ELSEVIER

COMPUTATIONAL
ANDSTRUCTURAL
BIOTECHNOLOGY
JOURNAL

journal homepage: www.elsevier.com/locate/csbj

QTL.gCIMapping.GUI v2.0: An R software for detecting small-effect and linked QTLs for quantitative traits in bi-parental segregation populations

Ya-Wen Zhang [a], Yang-Jun Wen [b], Jim M. Dunwell [c], Yuan-Ming Zhang [a,*]

[a] *Crop Information Center, College of Plant Science and Technology, Huazhong Agricultural University, Wuhan 430070, China*
[b] *State Key Laboratory of Crop Genetics and Germplasm Enhancement, Nanjing Agricultural University, Nanjing 210095, China*
[c] *School of Agriculture, Policy and Development, University of Reading, Reading RG6 6AR, United Kingdom*

ARTICLE INFO

Article history:
Received 21 July 2019
Received in revised form 10 November 2019
Accepted 15 November 2019
Available online 9 December 2019

Keywords:
GCIM
Small-effect QTL
Linked QTL
R
QTL.gCIMapping.GUI
QTL.gCIMapping

ABSTRACT

The methodologies and software packages for mapping quantitative trait loci (QTLs) in bi-parental segregation populations are well established. However, it is still difficult to detect small-effect and linked QTLs. To address this issue, we proposed a genome-wide composite interval mapping (GCIM) in bi-parental segregation populations. To popularize this method, we developed an R package. This program with two versions (Graphical User Interface: QTL.gCIMapping.GUI v2.0 and code: QTL.gCIMapping v3.2) can be used to identify QTLs for quantitative traits in recombinant inbred lines, doubled haploid lines, backcross and F_2 populations. To save running time, fread function was used to read the dataset, parallel operation was used in parameter estimation, and conditional probability calculation was implemented by C++. Once one input file with *.csv or *.txt formats is uploaded into the package, one or two output files and one figure can be obtained. The input file with the ICIM and win QTL cartographer formats is available as well. Real data analysis for 1000-grain weight in rice showed that the GCIM detects the maximum previously reported QTLs and genes, and has the minimum AIC value in the stepwise regression of all the identified QTLs for this trait; using stepwise regression and empirical Bayesian analyses, there are some false QTLs around the previously reported QTLs and genes from the CIM method. The above software packages on Windows, Mac and Linux can be downloaded from https://cran.r-project.org/web/packages/ or https://bigd.big.ac.cn/biocode/tools/7078/releases/27 in order to identify all kinds of omics QTLs.

1. Introduction

Linkage analysis has been the predominant statistical genetic mapping approach in the past decades. It has been widely used in the genetic dissection of quantitative traits in bi-parental segregation populations. Up to now numerous quantitative trait loci (QTLs) for quantitative traits have been identified, and some QTLs have been used to conduct marker-assisted selection in crop breeding and to clone genes in molecular biology. However, most of these QTLs have large effects.

At present the methodologies and software packages of mapping QTLs for quantitative traits in bi-parental segregation populations have been well established. At the early stage, Lander and Botstein [1] established a new framework, named interval mapping, for mapping QTLs in such populations, and its software is Mapmarker/QTL. However, this method does not consider the effect of other QTLs on the mapping of the current QTL. To overcome this issue, composite interval mapping (CIM) has been proposed [2–4]. Based on this CIM method, numerous software packages have been released, e.g., QTL PLABQTL [5], QGene [6], HAPPY [7], Map Manager [8], win QTL Cartographer [9], and QTL Express [10]. Similar to the CIM method, mixed linear model [11] and inclusive CIM (ICIM) [12] methods have been developed, and the corresponding software packages QTLnetwork [13] and QTL IciMapping [14] have also been released. As we know, multi-QTL mapping is now the state-of-the-art method for QTL mapping. Thus, a series of multi-QTL mapping approaches, including multiple-interval mapping [15], multi-marker analysis [16], Bayesian shrinkage estimation [17,18], penalized maximum likelihood [19], empirical Bayes [20], have been proposed, and some software packages have been released, for example, R/qtl [21,22], MapQTL [23], mpMap [24], win QTL Cartographer [25], and DOQTL [26]. These above methods and software packages have played an

[*] Corresponding author.
E-mail address: soyzhang@mail.hzau.edu.cn (Y.-M. Zhang).

https://doi.org/10.1016/j.csbj.2019.11.005
2001-0370/© 2019 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology.

important role in QTL mapping. However, there are still two shortcomings. First, it is difficult to detect small-effect QTLs. The reasons are as follows: the limited number of individuals or lines in the mapping population, large experimental error in field experiments, and low power of QTL mapping methods in the detection of small-effect QTLs. Secondly, it is also difficult to detect the closely linked QTLs. If there are two linked QTLs with effects in the same direction, one false QTL between the two linked QTLs may be detected, whereas if there are two linked QTLs with effects in the opposite directions, no QTL may be identified because the effects of the two linked QTLs may be canceled. To overcome the above two issues, Wang et al. [27] and Wen et al. [28] proposed genome-wide composite interval mapping (GCIM), respectively, in backcross and F_2 populations.

In the GCIM, first, a single-locus random-SNP-effect mixed linear model in genome-wide association studies (GWAS) is used to scan each putative QTL on the genome (File A.1). Here controlling polygenic background via selecting background markers in CIM is replaced by estimating polygenic variance in GWAS. Then, all the peaks in the negative logarithm P-value curve are viewed as the positions of multiple putative QTLs to be included in a multi-QTL model, the QTL effects are estimated by empirical Bayes, and non-zero effects are further evaluated by likelihood ratio test for true QTL detection. Our Monte Carlo simulation studies and real data analysis showed that the GCIM has high power in detecting QTL, high accuracy in estimating QTL parameters and low false positive rate as compared with the widely-used methods. To popularize the GCIM, it is necessary to develop its software package. In this article, we introduce the R packages QTL.gCIMapping (File A.2) or QTL.gCIMapping.GUI (File A.3).

2. Materials and methods

2.1. Genetic mapping population

There are two kinds of mapping populations available in this software package. The first ones, including backcross (BC), doubled haploid (DH) lines, recombinant inbred lines (RIL), F_2, and "immortalized F_2 (IF_2)", are primary segregation populations derived from the cross between two homozygous lines. The second ones are secondary segregation populations derived from the cross between two near iso-genic lines (introgressive lines or QTL isogenic lines). Generally speaking, the accuracy of QTL mapping in primary mapping populations is lower than that in secondary mapping populations. The sample size is frequently more than 250 in primary mapping and several thousands in fine mapping. All the dominant and co-dominant markers are available.

In the genetic mapping populations, linkage maps, marker genotypes and phenotypic values for quantitative traits are needed in the QTL mapping. In the GCIM software package, the above three kinds of information are incorporated into one input file. In addition, covariate information can be considered and incorporated into the input file.

2.2. 1000-grain weight in rice

In this study, four datasets for 1000-grain weight in rice RILs described in Yu et al. [29] are re-analyzed using the GCIM, CIM and ICIM methods, which were implemented, respectively, by the software packages QTL.gCIMapping.GUI v2.0, win QTL Cartographer v2.5 [25] and QTL IciMapping v4.1 [14]. The datasets Xing1997 and Xing1998 were downloaded from Xing et al [30], while the datasets Hua1998 and Hua1999 were downloaded from Hua et al. [31,32]. All the four datasets have the same genotypes and different phenotypes. There were 1619 bin markers for all the 210 RILs and these markers covered a length of 1625.509 cM on the genome (Table A.1).

2.3. Development of the QTL.gCIMapping.GUI v2.0 software

QTL.gCIMapping.GUI v2.0 is an R package with graphical user interface. Shiny is used to build interactive web applications in R by automatic reactive binding between inputs and outputs and extensive prebuilt widgets. The software mainly contains three modules: dataset inputting, parameter settings and plot redrawing. Once the dataset inputting and parameter settings are finished, users may run the program and the results will be saved in the directory that was set by the users. The plot redrawing is an additional module, and its function is to redraw the plot using the previously saved results, such as adjusting the color or thickness of the curve and the resolution of the plot.

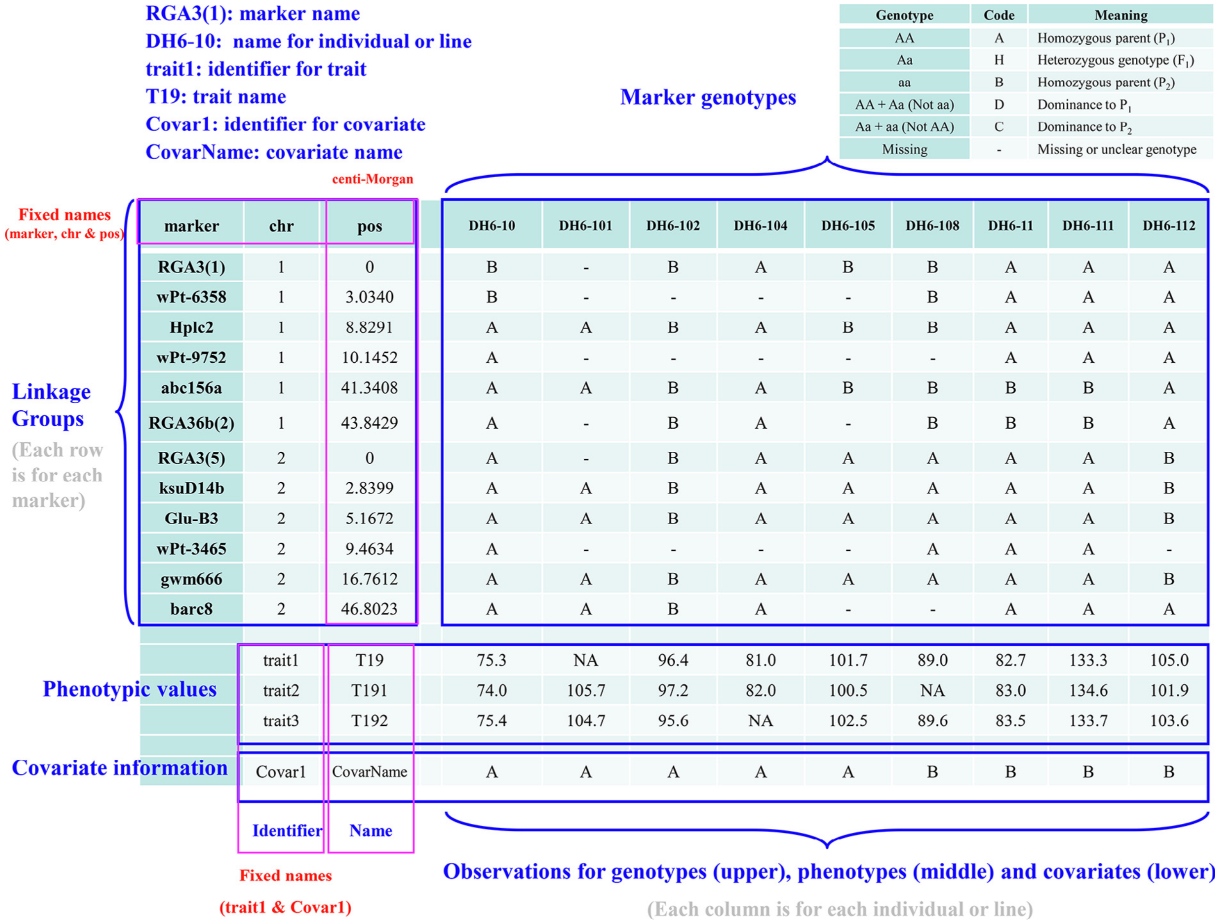
To reduce the running time, three approaches were adopted. First, conditional probability calculation for the genotypes of putative QTL was implemented by C++, and Rcpp [33,34] is used for the interface between R and C++. Then, parallel calculation was adopted in parameter estimation. Parallel is used to detect the number of CPU cores on the current host and create a set of copies of R running in parallel and communicating over sockets, and doParallel is used to register the parallel backend with the foreach package. Finally, fread function was used to read dataset.

Once the software is successfully installed in the R environment, users can operate the software by clicking the mouse. If users analyze the datasets on the server, the code version of this software, QTL.gCIMapping v3.2, is also available. In this situation, users need to write R script to call the function instead of clicking the mouse directly.

2.4. Preparation of the input file

There are three kinds of input file formats available in this software. The first one is the GCIM format with the *.csv or *.txt file. This file consists of four parts: linkage maps, marker genotypes, phenotypic values for quantitative trait, and covariates (Fig. 1). Linkage maps, including marker names, chromosome number and marker positions on chromosome (cM), are listed in the first three columns. The marker genotypes are listed from the fourth column. Each column represents one individual or line, the first row shows the name of individual or line, and the other rows on the right side of the linkage maps show marker genotypes. Each marker genotype is encoded by a given single capital letter, "A", "H" and "B" indicate genotypes "AA", "Aa" and "aa", respectively, while dominant marker genotypes "AA + Aa" or "Aa + aa" are represented by "D" or "C", respectively. "-" means the absence of marker genotypes. The phenotypic values for quantitative trait and covariates are presented below both linkage maps and marker genotypes. The phenotypic values for each trait are arranged in one row and distinguished by "trait*", where "*" is the serial number of the trait. The trait name is given on the next column with the same row, and the phenotypic value of an individual or line is placed at the intersection of the trait and the individual or line. If some phenotypic values are missing, these values are indicated by "NA". Each covariate is arranged into one row. The covariates are distinguished by "Covar*", where "*" is the serial number of the covariate. The covariate name is presented on the next column with the same row, and the covariate value of an individual or line is placed at the intersection of the covariate and the individual. If there are no covariates, this block is empty.

The second format for the input file is the Mt-MIM Control File format of QTL Cartographer, and the third format is the EXCEL 2007 format for BIP of QTL IciMapping. The sample files for the two formats are showed on Tables A.2 and A.3.


RGA3(1): marker name
DH6-10: name for individual or line
trait1: identifier for trait
T19: trait name
Covar1: identifier for covariate
CovarName: covariate name
Marker genotypes
centi-Morgan
Fixed names (marker, chr & pos)
Linkage Groups
(Each row is for each marker)
Phenotypic values
Covariate information
Identifier
Name
Fixed names (trait1 & Covar1)
Observations for genotypes (upper), phenotypes (middle) and covariates (lower)
(Each column is for each individual or line)

Genotype	Code	Meaning
AA	A	Homozygous parent (P_1)
Aa	H	Heterozygous genotype (F_1)
aa	B	Homozygous parent (P_2)
AA + Aa (Not aa)	D	Dominance to P_1
Aa + aa (Not AA)	C	Dominance to P_2
Missing	-	Missing or unclear genotype

marker	chr	pos	DH6-10	DH6-101	DH6-102	DH6-104	DH6-105	DH6-108	DH6-11	DH6-111	DH6-112
RGA3(1)	1	0	B	-	B	A	B	B	A	A	A
wPt-6358	1	3.0340	B	-	-	-	-	B	A	A	A
Hplc2	1	8.8291	A	A	B	A	B	B	A	A	A
wPt-9752	1	10.1452	A	-	-	-	-	-	A	A	A
abc156a	1	41.3408	A	A	B	A	B	B	B	B	A
RGA36b(2)	1	43.8429	A	-	B	A	-	B	B	B	A
RGA3(5)	2	0	A	-	B	A	A	A	A	A	B
ksuD14b	2	2.8399	A	A	B	A	A	A	A	A	B
Glu-B3	2	5.1672	A	A	B	A	A	A	A	A	B
wPt-3465	2	9.4634	A	-	-	-	-	A	A	A	-
gwm666	2	16.7612	A	A	B	A	A	A	A	A	B
barc8	2	46.8023	A	A	B	A	-	-	A	A	A
	trait1	T19	75.3	NA	96.4	81.0	101.7	89.0	82.7	133.3	105.0
	trait2	T191	74.0	105.7	97.2	82.0	100.5	NA	83.0	134.6	101.9
	trait3	T192	75.4	104.7	95.6	NA	102.5	89.6	83.5	133.7	103.6
	Covar1	CovarName	A	A	A	A	A	B	B	B	B

Fig. 1. The GCIM format of the input file for the software QTL.gCIMapping.GUI v2.0.

2.5. Statistical analysis

Stepwise regression is implemented by function “step” in R package “stats”. In the “step” function, we selected “both”, which does the forward and backward stepwise regression analysis. The introduction and deletion of a variable are based on the decrease of the AIC value for the model [35]. Once the AIC value no longer decreases, the stepwise regression analysis terminates and the optimal regression equation is output.

2.6. Installation of the software package

This software can be installed in two ways: online installation and offline installation. For online installation, users can install directly using the below command:

install.packages(“QTL.gCIMapping.GUI”)

then all the add-on packages and QTL.gCIMapping.GUI v2.0 will install automatically. For offline installation, users first open R GUI, select “Packages” — “Install package(s) from local files…”, and then find and install the add-on packages, which include 39 packages:

“cmprsk”,“corpcor”,“data.table”,“digest”,“doParallel”,“Epi”,“etm”,“fdrtool”,“foreach”,“GeneNet”,“glmnet”,“htmltools”,“httpuv”,“iterators”,“jsonlite”,“later”,“longitudinal”,“magrittr”,“MASS”,“mime”,“numDeriv”,“openxlsx”,“parcor”,“plyr”,“ppls”,“promises”,“QTL.gCIMapping”,“qtl”,“R6”,“Rcpp”,“shiny”,“sourcetools”,“stringi”,“stringr”,“testthat”,“utf8”,“xtable”,“zip”,“zoo”

Finally, users install QTL.gCIMapping.GUI v2.0 package, which has been downloaded on your computer (Fig. 2).

2.7. Implementation of the software package

Once the software is installed, users can run the software by two commands:

library(QTL.gCIMapping.GUI) and QTL.gCIMapping.GUI()

Once the software is restarted, users use the above two commands as well.

The installation for the code version is almost the same as that for the above GUI version. The difference is that users need to write the below R scripts:

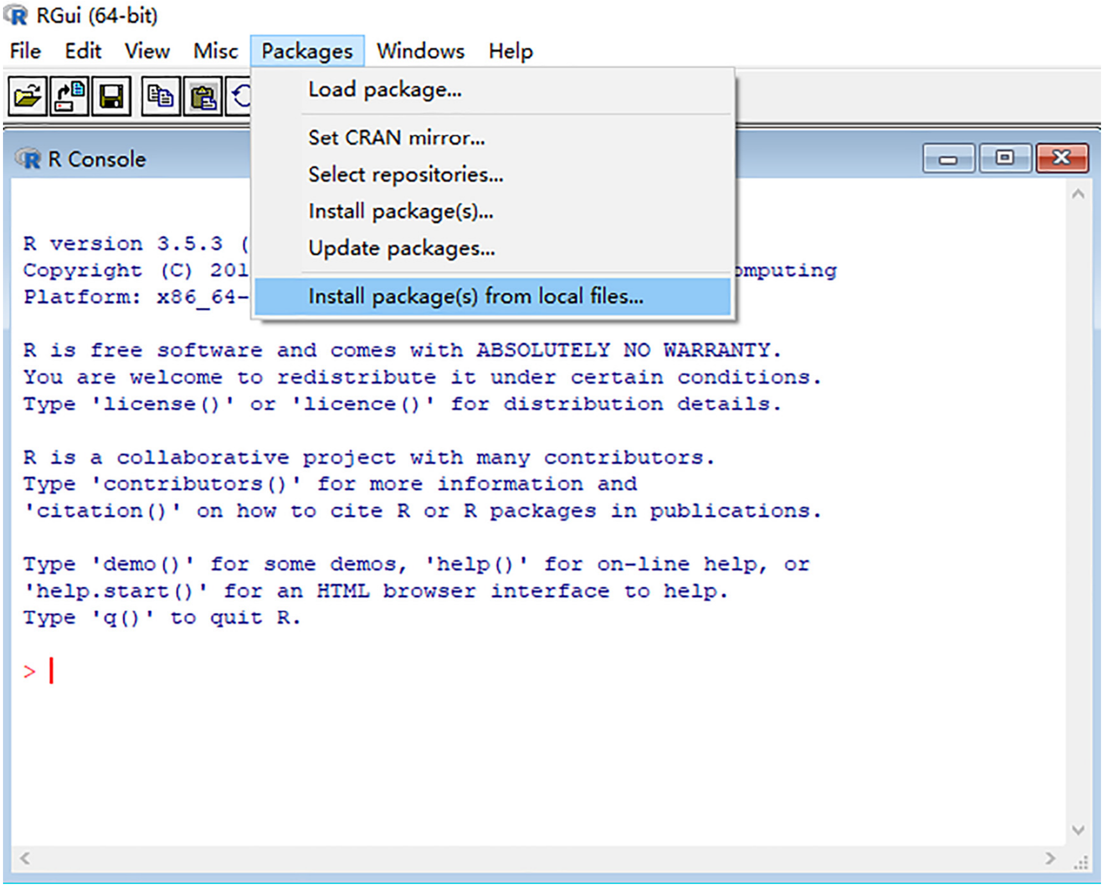
QTL.gCIMapping(file=“D:/Users/GCIM_Format_DH.csv”,fileFormat=“GCIM”,fileICIMcov = NULL,Population=“DH”,Model=“Random”,WalkSpeed = 1,CriLOD = 2.5,Likelihood=“REML”,flagrqtl=“FALSE”,DrawPlot=“TRUE”,PlotFormat=“png”,Resolution=“Low”,Trait = 1:1,dir=“D:/Users”)

The explanations for the above parameters are found in Table A.4. In the parameter setting module, users need to set fourteen parameters. Among these parameters, Likelihood and flagrqtl are specific to the F_2 population, and eight may be default or set by users, including fileFormat, fileICIMcov, Model, Likelihood, flagrqtl, DrawPlot, PlotFormat, and Resolution (File A.3).

3. Results

3.1. The description for the result files

If users run the code version of the software package, users will get one excel file with the *.csv format (named “*i*_GCIM result.csv”) and one plot file with the png or jpeg or pdf formats (named “*i*_res.tiff”) in the directory that is set by users, where “*i*” represents the



```
R version 3.5.3 (
Copyright (C) 201                                         omputing
Platform: x86_64-

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> |
```

Fig. 2. The installation of the software QTL.gCIMapping.GUI v2.0.

number of the trait. If users run the GUI version, users will obtain one additional file, named "i_resultforplot.xlsx". Users may upload this additional file to the plot redrawing module of the software in order to draw one new plot, if users need to change the plot parameters, e.g., resolution and color. It is easy to change these plot parameters by clicking the mouse.

In the "*i*_GCIM result.csv" file for DH, RIL, BC_1, BC_2, there are nine parameters for each significant QTL and two parameters for each mapping population (Table A.5). These parameters are described as below. "Trait" is the trait name analyzed; "Chr" is the abbreviation for chromosome and its result is indicated by an integer number; "Position (cM)" is the position (cM) of each significant QTL on the chromosome; "Additive Effect" is the additive effect for each significant QTL; "LOD" is the LOD score for each significant QTL; "Left_Marker" and "Right_Marker" are respectively the left and right flanking marker names around each significant QTL; "Var_Genet" is the genetic variance for each significant QTL; "r^2 (%)" is the proportion of phenotypic variance explained by single QTL. "Var_Error" is the residual variance for the multi-QTL genetic model, and "Var_Phen (total)" is the variance for the phenotypic values of the trait in the mapping population. In the "*i*_GCIM result.csv" file for F_2, the parameter "Additive Effect" for QTL effect is replaced by two parameters "Effect.a" and "Effect.d", where "Effect.a" is additive effect and "Effect.d" is dominant effect for each significant QTL.

In the "*i*_res.tiff" plot file for DH, RIL, BC_1, BC_2, horizontal (x) axis indicates marker position on the genome (cM); vertical axis on the right side means $-\log_{10}$(P-value) for all the putative QTLs on the genome, which are the results in the first step in GCIM, and vertical axis on the left side represents LOD scores for all the significant QTLs, which are the results in the second step in GCIM. In the "*i*_res.tiff" plot file for F_2, there are two $-\log_{10}$(P-value) curves. One is for an additive effect and another is for a dominant effect.

3.2. Real data analysis for 1000-grain weight in rice

Four datasets for 1000-grain weight in rice from Xing et al. [30] and Hua et al. [31,32] were re-analyzed by the GCIM, CIM and ICIM methods. All the results are listed in the Tables A.6–A.9. A total of 13, 11 and 10 stable QTLs across the four datasets were detected by GCIM, CIM and ICIM, respectively (Table 1). Among the 18 stable QTLs, 10 are detected commonly by at least two methods and have the average r^2 of 5.40% from the GCIM, while 8 are identified by a single method and have the average r^2 of 2.82% (Tables A.6–A.9). This means that common QTLs have large effects and unique QTLs for each method have small effects. Meanwhile, a total of 8, 8, and 8 previously reported QTLs and 8, 5, and 7 previously reported genes around the above detected QTLs were found by the GCIM, CIM and ICIM methods, respectively, to be associated with the 1000-grain weight (Tables A.10–A.16; Figs. 3 and A.1–A.3; File A.4).

To compare the above three methods, we established multiple regression equations of all the QTLs, detected by each QTL mapping method, for the trait (model I), and then conducted stepwise regression (model II) and empirical Bayes (model III) analyses. As a result, the GCIM generates consistent results across the three models. Among all the 41 QTLs detected by the CIM, 10, 8, 9, and 10 were found in the model II not to be associated with 1000-grain weight, respectively, in the Xing1997, Xing1998, Hua1998 and Hua1999 datasets; 11, 8, 7, and 10 were found in the model

Table 1
Stable QTLs for rice 1000-grain weight in multiple environments detected by composite interval mapping (CIM), genome-wide CIM (GCIM) and inclusive CIM (ICIM).

QTL	Chr	GCIM				CIM				ICIM			
		Posi (cM)	Effect	LOD	r^2 (%)	Posi (cM)	Effect	LOD	r^2 (%)	Posi (cM)	Effect	LOD	r^2 (%)
3	1	36.1 ~ 37.3	0.54 ~ 1.06	5.49 ~ 18.09	2.00 ~ 8.42	36.1 ~ 37.3	0.72 ~ 0.80	7.55 ~ 9.17	2.71 ~ 2.83	37.0 ~ 38.0	0.65 ~ 0.88	11.73 ~ 22.66	6.91 ~ 10.49
7	1	146.2 ~ 147.7	−0.61 ~ −0.52	6.59 ~ 7.97	2.05 ~ 3.54	146.2 ~ 148.2	−0.76 ~ −0.54	4.82 ~ 6.13	2.54 ~ 3.59	146.0 ~ 148.0	−0.56 ~ −0.47	6.82 ~ 8.07	2.95 ~ 5.29
8	2									147.0	0.39 ~ 0.47	4.74 ~ 7.76	2.55 ~ 3.04
9	2					171.1	0.41 ~ 0.45	2.90 ~ 3.28	2.55 ~ 2.64				
11	3	93.0 ~ 93.8	−1.21 ~ −0.98	15.4 ~ 27.12	7.23 ~ 16.64	93.0 ~ 93.8	−1.46 ~ −0.95	11.94 ~ 20.26	2.71 ~ 3.56	94.0	−1.00 ~ −0.89	15.33 ~ 23.77	12.94 ~ 15.91
14	3	122.1 ~ 122.4	−0.62 ~ −0.47	6.48 ~ 7.43	2.59 ~ 3.30	122.4	−0.67 ~ −0.41	2.54 ~ 5.22	2.57 ~ 3.18				
16	3	135.7 ~ 136.9	−0.79 ~ −0.53	7.78 ~ 10.14	2.16 ~ 6.04								
18	5	29.7	0.95 ~ 1.30	17.21 ~ 37.72	10.46 ~ 13.68	29.7	0.92 ~ 1.31	11.97 ~ 18.53	2.71 ~ 3.56	30.0	0.93 ~ 1.42	19.20 ~ 39.28	14.20 ~ 22.63
20	5									93.0 ~ 96.0	−0.53 ~ −0.35	3.70 ~ 6.26	1.64 ~ 4.82
21	5	100.5 ~ 103.4	−0.36 ~ −0.34	3.18 ~ 3.99	0.96 ~ 1.70								
23	6	7.3 ~ 12.4	0.36 ~ 0.81	2.98 ~ 10.32	1.40 ~ 4.44					8.0 ~ 12.0	0.38 ~ 0.75	4.37 ~ 12.91	2.50 ~ 5.61
27	6					70.4	−0.71 ~ −0.59	4.82 ~ 5.56	2.80 ~ 3.52	69.0	−0.89 ~ −0.51	9.20 ~ 18.16	3.59 ~ 8.41
28	6					75.3 ~ 77.2	−0.57 ~ −0.46	3.49 ~ 5.78	2.39 ~ 2.59				
29	6	81.2 ~ 82.1	−0.74 ~ −0.41	2.87 ~ 15.44	1.27 ~ 8.30								
33	7	3.9	−0.44 ~ −0.34	4.48 ~ 4.66	1.46 ~ 1.74								
36	7	54.7	−0.76 ~ −0.44	3.60 ~ 17.43	1.31 ~ 8.62	55.2	−0.72 ~ −0.41	2.63 ~ 7.74	2.67 ~ 2.71	55.0 ~ 56.0	−0.98 ~ −0.31	3.49 ~ 20.93	1.33 ~ 10.19
40	9	84.2 ~ 88.0	0.44 ~ 0.66	3.26 ~ 12.08	1.29 ~ 3.29	86.6 ~ 87.3	0.55 ~ 0.65	3.72 ~ 6.36	2.36 ~ 3.28				
47	11	53.0 ~ 53.3	−0.48 ~ −0.41	3.43 ~ 6.40	1.16 ~ 1.75	53.0 ~ 53.3	−0.56 ~ −0.49	3.72 ~ 6.36	2.36 ~ 3.28	53.0 ~ 57.0	−0.52 ~ −0.46	6.12 ~ 6.88	2.48 ~ 2.63

III not to be associated with 1000-grain weight, respectively, in the above four datasets (Tables A.17–A.24). Among all the 25 QTLs detected by the ICIM, 1, and 2 were found in the model III not to be associated with 1000-grain weight, respectively, in the Xing1998 and Hua1998 datasets (Tables A.17–A.24). The results indicated that all the QTLs identified by the GCIM are found to be associated with the trait of interest and some QTLs detected by the CIM and ICIM are found not to be associated with the trait of interest. In addition, we calculated the AIC values for the above models and found that all the minimum AIC values are always from the GCIM (Table 2).

Based on the above results, we found that more small-effect known genes or QTLs are detected by the GCIM, for example, *OsSec18* (r^2 = 0.96%) is detected only by the GCIM in the Hua1998 and Hua1999 datasets, *ONAC106* (r^2 = 1.42%) is detected commonly by the GCIM and ICIM methods in the Hua1999 dataset, and *RDD1* (r^2 = 1.44%) and gw7.1 (r^2 = 1.46%) are identified commonly by all the three methods in the Hua1998 dataset (Tables A.10–A.13). Meanwhile, we also found two linked QTLs (kgw1a and *RDD1*) by all the three methods in the Hua1998 dataset (Fig. 3).

Around the above detected QTLs, we found two new candidate genes: *LOC_Os12g44290* and *LOC_Os11g04860*. Based on the KEGG pathways at http://kobas.cbi.pku.edu.cn/ (KOBAS), these two genes are associated with steroid hormone biosynthesis, and some known genes for 1000-grain weight, such as *SLG* and *GW5* [36,37], can modulate brasssinosteroid homeostasis in rice.

4. Discussion

The software package developed in this study has some advantages over similar packages. First, the GCIM can detect small-effect and linked QTLs. Secondly, all the QTLs identified by the GCIM are found to be associated with the trait of interest, and some QTLs identified by the CIM are found not to be associated with the trait of interest. Thirdly, the GCIM detects more QTLs than does the ICIM. Finally, the current package considers the running time in three aspects. fread function is used to read the dataset, parallel operation is used in parameter estimation, and conditional probability calculation is implemented by C++.

As described by Kroymann and Mitchell-Olds [38] and Mackay et al. [39], detecting small and linked QTLs is a thorny issue in the genetic dissection of quantitative traits. However, this situation will change significantly since we propose the GCIM method and develop its software in this study. In the GCIM [27,28], there are two main improvements. On one hand, selecting background markers in the CIM and ICIM to control polygenic background has been replaced by estimating polygenic variance in GWAS. On the other hand, all the peaks in the negative logarithm P-value curve against genome position for each QTL effect are viewed as potential QTLs and these potential QTLs are placed into a multi-locus genetic model for true QTL identification. In our opinion, the difficulty of detecting small-effect and linked QTLs is due to their low LOD scores, although their peaks exist. Once all these peaks are viewed as potential QTLs in a multi-QTL genetic model, the power of detecting these QTLs increases. This is why the GCIM has higher power in identifying small-effect and linked QTLs than the other methods, including Bernardo [40] and Xu [41], which have the first improvement.

Based on previous and current studies, several peaks around one large-effect QTL are frequently observed when win QTL Cartographer (the CIM method) is used. In this case, we do not know whether there are multiple linked QTLs or only one single large-effect QTL. To address this issue, we optimized the regression equation of all the identified QTLs for the trait using stepwise regression and empirical Bayesian analyses. As a result, quite a

Table 2
AIC values for the regression model of all the QTLs of 1000-grain weight in rice detected by CIM, GCIM and ICIM on the trait under Xing1997, Xing1998, Hua1998 and Hua1999 datasets.

Method	Xing1997		Xing1998		Hua1998		Hua1999	
	AIC$_1$	AIC$_2$	AIC$_1$	AIC$_2$	AIC$_1$	AIC$_2$	AIC$_1$	AIC$_2$
GCIM	169.11	169.11	170.41	170.41	66.35	66.35	56.9	56.9
CIM	195.53	179.87	191.14	179.61	138.93	128.27	123.98	109.74
ICIM	203.36	203.36	263.37	263.37	269.77	269.77	171.08	171.08

CIM: composite interval mapping; GCIM: genome-wide CIM; ICIM: inclusive CIM; AIC$_1$ and AIC$_2$: the AIC value for the full and reduced models in the stepwise regression analysis, respectively.

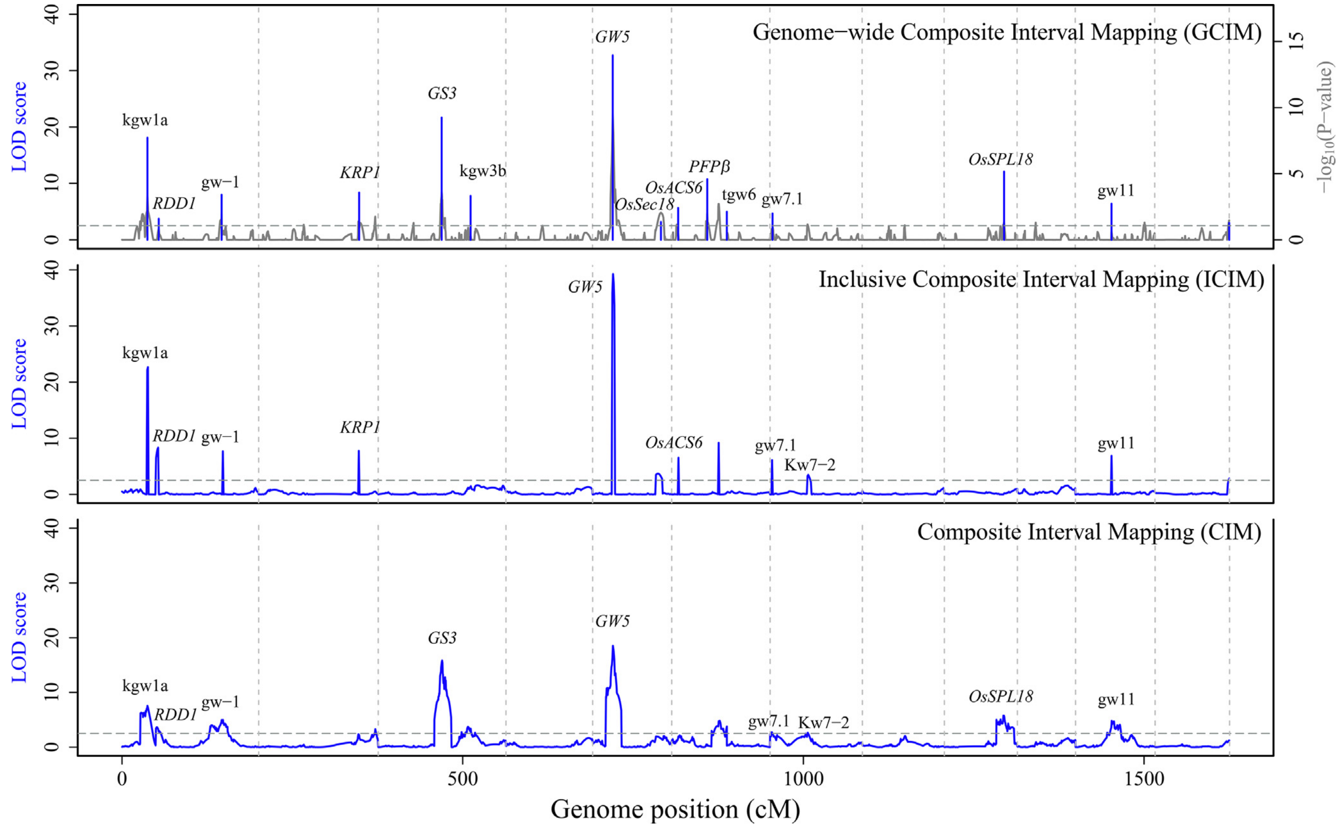


Fig. 3. Mapping QTLs for 1000-grain weight in rice detected by composite interval mapping (CIM), genome-wide CIM (GCIM) and inclusive CIM (ICIM) using the Hua1998 dataset.

number of QTLs around large-effect QTLs in the CIM are found not to be associated with the trait of interest in real data analysis. This phenomenon does not happen in the GCIM. In other words, all the QTLs detected by the GCIM are significant. This result may explain why the number of QTLs detected by the CIM is more than those by the GCIM in some real data analyses.

Based on the number of stable QTLs detected by the above-mentioned three methods in real data analysis, the GCIM generates the most, and ICIM gives the least. However, the AIC values from the GCIM are the minimum, more known genes or previously reported QTLs are identified by the GCIM, and some QTLs identified by the CIM are found not to be associated with 1000-grain weight in step-wise regression and empirical Bayes. Thus, the GCIM should be widely adopted in the near future.

At present the methodologies and software packages of mapping main-effect QTL, and QTL-by-environment and QTL-by-QTL interactions for quantitative traits in bi-parental segregation populations have been well established. However, only main-effect QTLs can be detected in the software packages QTL.gCIMapping v3.2 and QTL.gCIMapping.GUI v2.0. Thus, it is necessary to extend the methodologies for detecting QTL-by-environment and QTL-by-QTL interactions in the near future.

5. Authors' Contributions

YMZ conceived and designed the study. YWZ, YJW and YMZ wrote the codes, performed the data analyses, and wrote the draft. YMZ, JMD, and YWZ revised the manuscript.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (31571268, 31871242, 31701071), Huazhong Agricultural University Scientific & Technological Self-Innovation Foundation, China (2014RC020), and State Key Laboratory of Cotton Biology Open Fund, China (CB2019B01).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.csbj.2019.11.005.

References

[1] Lander ES, Botstein D. Mapping Mendelian factor underlying quantitative traits using RFLP linkage maps. Genetics 1989;121:185–99.

[2] Zeng ZB. Theoretical basis of separation of multiple linked gene effects on mapping quantitative trait loci. Proc Natl Acad Sci USA 1993;90:10972–6.

[3] Zeng ZB. Precision mapping of quantitative trait loci. Genetics 1994;136:1457–68.

[4] Jansen RC. Interval mapping of multiple quantitative trait loci. Genetics 1993;135:205–11.

[5] Utz HF, Melchinger AE. PLABQTL: a program for composite interval mapping of QTL. J Quant Trait Loci 1996;2(1):1–5.

[6] Nelson JC. QGene: software for marker-based genomic analysis and breeding. Mol Breed 1997;3(3):239–45.

[7] Mott R, Talbot CJ, Turri MG, et al. A method for fine mapping quantitative trait loci in outbred animal stocks. Proc Natl Acad Sci U S A 2000;97(23):12649–54.

[8] Manly KF, Cudmore Jr RH, Meer JM. Map Manager QTX, cross-platform software for genetic mapping. Mamm Genome 2001;12(12):930–2.

[9] Basten CJ, Weir BS, Zeng ZB. QTL Cartographer, Version 1.16. Raleigh, NC: Department of Statistics, North Carolina State University; 2002.

[10] Seaton G, Haley CS, Knott SA, et al. QTL Express: mapping quantitative trait loci in simple and complex pedigrees. Bioinformatics 2002;18(2):339–40.

[11] Zhu J. Mixed linear model method for mapping quantitative trait loci. Hereditas (Beijing) 1998;20(A1):137–8.

[12] Li HH, Ye GY, Wang JK. A modified algorithm for the improvement of composite interval mapping. Genetics 2007;175:361–74.

[13] Yang J, Hu C, Hu H, et al. QTLNetwork: mapping and visualizing genetic architecture of complex traits in experimental populations. Bioinformatics 2008;24(5):721–3.

[14] Meng L, Li H, Zhang L, et al. QTL IciMapping: integrated software for genetic linkage map construction and quantitative trait locus mapping in biparental populations. Crop J 2015;3(3):269–83.

[15] Kao CH, Zeng ZB, Teasdale RD. Multiple interval mapping for quantitative trait loci. Genetics 1999;152:1203–16.

[16] Broman KW, Speed TP. A model selection approach for the identification of quantitative trait loci in experimental crosses. J R Stat Soc 2002;64:641–56.

[17] Xu S. Estimating polygenic effects using markers of the entire genome. Genetics 2003;163:789–801.

[18] Wang H, Zhang YM, Li X, et al. Bayesian shrinkage estimation of quantitative trait loci parameters. Genetics 2005;170:465–80.

[19] Zhang YM, Xu S. A penalized maximum likelihood method for estimating epistatic effects of QTL. Heredity 2005;95:96–104.

[20] Xu S. An empirical Bayes method for estimating epistatic effects of quantitative trait loci. Biometrics 2007;63:513–21.

[21] Broman KW, Wu H, Sen S, et al. R/qtl: QTL mapping in experimental crosses. Bioinformatics 2003;19(7):889–90.

[22] Broman KW, Gatti DM, Simecek P, et al. R/qtl2: software for mapping quantitative trait loci with high-dimensional data and multiparent populations. Genetics 2019;211(2):495–502.

[23] van Ooijen JW. MapQTL®6, Software for the mapping of quantitative trait loci in experimental populations of diploid species *Kyazma BV*. Netherlands: Wageningen; 2009.

[24] Huang BE, George AW. R/mpMap: a computational platform for the genetic analysis of multiparent recombinant inbred lines. Bioinformatics 2001;27(5):727–9.

[25] Wang S, Basten C, Zeng ZB. Windows QTL Cartographer v2.5. Raleigh, NC: Department of Statistics, North Carolina State University; 2012.

[26] Gatti DM, Svenson KL, Shabalin A, et al. Quantitative trait locus mapping methods for diversity outbred mice. G3 (Bethesda) 2014;4(9):1623–33.

[27] Wang SB, Wen YJ, Ren WL, et al. Mapping small-effect and linked quantitative trait loci for complex traits in backcross or DH populations via a multi-locus GWAS methodology. Sci Rep 2016;6:29951.

[28] Wen YJ, Zhang H, Zhang J, et al. An efficient multi-locus mixed model framework for the detection of small and linked QTLs in F_2. Brief Bioinform 2018. https://doi.org/10.1093/bib/bby058.

[29] Yu H, Xie W, Wang J, et al. Gains in QTL detection using an ultra-high density SNP map based on population sequencing relative to traditional RFLP/SSR markers. PLoS One 2011;6(3):e17595.

[30] Xing Z, Tan F, Hua P, et al. Characterization of the main effects, epistatic effects and their environmental interactions of QTLs on the genetic basis of yield traits in rice. Theor Appl Genet 2002;105(2–3):248–57.

[31] Hua JP, Xing YZ, Xu CG, et al. Genetic dissection of an elite rice hybrid revealed that heterozygotes are not always advantageous for performance. Genetics 2002;162(4):1885–95.

[32] Hua J, Xing Y, Wu W, et al. Single-locus heterotic effects and dominance by dominance interactions can adequately explain the genetic basis of heterosis in an elite rice hybrid. Proc Natl Acad Sci USA 2003;100:2574–9.

[33] Eddelbuettel D, François R, Allaire J, et al. Rcpp: Seamless R and C++ integration. J Stat Softw 2011;40(8):1–18.

[34] Eddelbuettel D. Seamless R and C++ integration with Rcpp. New York: Springer; 2013.

[35] Venables WN, Ripley BD. Modern applied statistics with S. 4th Edition. Berlin: Springer; 2002.

[36] Feng Z, Wu C, Wang C, et al. *SLG* controls grain size and leaf angle by modulating brassinosteroid homeostasis in rice. J Exp Bot 2016;67(14):4241–53.

[37] Liu J, Chen J, Zheng X, et al. *GW5* acts in the brassinosteroid signalling pathway to regulate grain width and weight in rice. Nat Plants 2017;3:17043.

[38] Kroymann J, Mitchell-Olds T. Epistasis and balanced polymorphism influencing complex trait variation. Nature 2005;435(7038):95–8.

[39] Mackay TF, Stone EA, Ayroles JF. The genetics of quantitative traits: challenges and prospects. Nat Rev Genet 2009;10(8):565–77.

[40] Bernardo R. Genome wide markers as cofactors for precision mapping of quantitative trait loci. Theor Appl Genet 2013;126(4):999–1009.

[41] Xu S. Mapping quantitative trait loci by controlling polygenic background effects. Genetics 2013;195(4):1209–22.