.0	100.0	100.0	100.0	99.9	100.0	99.9	100.0	100.0	99.8	100.0	100.0	99.9	99.9	99.8	100.0	99.9
3	0.2	0.0		100.0	100.0	100.0	99.9	100.0	99.9	100.0	100.0	99.8	100.0	100.0	99.9	99.9	99.8	100.0	99.9
4	0.2	0.0	0.0		100.0	100.0	99.9	100.0	99.9	100.0	100.0	99.8	100.0	100.0	99.9	99.9	99.8	100.0	99.9
5	0.2	0.0	0.0	0.0		100.0	99.9	100.0	99.9	100.0	100.0	99.8	100.0	100.0	99.9	99.9	99.8	100.0	99.9
6	0.2	0.0	0.0	0.0	0.0		99.9	100.0	99.9	100.0	100.0	99.8	100.0	100.0	99.9	99.9	99.8	100.0	99.9
7	0.2	0.1	0.1	0.1	0.1	0.1		99.9	99.9	99.9	99.9	99.7	99.9	99.9	99.8	99.9	99.8	99.9	99.8
8	0.2	0.0	0.0	0.0	0.0	0.0	0.1		99.9	100.0	100.0	99.8	100.0	100.0	99.9	99.9	99.8	100.0	99.9
9	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.1		99.9	99.9	99.8	99.9	99.9	99.9	100.0	99.8	99.9	99.9
10	0.2	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1		100.0	99.8	100.0	100.0	99.9	99.9	99.8	100.0	99.9
11	0.2	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.0		99.8	100.0	100.0	99.9	99.9	99.8	100.0	99.9
12	0.3	0.2	0.2	0.2	0.2	0.2	0.3	0.2	0.2	0.2	0.2		99.8	99.8	99.8	99.8	99.7	99.8	99.8
13	0.2	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.0	0.2		100.0	99.9	99.9	99.8	100.0	99.9
14	0.2	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.0	0.2	0.0		99.9	99.9	99.8	100.0	99.9
15	0.1	0.1	0.1	0.1	0.1	0.1	0.2	0.1	0.1	0.1	0.1	0.2	0.1	0.1		99.9	99.9	99.9	99.9
16	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.0	0.1	0.1	0.2	0.1	0.1	0.1		99.8	99.9	99.9
17	0.0	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.3	0.2	0.2	0.1	0.2		99.8	99.8
18	0.2	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.1	0.0	0.0	0.2	0.0	0.0	0.1	0.1	0.2		99.9
19	0.2	0.1	0.1	0.1	0.1	0.1	0.2	0.1	0.1	0.1	0.1	0.2	0.1	0.1	0.1	0.1	0.2	0.1	

1-19 stood for col , bur, can, edi, hi , kn, po, ler, rsch, mt, sf , no, tsu, oy, wil, ws, wu, zu and ct, right upper region stood for the identity and the left bottom region stood for divergence

Table 3: The identity and divergence of *AT1G69340* and *AT2G40600* promoter region among 19 ecotypes.

Discussion

In the present study, we identified two Appr-1″-pases, AT1G69340 and AT2G40600, containing A1pp or MACRO domain in *Arabidopsis thaliana* genome with bioinformatics methods. Searching SMART database, we found AT1G69340 had another domain—SEC14 domain. The SEC14-containing proteins have different functions. The SEC14-only proteins are bona fide lipid transport proteins and the multi-domain SEC14-containing proteins have more complex functions in signal transduction, transport, and organelle biology [30]. Perhaps *AT1G69340* could also play more functions and involved in more biological processes because of the SEC-14 domain (Figure 1).

The previous research on yeast shows that Asn (80), Asp (90), His (145) are the functional sites of A1pp [2]. We compared amino acid sequence and 3D structure of conserved A1pp domian among YMX7_YEAST, AT1G69340 and AT2G40600 (Figures 2 and 1S). Our work showed that although the amino acid sequence of conserved region had major variations, the functional domains (α/β domains) were conserved. So they could have the similar function of ADP-ribose-1″-monophosphatase.

At present, the functions of *AT1G69340* and *AT2G40600* aren't very clear. To interpret their functional difference in different ecotypes, we analyzed the sequence of amino acids, CDS, UTR, promoters and the transcriptome of *AT1G69340* and *AT2G40600* in 19 *Arabidopsis thaliana* ecotypes. Together with all the results, we found there were a few variations in both genes' structures. The sequence of AT1G69340 at amino acid level had no changes among different ecotypes, whereas there were some single-base substitutions at CDS level: compared with other ecotypes, the 383th site A and 404th site A in his were turned into C and T, respectively; the 175th site C in ler, no, mt, ws, edi, wil was turned into T; the 175th site A in sf was turned into C. The CDS of *AT2G40600* had a few bases replaced in zu and edi. In zu and edi, the 143th site T was turned into G, the 388th site C was turned into A and the 705th site G was turned into A; as a result, the corresponding codon of the 388th site GCA (residue A) was turned into GAA (residue E), the 705th site GTT (residue V) was turned into ATT (residue I) (Tables 1S and 2, Supplement 1-4).

The alignment of 5'-UTR showed that 23 nt of *AT2G40600* had no variations among 19 ecotypes, and in 234 base pairs of AT1G69340, only -142 site G was replaced by C in ler compared with other ecotypes (Supplement 5,6). It suggested that both *AT2G40600* and *AT1G69340* had no difference in translation regulation among 19 ecotypes.

The analysis results of promoter showed high conservation among different ecotypes (Table 3, Supplement 7, 8): in *AT1G69340*, their identities were above 96.4%; in *AT2G40600*, their identities were above 99.7%. We predicted the potential transcription factors and transcription factor binding sites using TESS (Figure 4), the results showed that the number of cis-acting elements and trans-acting factors in *AT1G69340* promoter had no large difference in 19 ecotypes and so was in *AT2G40600*, which was consistent with the results of promoter sequence alignment of each gene among 19 ecotypes. The detailed analysis on the transcription factors demonstrated that among 19 ecotypes, the types of the transcription factors binding in promoter region of each gene were different. The result of RNA-seq showed the expression of *AT1G69340* and *AT2G40600* had some differences: in ws, the expression of *AT1G69340* was 1.4 times as in rsch; the expression of *AT2G40600* in bur was 2.6 times as in zu (Figure 6). These results may be ascribed to the small difference in cis-acting elements and the types of transcription factors binding in promoter region of each gene in 19 ecotypes. So we thought that the expression of the same gene in different ecotypes existed different regulation mechanism.

The data from AtGenExpress database showed the expression patterns of *AT1G69340* and *AT2G40600* in different organs were similar except in floral organs and seeds. In floral organs and seeds, the expression of *AT2G40600* was less than *AT1G69340* (Figure 5). The data of RNA-seq showed that among 19 ecotypes, the expression of both genes in seedlings also exhibited some differences. The expression of *AT2G40600* was more than *AT1G69340* in all other ecotypes except in zu (Figure 6), and in bur, ct and mt, the expression of *AT2G40600* was 2 times as that of *AT1G69340*. These results were perhaps because of the differences both in promoter diversities including the changes in the sequences and the number of cis-acting elements, and in the types of transcription factors between the two genes. Although the *AT2G4060* and *AT1G69340* were homologous genes, the regulation of their expression was complex.

Based on the above results, we proposed the optimal ecotypes used to gene function studies must meet two principles. Firstly, the expression of the gene in the ecotype was relatively high compared with other ecotypes; secondly, the expression difference of the homologous genes was obvious. So, we chose ct or bur to study the function of *AT2G40600* and chose zu or no to research the function of *AT1G69340*.

In our work, we got *AT1G69340* and *AT2G40600* containing

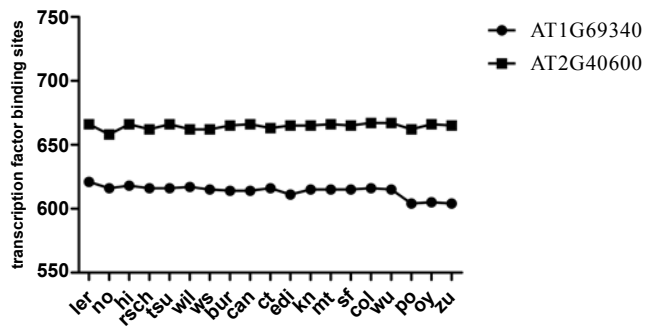

Figure 4: The number of transcription factor binding sites of both genes' promoter. The number of transcription factor binding sites in 19 ecotypes by TESS searching.

Figure 5: The expression pattern of *AT1G69340*, *AT2G40600* in different tissues in Col-0. We got the expression data from AtGenExpress database, and used software R 2.12.0 to make the expression profile.

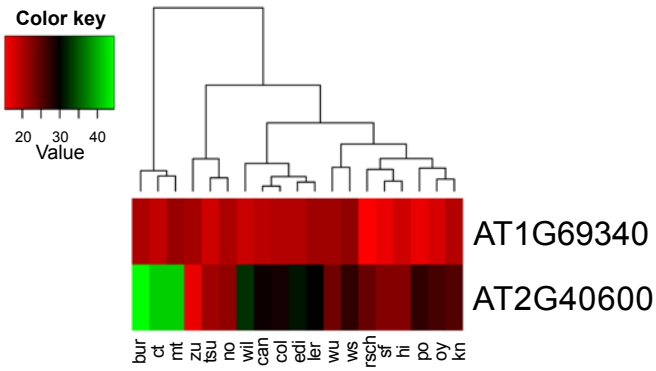

Figure 6: The expression pattern of *AT1G69340*, *AT2G40600* in 19 ecotypes during seedling stage. All the data were from the web site (http://mus.well.ox.ac.uk/19genomes) (all data was normalized by RPKM (Reads Per Kilobase per Million mapped reads)). "The color scale corresponds to the mean expression of a gene. Red represents low expression, while green represents high expression."

A1pp or MACRO domain in *Arabidopsis thaliana*. After analyzing the structures of both genes, we got the conclusion that the structures of the promoter, 5'-UTR and CDS of each gene had high identities among different ecotypes. We analyzed the relationship between gene structure and expression, including the same gene in different ecotypes and different genes in the same ecotype and found the expression profiles had some differences. The analysis of amino acid sequence showed high diversities in both proteins, but the 3D structures of their MACRO domains were very similar. So they could possess the similar functions. These works exposes the differences in the same gene in genomic structure and gene expression in different ecotypes and will provide a clue for further study of gene functions and for the candidate ecotype suitable to functional analysis.

Acknowledgement

This work was supported in part by the National Basic Research Program of China (Grant No. 2012CB114204), the Natural Science Foundation of Shandong Province, China (Grant No. ZR2010CM036), and by Science and Technology Development Planing of Shandong Province, China (Grant No. 2012GGB01136).

References

1. Dunstan MS, Barkauskaite E, Lafite P, Knezevic CE, Brassington A, et al. (2012) Structure and mechanism of a canonical poly(ADP-ribose) glycohydrolase. Nat Commun 3: 878.
2. Kumaran D, Eswaramoorthy S, Studier FW, Swaminathan S (2005) Structure and mechanism of ADP-ribose-1″-monophosphatase (Appr-1″-pase), a ubiquitous cellular processing enzyme. Protein Sci 14: 719-726.
3. Bowman S, Churcher C, Badcock K, Brown D, Chillingworth T, et al. (1997) The nucleotide sequence of Saccharomyces cerevisiae chromosome XIII. Nature 387: 90-93.
4. Culver GM, Consaul SA, Tycowski KT, Filipowicz W, Phizicky EM (1994) t-RNA splicing in yeast and wheat germ. A cyclic phosphodiesterase implicated in the metabolism of ADP-ribose 1",2"-cyclic phosphate. J Biol Chem 269: 24928-24934.
5. Hofmann A, Zdanov A, Genschik P, Ruvinov S, Filipowicz W, et al. (2000) Structure and mechanism of activity of the cyclic phosphodiesterase of Appr>p, a product of the tRNA splicing reaction. EMBO J 19: 6207-6217.
6. Martzen MR, McCraith SM, Spinelli SL, Torres FM, Fields S, et al. (1999) A biochemical genomics approach for identifying genes by the activity of their products. Science 286: 1153-1155.
7. Gabelli SB, Bianchet MA, Bessman MJ, Amzel LM (2001) The structure of ADP-ribose pyrophosphatase reveals the structural basis for the versatility of the Nudix family. Nat Struct Biol 8: 467-472.
8. Hassa PO, Haenni SS, Elser M, Hottiger MO (2006) Nuclear ADP-ribosylation reactions in mammalian cells: where are we today and where are we going? Microbiol Mol Biol Rev 70: 789-829.
9. Karras GI, Kustatscher G, Buhecha HR, Allen MD, Pugieux C, et al. (2005) The macro domain is an ADP-ribose binding module. EMBO J 24: 1911-1920.
10. Han W, Li X, Fu X (2011) The macro domain protein family: structure, functions, and their potential therapeutic implications. Mutat Res 727: 86-103.
11. Oshima T, Aiba H, Baba T, Fujita K, Hayashi K, et al. (1996) A 718-kb DNA sequence of the Escherichia coli K-12 genome corresponding to the 12.7-28.0 min region on the linkage map. DNA Res 3: 137-155.
12. Kawarabayasi Y, Sawada M, Horikawa H, Haikawa Y, Hino Y, et al. (1998) Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, Pyrococcus horikoshii OT3. DNA Res 5: 55-76.
13. Durkacz BW, Omidiji O, Gray DA, Shall S (1980) (ADP-ribose)n participates in DNA excision repair. Nature 283: 593-596.
14. Rouleau M, Aubin RA, Poirier GG (2004) Poly(ADP-ribosyl)ated chromatin domains: access granted. J Cell Sci 117: 815-825.
15. Kraus WL, Lis JT (2003) PARP goes transcription. Cell 113: 677-683.
16. Chang P, Jacobson MK, Mitchison TJ (2004) Poly(ADP-ribose) is required for spindle assembly and structure. Nature 432: 645-649.
17. Dynek JN, Smith S (2004) Resolution of sister telomere association is required for progression through mitosis. Science 304: 97-100.
18. Smith S, Giriat I, Schmitt A, de Lange T (1998) Tankyrase, a poly(ADP-ribose) polymerase at human telomeres. Science 282: 1484-1487.
19. Yu W, Ginjala V, Pant V, Chernukhin I, Whitehead J, et al. (2004) Poly(ADP-ribosyl)ation regulates CTCF-dependent chromatin insulation. Nat Genet 36: 1105-1110.
20. Chang W, Dynek JN, Smith S (2005) NuMA is a major acceptor of poly(ADP-ribosyl)ation by tankyrase 1 in mitosis. Biochem J 391: 177-184.
21. Chabert MG, Niedergang CP, Hog F, Partisani M, Mandel P (1992) Poly(ADPR) polymerase expression and activity during proliferation and differentiation of rat astrocyte and neuronal cultures. Biochim Biophys Acta 1136: 196-202.
22. Fontan-Lozano A, Suarez-Pereira I, Horrillo A, del-Pozo-Martin Y, Hmadcha A, et al. (2010) Histone H1 poly[ADP]-ribosylation regulates the chromatin alterations required for learning consolidation. J Neurosci 30: 13305-13313.
23. Marini M, Zunica G, Monti D, Cossarizza A, Ortolani C, et al. (1989) Inhibition of poly(ADP-ribosyl)ation does not prevent lymphocyte entry into the cell cycle. FEBS Lett 253: 146-150.
24. Shall S, Sugimura T (2006) What is new about ADP-ribosylation?. Bioessays 28: 97-99.
25. An NH, Han MK, Um C, Park BH, Park BJ, et al. (2001) Significance of ecto-cyclase activity of CD38 in insulin secretion of mouse pancreatic islet cells. Biochem Biophys Res Commun 282: 781-786.
26. Lodhi IJ, Clift RE, Omann GM, Sweeney JF, McMahon KK, et al. (2001) Inhibition of mono-ADP-ribosyltransferase activity during the execution phase of apoptosis prevents apoptotic body formation. Arch Biochem Biophys 387: 66-77.
27. Berger NA (1985) Poly(ADP-ribose) in the cellular response to DNA damage. Radiat Res 101: 4-15.
28. Gan X, Stegle O, Behr J, Steffen JG, Drewe P, et al. (2011) Multiple reference genomes and transcriptomes for Arabidopsis thaliana. Nature 477: 419-423.
29. Schmid M, Davison TS, Henz SR, Pape UJ, Demar M, et al. (2005) A gene expression map of Arabidopsis thaliana development. Nat Genet 37: 501-506.
30. Saito K, Tautz L, Mustelin T (2007) The lipid-binding SEC14 domain. Biochim Biophys Acta 1771: 719-726.