ABSTRACT

Summary: Marker2sequence (M2S) aims at mining quantitative trait loci (QTLs) for candidate genes. For each gene, within the QTL region, M2S uses data integration technology to integrate putative gene function with associated gene ontology terms, proteins, pathways and literature. As a typical QTL region easily contains several hundreds of genes, this gene list can then be further filtered using a keyword-based query on the aggregated annotations. M2S will help breeders to identify potential candidate genes for their traits of interest.

Availability: Marker2sequence is freely accessible at http://www.plantbreeding.wur.nl/BreeDB/marker2seq/. The source code can be obtained at https://github.com/PBR/Marker2Sequence.

Contact: richard.finkers@wur.nl

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on January 23, 2012; revised on April 13, 2012; accepted on May 15, 2012

1 INTRODUCTION

Quantitative trait loci (QTLs) are regions on the genome statistically associated with a phenotype. Plant breeders aim to introgress these regions from a donor parent to improve a cultivar (Zamir, 2001). However, a typical QTL region may contain hundreds of genes, including genes negatively influencing the breeding goals. Complete genome sequences of many crop plants are becoming available, including the genome of important food crops such as tomato (the International Tomato Genome Sequencing Consortium, 2012) and potato (the Potato Genome Sequencing Consortium, 2011). The availability of structural and functional genome annotations makes it possible to investigate the QTL region for genes positively or negatively influencing the trait of interest. A tool that offers this functionality is GBrowse (Stein et al., 2002). However, exploring several hundreds of putative candidate genes with GBrowse will still be a lot of work, as limited information is available for each gene.

Identification of putative candidate genes can be improved using data integration approaches. Biologically relevant knowledge about, for example, gene ontology (GO; The Gene Ontology Consortium, 2000), protein functions or metabolic pathways can be combined with expert knowledge about the trait under investigation. The basic principle of Semantic Web technology is to integrate different types of data from different sources using standardized ontologies (Berners-Lee et al., 2001). Important resources such as UniProt (Redaschi and UniProt Consortium, 2009) and GO have become available in a Resource Description Framework (RDF; http://www.w3.org/TR/rdf-primer/) format allowing data integration against these resources.

Our research aims to develop a tool, called Marker2sequence (M2S), which plant breeders can use to identify the putative candidate gene for their QTL. We describe how M2S uses semantic data integration approaches to obtain available information for each gene model and combine this with a keyword-based search function to mine for the appropriate candidate gene. M2S, by design, can work with any genome annotation, but we show the functionality of M2S, using the tomato genome annotation.

2 DESIGN AND IMPLEMENTATION

M2S is a web-based tool using Java EE 6 and the Struts framework (v1.3.10). It runs on a Glassfish (v3.1) application server. A utility, gff2RDF, has been developed to perform the conversion of the tomato, potato and Arabidopsis genome annotation and linkage maps to the RDF format. It is available at https://github.com/PBR/gff2RDF. This utility extracts for all genes their location, human readable description, associated GO term identifier and associated UniProt protein identifiers (Jain et al., 2009) from the annotation. For Tomato, the EXPEN 2000 map was used as reference linkage map (Fulton et al., 2002). The Jena library (Carroll et al., 2004) is used to build the RDF model and write it to disk. These graphs were loaded into a Virtuoso open-source edition (version 6.1.3) (Erling and Mikhailov, 2007) triple store together with the GO (version 2011_11_03) and UniProt (version 2011_10). Any triple store with a SPARQL endpoint can be used with M2S.

M2S can handle three types of inputs (Supplementary Fig. S1). The first two inputs are two markers flanking the QTL region or a list of markers spanning the QTL region. These markers should have a physical position on the genome sequence or a position on the reference linkage map. The third input requires a genomic region using the format Chr:start..stop. Either input leads to a summary page divided into three sections; the top section shows the alignment of the reference genetic map with the genomic information, which helps to identify problems in the genetic map or the assembly of the genome. The lower section consists of three tabs. The first tab contains a list of all genes in the specified
The gene Solyc06g074240.1.1 is associated with the GO term 'carotenoid biosynthetic process', the pathway 'carotenoid biosynthesis', and more specifically, it is part of 'beta-carotene beta-hydroxylase' and the pathway for 'carotenoid biosynthesis'. Information for each gene can be quickly mined using M2S and both genes are candidates for our trait of interest.

3 EXAMPLE

β-Carotene content is a trait influencing the color of tomatoes (Lincoln and Porter, 1949). Based on our QTL analysis, using data from the Solanum lycopersicum × Solanum galapagense LA0483 RIL population (Paran et al., 1995), this compound has QTLs on chromosomes 3 (between TG130 and TG74) and 6 (between TG253 and TG314). M2S identified 2003 genes on chromosome 3 and 988 genes on chromosome 6. A query with the keyword ‘beta-carotene’ returns the gene Solyc03g007960.1.1 on chromosome 3 and the gene Solyc06g074240.1.1 on chromosome 6. The gene Solyc03g007960.1.1 has the description ‘carotene beta-hydroxylase’ and is associated with the GO term ‘carotenoid biosynthetic process’, the pathway ‘carotenoid biosynthesis’ and more specifically, it is part of ‘beta-carotene biosynthesis’. Information for each gene can be quickly mined using M2S and both genes are candidates for our trait of interest.

4 CONCLUSIONS

M2S provide plant breeders with a way to obtain all annotated gene models in a QTL region, to query, over multiple databases, within the QTL region of interest and an extensive summary for each gene model. M2S will help breeders to identify potential candidate genes for their traits of interest.

ACKNOWLEDGEMENTS

The authors acknowledge the International Tomato Genome Consortium (http://solgenomics.net/tomato/) for their work on the tomato genome sequence and genome annotation.

Funding: This project was co-financed by the sixth framework EU project ‘High Quality Solanaceous crops for consumers, processors and producers by exploration of Biodiversity’, contract number: FOOD-CT-2006-016214, Wageningen UR Plant Breeding and the Centre for BioSystems Genomics (CBSG) which is part of the Netherlands Genomics Initiative/Netherlands Organization for Scientific Research’.

Conflict of Interest: none declared.

REFERENCES


Jain, E. et al. (2009) Infrastructure for the life sciences: design and implementation of the UniProt website. BMC Bioinformatics, 10, 136.


