User Manual for

DeepBSA

(Version 1.4)

MASKED BECAUSE OF DOUBLE BLIND PEER REVIEW

Last updated on November 15, 2022

Introduction

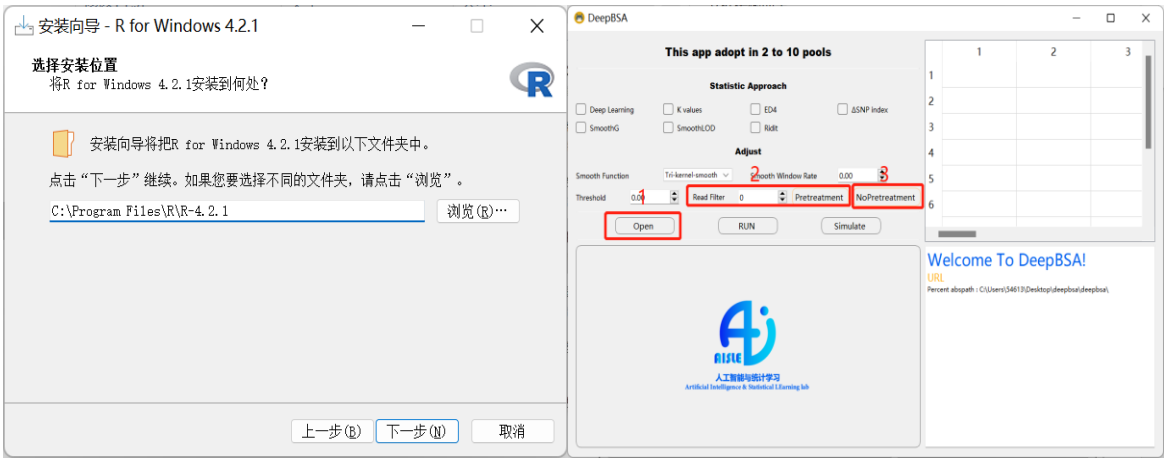
DeepBSA is a novel bulked segregant analysis (BSA) software for the dissection of complex traits. Two brand-new algorithms are developed in DeepBSA named deep learning (DL) and k-value (K), which can be applied on different number (at least 2) of bulked pools. DeepBSA also integrates five widely used algorithms - ED^4, G', ΔSNP-index, SmoothLOD and Ridit, and DL performs better than them with absolute bias and signal-noise-ratio in our simulation. Overall, DeepBSA provides a user-friendly, OS-compatible, and all-in-one pipeline, which do not need sophisticated bioinformatics skills for BSA.

Operation process

1. Installation

DeepBSA is available for both Windows and Linux, and the download link is: http://zeasystemsbio.hzau.edu.cn/Tools. The following process are offered for DeepBSA in windows.

R is required for DeepBSA running. After DeepBSA_windows_v1.4 is downloaded and unzipped, users can enter the folder named *deepbsa* and double-click *R-4.2.1-win.exe* to install R in "C:\Program Files\". Then double-clicking *deepbsa.exe*, the interface of software will be as follows if R is installed properly.



2. Input

Users can click **Open** to load the input file (step 1). Currently, only standard VCF and corresponded CSV are supported (The example input files are provided in the folder named *Demo*).

Note: (1) The VCF is available without header, but it is recommended that only chromosomes are retained, and scaffolds should be removed. (2) Only mixed pool data is required in VCF, parental data should be removed. Meanwhile, if there are multiple pools, they should be ordered according to phenotype. (3) Compressed

file is unavailable.

3. Data pretreatment

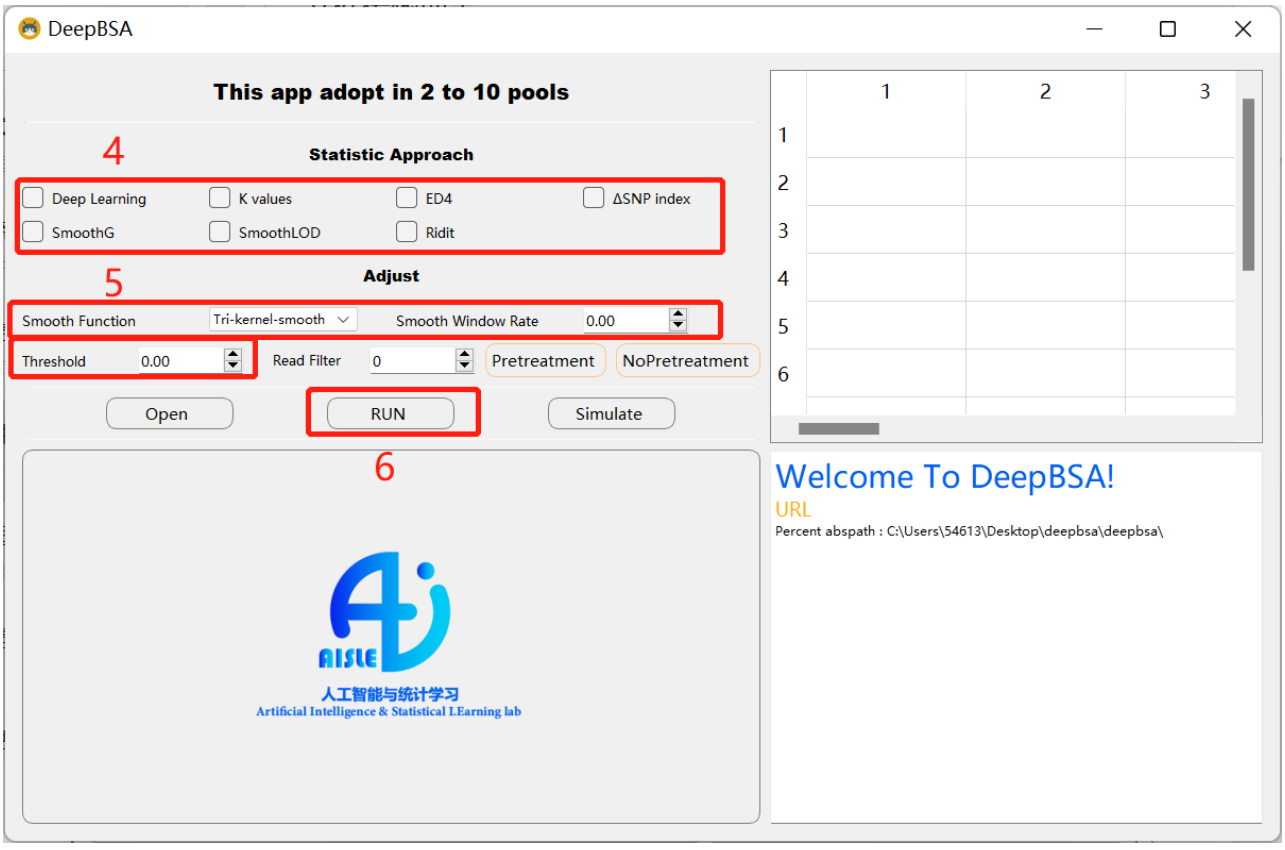
The software provides **pretreatment** for input data. Preprocessing contains built-in programs to remove low-quality SNPs, and to remove SNPs whose read number is below the custom number (custom number is recommended at half the sequencing coverage) (step 2). This process will take some time, which depends on the data size and computer performance. Meanwhile, **nopretreatment** is available (step 3).

4. Methods selection and running

Seven algorithms are provided for QTL detection, users can select one method at once (step 4). Meanwhile, three custom parameters are available - **Smooth Function, Smooth Window Rate** and **Threshold** (step 5).

- Smooth Function: three fitting methods are available - Tri-kernel-smooth, LOWESS and Moving Average. The default is Tri-kernel-smooth.
- Smooth Window Rate: window size for fitting. The default is set to 0, which represents the best parameter by automatic tuning with loess.as (). An optional value of 0-1, such as 0.5, represents using 50% SNPs for local fitting. 0.1 is an experience value while the result of 0 is not good.
- Threshold: the candidate QTL regions will be extracted where the fitted line exceeds this value. The default is set to 0, which represents the threshold is three standard deviations above the genome-wide median.

After setting, click **RUN** to map QTLs (step 6).

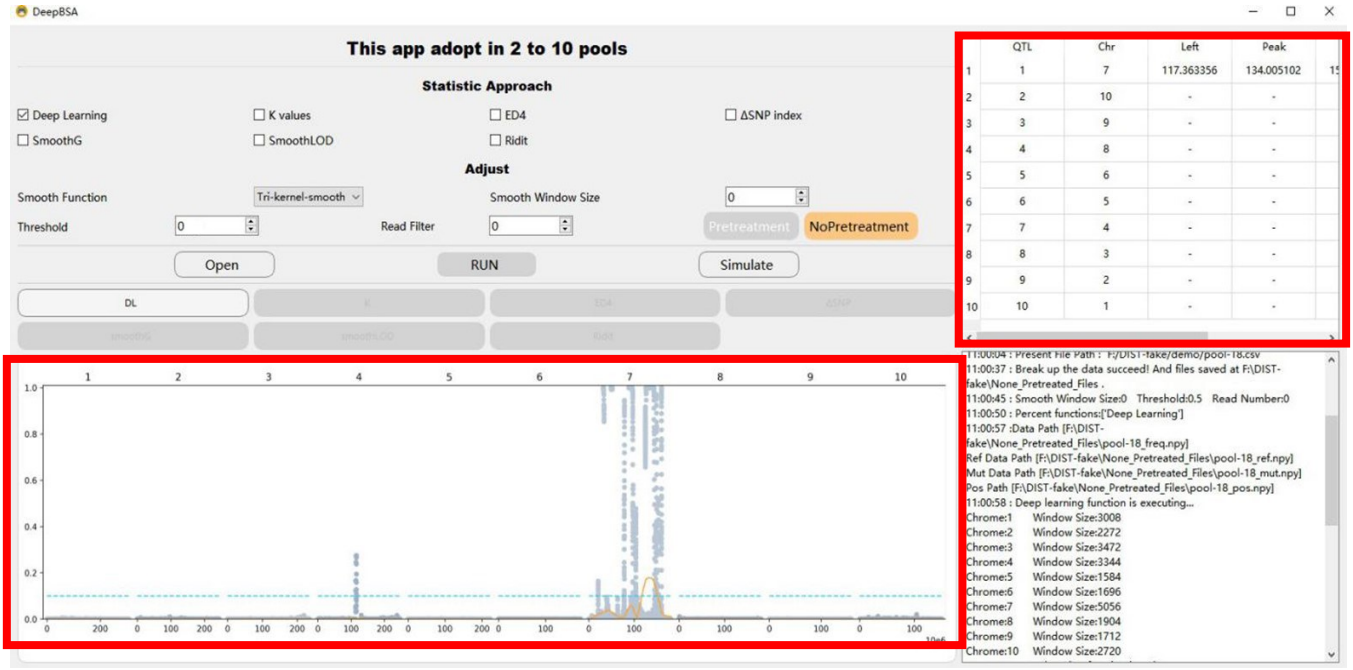


Note: Only DL and Ridit can apply to multi-pool data, so the other methods

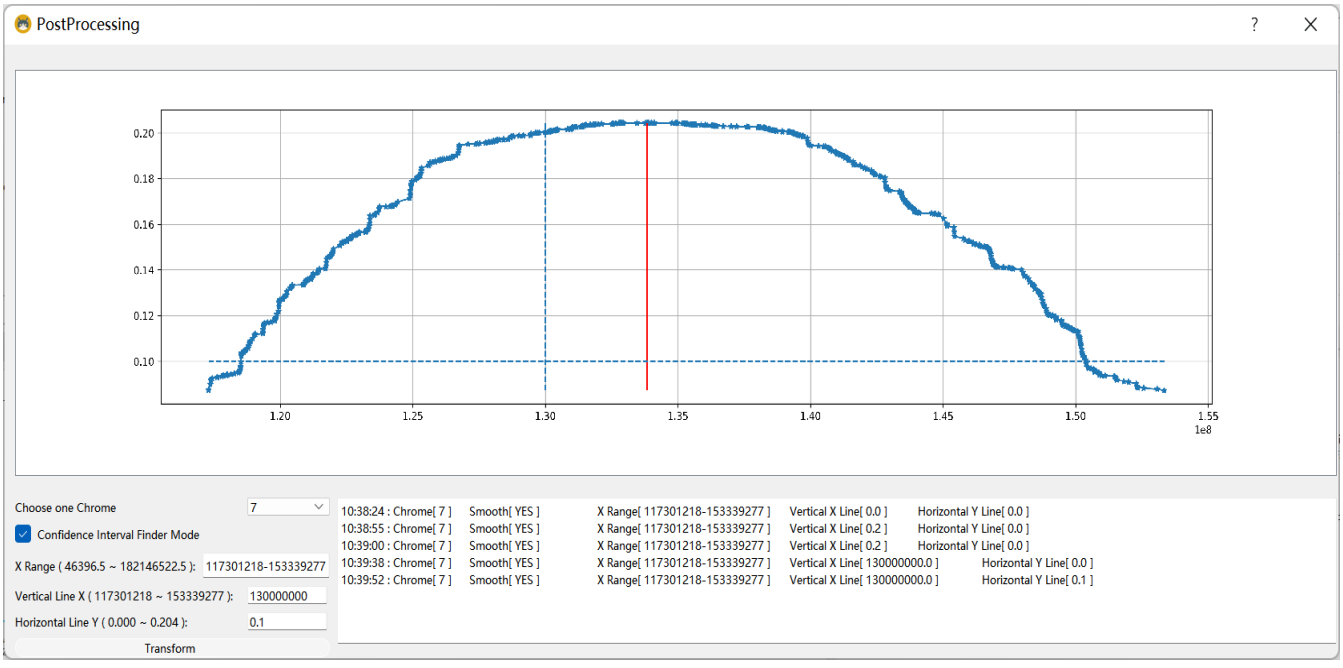
only take information from the first and last pools in the VCF file when running data from multiple pools.

5. Output

The mapping results are presented in two ways as follows - mapping figure and the candidate regions.



A detailed plate of figure can be visualized by clicking on the figure, where to visualize more details.



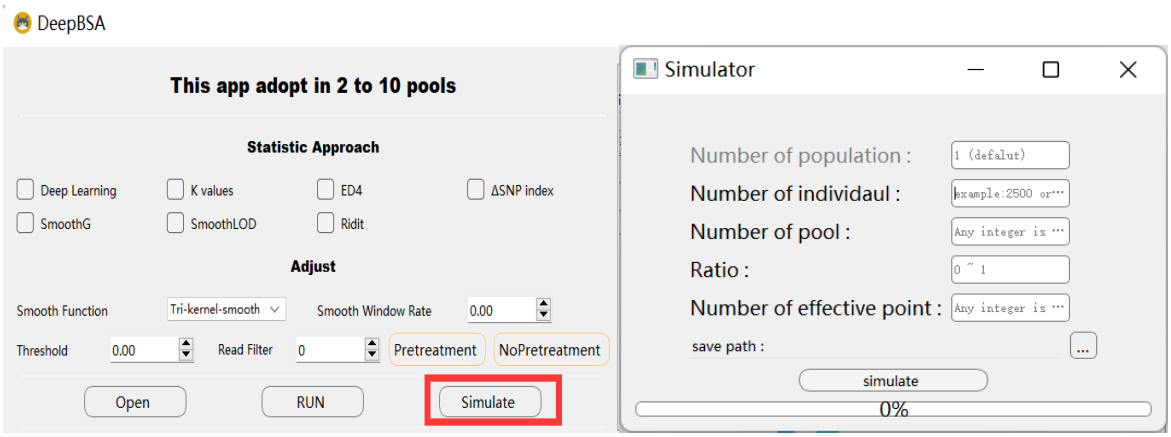
Meanwhile, the results are saved locally in deepbsa/**_Visualize_Results/. The figure is saved as PNG and PDF, and the QTL information is saved as CSV. We also provide a TXT file containing the location and method value of all SNPs, which can be further analyzed and plotted.

6. Other function

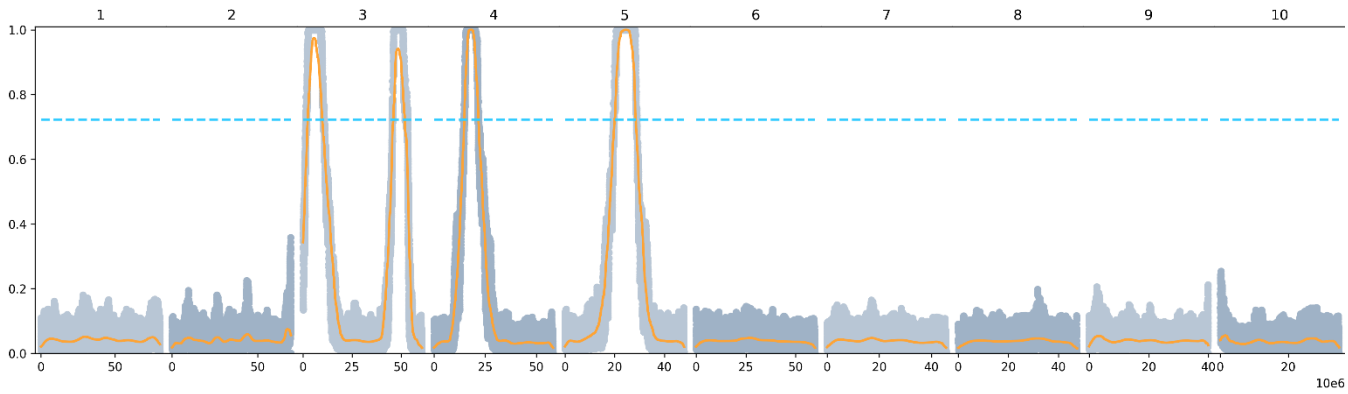
The software provides a simulator to generate simulation data by clicking the button named **Simulate**. Four parameters can be customized as follows.

- Number of individuals: the simulated individuals in population, which represent the population size for next field experiments.
- Number of pools: the simulated pools for bulking, which ranges from 2 to 10.
- Ratio: the ratio of plants for each pool, so the following conditions should be met:

  Ratio * Number of pools < 1
- Number of effective points: the simulated number of QTLs, whose phenotype variant effect (PVE) is assumed to be 0.1.



The simulated data is saved in deepbsa/ *-*-*/, which can be loaded as input data for next mapping directly. The result figure is as follows:



6. Cite

Li Z., Chen X., Shi S., Zhang H., Wang X., Chen H., Li W., and Li L. (2022). DeepBSA: A deep-learning algorithm improves bulked segregant analysis for dissecting complex traits. Mol. Plant. doi: https://doi.org/10.1016/j.molp.2022.08.004.

https://www.cell.com/molecular-plant/abstract/S1674-2052(22)00267-2