User Manual for

QTL.gCIMapping

QTL genome-wide Composite Interval Mapping

(version 3.4)

Zhou Ya-Hui, Zhang Ya-Wen, Zhang Yuan-Ming

(soyzhang@mail.hzau.edu.cn)

Last updated on December 2021

Disclaimer: While extensive testing has been performed by Yuan-Ming Zhang's Lab (Statistical Genomics Lab) at Huazhong Agricultural University, the results are, in general, reliable, correct or appropriate. However, results are not guaranteed for any specific datasets. We strongly recommend that users validate the GCIM results with other software packages, such as Windows QTL Cartographer V2.5_011 (https://brcwebportal.cos.ncsu.edu/qtlcart/WQTLCart.htm), QTL IciMapping V4.2 (https://www.isbreeding.net/software/?type=detail&id=28) and QTLNetwork 2.1 (http://ibi.zju.edu.cn/software/).

Download website:

https://cran.r-project.org/web/packages/QTL.gCIMapping/index.html

References

1 Wang Shi-Bo, Wen Yang-Jun, Ren Wen-Long, Ni Yuan-Li, Zhang Jin, Feng Jian-Ying, Zhang Yuan-Ming*. Mapping small-effect and linked quantitative trait loci for complex traits in backcross or DH populations via a multi-locus GWAS methodology. ***Scientific Reports*** 2016, 6: 29951.

2 Wen Yang-Jun, Zhang Ya-Wen, Zhang Jin, Feng Jian-Ying, Jim M. Dunwell, Zhang Yuan-Ming*. An efficient multi-locus mixed model framework for the detection of small and linked QTLs in F_2. ***Briefings in Bioinformatics*** 2019, 20(5): 1913-1924.

3 Zhang Ya-Wen, Wen Yang-Jun, Jim M. Dunwell, Zhang Yuan-Ming*. QTL.gCIMapping.GUI v2.0: An R software for detecting small-effect and linked QTLs for quantitative traits in bi-parental segregation populations. ***Computational and Structural Biotechnology Journal*** 2020, 18: 59-65.

4 Wen Yang-Jun, Zhang Ya-Wen, Zhang Jin, Feng Jian-Ying, Zhang Yuan-Ming*. The improved FASTmrEMMA and GCIM algorithms for genome-wide association and linkage studies in large mapping populations. ***The Crop Journal*** 2020, 8(5): 723-732.

5 Zhou Ya-Hui, Li Guo, Zhang Yuan-Ming*. A compressed variance component mixed model framework for detecting small and linked QTL-by-environment interactions. ***Briefings in Bioinformatics*** 2021, Accepted, doi:10.1093/bib/bbab596.

Quantitative Trait Loci

Genome-wide Composite Interval Mapping

Funding

The works were supported by the National Natural Science Foundation of China (32070557, 31571268, 31871242, 31701071 and U1602261), Huazhong Agricultural University Scientific & Technological Self-innovation Foundation (Program No. 2014RC020), and State Key Laboratory of Cotton Biology Open Fund (CB2021B01).

INTRODUCTION

1.1 Why GCIM?

QTL.gCIMapping v3.4 (**QTL G**enome-wide **C**omposite **I**nterval **M**apping) is an R package, which is used to identify main-effect QTL and QTL-by-environment interaction (QEI) for quantitative traits in backcross (BC), doubled haploid (DH) lines, recombinant inbred lines (RIL), F_2, immortalized F_2 (IF_2), and $F_{2:3}$ design. QTL.gCIMapping v3.4 works well on the R environment on Windows, Linux (desktop) and MacOS.

1.2 Getting started

The software package QTL.gCIMapping v3.4 can be freely downloaded from https://cran.r-project.org/web/packages/QTL.gCIMapping/index.html, or request from the maintainer, Dr Yuan-Ming Zhang at Crop Information Center, College of Plant Science & Technology, Huazhong Agricultural University, Wuhan 430070, China (soyzhang@mail.hzau.edu.cn).

1.2.1 One-Click online installation

On R environment and network connection, the command,

```
install.packages(pkgs="QTL.gCIMapping")
```

is used to directly install the software package QTL.gCIMapping v3.4.

1.2.2 Step-by-step offline installation

1.2.2.1 Install the add-on packages

First, users download sixteen R packages, including
"data.table","doParallel","foreach","glmnet","iterators","magrittr","openxlsx","qtl","Rcpp","shape","stringi","stringr","zip","MASS","lars","readxl"
from CRAN, github (https://github.com/), or google search.

On the R environment, then, users select all the 16 packages and install them offline.

1.2.2.2 Install QTL.gCIMapping v3.4

On R GUI environment, users first select "Packages"—"Install package(s) from local files…", then find the software package QTL.gCIMapping v3.4 on user's desktop

computer or mobile device, and launch QTL.gCIMapping v3.4.

1.2.3 Run QTL.gCIMapping v3.4

Once the software package QTL.gCIMapping v3.4 is installed, users may run it using two commands:

```
library(QTL.gCIMapping)
QTL.gCIMapping(***)      (*** are the parameter list: please see § 2 Example)
```

If users re-use the software QTL.gCIMapping v3.4, users use the above two commands as well.

User Manual file Users can decompress the QTL.gCIMapping package and find the User Manual file (name: **Instruction.pdf**) in the folder of "…/QTL.gCIMapping/inst/doc".

2. Parameter settings

Parameter	Meaning	File format	Note
file	File path & name in your computer, i.e., file="D:/Users/GCIM_Format_DH.csv"	*.csv; *.txt	Table 1
fileFormat	Format for input file: GCIM (QTL.gCIMapping), ICIM (QTL IciMapping), MCIM (QTLNetwork) and Cart (WinQTLCart)		
filecov	File that requires additional input due to the absence of covariates in the input information of ICIM and MCIM, i.e., filecov="D:/Users/cov_file.csv" or filecov=NULL	*.csv; *.txt (Covariate values: **Row:** individual**; Column**: covariate name)	Table 3
Population	BC1 (F1×P1), BC2 (F1×P2), DH, RIL, F_2, i.e., Population="F_2"		
method	GCIM for main-effect QTL detection or GCIM-QEI for QEI detection are available, i.e., method="GCIM-QEI"		
MultiEnv	This parameter is specific to GCIM-QEI. If multiple environment datasets are analyzed, MultiEnv=TRUE. If not, MultiEnv=FALSE		
Model	Random (random model) or Fixed (fixed model) for QTL or QEI effects, i.e., Model="Random"		
WalkSpeed	Walk speed for Genome-wide Scanning (centi-Morgan, cM), should be less than 5 cM when setting, i.e., WalkSpeed=1		
CriLOD	The LOD score threshold for significant QTL or QEI. CriLOD=3: the LOD score threshold for significant QTL is set at 3		
CriDis	This parameter is specific to GCIM-QEI. The default is CriDis=5, which means that the significant QTLs and QEIs are optimized within the range of ≤ 5cM.		
Likelihood	This parameter is only for GCIM in F_2 population, including restricted maximum likelihood (REML) and maximum likelihood (ML). Likelihood="REML" or Likelihood="ML"		
SetSeed	This parameter is only for GCIM in F_2 population, in which the cross-validation experiment is needed		
flagrqtl	This parameter is only for GCIM in F_2 population, flagrqtl="FALSE" in the first running. If the other software detects only one QTL in a neighborhood but this software finds two linked QTLs (one with additive effect and another with dominant effect) in the region, let flagrqtl="TRUE"		
DrawPlot	This parameter is for all the populations, including FALSE and TRUE. DrawPlot=FALSE indicates no figure output; DrawPlot=TRUE indicates the LOD score [or $-\log_{10}(P\text{-}value)$] figure against genome position.		

Plotformat	*.jpeg, *.png, *.tiff and *.pdf. For example, Plotformat="jpeg" indicates the *.jpeg format of the figure file.
Resolution	Low or High. Resolution="Low" indicates the low resolution of the figure file.
Trait	Trait=1:3 indicates the analyses from the first trait to the third trait.
dir	Save path in your computer, i.e.,"D:/Users"
CLO	The number of CPU occupied by running. The default is the number of CPUs on the computer minus 1, and doesn't exceed 10.
MCIMmap	If the format for input file is MCIM, please input the map file, i.e., MCIMmap="D:/Users/SimF2_MCIM_Format.map"

Example

The GCIM-QEI method of identifying QTL-by-environment interactions

The full codes

QTL.gCIMapping(file="D:/Users/GCIM_Format_F2.csv",fileFormat="GCIM",filecov=NULL,Population="F2",method="GCIM-QEI",MultiEnv=TRUE,Model="Random",WalkSpeed=1,CriLOD=3,CriDis=5,DrawPlot=TRUE,PlotFormat="tiff",Resolution="Low",Trait=1:2,dir="D:/Users",CLO=NULL,MCIMmap=NULL)

The reduced codes

QTL.gCIMapping(file="D:/Users/GCIM_Format_F2.csv",Population="F2",MultiEnv=TRUE,dir="D:/Users")

Note

It should be noted that users must set "file", "Population", "MultiEnv", and "dir" for GCIM-QEI method, and the other parameters can be default in this function, including method="GCIM-QEI", Model="Random", WalkSpeed=1, CriLOD=3, CriDis=5, DrawPlot=TRUE, PlotFormat="tiff", Resolution="Low",CLO=NULL. If Population="F2", the parameter of "method" may be defaulted, and the default is method="GCIM-QEI". If the parameter of "Trait" is defaulted, all traits are analyzed by default, and this default is only available in the GCIM-QEI method.

The GCIM method of identifying main-effect QTLs

The full codes

QTL.gCIMapping(file="D:/Users/GCIM_Format_DH.csv",fileFormat="GCIM",filecov=NULL,Population="DH",method="GCIM",Model="Random",WalkSpeed=1,CriLOD=2.5,Likelihood="REML",SetSeed=11001,flagrqtl="FALSE",DrawPlot="TRUE",PlotFormat="png",Trait=1:1,dir="D:/Users")

The reduced codes

QTL.gCIMapping(file="D:/Users/GCIM_Format_F2.csv",Population="F2",method="GCIM",WalkSpeed=1,CriLOD=2.5,Trait=1,dir="D:/Users")

Note

It should be noted that users must set "file", "Population", "method", "WalkSpeed", "CriLOD", "Trait" and "dir", and the other parameters may be defaulted, including Likelihood="REML"; SetSeed=11001 and flagrqtl="FALSE" only for F2 population; DrawPlot=TRUE; Plotformat= "jpeg"; Resolution= "Low". Generally speaking, the random seed in the cross-validation experiment was set as 11001. If some known genes aren't identified by the

seed, users may try to use some new random seeds. At this case, one better result may be obtained.

Dataset format

GCIM format for Dataset The first three columns, named "**marker**", "**chr**" and "**pos**", stand for marker name, chromosome and marker position (cM) on the chromosome, respectively. Among the remaining columns, each column lists all the genotypes of one individual or line, while the first row shows the name of one individual or line. For the genotypes of each marker, the coding criteria are shown as Table 1.

Table 1. Coding criteria for GCIM format

Genotype	Code	Meaning
AA	A	Homozygous genotype (P_1)
Aa	H	Heterozygous genotype (F_1)
aa	B	Homozygous genotype (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

The genotypic, phenotypic and covariate datasets are located on the upper, middle, lower sections, and each covariate or trait is presented on one row. On each row, the first column is empty (single environment analysis using GCIM) or "**Env1**" (multiple environment analysis using GCIM-QEI), followed by "**trait1**", "the name of quantitative trait", and "phenotypic values for all the individuals or lines". If there are multiple traits, these traits occupy multiple lines. **"NA" indicates the missing or unknown phenotypes.** If there are covariates, the content lies below the trait dataset. The format is seen in Table 2. If there is no covariate, users should delete the last row in Table 2.

The format of ICIM dataset If users have the QTL IciMapping dataset, these files are also available in our software. Details can be found in the folder of "…/QTL.gCIMapping/inst/extdata", i.e., WheatDH_QTLIciMapping_Format.xlsx.

The format of MCIM dataset If users adopted the QTLNetwork software to analyze the dataset, the files with QTLNetwork format are also available by

GCIM-QEI in our software. See folder "…/QTL.gCIMapping/inst/extdata" for details, i.e., SimF2_MCIM_Format.txt and SimF2_MCIM_Format.map.

The format of WinQTLCart dataset If users adopted the WinQTLCart software to analyze the dataset, its file with WinQTLCart format is also available in our software. Details can be found in the folder of "…/ QTL.gCIMapping/inst/extdata", i.e., env1-jun3_WinQTLCart_Format.mcd. It should be noted that WinQTLCart only treats multiple environments as multiple traits when analyzing multi-environment data. Therefore, when inputting this format, the trait name needs to be written as follows:

```
 -start traits
t1E1    1.370946    0.539544
t1E2    -0.326883   1.082551
t2E1    4.112838    -1.618632
t2E2    3.181120    -1.220501
 -stop traits
```

Note that "t1" is "the first trait" and "t2" is "the second trait", while "E1" is "the first environment" and "E2" is "the second environment".

Table 2. The GCIM format of the dataset

marker	chr	pos	DH6-10	DH6-101	DH6-102
RGA3(1)	1	0	B	-	B
wPt-6358	1	3.034	B	-	-
Hplc2	1	8.8291	A	A	B
abc156a	1	41.3408	A	A	B
⋮	⋮	⋮	⋮	⋮	⋮
gwm437	21	162.5218	A	B	-
wmc157	21	197.9196	A	B	A
*stm1actc	21	200.4216	-	-	-
Env1	**trait1**	T19	10.27	15.68	9.98
Env2	**trait1**	T19	11.55	18.63	5.66
Env1	**trait2**	T191	74	105.68	97.16
Env2	**trait2**	T191	75.37	104.67	95.55
Env3	**trait2**	T191	75.33	105	96.33
	Covar1	CovarName	A	B	B

The format of ICIM and MCIM covariate dataset If users adopted the ICIM or MCIM methods to analyze the dataset, and the covariates exist, users should input a covariate file. In the file, the first column indicates individual name and the second column is the covariate information (Table 3). In Table 3, the covariate values are indicated by such as A, B, and C.

Table 3. The covariate file format

Individual ID	Covariate
DH6-10	A
DH6-101	A
DH6-102	A
DH6-104	A
DH6-164	B
DH6-165	B
DH6-166	C
DH7-124	C

3. Result

Once the software running is finished, “results” files are appeared on the Directory, which was set up by users before running the software. When GCIM is adopted, three files are outputted, including “*_GCIM result.csv”, “*_resultforplot.csv”, and a plot. When GCIM-QEI is adopted, three files are outputted, including “*_GCIM-QEI result.csv”, “*_resultforplot.csv”, and one GCIM-QEI plot.

In the *_GCIM result.csv file, there are ten columns for BC1, BC2, DH, and RIL populations, as shown below.

Trait: The name of trait analyzed

Chr: Chromosome, represented by an integer number

Position (cM): The QTL position (cM) on the chromosome

Additive Effect: Additive effect for significant QTL

LOD: LOD score for significant QTL

Left_Marker: the name of Left flanking marker around significant QTL

Right_Marker: the name of Right flanking marker around significant QTL

Var_Genet: Genetic variance for each significant QTL

r^2 (%): Proportion of phenotypic variance explained by single QTL

Var_Error: residual variance under the full model

Var_Phen (total): Phenotypic variance in the analyzed population

In main-effect QTL detection in F_2, the **Results** file includes twelve columns. Trait, Chr, Position (cM), Left_Marker, Right_Marker, Var_Genet, LOD, r^2 (%), Var_Error, and Var_phen are same as those in the above populations, while QTL effects include additive (**Effect.a**) and dominant (**Effect.d**) effects.

In the GCIM plot of main-effect QTL detection, the $-\log_{10}(P)$ values are indicated by a curve. If the genetic populations analyzed are F_2, there will be two curves in the GCIM plot, one is for additive effect and another is for dominant effect. All the significant QTLs identified are indicated by the vertical lines (**Figure 1**).

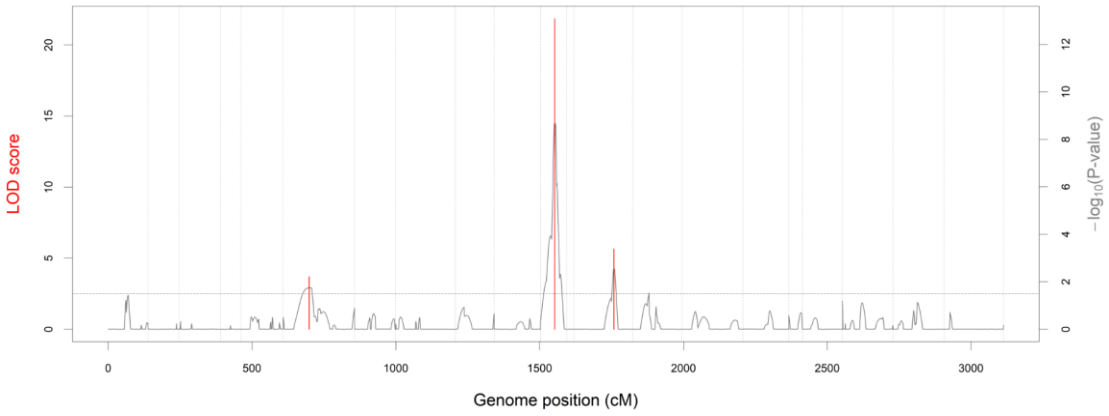

Figure 1. Genome-wide composite interval mapping plot

In QEI detection in IMF_2 or $F_{2:3}$ design with multi-environment datasets, QEI effects include additive-by-environment interaction effects (**Effect.aE1, Effect.aE2, ⋯**), and dominant-by-environment interaction effect (**Effect.dE1, Effect.dE2, ⋯**). There are three Var_Genet columns: Var_Genet (total), Var_Genet_QTL (main-effect QTL) and Var_Genet_QEI (QEI); three LOD score columns: LOD (total), LOD_QTL

(main-effect QTL) and LOD_QEI (QEI); three r2(%) columns: r2(%) (total), r2_QTL(%) (main-effect QTL), r2_QEI(%) (QEI).

In the GCIM-QEI plot, if single environment analysis of the GCIM-QEI method is selected, the parameter is set as "MultiEnv=FALSE", the -$\log_{10}$(P) values of main-effect QTL detection are indicated by a gray curve (Figure 2), and all the significant main-effect QTLs (red) identified are indicated by the red vertical lines.

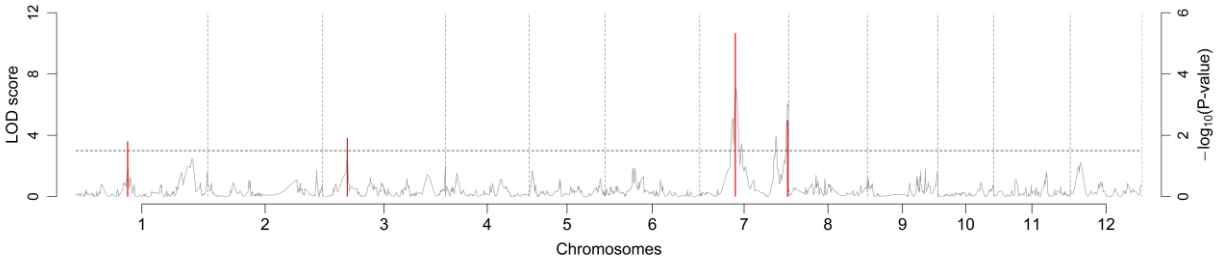

Figure 2. Main-effect QTL detection in single environment analysis using the GCIM-QEI method

If multiple environment analysis of the GCIM-QEI method is selected, the parameter is set as "MultiEnv=TRUE", the -$\log_{10}$(P) values of main-effect QTL and QEI detection are indicated by a gray curve and a light blue curve, respectively (Figure 3), and all the significant QTLs (red) and QEIs (pink) identified are indicated by the red vertical and pink vertical lines, respectively.

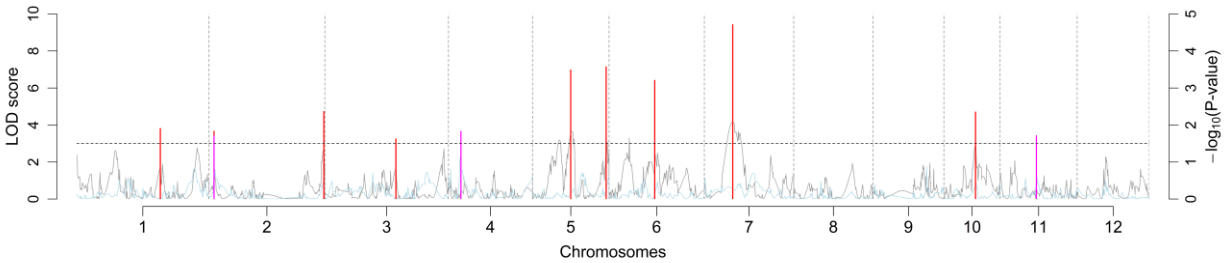

Figure 3. Main-effect QTL and QEI detection in multiple environment analysis using GCIM-QEI method