User Manual for

dQTG.seq

(version 1.0.1)

Pei Li and Yuan-Ming Zhang

(soyzhang@mail.hzau.edu.cn)

Last updated on April 2022

Disclaimer: While extensive testing has been performed by Yuan-Ming Zhang's Lab (Statistical Genomics Lab) at College of Plant Science and Technology of 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 dQTG-seq results with other software programs.

Download website:

https://cran.r-project.org/web/packages/dQTG.seq/index.html

Citation:

Method	References
dQTG-seq1 **dQTG-seq2**	Li P, Li G, Zhang YW, Zuo JF, Liu JY, Zhang YM. A combinatorial strategy to identify various types of QTLs for quantitative traits using extreme phenotype individuals in F_2. ***Plant Communications*** 2022, 3: 100319. doi: 10.1016/j.xplc.2022.100319
smoothLOD	Zhang H, Wang X, Pan Q, Li P, Liu Y, Lu X, Zhong W, Li M, Han L, Li J, Wang P, Li D, Liu Y, Li Q, Yang F, Zhang YM, Wang G, Li L. QTG-seq accelerates QTL fine mapping through QTL partitioning and whole-genome sequencing of bulked segregant samples. ***Molecular Plant*** 2019; 12(3): 426-437.
G'	Magwene PM, Willis JH, Kelly JK. The statistics of bulk segregant analysis using next generation sequencing. ***PLoS Comput Biol*** 2011, 7: e1002255
ΔSNP index	Takagi H, Abe A, Yoshida K, Kosugi S, Natsume S, Mitsuoka C, Uemura A, Utsushi H, Tamiru M, Takuno S, Innan H, Cano LM, Kamoun S, Terauchi R. QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. ***Plant J*** 2013, 74: 174-183
ED	Hill JT, Demarest BL, Bisgrove BW, Gorsi B, Su YC, Yost HJ. MMAPPR: mutation mapping analysis pipeline for pooled RNA-seq. ***Genome Res*** 2013, 23: 687-697

Note: These references are listed in section of References

Funding

This work was supported by the National Natural Science Foundation of China (32070557 and 31871242), the Fundamental Research Funds for the Central Universities (2662020ZKPY017), and Huazhong Agricultural University Scientific & Technological Self-Innovation Foundation (2014RC020).

1. Introduction

1.1 Why dQTG.seq?

dQTG.seq is an R package for quickly detecting quantitative trait gene (QTG), especially, **dQTG.seq2 may be used to identify extremely over-dominant and small-effect genes in F_2**. At present this software package (v1.0.1) includes six methods: dQTG-seq1, dQTG-seq2, smoothLOD, G', ΔSNP index and ED. dQTG.seq v1.0.1 is able to work on the Windows, Linux (desktop), and MacOS platforms.

1.2 Getting started

dQTG.seq is a package that runs in the R software environment, which can be freely downloaded from *https://cran.r-project.org/web/packages/dQTG.seq/index.html*, or requested from maintainer, Dr Yuan-Ming Zhang at College of Plant Science and Technology of Huazhong Agricultural University (soyzhang@mail.hzau.edu.cn).

1.2.1 One-Click installation

Within R environment, the dQTG.seq software can be installed directly using the below command:

```
install.packages("dQTG.seq")
```

1.2.2 Step-by-step installation

1.2.2.1 Install the add-on packages

First, users download seven R packages, including "data.table", "BB", "doParallel", "openxlsx", "qtl", "stringr", "vroom" and "writexl" from CRAN (https://cran.r-project.org/), github (https://github.com/), or google search.

Under the R environment, then, users find "Packages"—"Install package(s) from local files…", select all the above seven packages, and install them offline.

1.2.2.2 Install dQTGseq

Open R GUI, select "Packages"—"Install package(s) from local files…" and then find the dQTG.seq package which you have downloaded on your desktop. Within R environment, launch the dQTGseq by command:

```
library (dQTG.seq)
```

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

1.2.3 Run dQTGseq

Once the software **dQTG.seq** is installed, users may run it using two commands:

*library(**dQTG.seq**)*

dQTG.seq(***) (***: please see § 2.1.2 Example)

If users re-use the dQTG.seq software, users use the above two commands as well.

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

2. Function

2.1 Function dQTG.seq()

2.1.1 Parameter settings

Table 1. Main parameters and their settings in the R function *dQTG.seq()*

Parameter	**Meaning**
dir	Path of inputting and outputting files in your computer: dir="D:/users"
filegen	The name of input data file. filegen="D:/users/BSA.csv"
chr	chr="all": to output the results of all the chromosomes; chr=c(7,8): to output the results of chromosomes 7 and 8.
color	color=c("blue","red"): the blue and red points of smooth values in adjacent chromosomes.
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, CLO=NULL; Users can set own parameters, if CLO is greater than 10, the value is 10, CLO=2.

2.1.2 Example

In Windows platform:

dQTG.seq(dir="D:/users", filegen=" D:/users/BSA.csv", chr="all", color = c("blue","red"), CLO=NULL)

In server:

dQTG.seq(dir="/home", filegen="/home/BSA.csv", chr="all", color=c("blue","red"), CLO=NULL)

In server platform, at present the *.tiff format of plot file is not available.

2.2 Dataset format

2.2.1 Parameter settings

Table 2. Main parameters and their settings in dataset file

Parameter	**Meaning**
Species	The transformation between genetic distance (cM) and genome length (Mb) varies across species, at present fifteen species are directly available once users input the name of species, e.g., Rice. Please see 2.2.2 section.
Data-file-format	Five dataset file formats are available: BSA, Extreme individuals, CIM, ICIM, and GCIM. Please see 2.2.3 section.
Sample-size	The number of individuals in F_2 population, e.g., 1000.
Population-type	Four population types are available: F_2, BC (backcross), DH (doubled haploid), and RIL (recombinant inbred line).
Sampling-fraction (%)	The proportion of extreme low and high individuals, e.g., 20%.
Smooth-method	Four smoothing methods are available: None, AIC, Window size, and Block, while the default is AIC. Please see 2.2.4 section. If users want to change the default value of window size or Block, users may use “**smooth-method=**Window size, 0.5” or “**smooth-method=**Block, 30”. In detail, smoothing method is Window size or Block, and the window size is changed into 0.5 or 30 (the number of markers in a block) (see Table 4). The default number of markers in Block is 20 (see Table 5).
No. of permutations	The number of permutation experiments when determining the threshold of significant QTN, and the default is 300. Please see Table 4.
Figure	**Figure**=False: no figure output; **Figure**=True: the output of figures from different methods. The default is True. Please see Table 4.
Figure-resolution	Figure-resolution=Low: the figure with low resolution; Figure-resolution=High: the figure with high resolution. The default is High. Please see Table 4.
Figure-file-format	**Figure-file-format=** jpeg, png, tiff, or pdf. jpeg indicates the *.jpeg format of figure file. Please see Table 4.

2.2.2 Species and window size

The transformation between genetic distance (cM) and genome length (Mb) varies across species. In this software, the genome lengths (Mb) per genetic distance (cM) in fifteen species are listed in the below table, and can be directly available once users input the name of species. This value is regarded as window size in BSA. For example, in *Arabidopsis*, 1 cM is approximately equivalent to 0.2083 Mb on average, so its default of the window size is 0.2083 Mb.

Table 3. Total linkage genetic distances (cM), genome sizes and their relationships in fifteen species

No.	**Species**	**Total genetic distance (cM)**	**Total genome length (Mb)**	**Mb/cM**
1	*Arabidopsis*	600 (Meinke et al., 2003)	125 (Arabidopsis genome Initiative., 2000)	0.2083
2	Cucumber	1384 (Zhou et al., 2015)	350 (Huang et al., 2009)	0.2529
3	Maize	1879 (Pan et al., 2016)	2106 (Jiao et al., 2017)	1.1208
4	*Brassica juncea*	2515 (Raman et al., 2014)	955 (Yang et al., 2016)	0.3797
5	*Brassica napus*	2515 (Raman et al., 2014)	850 (Chalhoub et al., 2014)	0.3380
6	Rice	2840 (Jiang et al., 2020)	390 (Zhang et al., 2018)	0.1373
7	Tobacco	3270 (Bindler et al., 2011)	2700 (Sierro et al., 2014)	0.8257
8	Tomato	1467 (Shirasawa et al., 210)	1200 (Bolger et al., 2014)	0.8180
9	Wheat	2780 (Yang et al., 2018)	17000 (Brenchley et al., 2012)	6.1151
10	Yeast	4900 (Bloom et al., 2015)	12 (Foury et al., 1998)	0.0024
11	*Glycine soja*	2623 (Lee et al., 2020)	939 (Kim et al., 2010)	0.3580
12	*Glycine max*	2524 (Zuo et al., 2019)	1100 (Schmutz et al., 2010)	0.4358
13	*Gossypium hirsutum* L	3854 (Huang et al., 2017)	2186 (Huang et al., 2017)	0.5672
14	*Gossypium barbadense*	2727 (Chang et al., 2008)	2120 (Wang., et al 2019)	0.7774
15	*Brassica pekinensis*	1688 (Liu et al., 2019)	529 (Wang., et al 2011)	0.3134

1. Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 2000, 408(6814): 796-815.
2. Bindler G, Plieske J, Bakaher N, Gunduz I, Ivanov N, Van der Hoeven R, Ganal M, Donini P. A high density genetic map of tobacco (*Nicotiana tabacum* L.) obtained from large scale microsatellite marker development. Theor Appl Genet 2011, 123(2): 219-230.
3. Bloom JS, Kotenko I, Sadhu MJ, Treusch S, Albert FW, Kruglyak L. Genetic interactions contribute less than additive effects to quantitative trait variation in yeast. Nat Commun 2015, 6: 8712.
4. Bolger A, Scossa F, Bolger ME, Lanz C, Maumus F, Tohge T, Quesneville H, Alseekh S, Sørensen I, Lichtenstein G, et al. The genome of the stress-tolerant wild tomato species *Solanum pennellii*. Nat Genet 2014, 46(9): 1034-1038.
5. Brenchley R, Spannagl M, Pfeifer M, Barker GL, D'Amore R, Allen AM, McKenzie N, Kramer M, Kerhornou A, Bolser D, et al. Analysis of the bread wheat genome using whole-genome shotgun sequencing. Nature 2012, 491(7426): 705-710.
6. Chalhoub B, Denoeud F, Liu S, Parkin IA, Tang H, Wang X, Chiquet J, Belcram H, Tong C, Samans B, et al. Plant genetics. Early allopolyploid evolution in the post-Neolithic *Brassica napus* oilseed genome. Science 2014, 345(6199): 950-953.
7. Chang Yang, Wangzhen Guo, Guoying Li, Feng Gao, Shunshun Lin, Tianzhen Zhang. QTLs mapping for verticillium wilt resistance at seedling and maturity stages in *Gossypium barbadense* L. Plant Sci 2008, 174(2008): 290-298.
8. Foury F, Roganti T, Lecrenier N, Purnelle B. The complete sequence of the mitochondrial genome of saccharomyces cerevisiae. FEBS Lett 1998, 440(3): 325-31.
9. Huang C, Nie X, Shen C, You C, Li W, Zhao W, Zhang X, Lin Z. Population structure and genetic basis of the agronomic traits of upland cotton in China revealed by a genome-wide association study using high-density SNPs. Plant Biotechnol J 2017, 15(11): 1374-1386.
10. Huang S, Li R, Zhang Z, Li L, Gu X, Fan W, Lucas WJ, Wang X, Xie B, Ni P, et al. The genome of the cucumber, Cucumis sativus L. Nat Genet 2009, 41(12): 1275-1281.
11. Jiang S, Yang C, Xu Q, Wang L, Yang X, Song X, Wang J, Zhang X, Li B, Li H, Li Z, Li W. Genetic dissection of germinability under low temperature by building a resequencing linkage map in *japonica* rice. Int J Mol Sci 2020, 21(4): 1284.
12. Jiao Y, Peluso P, Shi J, Liang T, Stitzer MC, Wang B, Campbell MS, Stein JC, Wei X, Chin CS, et al. Improved maize reference genome with single-molecule technologies. Nature 2017, 546(7659): 524-527.
13. Kim MY, Lee S, Van K, Kim TH, Jeong SC, Choi IY, Kim DS, Lee YS, Park D, Ma J, et al. Whole-genome sequencing and intensive analysis of the undomesticated soybean (*Glycine soja* Sieb. and Zucc.) genome. Proc Natl Acad Sci U S A 2010, 107(51): 22032-22037.
14. Lee K, Kim MS, Lee JS, Bae DN, Jeong N, Yang K, Lee JD, Park JH, Moon JK, Jeong SC. Chromosomal features revealed by comparison of genetic maps of *Glycine max* and *Glycine soja*. Genomics 2020, 2(2): 1481-1489.
15. Liu S, Wang R, Zhang Z, Li Q, Wang L, Wang Y, Zhao Z. High-resolution mapping of quantitative trait loci controlling main floral stalk length in Chinese cabbage (*Brassica rapa* L. ssp. pekinensis). BMC Genomics 2019, 20(1): 437.
16. Meinke DW, Meinke LK, Showalter TC, Schissel AM, Mueller LA, Tzafrir I. A sequence-based map of Arabidopsis genes with mutant phenotypes. Plant Physiol 2003, 131(2): 409-418.
17. Pan Q, Li L, Yang X, Tong H, Xu S, Li Z, Li W, Muehlbauer GJ, Li J, Yan J. Genome-wide recombination dynamics are associated with phenotypic variation in maize. New Phytol 2016, 210(3): 1083-1094.
18. Raman H, Dalton-Morgan J, Diffey S, Raman R, Alamery S, Edwards D, Batley J. SNP markers-based map construction and genome-wide linkage analysis in *Brassica napus*. Plant Biotechnol J 2014, 12(7): 851-860.
19. Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, et al. Genome sequence of the palaeopolyploid soybean. Nature 2010, 463(7278): 178-183.
20. Shirasawa K, Isobe S, Hirakawa H, Asamizu E, Fukuoka H, Just D, Rothan C, Sasamoto S, Fujishiro T, Kishida Y, et al. SNP discovery and linkage map construction in cultivated tomato. DNA Res 2010, 17(6): 381-391.
21. Sierro N, Battey JN, Ouadi S, Bakaher N, Bovet L, Willig A, Goepfert S, Peitsch MC, Ivanov NV. The tobacco genome sequence and its comparison with those of tomato and potato. Nat Commun 2014, 5: 3833.
22. Wang M, Tu L, Yuan D, Zhu D, Shen C, Li J, Liu F, Pei L, Wang P, Zhao G, et al. Reference genome sequences of two cultivated allotetraploid cottons, *Gossypium hirsutum* and *Gossypium barbadense*. Nat. Genet. 2019, 51(2): 224-229.
23. Wang X, Wang H, Wang J, Sun R, Wu J, Liu S, Bai Y, Mun JH, Bancroft I, Cheng F, et al. The genome of the mesopolyploid crop species *Brassica rapa*. Nat Genet 2011, 43(10): 1035-1039.
24. Yang J, Liu D, Wang X, Ji C, Cheng F, Liu B, Hu Z, Chen S, Pental D, Ju Y, et al. The genome sequence of allopolyploid *Brassica juncea* and analysis of differential homoeolog gene expression influencing selection. Nat Genet 2016, 48(10): 1225-1232.
25. Yang Q, Yang Z, Tang H, Yu Y, Chen Z, Wei S, Sun Q, Peng Z. High-density genetic map construction and mapping of the homologous transformation sterility gene (*hts*) in wheat using GBS markers. BMC. Plant Biol 2018, 18(1): 301.
26. Zhang Q, Liang Z, Cui X, Ji C, Li Y, Zhang P, Liu J, Riaz A, Yao P, Liu M, et al. N6-Methyladenine DNA methylation in *Japonica* and *Indica* rice genomes and its association with gene expression, plant development, and stress responses. Mol. Plant 2018, 11(12):

1492-1508.

27. Zhou Q, Miao H, Li S, Zhang S, Wang Y, Weng Y, Zhang Z, Huang S, Gu X. A sequencing-based linkage map of cucumber. Mol Plant 2015, 8(6): 961-963.
28. Zuo JF, Niu Y, Cheng P, Feng JY, Han SF, Zhang YH, Shu G, Wang Y, Zhang YM. Effect of marker segregation distortion on high density linkage map construction and QTL mapping in soybean (*Glycine max* L.). Heredity 2019, 123(5): 579-592.

In the parameter setup of smoothing method, if users select "Window size" or "Block", the default value of the window size in *Arabidopsis* is 0.2083. Of course, other values may be set up, e.g., 0.50.

2.2.3 dataset formats

BSA format for the dataset file (filegen; Table 4**)** The parameters in lines 1 to 10 can be found in Table 4 (2.2.1 section). In line 11, the first column, named "Marker", presents marker name; the second column, "Chromosome" presents chromosome number (number) of the marker; the third column, named "Position (bp)", stands for the positions (bp) of markers on its chromosome; the fourth to seventh column, named "AL", "aL", "AH" and "aH", stands for the numbers of alleles A and a in low (L) and high (H) pools, respectively. Details can be found in the BSA.csv file in the folder of ".../dQTGseq/inst/extdata" or Table 4. The alleles of parent P$_1$ or reference genome are defined as A in our software.

Table 4. The BSA format of dataset file to be input

Species	Maize					
Data-file-format	BSA					
Sample-size	3120					
Population-type	F2					
Sampling-fraction (%)	20					
Smooth-method	Window size, 0.5					
No. of permutations	300					
Figure	TRUE					
Figure-resolution	High					
Figure-file-format	png					
Marker	Chromosome	Position (bp)	AL	aL	AH	aH
1	1	35496	105	73	63	97
2	1	55610	110	66	51	122
3	1	118174	108	67	57	94
⋮	⋮	⋮	⋮	⋮	⋮	⋮
4	20	150980515	120	69	75	118

Extreme individuals format for the dataset file (filegen; Table 5**)** The parameters in lines 1 to 10 can be found in Table 5 (2.2.1 section). In line 11, the first column, named "Marker", presents marker name; the second column, "Chromosome" presents chromosome number of the marker; the third column, named “Position (bp)”, stands for the positions (bp) of markers on its chromosome; the fourth column to the end are individual names, where the numbers 2, 1, and 0 stand for the genotypes AA, Aa and aa in the extreme low and high pools, respectively; “-” indicates the missing or unknown genotypes. The phenotypic values are located on the trait row, and each trait is presented on one row. On each row, the first and second columns are empty, followed by “trait”, phenotypic values for all the individuals. Details can be found in the Extreme individuals.csv file in the folder of “…/dQTGseq/inst/extdata” or Table 5.

Table 5. The Extreme individual format of dataset file to be input

Species	Rice								
Data-file-format	Extreme individuals								
Sample-size	246								
Population-type	F2								
Sampling-fraction (%)	20								
Smooth-method	Block 30								
No. of permutations	300								
Figure	TRUE								
Figure-resolution	High								
Figure-file-format	png								
Marker	Chromosome	Position (bp)	Low1	Low2	Low3	⋯	High47	High48	High49
Bin1	1	0	1	2	0	⋯	1	1	1
Bin2	1	565000	1	2	0	⋯	1	1	1
Bin3	1	599000	1	2	-	⋯	1	1	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Bin4	12	27016000	1	2	1	⋯	1	1	0
		trait	45.170	62.730	67.373	⋯	146.102	150.858	152.878

CIM format for the dataset file (filegen; Table 6) If users have adopted composite interval mapping (CIM) to identify QTLs for quantitative traits in F_2 via the WinQTLCart software, its dataset file with the *.csv format and minor modifications is available in our software, and the details can be found in Table 6. In details, the information of main parameters in lines 1 to 9 should be added, which can be found in Table 6 or 2.2.1 section. The line 10, named “Marker”, presents marker names; the line 11, named “Position(bp)”, stands for the positions (bp) of markers on chromosome; the line 12, “Chromosome” presents chromosome number for markers; in the line 13, the first column is the individual name, the other columns present marker genotypes AA, Aa, and aa in low and high pools, in which their numbers are 2, 1, and 0, respectively. The phenotypic values are located on the rightmost columns, and each trait is presented on one column. In each trait column, the first row is “trait”, which is in the same row of “Chromosome”, the second to last rows are phenotypic values of quantitative trait, such as 3.9308. Here missing marker genotypes are indicated by “-1”, while missing trait values are indicated by “.” (dot). Details can be found in the CIM_Format_F2.csv file in the folder of “…/dQTGseq/inst/extdata” or Table 6.

ICIM format for the dataset file (filegen; Table 7) If users have adopted inclusive composite interval mapping (ICIM) to identify QTLs for quantitative traits in F_2 via the QTL IciMapping software, its dataset file with the *.xlsx format and minor modifications is available in our software, and the details can be found in Table 7. In this table, there are four sheets: “GeneralInfo”, “LinkageMap”, “Genotype”, and “Phenotype”. In the first sheet (Table 7.1), origin information has been deleted, and new information, such as main parameters in lines 1 to 9, need to be added, which can be found in Table 7.1.

Table 6. The CIM format of dataset file to be input

Species	Rice								
Data-file-format	CIM								
Population-type	F2								
Sampling-fraction (%)	20								
Smooth-method	AIC								
No. of permutations	300								
Figure	TRUE								
Figure-resolution	High								
Figure-file-format	png								
Marker	Bin1	Bin2	Bin3	Bin4	Bin5	Bin6	···	Bin1619	
Position(bp)	0	565000	599000	922000	1075000	1147000	···	27016000	
Chromosome	1	1	1	1	1	1	···	12	trait
F001	2	2	2	-1	2	2	···	1	3.9308
F002	0	0	0	0	0	0	···	0	-0.9122
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
F360	1	1	2	2	2	2	···	1	.

Table 7.1. The first sheet of ICIM dataset to be input

Species	Rice
Data-file-format	ICIM
Population-type	F2
Sampling-fraction (%)	20
Smooth-method	AIC
No. of permutations	300
Figure	TRUE
Figure-resolution	High
Figure-file-format	png

In the second sheet “LinkageMap”, the details are listed in Table 7.2. The first column, named “Marker”, presents marker names; the second column, named “Chromosome” presents chromosome number of markers; the third column, named “Position(bp)”, stands for the physical positions (bp) of markers on genome.

In the third sheet "Genotype", the details are listed in Table 7.3. The first column, named "Marker" (the first row), presents marker names (other rows); the second to last columns, are individual's genotypes, the first row is individual's names, such as "F001", the other rows are individual's genotypes, the numbers 2, 1, 0 stand for genotypes AA, Aa, and aa, respectively, and missing marker genotypes are indicated by "-1".

Table 7.2. The second sheet of ICIM dataset to be input

Marker	Chromosome	Position(bp)
Bin1	1	0
Bin2	1	565000
Bin3	1	599000
⋮	⋮	⋮
Bin1619	12	27016000

Table 7.3. The third sheet of ICIM dataset to be input

Marker	F001	F002	F003	F004	···	F360
Bin1	2	0	0	1	···	1
Bin2	2	0	0	1	···	1
Bin3	2	0	0	-1	···	2
⋮	⋮	⋮	⋮	⋮	···	⋮
Bin5	2	0	0	1	···	2

In the four sheet "Phenotype", the details are listed in Table 7.4. Each trait is presented on one row. In each row, the first column is "trait", the others are phenotypic values of quantitative trait, and "NA" specifies missing phenotypic value.

Table 7.4. The fourth sheet of ICIM dataset to be input

trait	118.97	101.33	109.5	NA	···	97.494

GCIM format for the dataset file (filegen; Table 8) If users have adopted genome-wide composite interval mapping (GCIM) to identify QTLs for quantitative traits in F_2 via the QTL.gCIMapping software, its dataset file with the *.csv format and minor modifications is available in our software, and the details can be found in Table 8. In details, the information of main parameters in lines 1 to 9 in the first and second columns should be added, which can be found in Table 8. The line 10 in the first column, named "Marker", presents marker names; and the eleventh to last lined in the first column stand for marker names, such as "Bin1". The tenth to last rows in the second column are chromosome information, the tenth row is "Chromosome", and the others are chromosome number, such as "12". The tenth to last rows in the third column are the information of marker positions (bp) on genome, the tenth row is "Position(bp)", the others except for the last one are marker positions (bp) on genome, such as "599000", and the last one is "trait". The tenth to last rows in the fourth to last columns are genotypic and phenotypic values of each individual, the tenth row is individual's name, such as "F001", the others, except for trait rows, are genotypic number, where the numbers 2, 1, 0 stands for genotypes AA, Aa, and aa, respectively, and the values in trait rows are phenotypic values. The missing or unknown genotypes are indicated by "-", while missing phenotypic values are indicated by "NA". Details can be found in Table 8 or the GCIM_Format_F2.csv file in ".../dQTGseq/inst/extdata" folder.

Table 8. The GCIM format of GCIM dataset file to be input

Species	Rice							
FileFormat	GCIM							
Population	F2							
Sampling fraction (%)	20							
SmoothMethod	AIC							
Permutation times	300							
DrawPlot	TRUE							
Resolution	High							
Plotformat	png							
Marker	**Chromosome**	**Position(bp)**	**F001**	**F002**	**F003**	**F005**	**…**	**F360**
Bin1	1	0	1	2	2	0	…	0
Bin2	1	565000	1	2	-	0	…	0
Bin3	1	599000	1	2	2	0	…	1
Bin4	1	922000	1	2	2	0	…	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Bin1619	12	27016000	0	2	0	0	…	0
		trait	13.16	1.89	9.08	NA	…	NA

2.2.4 Population

The dQTG.seq package can analyze datasets in F_2, BC, DH, and RIL populations. If the genotypes of extreme low and high individuals in F_2 are unknown, the allelic datasets in low and high pools may be analyzed using dQTG-seq1, SmoothLOD, G', ΔSNP, and ED methods. If the genotypes of extreme low and high individuals in F_2 are known, such as QTL mapping in F_2, the allelic and genotypic datasets can be analyzed using dQTG-seq2, SmoothLOD, G', ΔSNP, and ED methods. In BC, DH, and RIL populations, the allelic datasets can be analyzed using SmoothLOD, G', ΔSNP, and ED methods.

2.2.5 Functions or methods in our dQTG.seq software

In the dQTG.seq software package, there are six methods available, including dQTG-seq1, dQTG-seq2, SmoothLOD, G', ΔSNP, and ED, which are in details described as below.

The dQTG-seq1 method

The observed numbers of read counts of marker alleles (Q (j=1) and q (j=2)) in extreme low (i=1) and high (i=2) pools are used by the dQTGseq1 method to predict the numbers of read counts of marker genotypes (QQ (j=1), Qq (j=2), and qq (j=3)) in the two pools. Thus, the observed allele and predicted genotype numbers are used to calculate a new statistic G_w

$$G_w = \frac{G_1}{G_1 + G_2} \times G_1 + \frac{G_2}{G_1 + G_2} \times G_2 \tag{1}$$

where G_1 and G_2 are standard G statistic of Magwene et al. (2011), $G_1 = \sum_{i=1}^{2}\sum_{j=1}^{2} n_{Qij} \ln\left[n_{Qij} / E(n_{Qij}) \right]$ is from allelic read numbers n_{Qij} and their expectations $E(\bullet)$, and $G_2 = \sum_{i=1}^{2}\sum_{j=1}^{3} n_{Gij} \ln\left[n_{Gij} / E(n_{Gij}) \right]$ is from genotypic read counts n_{Gij} and their expectations $E(\bullet)$. The G_w statistic is used to identify QTL for quantitative trait, where "d" indicates dominant effect of the QTL. In this method, the default smooth technique is the "Windowsize" method of Magwene et al. (2011). The file format of the dQTG-seq1 method is the BSA dataset format in Table 4 (see 2.2.3 section).

For dQTG-seq2 method

Although dQTG-seq1 has relatively higher power in QTL detection than existing BSA methods, it may fail in the detection of high degree of dominance and/or small effect QTLs. In other words, no loci are detected in secondary F$_2$ and no ideal loci are detected in primary F$_2$. In this case, the dQTG-seq2 method uses the observed numbers of marker alleles and genotypes in extreme low and high pools to calculate G_w. If the genotypes of each extreme phenotypic individual are known, the file format of the dQTG-seq2 method is Extreme individuals format in Table 5 (see 2.2.3 section). In this method, the "Windowsize" method of Magwene et al. (2011) is recommended to smooth G_w. If the genotypes of all the F$_2$ individuals are known, the file format of the dQTG-seq2 method should be the CIM, ICIM, and GCIM formats (Tables 6-8; see 2.2.3 section).

The SmoothLOD method

The statistic "*SmoothLOD*" is proposed in Zhang et al. (2019). The symbols for the numbers of read counts of marker alleles "a" and "A" at low and high pools are defined in the dQTGseq1 method of 2.2.5 section, let $n_L = n_{Q11} + n_{Q12}$, $\hat{f}_L = n_{Q11}/n_L$, $n_H = n_{Q21} + n_{Q22}$, and $\hat{f}_H = n_{Q21}/n_H$. To test no QTL at the putative locus, $H_0 : f_L = f_H = 0.5$ and the LOD statistic is defined as

$$LOD = \log_{10} \frac{C_{n_L}^{n_{Q11}} \left(\hat{f}_L\right)^{n_{Q11}} \left(1-\hat{f}_L\right)^{n_{Q12}} C_{n_H}^{n_{Q21}} \left(\hat{f}_H\right)^{n_{Q21}} \left(1-\hat{f}_H\right)^{n_{Q22}}}{C_{n_L}^{n_{Q11}} C_{n_H}^{n_{Q21}} \left(1/2\right)^{n_L+n_H}} \tag{2}$$

(Zhang et al. 2019), where $C_{n_L}^{n_{Q11}} = \frac{n_L\,!}{n_{Q11}\,!\,n_{Q12}\,!}$ and $C_{n_H}^{n_{Q21}} = \frac{n_H\,!}{n_{Q21}\,!\,n_{Q22}\,!}$. In this method, the default smooth technique is the "Windowsize" method of Magwene et al. (2011). The file format of the SmoothLOD method is BSA format (Table 4; see 2.2.3 section).

The G' method

Magwene et al. (2011) proposed the statistic G' in BSA. The symbols for the numbers of read counts of marker alleles "a" and "A" at low and high pools has been defined in the dQTGseq1 method (see 2.2.5 section). Thus, standard *G* statistic is defined as

$$G = \sum_{i=1}^{2}\sum_{j=1}^{2} n_{Qij} \ln\left[n_{Qij} / E(n_{Qij})\right] \tag{3}$$

using allelic read numbers n_{Qij} and their expectations $E(\bullet)$. Using all the *G* statistic values in one window, smooth statistic *G'* is calculated via the "Windowsize" method (Magwene et al. 2011). The file format of G' method is BSA format (Table 4; see 2.2.3 section).

The ΔSNP method

The △SNP index method is presented in Abe et al. (2012) and Takagi et al. (2013). The statistic, Δ(SNP-index), is allele frequency difference between low and high pools. Using this method, candidate genes for trait of interest may be located around the peaks on the curve of allele frequency difference on genome positions. In this method, the "Block" method is recommended by *** to smooth △SNP index. The dataset file format is BSA format (Table 4) or in 2.2.3 section.

The ED method

The Euclidean distance (ED) statistic is defined as

$$ED = \sqrt{(A_L - A_H)^2 + (C_L - C_H)^2 + (G_L - G_H)^2 + (T_L - T_H)^2} \tag{4}$$

in mutation mapping analysis pipeline for pooled RNA-seq (MMAPPR) (Hill et al. 2013), where A_L and A_H are the frequencies of allele A (base of SNP) in extreme low and high pools, respectively, and the others have similar meanings. Here two key techniques should be noticed in real data analysis. One is to find the best index number of the ED statistic, and another is to conduct Loess regression analysis (see Smooth method). The dataset file format is BSA format (Table 4) or in 2.2.3 section.

2.2.6 Smooth-method

Once users calculate each statistic values in genome-wide scanning via the above-mentioned methods, the peaks of the curve may be not such obvious. In this case, it is necessary to conduct smoothing analysis. Actually, each BSA method has itself smooth technique, such as Loess regression + AIC for the ED method, "Block" for the △SNP index method, and "Windowsize" for the G', SmoothLOD and dQTG-seq (G_w) methods, which are the default for these methods. However, users may select other smooth approaches in real data analysis. Thus, in our software there are several options available.

"None": no smoothing.

"Default": the default smooth techniques described above.

"AIC": in the smoothing methods, Loess regression is first used, and then AICc is used to select the best model of optimizing Loess fit curves (Hurvich et al., 1998).

"Window size": The smooth statistic is calculated as below.

Firstly, for the focal marker X_0, there is one window with the intervals $(X_0 - x,\ X_0 + x)$, where x is obtained from Table 3. D_j for the *j*th marker in the window is a standardized distance, which is calculated from

$$D_j = \left|X_j - X_0\right| / x \qquad X_j \in (X_0 - x,\ X_0 + x) \text{ ....(5)}$$

Clearly, D_j is 0 at the position of focal marker, while D_j is 1 at the edge of the window.

Based on Nadaraya-Watson kernel regression of Cleveland et al. (1979), secondly, the weight for the *j*th marker in the window, k_j, is calculated from

$$k_j = \left(1 - D_j^3\right)^3 / S_W \tag{6}$$

where $S_W = \sum_j \left(1 - D_j^3\right)^3$.

Finally, smooth statistic (*SmoothG*$_w$) value of the focal marker X_0 is a weighted average of original statistic (G_w) values for all the SNPs in the window, which is indicated by

$$\mathrm{S}moothG_{w(X_0)} = \sum_{j \text{ in a window}} k_j G_{w(j)} \tag{7}$$

(Cleveland et al. 1979). The same is true for statistics *LOD*, *G*, *△SNP* index, and ED^k.

In general, window size depends on the physical distance of a 1 cM long in the species, for example, 0.44 Mb/cM for *Glycine max*.

"Block": In a block with a certain number of markers, the default number of markers (*n*) is 20, the values of the statistic for *n* adjacent markers in the block are averaged. The number *n* varies across species (Pool et al. 2016).

2.3 Permutation test

The critical values of various methods for significant QTLs at the 0.10 probability levels were determined by permutation experiments. Various population datasets (F_2, BC, DH, and RIL) are simulated by the R package "qtl", sample sizes and sampling fraction are obtained from raw data.

2.4 Result

The format of dataset to be output. Once the running of software dQTGseq ended, the "results" files would appear on the directory, which was set up by users before running the software. The results for each trait include three files: "all_result.csv", "significant_result.csv", and a plot.

In the all_result.csv file, there are thirteen columns, as shown below.

Marker: marker name.

Chromosome: Chromosome, an integer number.

Position: The position (bp) of markers on the chromosome.

G_w: The value of statistic G_w calculated by the dQTGseq1 or dQTGseq2 method.

Smooth_G_w: smooth G_w value of one marker via the windowsize method.

LOD: The value of statistic *LOD* calculated by the smoothLOD method.

Smooth_LOD: smooth *LOD* value of one marker via the windowsize method.

G: The value of standard *G* statistic.

G': smooth G statistic value from all the markers in one window.

deltaSNP: statistic *deltaSNP* value calculated by deltaSNP index method.

Smooth_deltaSNP: smooth *deltaSNP* index value.

ED: The ED statistic value calculated by the ED method.

Smooth_ED: smooth *ED* statistic value.

In the significant_result.csv file, there are five sheets, and each sheet lists the results for one method (see 2.2.4 section). In each sheet, there are sixth columns, including "Marker", "Chromosome", "Position(bp)", the statistic values, smooth statistic values, and the critical value of significant QTL for the dQTGseq1 method.

In the output plot, there are five sub-plots from five methods. In each sub-plot, users may modify some parameters, such as colors (see 2.1.2 section). The figure file format can be found in 2.2.1 section.

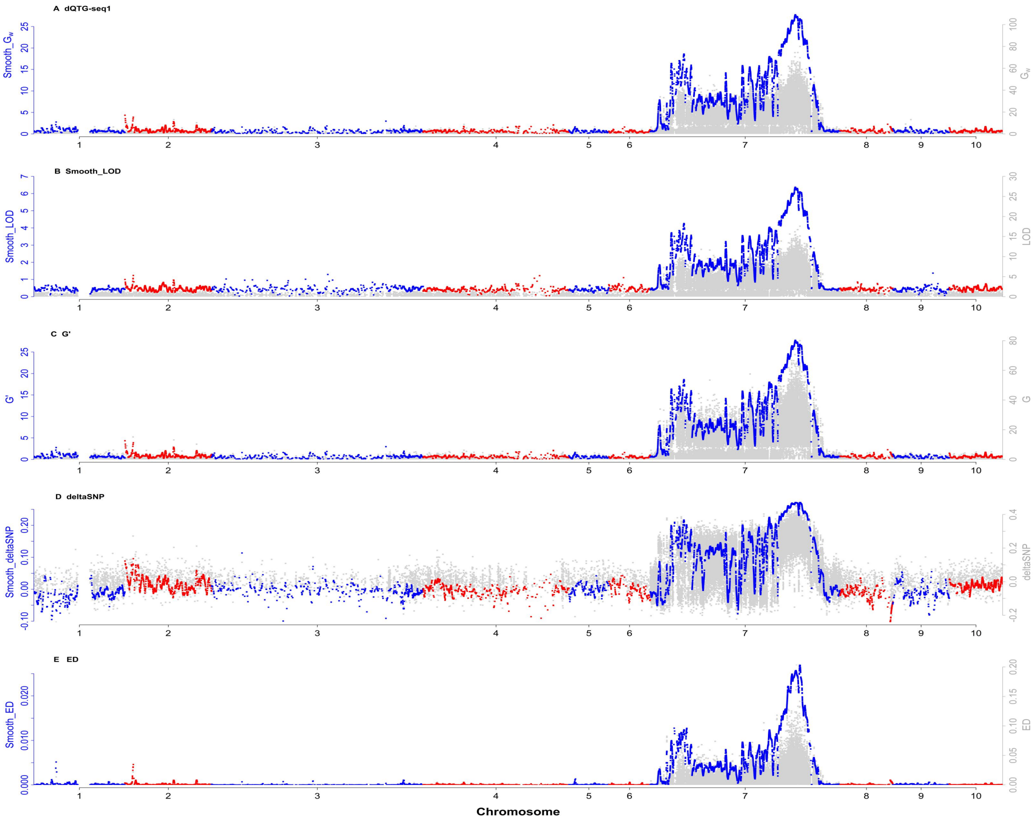

Figure 1. Five sub-plots in an output figure file

3. References

1. Abe A, Kosugi S, Yoshida K, Natsume S, Takagi H, Kanzaki H, Matsumura H, Yoshida K, Mitsuoka C, Tamiru M, Innan H, Cano L, Kamoun S, Terauchi R. Genome sequencing reveals agronomically important loci in rice using MutMap. ***Nat Biotechnol*** 2012, 30: 174–178.
2. Cleveland WS. Robust locally weighted regression and smoothing scatterplots. ***J. Amer. Stat. Assoc.*** 1979, 74: 829–826.
3. Hill JT, Demarest BL, Bisgrove BW, Gorsi B, Su YC, Yost HJ. MMAPPR, mutation mapping analysis pipeline for pooled RNA-seq. ***Genome Res*** 2013, 23: 687–697.
4. Li P, Li G, Zhang YW, Zuo JF, Liu JY, Zhang YM. A combinatorial strategy to identify various types of QTLs for quantitative traits using extreme phenotype individuals in F_2. ***Plant Commun*** 2022, 3: 100319. https://doi.org/10.1016/j.xplc.2022.100319
5. Magwene PM, Willis JH, Kelly JK. The statistics of bulk segregant analysis using next generation sequencing. ***PLoS Comput Biol*** 2011, 7: e1002255.
6. Pool JE. Genetic mapping by bulk segregant analysis in *Drosophila*: experimental design and simulation-based inference. ***Genetics*** 2016, 204(3): 1295–1306.
7. Takagi H, Abe A, Yoshida K, Kosugi S, Natsume S, Mitsuoka C, Uemura A, Utsushi H, Tamiru M, Takuno S, Innan H, Cano LM, Kamoun S, Terauchi R. QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. ***Plant J*** 2013, 74: 174–183.
8. Zhang H, Wang X, Pan Q, Li P, Liu Y, Lu X, Zhong W, Li M, Han L, Li J, Wang P, Li D, Liu Y, Li Q, Yang F, Zhang YM, Wang G, Li L. QTG-Seq accelerates QTL fine mapping through QTL partitioning and whole-genome sequencing of bulked segregant samples. ***Mol Plant*** 2019, 12: 426–437.