Data and text mining

MSQT for choosing SNP assays from multiple DNA alignments

Norman Warthmann, Joffrey Fitz and Detlef Weigel*
Max Planck Institute for Developmental Biology, Department of Molecular Biology, 72076 Tübingen, Germany

Received on June 26, 2007; revised on July 27, 2007; accepted on August 16, 2007
Advance Access publication September 4, 2007
Associate Editor: John Quackenbush

ABSTRACT

Motivation: One challenging aspect of genotyping and association mapping projects is often the identification of markers that are informative between groups of individuals and to convert these into genotyping assays.

Results: The Multiple SNP Query Tool (MSQT) extracts SNP information from multiple sequence alignments, stores it in a database, provides a web interface to query the database and outputs SNP information in a format directly applicable for SNP-assay design. MSQT was applied to Arabidopsis thaliana sequence data to develop SNP genotyping assays that distinguish a recurrent parent (Col-0) from five other strains. SNPs with intermediate allele frequencies were also identified and developed into markers suitable for efficient genetic mapping among random pairs of wild strains.

Availability: The source code for MSQT is available at http://msqt.weigelworld.org, together with an online instance of MSQT containing data on 1214 sequenced fragments from 96 ecotypes (wild inbred strains) of the reference plant A.thaliana. All SNP genotyping assays are available in several formats for broad community use.

Contact: weigel@weigelworld.org

Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

A major focus of current genetic research is the mapping and characterization of loci that underlie phenotypic variation between individuals, including association studies that link phenotypic to genotypic variation. Single nucleotide polymorphisms (SNPs) are the genotyping markers of choice. SNPs are readily detected with various methods and well suited for medium- to high-throughput genotyping (Engle et al., 2006; Syvanen, 2001). Nonetheless, identification of informative markers is often a challenge, especially when applying medium-throughput methods that incur a cost for every attempted genotyping assay. A starting point for identifying such markers is typically the comparison of DNA sequences obtained by PCR-based re-sequencing of whole-genome shotgun sequences. The ‘Multiple SNP Query Tool’ (MSQT) was designed to facilitate marker detection, selection, storage and especially assay design from sequence alignments.

SNP genotyping technologies typically require sequence information surrounding a targeted SNP for placement of primers and/or probes used in the genotyping assay. Sequence variation in flanking sequence can be a major cause of genotyping error and assay failure (Pompanon et al., 2005). MSQT therefore provides an output that includes polymorphism information for flanking sequences, to improve the success rate of SNP assay development. We describe the application of MSQT to genotyping tasks in Arabidopsis thaliana and provide the genotyping assays and results.

2 SYSTEM AND METHODS

2.1 Features

MSQT is a client-server application and its architecture is depicted in Figure 1. The data is administered and queried via a web interface. All data are stored in a relational database management system (RDBMS), retrieved and processed by the ‘distinct SNPs’ module and delivered to the web applications in XML format.

Dataset administration, including data upload, quality control, SNP detection and database population, is performed with the MSQT/Admin front-end, which can be password protected. MSQT/SBE (SNPs Between Ecotypes) quickly identifies SNPs between any groups of user-specified individuals. The user can inspect all original sequence alignments with a built-in alignment viewer.

With the MSQT/Admin module ‘Snipe4SNPs’ each SNP in the dataset is automatically analyzed and SNPs can later be selected in batch according to allele frequency, position, minimum distance to neighboring SNPs, etc. using the MSQT/SNIPED! interface. With MSQT/E, the Expert mode, all database tables can be queried directly using Structured Query Language (SQL).

For SNPs selected by MSQT/SBE, MSQT/SNIPED! or MSQT/E, MSQT/ADF computes and returns the SNP and surrounding sequence in Assay-Development-Format. ADF is a sequence representation that can be used directly to design SNP-detection assays: the alignment- and SNP-information between groups of individuals is represented in a single line of text, facilitating the placement of primers and probes. An example is given in Figure 2.

2.2 Availability and requirements

A PostgreSQL database as back-end is linked with a CGI-driven web application written in Perl, deployed on a standard web server. Additional dependencies are only Perl modules available at CPAN (http://www.cpan.org). An online installation of MSQT containing data for 1214 fragments from 96 ecotypes/accessions (wild inbred strains) of the reference plant A.thaliana (Clark et al., 2007; Nordborg et al., 2005) and the source code for the entire software-suite with user- and administrator-documentation is available at

*To whom correspondence should be addressed.
The 'sniped' table. Each entry in the 'sniped' table then contains SNP by a specified distance (default is one base) is recorded in of individuals as input. Every SNP that is separated from a neighboring polymorphic position the two most common, unambiguous alleles (target) at all loci and 'snps' contains allele information for all individuals at positions where at least one individual has a polymorphism compared to the reference sequence. The fourth table, 'sniped' is created by parsing the 'snps and 'target_sequences' tables: at each SNP position the two most common, unambiguous alleles are identified and the individuals are placed in two separate databases. In MSQT/SBE (SNPs Between Ecotypes), all individuals in the database will be presented to the user in each of three select boxes (groups). Upon selection of individuals, the program will return a list of all distinct SNPs between individuals in groups 1 and 2 in the requested position range. For each SNP, links to a multiple alignment view of the original file and the ADF output are provided. The user can thus instantly review any SNP in the original data and extract the information necessary for genotyping assay development.

Main outputs are lists of suitable SNPs and the Assay-Development-Format, ADF, an example of which is given in Figure 2. In the context of the target sequence, alleles of the primary SNP are inserted in square brackets and all other polymorphisms found between and within selected groups of individuals are annotated in curly brackets. This one line of text therefore incorporates all (known) information necessary for assay design.

**2.3 Data input and output**

Input is typically in form of alignments of DNA sequences from multiple individuals at multiple loci in FASTA format. Each file must contain one alignment, with one sequence defined as the reference sequence (‘target’). SNP data may, however, also be loaded into the database by other means as long as the data and table specifications are met.

Main outputs are lists of suitable SNPs and the Assay-Development-Format, ADF, an example of which is given in Figure 2. In the context of the target sequence, alleles of the primary SNP are inserted in square brackets and all other polymorphisms found between and within selected groups of individuals are annotated in curly brackets. This one line of text therefore incorporates all (known) information necessary for assay design.

**2.4 Parsing and data storage**

MSQT parses multiple sequence alignments and compares the sequence of each individual to a previously defined reference sequence (‘target’). At any given position, an individual has either the same or a different base than the target. For every position where one individual differs from the target, the DNA base information for all individuals is stored in the database. MSQT can handle completely sequenced genomes, where a SNP will be identified by its absolute position in the genome (chromosome, position), as well as species with limited genome sequence information, in which case unique SNP identifiers are composed of the name of the locus and the position within the locus (position-name).

An MSQT instance can host many datasets. Each dataset is stored in a separate database schema with four tables. Table ‘ecotypes’ stores all distinct individuals belonging to one dataset; ‘target_sequences’ contains the complete nucleotide sequence of the reference sequence (target) at all loci and ‘snps’ contains allele information for all individuals at positions where at least one individual has a polymorphism compared to the reference sequence. The fourth table, ‘sniped’ is created by parsing the ‘snps and ‘target_sequences’ tables: at each polymorphic position the two most common, unambiguous alleles (A, T, C, G, –) are identified and the individuals are placed in two groups according to the specific allele they have at this position. The ADF output is then created taking the position and the two groups of individuals as input. Every SNP that is separated from a neighboring SNP by a specified distance (default is one base) is recorded in the ‘sniped’ table. Each entry in the ‘sniped’ table then contains the SNP position, the two main alleles and their frequencies, the list of individuals for each allele, the ADF output and information about the distances to neighboring polymorphisms.

**2.5 Usage**

The MSQT web application offers several ways of interacting with the databases. In MSQT/SBE (SNPs Between Ecotypes), all individuals in the database will be presented to the user in each of three select boxes (groups). Upon selection of individuals, the program will return a list of all distinct SNPs between individuals in groups 1 and 2 in the requested position range. For each SNP, links to a multiple alignment view of the original file and the ADF output are provided. The user can thus instantly review any SNP in the original data and extract the information necessary for genotyping assay development. Individuals in group 3 (optional) are not considered in the SNP selection process, but all sequence changes additionally found in these individuals will also be annotated in the ADF output in curly brackets.

The MSQT/SNIPED program is a selection interface for the pre-computed ‘snipped’ table. The user can choose the dataset to be used and set limits for the sum and the difference of allele frequencies of the SNPs to be returned. Since at any position only the two most frequent unambiguous alleles were considered, the ‘allele frequency sum’ is a measure for the number of individuals in which the marker is informative, and the ‘allele frequency difference’ is a measure for relative group sizes. For example, if a dataset consists of 100 individuals, and at a given position 40 individuals are ‘A’, 35 individuals are ‘T’, 4 are heterozygous (‘W’) and 21 have missing data, the allele frequency sum for this SNP will be (40+35)/100=0.75 and the allele frequency difference (40−35)/100=0.05. A large allele frequency sum indicates that a SNP assay will be informative in many individuals, and a small allele frequency difference indicates that the two alleles are of similar, intermediate frequency. In addition, the user can specify how many bases upstream (5) AND/OR downstream (3) of the SNP of interest must not contain any sequence change. This requirement will differ for different genotyping technologies.

The MSQT/E (Expert-Mode) provides a text field where users can directly query the database with SQL statements. A detailed description of the database- and table-structures together with usage examples can be found in the documentation and online at http://msqt.weigelworld.org.
3 IMPLEMENTATION

We have used MSQT to mine publicly available sequence data from 96 Arabidopsis strains (the ‘2010’ dataset, Clark et al., 2007; Nordborg et al., 2005). This dataset consists of 1214 dideoxy-sequenced PCR fragments spaced throughout the genome from 96 individuals. The alignment of sequences from 96 individuals comprises ~690 000 bp, or 0.5% of the Arabidopsis reference genome, and are in FASTA format. On average 611 554 bases were available per individual, with an average of 2611 SNPs and an additional 2090 bases covered by insertion/deletions compared to the reference sequence, Col-0. We loaded this dataset into MSQT and extracted candidate SNPs for three sets of markers suitable for the MassARRAY® system, which combines proprietary primer extension chemical reactions with MALDI-TOF mass spectrometry analysis (Ragoussis et al., 2006; Tang et al., 1999).

The first marker set of 289 SNPs that we selected is enriched for SNPs for which the standard laboratory strain Columbia (Col-0) has the rare allele. The set is optimized for distinguishing alleles in the strains Est-1, Kin-0, Mr-0, Nd-1 and Van-0 from Col-0, with each SNP being informative in at least four comparisons. A task like this requires six sequential queries using MSQT/SBE: all five strains and all five possible combinations of four of these five strains queried in ‘ecotype group one’ versus Columbia (Col-0) in ‘ecotype group two’, followed by manual removal of redundant SNPs. Genaissance Pharmaceuticals (now Cogenics, Inc.) designed SNP detection assays for the MassARRAY® system based on the MSQT/ADF output. The distribution of the 289 resulting markers along the Arabidopsis chromosomes is shown in Supplementary Figure 1 and the assay details are given in Supplementary Table 1. All assays where tested by genotyping the reference accession Col-0 and an additional 80 accessions, including Est-1, Kin-0, Mr-0, Nd-1 and Van-0. The complete list of strains and genotyping results are available as Supplementary Tables 6 and 2. Evaluation of the results indicated that the majority of markers (on average, 196, disregarding heterozygous calls and failed genotyping reactions) would be informative in a random cross of Col-0 with another strain, even though the markers were initially only selected for crosses of Col-0 to five specific strains, for which an average of 213 markers were informative (Fig. 3B). An exception was M7323S (accession number: N6184), which differed from Col-0 at only 78 markers, reflecting the fact that Arabidopsis is not without population structure (Fig. 3B). As expected, markers are less informative for crosses that do not involve Col-0 or M7323S. Population structure (Fig. 3B). As expected, markers are less informative for crosses that do not involve Col-0 or M7323S. In most random pairwise comparisons, between 50 and 100 assays were informative (Fig. 3A).

Our next goal was to create a marker set suitable for mapping in any cross, and for fingerprinting or genotyping of any accession. Using MSQT/SNIPED! we selected SNPs from the ‘2010’ dataset limiting our choice to SNPs with intermediate allele frequency and sufficient distance to neighboring polymorphisms by setting ‘frequency sum’ >0.6 AND ‘frequency difference’ <0.5 AND ‘left neighborhood length’ >30 bp OR ‘right neighborhood length’ >30 bp. We identified nearly 1000 SNPs that satisfied these criteria. Because of linkage disequilibrium, most SNPs within the same sequence fragment contain redundant information for mapping purposes. We therefore chose one SNP per fragment based on optimal (i.e. as close to 0.5 as possible) allele frequency and completeness of the data. Of 683 remaining candidate SNPs, we selected the 258 that most evenly covered the genome. To further improve the coverage of the genome, we included 71 additional SNPs that were not parsed from this dataset, but had been identified in a separate project (Clark et al., 2007). Of 329 assays that were tested on DNA from 60 individuals, some of which had previously been phenotypically or genotypically characterized (Clark et al., 2007; Lempe et al., 2005; Schmuths et al., 2004), 311 assays yielded genotyping results in more than 80% of individuals. The average distance between markers is 384 kb, with only 11 neighboring markers further apart than 1 Mb (Supplementary Figure 2 and Supplementary Table 3). The average distance from the chromosome ends is 162 kb, and from the annotated centromere boundaries 352 kb.

Of the 60 individuals genotyped at the 311 markers, 41 individuals comprised a worldwide sample and 19 individuals were originally collected from Central Asia (Schmuths et al., 2004). All genotyping results are provided in Supplementary Table 4. On average, 138 markers were informative between individuals in the worldwide sample (Fig. 3C, light gray), including pairwise comparisons to the Col-0 reference (Fig. 3D). Between the individuals from Central Asia, previously shown to be closely related (Schmuths et al., 2004), on average 49 markers were informative (Fig. 3C, dark gray). To suit the needs of particular crosses and avoid genotyping with uninformative markers, these 311 genotyping assays are designed such that customized subsets can be used for genotyping.

To improve the cost effectiveness, we submitted 415 candidate SNPs for incorporation into Sequenom iPLEX.

![Fig. 3. Informative marker pairs based on actual genotyping results.](image-url)
Gold assays (Sequenom, 2006). From the 415 potential assays, 335 SNP assays were grouped into 11 assay pools (W1–W11) with a maximum of 38 reactions in pool W1. The assay details and the marker distribution across the genome are provided in Supplementary Table 5 and Supplementary Figure 3. Four of these SNP assay pools (W1–W4) are currently being used to genotype large cohorts of stock center accessions as well as wild collections of *A. thaliana* (J.O. Borevitz, personal communication). All assays are available at Sequenom, Inc., for community use.

4 DISCUSSION

We designed the Multiple SNP Query Tool (MSQT) to parse SNPs from sequence alignments and to store and select SNPs based on a variety of criteria. Primary criteria are allelic differences between groups of individuals, allele frequencies and distance to neighboring SNPs within a dataset. The input format, FASTA, is flexible, and can represent alignments of sequences generated with traditional dideoxy-sequencing, next generation sequencing-by-synthesis (Seo et al., 2005), or array-based re-sequencing technologies. The data presentation is separated from the computation, which returns the results in XML format. This should enable integration into third party software (e.g. Web Services, command line clients). MSQT offers an output format directly suitable for genotyping assay design. Since the quality of the output directly depends on the quality of the input data, MSQT provides the opportunity to quickly inspect the original alignments in a built-in viewer in order to identify potential inconsistencies/misalignments.

We used MSQT to parse and analyze a publicly available dataset of *A. thaliana* sequences (Nordborg et al., 2005) and to extract target SNPs in a format suitable for direct submission to commercial genotyping companies. Three sets of SNP genotyping assays were designed for community use. The first set of 289 assays is ideal for mapping crosses where the second set consists of 311 assays for SNPs with intermediate allele frequency markers are useful for genetic mapping between many pairs of *A. thaliana* strains and for fingerprinting wild strains, e.g. when assessing genetic diversity present in local populations or ascertaining the identity of strains obtained from stock centers. All assays as well as all genotyping data are publicly available for community use.

ACKNOWLEDGEMENTS

We are thankful to Stephan Ossowski and Richard M. Clark for providing code, Richard M. Clark for sharing unpublished results, discussion and comments that improved the manuscript, Justin O. Borevitz for suggesting the implementation of an important feature, Sureshkumar Balasubramanian for selecting many of the *A. thaliana* strains and providing DNAs, and Richard M. Clark, Sureshkumar Balasubramanian, Janne Lempe, Heike Schmuths, Christopher Schwartz and Yasushi Kobayashi for seeds. This work was supported by BMBF funded ERA-PG ARABRAS project and the Max Planck Society, of which Detlef Weigel is a Director.

Conflict of Interest: none declared.

REFERENCES


