ISSN : 0974 - 7435 Volume 8 Issue 2

BioTechnology
An Indian Journal

FULL PAPER

BTAIJ, 8(2), 2013 [270-274]

ESAP: A new pipeline for genome wide identification of poly(A)-site-modifying SNPs in Arabidopsis

Lingyan Wang, Jinting Guan, Jingyi Fu, Guoli Ji, Xiaohui Wu*
Department of Automation, Xiamen University, Xiamen 361005, Fujian, (CHINA)
E-mail: xhuister@xmu.edu.cn

ABSTRACT

Single nucleotide polymorphisms (SNPs) can change polyadenylation signals. Poly(A) signal (PAS) plays an important role during polyadenylation process. Therefore, the selection of poly(A) sites may be affected by the appearance of SNPs. With the rapid development of the next-generation high-throughput DNA sequencing techniques, more and more SNPs are discovered, but not all the SNPs can lead to the change of the poly(A) sites. Here a pipeline named ESAP (extract SNPs associating with polyadenylation) was designed that can extract the SNPs affecting poly(A) sites. ESAP uses poly(A) prediction program PASS to compute the base prediction scores which are used to identify the difference of poly(A) sites because of the existence of SNPs. Finally, the SNPs are classified into "likely", "probable", "unlikely" according to their effect on poly(A) sites. 569,859 SNPs from Arabidopsis…Bur-0 and 40,026 poly(A) site clusters (PACs) from Arabidopsis Thaliana are analyzed. Total 160 SNPs that can affect PAS (PAS-SNPs) were found, including 84 "likely", 38 "probable".

KEYWORDS

SNPs;
Poly (A) site;
Classification;
Association analysis;
Arabidopsis.

INTRODUCTION

SNP is an abundant form of genome variation, distinguished from rare variation by a requirement for the least abundant allele to have a frequency of 1% or more[1]. SNPs may be functionally responsible for specific traits or phenotypes, or they may be informative for tracing the evolutionary history of a species or the pedigree of a variety[2]. The most significant function of SNPs is their strong relevance to diseases, such as a SNP in the APOE gene increases the risk for developing Alzhei-mer disease[3]. SNPs can create or disrupt polyadenylation signals which may cause alternative polyadenylation (APA)[4].

Many studies on SNPs according to their potential effects to human health have been designed[5]. An integrative scoring system for classifying SNPs contains a collection of previously evaluated SNPs which can be queried by SNPs id, disease or chromosomal region. These SNPs are analyzed and scored according to the location of the SNPs like splice site, ESE, TFBS, coding region and putative deleterious effects on human genes[5]. Here a pipeline named ESAP (extract SNPs associating with polyadenylation) was designed to identify SNPs which can make some change on poly(A) sites. Firstly, the SNPs that can affect PAS were iden-

tified. Secondly, PASS[6,7] is used to compute the prediction scores whose difference are an important evaluation of the classification. Finally, SNPs are classified into "likely", "probable", "unlikely".

MATERIALS AND METHODS

The datasets

The reliable published datasets of Arabidopsis SNPs from the 1001 Genomes (http://www.1001genomes.org/) which consists of 569,859 SNPs for ecotype Bur-0 and 501,399 SNPs for ecotype Tsu-1 were used[8]. Poly(A) sites of Arabidopsis were analyzed and discovered by mapping polyadenylated ESTs and full-length cDNA sequences to the reference genomes[9]. Because of poly (A) site microheterogeneity in plants, poly(A) sites that locate within 24 nt of each other in the same gene were clustered to a poly (A) site cluster (PAC). The total number of PACs we used is 40,026 in the correlation analysis of SNPs and poly(A) sites[10]. Moreover, the DNA reference sequences are downloaded from TAIR (http://arabidopsis.org/).

The dataflow of the pipeline ESAP

The pipeline was designed to analyze SNPs provided within the context of a DNA sequence. ESAP takes a file which specifies all available options as input. The command line version of ESAP is written in C++ and the classification is done by R. We first extracted the SNPs that can affect the poly(A) signals (PAS-SNPs) based on the existing PACs data. Second, the prediction scores were computed by the program PASS[6,7] whose input data are the sequences containing 301 nt upstream and 99 nt downstream of the PACs that can map to PAS-SNPs. Third, a rule system classified the SNPs into "likely", "probable", "unlikely" (Figure 1).

Extracting PAS-SNPs

The PAS is AAUAAA during our associate analysis of SNPs and poly(A) sites. Using the cleavage site (CS) as a reference point, the NUE region is located 10 to 40 nt upstream[11]. Therefore, the SNPs which are located 50 nt upstream or downstream from the PACs were extracted according to the strand of PACs. These SNPs were referred to as our candidate SNPs and name them as PAS-SNPs. Based on their effect on PAS, a classification of the PAS-SNPs into two categories is given: "delete" that changes the signal to another motif, "create" that changes no signal to AAUAAA in the locus.

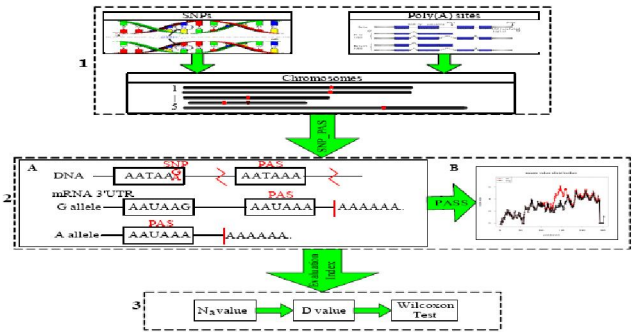

Figure 1 : Overview of the dataflow of the pipeline ESAP. (1) In the first step, SNPs and poly(A) sites are mapped to the DNA sequences. (2) In the second step, as shown in part A, for the G allele, the right cleavage site (CS) is used. For the A allele, the left CS is used, because the PAS including a SNP is functional. The consequence of the choice is that the transcription may be shorter. As shown in part B, we extracted the PAS-SNPs and computed prediction scores by PASS [6, 7]. (3) The last step is the classification of SNPs.

The procedure of extracting PAS-SNPs includes three sections. First, for a given PAC, the SNPs that locate in 50 nt upstream from the PAC are chosen when the strand of PAC is "+". On the contrary, the SNPs chosen should locate in 50 nt downstream from the PAC when the strand is "-". Second, the SNPs in potential poly(A) signal AATAAA (in DNA sequence) are detected by a motif search using a sliding window. The motif search looks a given motif in DNA sequence consisting of the SNP and its flanking sequences (5 nucleotides upstream and downstream) for each allele. The DNA sequence should be reversed and paired when the strand of the PAC that relates to the SNP is "-".

Comparing the prediction scores of the DNA sequences affected by SNPs

To investigate whether PAS-SNPs can have an impact on poly(A) sites, the prediction scores of sequences containing 301 nt upstream and 99 nt downstream of the PACs that were computed by PASS[6,7] are analyzed and the whole difference in scores between two alleles of one SNP (a group) were regarded as a criterion. In order to reflect this influence intuitively,

FULL PAPER

a D value were defined to represent it. The D value is computed as the following equation.

$$D = \frac{\sum_{i=0}^{N-1} (S_{1i} - S_{2i})^2}{N \cdot |E_{min}|} \quad (1)$$

In the equation, N, S_{1i}, S_{2i}, E_{min} are respectively the number of the scores compared in a group, the i-th of N prediction scores for reference alleles, the i-th of N prediction scores for non-reference alleles, the smaller one between the mean of N prediction scores for reference alleles and the mean of N prediction scores for non-reference alleles.

The prediction scores of nucleotides located in the region N/2 nt upstream and N/2 nt (N=23) downstream of the PACs are compared to indicate the influence of the SNPs on poly(A) sites. The S_i were referred to as the difference between the prediction scores of i-th nucleotide for reference alleles and non-reference alleles, and $S_{D\max}$ as the maximum value among the S_i values. Then, the S_{1i}, S_{2i} that correspond to the that is greater than 0.6 when is greater than 1, and the S_{1i}, S_{2i} that correspond to the S_i that is greater than $0.6 S_{D\max}$ when $S_{D\max}$ is greater than 1, and the S_{1i}, S_{2i} that correspond to the S_i that is greater than $0.8 S_{D\max}$ when $S_{D\max}$ is less than 1, are defined as a set of prediction scores varying significantly. The number of these prediction scores is represented by N_a.

Classification of PAS-SNPs

The process of classification of PAS-SNPs is divided into three steps. First, those whose N_a values are greater than 30 and less than 41 are defined as a set "candidate" and the others as "unlikely". Second, since more than half of the D values are greater than 0.1, those whose D values are greater than 0.1 are classified as "likely" and the others are classified as "probable" in the set of "candidate". Third, after correcting for rank-sum test (Wilcoxon's test) and FDR in "likely" and "probable", those whose p-values are greater than 0.05 are grouped as "unlikely" in the "likely" and "probable".

Therefore, the "likely" includes those PAS-SNPs whose N_a values are greater than 30 and less than 41 and D values are greater than 0.1, and the "probable" includes those PAS-SNPs whose N_a values are greater than 30 and less than 41 and D values are less than 0.1, and the p-values of SNPs in the two classes should be less than 0.05. The "unlikely" not only includes those PAS-SNPs whose N_a values are less than or equal to 30 and p-values are greater than or equal to 0.05, but also includes those SNPs that can't affect PAS.

RESULTS

The SNPs distribute unevenly around the PACs

The distribution of SNPs within different gene region as well as the region around PACs sites is fairly nonuniform which is correlated with the conservation of genome sequences[12,13]. In Arabidopsis, the SNPs rates around PACs sites in 3'UTR are also nonuniform and show that the number of SNPs that locate upstream of PACs are larger than that locate downstream of *PACs* (Figure 2.A). Furthermore, the numbers of PAS-SNPs located 50 nt upstream or downstream of PACs in different gene region are 149, 3, 8 in 3'UTR, CDS, intron respectively. The distances of PAS-SNPs and PACs also distribute unevenly which are generally 16 nt or 20 nt (Figure 2. B).

The impacts on PACs caused by SNPs in gene regions are different

There are 112 of 148 D values greater than 0.2 and the biggest one is 1.159 in the 3'UTR. However, the total number of D values greater than 0.2 is 6 and the biggest one is 0.277 in the intron. The D values in the CDS are generally larger than those in the two other gene regions which are all greater than 0.2. What's more, the biggest D value in the CDS is 2.076 that is almost two times of the biggest D value in the 3'UTR and ten times of the biggest one in the intron. Therefore, the influence of SNPs on PACs is most significant in the CDS. According to the distribution of D values that predict the potential effect that a SNP have on poly(A) site in different ranges, we found that there is not a sharp distinction between the influence of "create" SNPs and "delete" SNPs for PACs (Figure 3. A). Different SNPs have different impacts on PACs, thus their N_a also distributes asymmetrical that are usually located in the scope of 31 to 44 and there are 89 groups whose N_a are 33 (Figure 3. B).

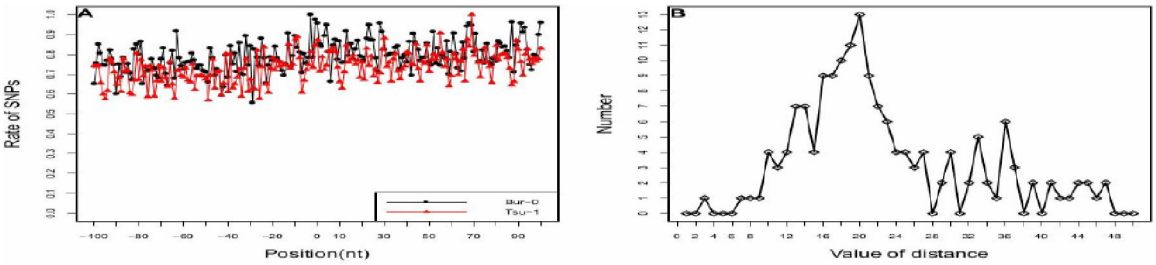

Figure 2 : Panel (A) shows the distribution of SNPs that locate in the region 100 nt upstream and 100 nt downstream of PACs of ecotype Bur-0 and Tsu-1 in Arabidopsis. There is slow upward trend from upstream to downstream. Panel (B) shows the distribution of distances between PAS-SNPs and related PACs, the horizontal axis represents the distance between SNPs and their related PACs and the vertical axis represents the number of the distances.

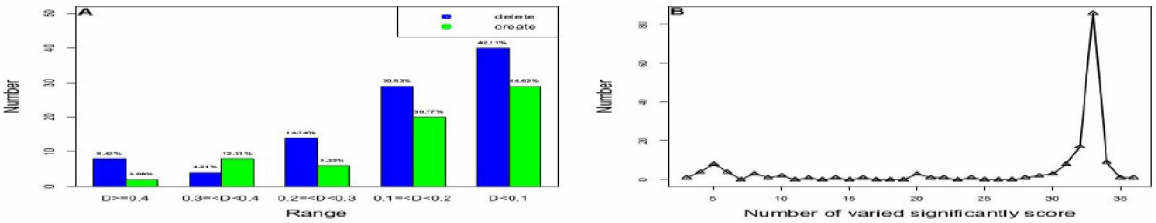

Figure 3 : The distribution of D values in class "delete" and "create" is shown by the Panel (A). The D values are mostly less than 0.2 and show little difference in two classes. The distribution of significantly affected score number is shown by the Panel (B) and there is a peak at the position of 33.

Extract SNPs in different gene regions with ESAP

The pipeline ESAP were used to extract SNPs affected poly(A) sites in different gene regions for the Bur-0. The result shows that 76, 36 SNPs were classified into "likely" and "probable" in the 3'UTR, while there are only 3 "likely ", 0 "probable" SNPs in the CDS and 5 "likely", 2"probable" SNPs in the intron (TABLE 1). The prediction scores of DNA sequences where the two alleles of one "likely" SNP located are shown in figure 4. Only 0.021% of 569,859 SNPs were classified as likely or probable to cause a change in poly(A) sites.

TABLE 1 : The distribution of different classes of SNPs in different gene regions.

Classes Regions	Likely	Probable	Unlikely	All
3'UTR	76	36		
Intron	5	2	569,737	569,859
CDS	3	0		

DISCUSSION

To identify SNPs which modify the poly(A) sites through changing PAS, the pipeline ESAP were designed. Unlike many other SNP analysis tools, our pipeline is aimed at analyzing the effect of SNPs on polyadenylation. It have shown with 569,859 SNPs and 40,026 PACs in Arabidopsis that the pipeline is able to divide SNPs into three classes on the basis of their effects on poly(A) sites. What's more, some conclusions are drawn during the analysis. The SNPs rate in the upstream of PACs is less than that in the downstream of PACs in Arabidopsis. The reason of this mainly owes to the genomic sequence conservation[13]. It found that the distances between PAS-SNPs and PACs are usually between 16 nt to 20 nt. The number of the prediction scores of DNA sequences where the two alleles of SNPs located varying significantly is 33.

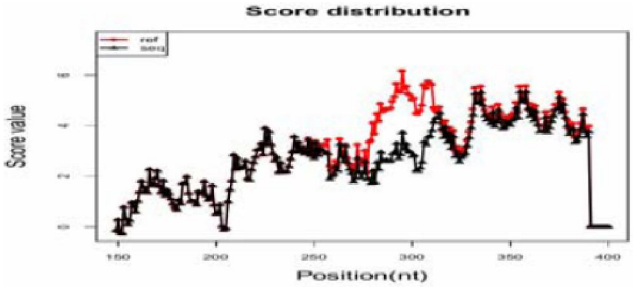

Figure 4 : The distribution of the prediction scores of DNA sequences where the two alleles of one "likely" SNP located. The red curve with small circles represents the scores of reference sequence and the black curve with small triangles represents the scores of non-reference sequence. The position of the PAC is 301 in the horizontal axis.

Not only the influence of SNPs for different poly(A) sites is different, but also the effect of SNPs on poly(A) site in different gene regions is different which is biggest in the CDS. But, SNPs which can create PAS show a same result as those that can delete PAS, which is seen from the difference of prediction scores.

The SNPs of the same haplotype together were not considered. Thus, ESAP treats all SNPs as independent. Only one SNP is taken into account even though the input sequences contain several SNPs. That's why that the combined effects if multiple SNPs occur are missed.

ACKNOWLEDGEMENTS

This project was funded by the National Natural Science Foundation of China (Nos. 61174161, 61304141, 61375077, 61201358 and 61203176), the Natural Science Foundation of Fujian Province of China (No. 2012J01154), the specialized Research Fund for the Doctoral Program of Higher Education of China (Nos. 20130121130004 and 20120121120038), the Key Research Project of Xiamen City of China (No. 3502Z20123014), and the Fundamental Research Funds for the Central Universities in China (Xiamen University: Nos. 2013121025, 201212G005 and CBX2013015).

REFERENCES

[1] A.J.Brookes; The essence of SNPs, Gene, **234**, 177-186, 7/8/ **(1999)**.

[2] S.R.McCouch, K.Zhao, M.Wright, C.W.Tung, K.Ebana, M.Thomson et al.; Development of genome-wide SNP assays for rice, Breeding Science, **60**, 524-535 **(2010)**.

[3] A.Chakravarti; Single nucleotide polymorphisms: ...to a future of genetic medicine, Nature, **409**, 02/ 15 **(2001)**.

[4] L.F.Thomas, P.Sætrom; Single Nucleotide Polymorphisms Can Create Alternative Polyadenylation Signals and Affect Gene Expression through Loss of MicroRNA-Regulation, PLoS Comput Biol, **8**, e1002621 **(2012)**.

[5] P.H.Lee, H.Shatkay; An integrative scoring system for ranking SNPs by their potential deleterious effects, Bioinformatics, **25**, 1048-1055, April 15, **(2009)**.

[6] G.Ji, J.Zheng, Y.Shen, X.Wu, R.Jiang, Y.Lin et al.; Predictive modeling of plant messenger RNA polyadenylation sites, BMC Bioinformatics, **8**, 43 **(2007)**.

[7] G.Ji, X.Wu, Y.Shen, J.Huang, Q.Quinn Li; A classification-based prediction model of messenger RNA polyadenylation sites, J Theor Biol, **265**, 287-96, Aug 7 **(2010)**.

[8] S.Ossowski, K.Schneeberger, R.M.Clark, C.Lanz, N.Warthmann, D.Weigel; Sequencing of natural strains of Arabidopsis thaliana with short reads, Genome Res, **18**, 2024-33, Dec **(2008)**.

[9] E.S.Ho, S.I.Gunderson, S.Duffy; A multispecies polyadenylation site model, BMC Bioinformatics, **14(2)**, S9, **(2013)**.

[10] X.Wu, M.Liu, B.Downie, C.Liang, G.Ji, Q.Q.Li et al.; Genome-wide landscape of polyadenylation in Arabidopsis provides evidence for extensive alternative polyadenylation, Proc Natl Acad Sci U S A, **108**, 12533-8, Jul 26 **(2011)**.

[11] J.C.Loke, E.A.Stahlberg, D.G.Strenski, B.J.Haas, P.C.Wood, Q.Q.Li; Compilation of mRNA polyadenylation signals in Arabidopsis revealed a new signal element and potential secondary structures, Plant Physiol, **138**, 1457-68, Jul **(2005)**.

[12] X.J.Mu, Z.J.Lu, Y.Kong, H.Y.Lam, M.B.Gerstein; Analysis of genomic variation in non-coding elements using population-scale sequencing data from the 1000 Genomes Project, Nucleic Acids Res, **39**, 7058-76, Sep 1 **(2011)**.

[13] J.C.Castle; SNPs Occur in Regions with Less Genomic Sequence Conservation, PLoS ONE, **6**, e20660 **(2011)**.