Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Applications Note

Genetic and population analysis

QTLseqr: An R package for bulk segregant analysis with next-generation sequencing

Ben N. Mansfeld[1,*] and Rebecca Grumet[1]

[1]Plant Breeding, Genetics and Biotechnology Program, Department of Horticulture, Michigan State University, East Lansing, MI, USA

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: QTL-seq is a relatively new and rapid way for performing bulk segregant analysis using next-generation sequencing data. However, no easy-to-use and multi-system compatible algorithms for performing this analysis are readily available. We developed the QTLseqr R package, that implements two methods, Δ*SNP-index* and *G'*, for quick identification of genomic regions associated with traits of interest in next-generation sequencing, bulk segregant analysis experiments.

Availability and Implementation: QTLseqr is available at https://github.com/bmansfeld/QTLseqr

Contact: mansfeld@msu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Since the early 1990's, Bulk Segregant Analysis (BSA) has been a valuable tool for rapidly identifying markers in a genomic region associated with a trait of interest (Giovannoni *et al.*, 1991; Michelmore *et al.*, 1991). BSA is amenable to any type of codominant markers, including single nucleotide polymorphism (SNP) markers. This has allowed for the adaptation of this technology for use with next-generation sequencing (NGS) reads. The recent reduction in cost of NGS has further contributed to the increased the use and development of this and similar methods [thoroughly reviewed by Schneeberger (2014)].

The BSA procedure is performed by establishing and phenotyping a segregating population and selecting individuals with high and low values for the trait of interest. DNA from these individuals is pooled in to high and low bulks which are subject to sequencing and SNP calling, thus mitigating a need to develop markers in advance. In bulks selected from F_2 populations, SNPs detected in reads derived from regions not linked to the trait of interest should be present in ~50% of the reads. However, SNPs in reads aligning to genomic regions closely linked to the trait should be over- or under-represented depending on the bulk. Thus, comparing relative allele depths, or *SNP-indices* (defined as the number of reads containing a SNP divided by the total sequencing depth at that SNP) between the bulks can allow QTL identification.

In plant breeding research, the main pipeline used for BSA, termed QTL-seq, was developed by Takagi *et al.* (2013) and has been widely used in several crops for many traits. Takagi and colleagues define the Δ*SNP-index* for each SNP, as the difference of the low value bulk *SNP-index* from the high value bulk *SNP-index*. They suggest averaging and plotting Δ*SNP-indices* over a sliding window. Regions with a Δ*SNP-index* that pass a confidence interval threshold, as calculated by a statistical simulation, should contain QTL. The algorithm described by Takagi et al. was released as a complete pipeline written in a combination of bash, pearl and R scripts, meant to perform all tasks from processing and cleaning raw reads to plotting Δ*SNP-index* plots. One drawback to this pipeline is the lack of configurability, specifically in SNP calling and filtering. Furthermore, as the pipeline has not been updated in several years, software and version incompatibility issues have arisen, limiting the widespread utilization of this otherwise well-designed pipeline.

An alternate approach to evaluate statistical significance of QTL from NGS-BSA was proposed by Magwene *et al.* (2011) – calculating a modified *G* statistic for each SNP based on the observed and expected allele depths and smoothing this value using a tricube, or Nadaraya-Watson smoothing kernel (Nadaraya, 1964; Watson, 1964). This method weights neighboring SNPs' *G* statistic by their relative distance from the focal SNP such that closer SNPs receive higher weights. Using the smoothed *G* statistic, or *G'*, Magwene *et al.* allow for noise reduction while also addressing linkage disequilibrium between SNPs.

Furthermore, as *G'* is close to being log normally distributed, p-values can be estimated for each SNP using non-parametric estimation of the null distribution of *G*'. This provides a clear and easy-to-interpret result as well as the option for multiple testing corrections.

We present QTLseqr, an R package for NGS-BSA, that incorporates both methods described above. QTLseqr, can quickly import and filter SNP data from Genome Analysis Tool Kit (GATK) (Van der Auwera *et al.*, 2013) pipelines, then calculate and plot SNP distributions, the tricube-smoothed *G'* values and Δ*SNP-indices*, as well as -$\log_{10}$(p-values) allowing for easy identification of QTL regions.

2 Methods and features

QTLseqr imports SNP data, from GATK's *VariantsToTable* function (Van der Auwera *et al.*, 2013), as a data frame where each row is a SNP and each column is a descriptive field. For each SNP, the total reference allele frequency, *SNP-index* and Δ*SNP-index* are calculated. The *filterSNPs()* function, allows for filtering SNPs based on reference allele frequency, total read depth, per bulk read depth and genotype quality score. The initial number of SNPs, number of SNPs filtered per step, total number of SNPs filtered, and remaining number are reported.

The primary analysis steps are performed by *runGprimeAnalysis()* which initially calculates *G* for each SNP. It then counts the number of SNPs within the set window bandwidth (Mb), and estimates the *G'* value of each SNP by local constant regression within that window. Subsequently, p-values and genome-wide Benjamini-Hochberg adjusted p-values are calculated for each SNP.

P-values are estimated from the null distribution of *G'* (i.e. assuming no QTL). To this end, *G'* values from QTL regions are temporarily removed, so that mean and variance of the null distribution of *G'* may be estimated. Magwene *et al.* (2011) suggest using Hampel's rule to filter out these regions. However, with the data we tested this method failed to filter QTL regions and resulted in inflated p-values. Yang *et al.* (2013) first manually removed putative QTL, and then randomly selected *G'* values from the remaining regions in the genome for estimating the null distribution parameters. We find that utilizing the Δ*SNP-index* as a data driven method for identifying and filtering potential QTL is successful for estimating p-values. QTLseqr offers this, alongside Hampel's rule, as an option for p-value calculation.

QTLseqr has two main plotting functions for quality control and data visualization. The *plotGprimeDist()* function can be used to plot the *G'* distribution as a check to assess the validity of the analysis (Supplementary Fig. 1). The *plotQTLStats()* function is used for plotting the number of SNPs/window, the tricube-weighted Δ*SNP-index* and *G'* values, or the -$\log_{10}$(p-value) (Fig. 1). Other QTLseqr functions are available for extracting, summarizing and reporting of significant QTL regions.

3 Results

We tested QTLseqr on data available from Yang *et al.* (2013), a BSA study which identified loci for seedling cold tolerance in rice. Raw reads were downloaded from the NCBI Short Read Archive, aligned to the v7 Nipponbare genome (http://rice.plantbiology.msu.edu/) and SNPs were called as described in the GATK "Best Practices" (https://software.broadinstitute.org/gatk/best-practices/). QTLseqr was successful at reproducing the analysis performed by Yang and colleagues, confirming QTL on chromosomes 1, 2, 8, and 10 (Supplementary Fig. 2). Figure 1 shows a putative QTL identified on rice chromosome 8, as output by the *plotQTLStats()* function.

Acknowledgements and Funding

We thank the members of the Michigan State University, Horticulture Department's *Joint Lab* group for critical reading of the manuscript as well as Dr. Robert VanBuren for other feedback and advice. This work was in part supported by the National Institute of Food and Agriculture, US Department of Agriculture, under award number 2015-51181-24285 and MSU Project GREEEN.

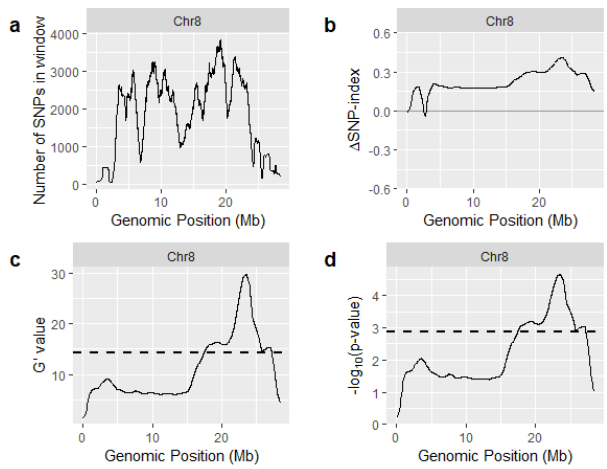

Fig. 1. Quantitative trait locus for rice seedling cold tolerance on chromosome 8 identified by QTLseqr. Plots produced by the *plotQTLStats()* function with a 1 Mb sliding window: Distribution of SNPs in each smoothing window (a). The tricube-smoothed ΔSNP-index (b). The tricubed-smoothed G' value, genome-wide FDR rate of 0.01 indicated by the dashed line (c). Another, more familiar way to display QTL, is using the -$\log_{10}$(p-value) which is derived from the *G'* value (d).

Conflict of Interest: none declared.

References

Van der Auwera,G.A. *et al.* (2013) From FastQ data to high-confidence variant calls: The Genome Analysis Toolkit best practices pipeline. In, *Current Protocols in Bioinformatics*. John Wiley & Sons, Inc., Hoboken, NJ, USA, p. 11.10.1-11.10.33.

Giovannoni, J.J. *et al.* (1991) Isolation of molecular markers from specific chromosomal intervals using DNA pools from existing mapping populations. *Nucleic Acids Res*, **19**, 6553–6568.

Magwene,P.M. *et al.* (2011) The statistics of bulk segregant analysis using next generation sequencing. *PLoS Comput Biol*, **7**, e1002255.

Michelmore,R.W. *et al.* (1991) Identification of markers linked To disease-resistance genes by bulked segregant analysis - a rapid method to detect markers in specific genomic regions by using segregating populations. *Proc Natl Acad Sci U S A*, **88**, 9828–9832.

Nadaraya,E.A. (1964) On estimating regression. *Theory Probab Its Appl*, **9**, 141–142.

Schneeberger,K. (2014) Using next-generation sequencing to isolate mutant genes from forward genetic screens. *Nat Rev Genet*, **15**, 662–76.

Takagi,H. *et al.* (2013) QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. *Plant J*, **74**, 174–83.

Watson,G.S. (1964) Smooth regression analysis. *Sankhya*, **26**, 359–372.

Yang,Z. *et al.* (2013) Mapping of quantitative trait loci underlying cold tolerance in rice seedlings via high-throughput sequencing of pooled extremes. *PLoS One*, **8**, e68433.