An efficient computational method for screening functional SNPs in plants

Ardashir Kharabian*

Centre for Plant Conservation Genetics, Southern Cross University, PO Box 157, Lismore Campus, NSW 2480, Australia

ARTICLE INFO

ABSTRACT

Granule Bound Starch Synthase I (GBSSI), which influences the grain quality of cereals and, particularly, rice, is one of the most important plant genes. Using GBSSI as a model plant gene in this study, we examined a number of different computational algorithm tools and programs to explore the functional SNPs of this important rice gene and the possible relationships between genetic mutation and phenotypic variation. A total of 51 SNPs/indels were retrieved from databases, including three important coding non-synonymous SNPs, namely those in exons 6, 9 and 10. Sorting Intolerant from Tolerant (SIFT) results showed that a candidate [C/A] SNP (ID: OryzaSNP2) in exon 6 (coordinate 2494) is the most important non-synonymous SNP with the highest phenotypic impact on waxy protein. This SNP can alter a tyrosine to serine residue at position 224 of waxy protein. Computational simulation of GBSSI protein with the Geno3D suggested that this mutant SNP creates a bigger loop on the surface of waxy protein and results in a shape different from that of native GBSSI. Here, we suggest a potential transcriptional binding factor site (TBF8) which has one [C/T] SNP [rs53176842] at coordinate 2777 in boundary site of intron 7/exon 8, according to Transcriptional Factor (TF) search analysis. This SNP might potentially have a major effect on regulation and function of GBSSI.

Combining SNP mining data and in silico structural analysis of waxy protein led us to prepare an efficient computational pathway which can be applied for other plant genes.

1. Introduction

Single nucleotide polymorphisms (SNPs) are the most common and simplest type of genetic variation in organisms. SNPs occur at a frequency of approximately one in a thousand base pairs in the human genome (Brookes, 1999) and one in every 170 bp in rice (Yu et al., 2002). SNPs can be found everywhere throughout the genome, such as the promoter region, coding and intronic sequences, but most of them are probably located in intergenic regions which are believed to be stable without any deleterious effect on organisms. In fact, the occurrence of human disease and evolution (Shastry, 2002), as well as many important traits in plants (Bryan et al., 2000; Kennedy et al., 2006; Edwards et al., 2007), can be attributed to the presence of SNPs and their variations.

SNPs can be categorized and named based on their location and function. For example, SNPs within the coding regions (cSNPs) of functional genes which introduce biological variations directly into the amino acids are called non-synonymous SNPs (ns SNPs) and are of major interest. Those SNPs which occur in the coding sequences, but do not change the amino acids, are called synonymous SNPs. However, most SNPs occur in the intronic regions. Study of these SNPs is also important because of their influence on gene expression which can be occurred through different molecular pathways such as changing regulatory elements, splicing patterns, up- and downregulation of exonic splice enhancers (ESE), intronic splice enhancers (ISE) and so forth. (ElSharawy et al., 2006).

Understanding the functional effect of SNPs is a major challenge. SNPs that lead to a single amino acid substitution, stop codon or frame shift mutation are normally recognized as functional and can be detected by regular experimental assays. It is thought that an experimental-based approach can provide the strongest evidence for the functional role of genetic variations. Consequently, many different types of SNP assays have been applied for experimental prioritization of SNPs (Chen and Sullivan, 2003). However, owing to the lack of reliable genotype and phenotypic data, these experiments are not always easy to set up for characterizing the real effect of SNPs. For example, functional analysis of SNPs in important plant genes needs a perfect segregating population or breeding lines, such as near iso-genic lines (NILs) (Umemoto et al., 2008; Mikami et al., 2008). On the other hand, many genes may have a vast number of intronic SNPs that cannot be easily associated with in vivo variation of plant populations.

Previous studies have focused on non-synonymous SNPs of human disease genes (George Priya Doss et al., 2008; Rajasekaran

et al., 2008), and among a huge number of available computational algorithms designed for human genetics, only a few of them is appropriate for plants studies.

In this research, we have developed a plant-relevant computational pipeline which covers most of important functional elements at DNA level, rather than non-synonymous SNPs, to characterize and prioritize substitutions in one of the most fully documented rice genes, Granule Bound Starch Synthase I (GBSSI), a major gene affecting the amylose content of rice grain (Chen et al., 2004). Different computational algorithm tools, such as Sorting Intolerant from Tolerant (SIFT), Exonic Splicing Enhancer Finder (ESE Finder), Transcriptional Factor search (TF search) and Exonic Splicing Silencer Search (FAS-ESS), were used to prioritize the candidate SNPs most likely to affect the encoded protein and, subsequently, amylose content and rice grain quality.

2. Materials and methods

2.1. GBSSI gene as a case study

We analysed the Granule Bound Starch Synthase (GBSSI) as an example gene to provide a guide for further confirmatory experimental studies for this or other important genes. This gene is well known in most cereals by its major affect on amylose content (Chen et al., 2008a), pasting properties (Chen et al., 2008b) and eating quality (Umemoto et al., 2008).

2.2. Sequence alignment

GBSSI coding DNA and mRNA sequences for *Oryza sativa* L. were downloaded from GenBank (http://www.ncbi.nlm.nih.gov/nuccore/297423). (GenBank locus number for genomic DNA is X65183.1 and NM_001063239.1 for mRNA.).

Nucleotide coordinates of 1765922–1769401 on chromosome 6 (LOC_Os06g04200) were extracted from the Rice Genome Annotation Project at Michigan State University (http://rice.plantbiology.msu.edu/LocusNameSearch.shtml).

The total sequence lengths of 5035 bp, 1830 bp and 609 amino acids were recognized in genomic, cDNA and GBSSI protein, respectively.

2.3. SNP dataset

SNP dataset for the GBSSI gene was retrieved from the NCBI database (Sherry et al., 2001) at 〈http://www.ncbi.nlm.nih.gov/sites/entrez?db=snp&TabCmd=Limits〉 for the relevant chromosome range (gene coordinates) and then checked with the SNP dataset in Oryza SNP Consortium 〈http://www.oryzasnp.org/cgi-bin/find_snps_in_genes.pl〉 using the following TIGR gene ID: LOC_Os06g04200. The extra DNA length of 2 Kbp from each end of the coding region was also searched for the possible existence of SNPs in the 3′ and 5′ UTRs. Final alignment was carried out by Sequencer 4.6 software (Ann Arbor, MI) and ClustalW2 (http://www.ebi.ac.uk/Tools/clustalw2/index.html) to identify exact location of SNPs in UTR or intronic/exonic regions.

Five different functional classes of SNPs were selected to cover the entire gene region, as follows: (1) non-synonymous coding, (2) intronic, (3) coding synonymous, (4) locus region and (5) 5′ and 3′ UTRs.

2.4. Computational tools for SNP analysis

Several computational software programs were applied to predict the actual or possible impact of SNPs on plant phenotypes,

as follows: (1) UTRscan, (2) TF search, (3) SIFT: Sorting Intolerant from Tolerant, (4) GeneSplicer, (5) SEE-ESE, (6) FAS-ESS, (7) Geno3D, (8) PDB viewer and (9) RasMol.

2.5. 3D Modelling of GBSSI and comparative study

We modelled the native and mutant structure of GBSSI protein by Geno3D software available at: 〈http://geno3d-pbil.ibcp.fr/cgi-bin/geno3d_automat.pl?page=/GENO3D/geno3d_home.html〉.

This is a highly efficient program to predict 3D structures of proteins and enzymes based on amino acid sequence of genes. This program is capable of extracting 3D structures of very similar proteins from different databases (specifically, PDB) and then modelling the query sequence using available structure, which, for the GBSSI gene, has the PDB identification number of 3D1J. The modelled structure can be validated by PROCHECK (Laskowski et al., 1993). The comparative studies were also carried out by Swiss PDB viewer (http://spdbv.vital-it.ch/download.html) and RasMol (http://openrasmol.org) based on superimposed structure and homology analysis of native and mutant protein (Rajesh et al., 2008).

2.6. Functional flow chart

A flow chart was prepared for computational analysis and prioritization of SNPs based on their functionality and possible impact on plant phenotypes Fig. 2.

3. Results

3.1. SNPs in GBSSI gene and comparative study

A total of 51 SNPs, including in/dels, were extracted from databases consisting of the following: one in the 5′ UTR, three in coding non-synonymous, seven in coding synonymous and 40 in the intronic sequences (Table 1) (Fig. 1).

3.2. Computational algorithm tools

The following computational tools were used consecutively for comprehensive functional analysis of the GBSSI gene:

3.2.1. UTR scan

The UTR scan server identifies patterns of regulatory region motifs from the UTR database and gives information about important elements in the 5′ and 3′ UTRs, (Pesole and Liuni, 1999). This program can be accessed at: 〈http://www.ba.itb.cnr.it/BIG/UTRScan/〉.

One regulatory element was found at the 3′ UTR end of the gene. No critical element was recognized by UTR scan in the 5′ UTR region of the GBSSI gene. One ′A/C′ SNP [rs20225948] was found in the non-regulatory element of 5′ UTR. Since the number of SNPs in the UTR regions of the GBSSI gene was limited to only one, with none in the regulatory regions, it may be presumed that GBSSI UTRs do not change the expression level of protein.

3.2.2. TF search

Two of the most important functional elements in plant genomes are transcriptional factors (TFs) and transcriptional factor binding sites (TFBSs). These sites are usually short DNA sequences, around 5–15 bp, where the TF elements bind to them to begin transcriptional process involving RNA polymerase and the promoter. Occurrence of any mutation in these regions can alter motifs and possibly transcriptional patterns (Bulyk, 2004).