PIP-SNP User Manual

01/26/2021

PIP-SNP Development Team

Noble Research Institute and University of California at Riverside.

Description of PIP-SNP

GWAS(Genome wide association studies) application essentially favor to the large scale sample data with high dimensional SNP markers. However, high dimensional SNP(single-nucleotide polymorphism) markers always means huge amount of calculating. In epistatic GWAS analysis, it's not direct SNP markers but the marker pairs are involved, which make it unendurable to handle only several thousands of SNPs.

Next Generation Sequencing (NGS) technology can provide a very cheap and high-throughput sequencing, which, in theory, make it possible to call and genotype highly dense SNP data and furtherly achieve a higher GWAS analysis resolution. However, most of the NGS technology users usually targeted for a lower cost and chose for low-coverage sequencing, which consequentially increase the difficulty in efficient alignment, and the inaccuracy in SNP calling and the higher ratios of missing values after genotype calling.

In short, there are two challenges that GWAS technology facing: one is high dimensional SNP data and the other is the incompleteness of genotype data. It's necessary to develop some method and tools to solve the two challenges.

The SNPs are not independent, and the correlation of nearby SNPs is known as linkage disequilibrium (LD), which can be used for LD conceptual SNP bin mapping, the missing genotype inferencing and the SNP's dimension reduction. Further, the SNPs in one LD bin can be synthesized as a representative marker to reduce the whole marker dimension. **Figure .1** provide the basic concept of the two challenges and the solutions.

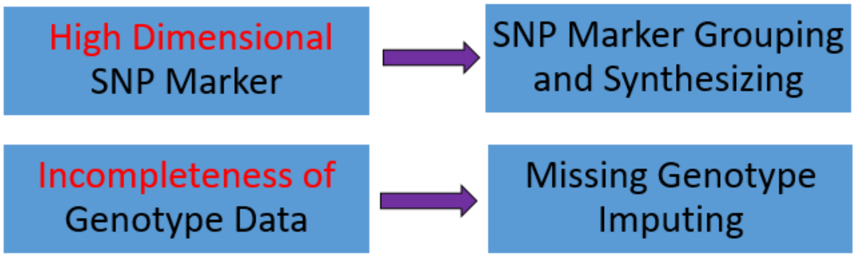

High Dimensional SNP Marker
SNP Marker Grouping and Synthesizing
Incompleteness of Genotype Data
Missing Genotype Imputing

Figure. 1 The concept of two challenges that GWAS face and the solutions

Although there are some research talking the LD selecting and Missing value imputing. These methods are individually proposed in theoretical part and have not been delivered into a easy-to-use tools or platform. To the best of our knowledge, there are no practical tools available to resolve both challenges in one-stop processing. To efficiently preprocess the SNP data and solve the two challenges of GWAS analysis, we referred to and modified the related methods, and also proposed some practical methods, by which PIP_SNP was developed.

PIPS-SNP, as a web platform (https://bioinfo.noble.org/PIP_SNP/), essentially is a SNP data preprocessing pipelines, which seamed several processing modules including LD Bin Mapping, Missing Genotype value Imputing and Marker Synthesizing. **Figure. 2** illustrate the main processing modules.

In the following sections, we will describe the basic principle of each of the processing modules, and then provide the typical applications to use PIP-SNP to preprocess your SNP data.

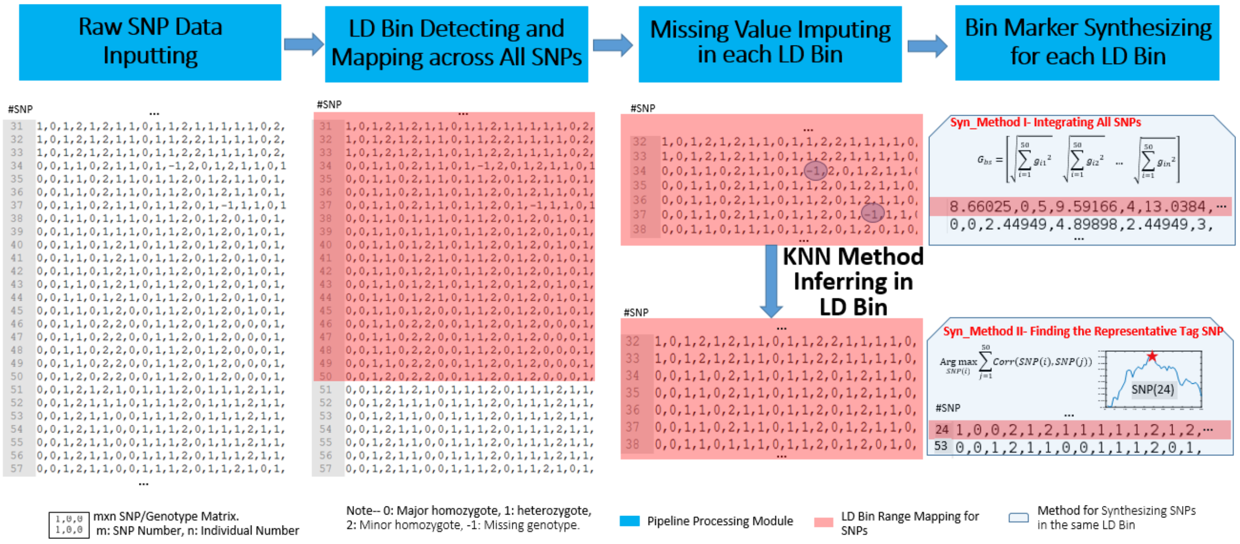

Raw SNP Data Inputting
LD Bin Detecting and Mapping across All SNPs
Missing Value Imputing in each LD Bin
Bin Marker Synthesizing for each LD Bin
#SNP
KNN Method Inferring in LD Bin
Syn_Method I- Integrating All SNPs
8.66025,0,5,9.59166,4,13.0384,…
0,0,2.44949,4.89898,2.44949,3,
Syn_Method II- Finding the Representative Tag SNP
SNP(24)
24 1,0,0,2,1,2,1,1,1,1,1,2,1,2,…
53 0,0,1,2,1,1,0,0,1,1,1,2,0,1,
mxn SNP/Genotype Matrix.
m: SNP Number, n: Individual Number
Note-- 0: Major homozygote, 1: heterozygote, 2: Minor homozygote, -1: Missing genotype.
Pipeline Processing Module
LD Bin Range Mapping for SNPs
Method for Synthesizing SNPs in the same LD Bin

Figure. 2 The PIP_SNP' main preprocessing modules including LD bin mapping, missing genotype imputing and representative marker synthesizing.

Module description

Correlation analysis to describe SNPs

Due to LD, nearby SNPs are correlative and can be well described by a stochastic process. Let the SNP genotypes be represented by an $M \times N$ matrix, where M and N are the SNP number and sample size, respectively. For a single SNP signal, its genotypes can be represented by a vector with length N. The Pearson correlation coefficient between SNP_i and SNP_j is expressed by

$$Corr(SNP_i, SNP_j) = \frac{\sum_{n=1}^{N}(SNP_i(n)-\overline{SNP_i})(SNP_j(n)-\overline{SNP_j})}{\sqrt{\sum_{n=1}^{N}(SNP_i(n)-\overline{SNP_i})^2}\sqrt{\sum_{n=1}^{N}(SNP_j(n)-\overline{SNP_j})^2}} \quad (1)$$

Base on the above correlation coefficient of two SNPs, we can get a correlation measure of two neighbor SNPs as $NR(i)$.

$$NR(i) = Corr(SNP_i, SNP_{i+1}) \;\; i=1,2,3,\ldots \quad (2)$$

LD bin detecting and mapping across all SNPs

The SNPs in one LD bin have a relatively high information redundancy, which can be measured by the above correlation metrics. We developed some method to detect all LD bins to map across all SNPs. One LD Bin is characterized by its left boundary and right boundary, while the left boundary is initially fixed but the right boundary will need to be tentatively test and finally determined until some criteria can be met. If the $i-1^{th}$SNP is the right boundary, then, the criteria can be defined as the comparison between the correlation coefficient of the left and/or right boundary of the bin with the i^{th} SNP against a preset threshold R_th. The i^{th} SNP cannot be clustered into the current bin, but is considered as a breakthrough point. **Figure.3** illustrate the

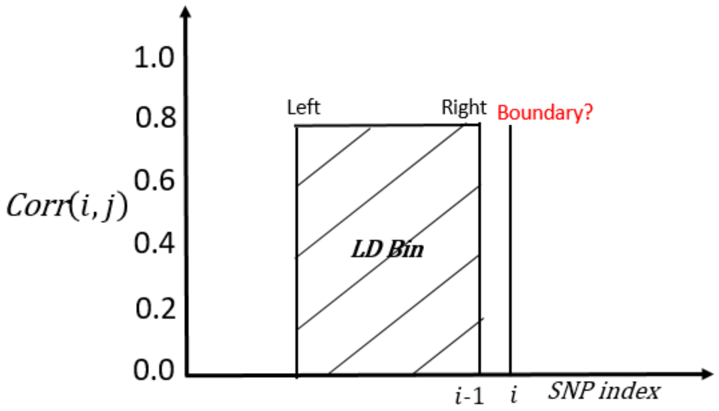


Figure. 3 LD Bin is detected by it’s Left boundary and Right boundary determined at a detected breakthrough point.

Method I: $Corr(Bin.Left, SNP(i)) < R\_th$

Method II: $Corr(Bin.Right, SNP(i)) < R\_th$

Method III: $Corr(Bin.Left, SNP(i)) < R\_th$ **&&** $Corr(Bin.Right, SNP(i)) < R\_th$

Method IV: $Corr(Bin.Left, SNP(i)) < R\_th$ || $Corr(Bin.Right, SNP(i)) < R\_th$

Figure. 4 Four methods were used to detect a boundary, which are based on the comparison of the correlation between the SNP(i) located at an assumed breakthrough point with a Bin’s Left and/or Right

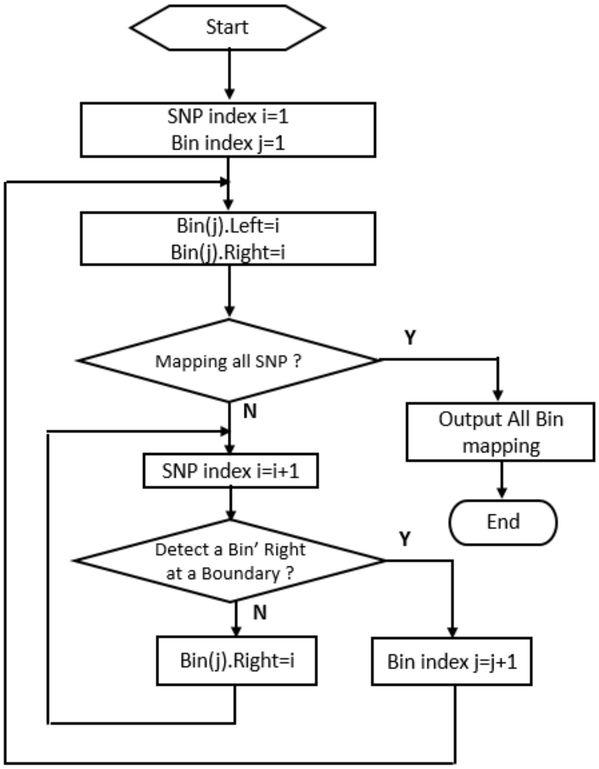


Figure. 5 The flowchart of detecting all LD bins and mapping across all of SNPs.

concept of a LD bin and **Figure.4** listed all of the four methods to detect the right boundary. Once the right boundary of a bin is determined, the breakthrough point can be considered as the left boundary of a new bin. We can continue this procedure until it map all of SNPs. **Figure.5** gives the flowchart of detecting all LD bins and mapping across all SNPs.

kNN based missing value imputing in each detected LD bin

Suppose a $b \times N$ matrix G_b represent a detected LD bin containing b SNP (genotype) markers, the distance of two individual sample s_1 and s_2 can define as the following equation

$$d_b(s_1, s_2) = \frac{1}{nb} \sum_{p \in P} |G_b(p, s_1) - G_b(p, s_2)| \tag{3}$$

Where P is the set of all possible SNP marker index, and $G_b(p, s)$ is the corresponding genotype values. It's possible that either or both of $G_b(p, s_1)$ or $G_b(p, s_2)$ are missing, in which case the difference of genotype value should be ignored in the summation. Here, we only consider the cases with two known corresponding genotype and normalize the summation by known genotype count number nb.

Based on the above defined distance between samples, we can use kNN (k-nearest neighbor) to impute the missing genotype value in the detected LD bin. To an individual sample s_i with missing genotype to impute, we firstly calculate all the $N - 1$ sample pair distance and sort them in ascending order, finally select the top k 'neighbor' samples to infer the missing values. The selected k individual samples are called as sample set *NS*, which are not connected neighbor samples in the genotype matrix but have the comparatively small distance to the specific sample s_i.

Once the k 'neighbor' samples are picked, we can use the following equation to infer the missing genotype $G_b(p_j, s_i)$

$$G_b(p_j, s_i) = \underset{a \in \{0,1,2\}}{\arg max} \sum_{s \in NS} \frac{1}{d_b(s_i, s)} I(G_b(p_j, s) = a) \tag{4}$$

Where *NS* is the picked neighbor sample set, $I(G_b(p_j, s) = a)$ is indicator function that take the value 1 if $G_b(p_j, s) = a$ and 0 otherwise.

LD bin marker synthesizing

Once the LD bins have been mapped across all SNPs and the missing genotypes in each LD bin have been imputed. We can develop some method to generate one synthesized marker to represent the bin, which can reduce the marker dimension size.

Supposing one bin containing b SNPs is represented as a $b \times N$ matrix, where the corresponding numerical genotype value is represented as $g_{i,n}$, we can use formula 5 or 6 to calculate the Euclidean norm as the integrated marker or find the optimal SNP as the Tag SNP, respectively.

$$G_{bs} = \left[\sqrt{\sum_{i=1}^{b} g_{i,1}^2} \quad \sqrt{\sum_{i=1}^{b} g_{i,2}^2} \quad \dots \quad \sqrt{\sum_{i=1}^{b} g_{i,N}^2}\right] \tag{5}$$

$$Tag\_SNP = \underset{SNP_i}{\operatorname{Arg\,max}} \frac{1}{b}\sum_{j=1}^{b} Corr(SNP_i,\, SNP_j) \tag{6}$$

Figure.6 illustrate the concept to find the optimal Tag SNP. We scan all the SNPs in the bin and calculate all of the $\overline{R^2}$ for one selected SNP across all other SNPs. Finally, we selected the SNP with the maximum $\overline{R^2}$ as the optimal Tag SNP to represent the detected LD bin.

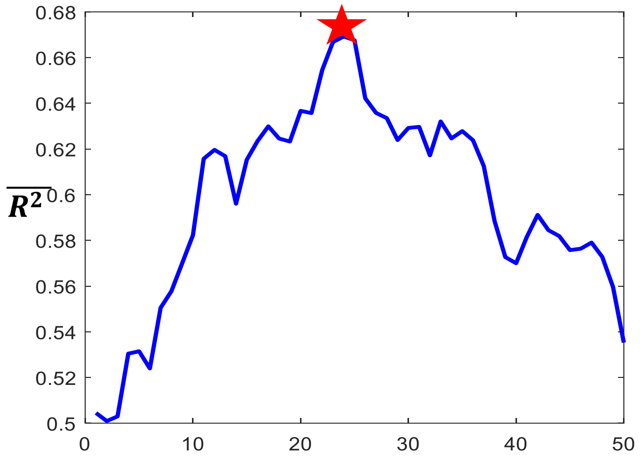


Figure. 6 The optimal Tag SNP determined by the maximal $\overline{R^2}$ across all SNPs in the detected LD bin

Regarding the binary genotype coding such as 0, 1, 2, the integrated marker will become continuous float format not following the original binary format. In addition, the synthetic marker comprehensively integrates the genetic information of the whole bin, but the resolution of the marker's position in the chromosome will decrease into a bin. Comparably, the Tag SNP still follows the same binary format and conserves the resolution of the marker's position.

Deep synthesizing

Due to the LD existing, there are information redundancy among the neighbor SNPs, and the whole genome can be partitioned into LD bins. We investigated two distinct SNP data from RIL(Recombinant Inbred Lines) population and random HapMap population. In the correlation analysis profile(**Figure. 7**), the RIL SNP data show us a conservatively stable profile while the random HapMap SNP data show us violent vibration and many spikes. Therefore, it is more challenging to process SNP data from the random population. However, the acute spike autocorrelation patterns indicate that several types of SNPs are closely entangled in a local region. The method to group and synthesize the similar type of SNPs should consider not only the neighbor joint SNPs but also the neighbor skip SNPs. We developed a special two-step method that includes an initial shallow synthesizing and an aggressive deep synthesizing. **Figure 8** illustrates the concept of the two related steps, with the shallow synthesizing as the first step aims at clumping the neighbor joint SNPs, while the deep synthesizing as a further step reaches out to merge the neighbor skip SNPs.

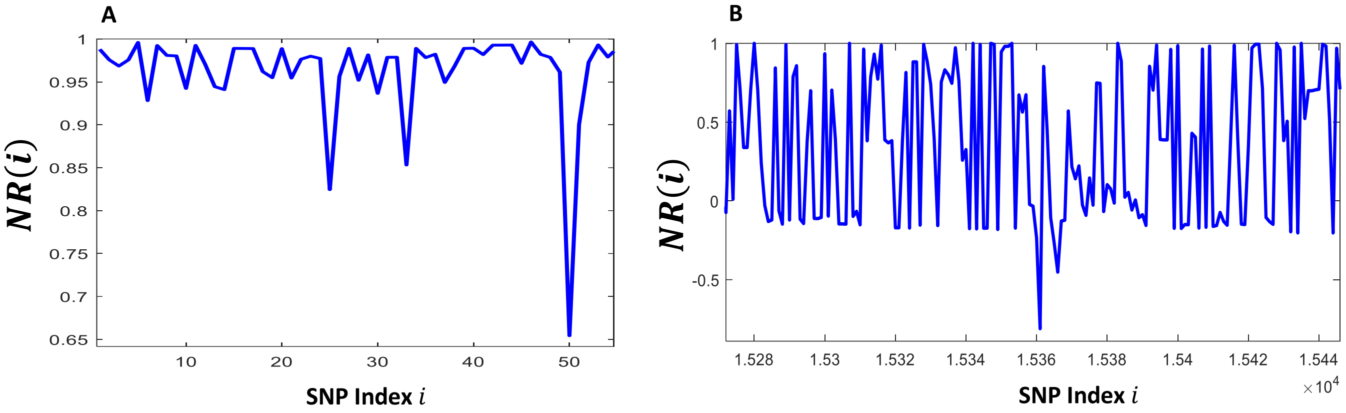


Figure. 7 The correlation analysis profile of (**A**) RIL SNP data and (**B**) random Hapmap SNP data

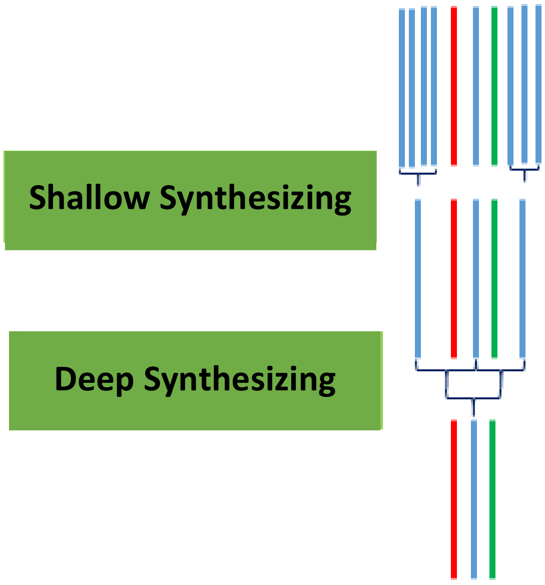


Figure. 8 The concept illustration of shallow synthesizing and aggressive deep synthesizing

According to the criteria of correlation, Deep Synthesising aim to group the neighbor continuous and or skipped SNPs. Essentially, the detection of a breakthrough boundary of marker bin move on until there is two more SNPs that can not be grouped at any cases. During a the detection proceed, all the SNPs are sifted and moved into a continuous and an un-continuous (skipped) marker buffers, which will be furthered synthesized into several bins. **Figure.9** illustrate the whole procedure of function module- Deep Synthesising.

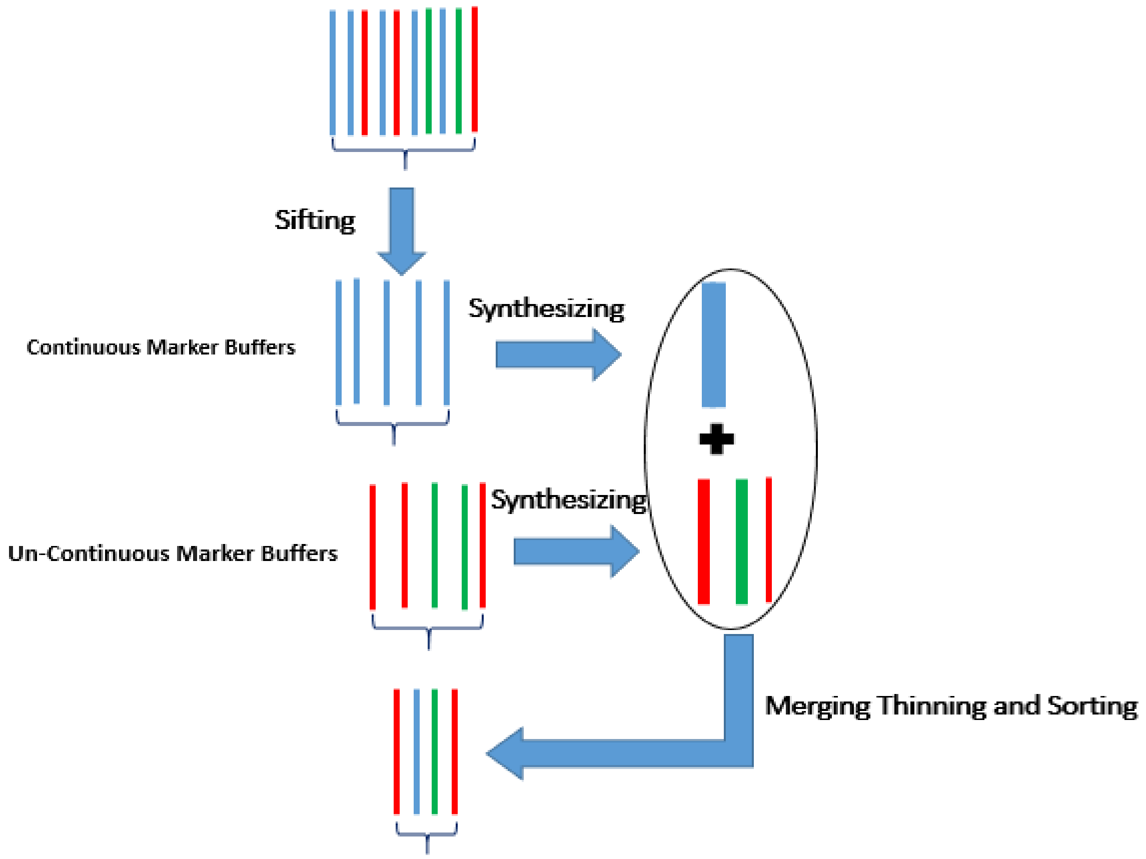

Sifting
Continuous Marker Buffers
Synthesizing
Un-Continuous Marker Buffers
Synthesizing
Merging Thinning and Sorting

Figure. 9 The procedure to implement deep synthesizing

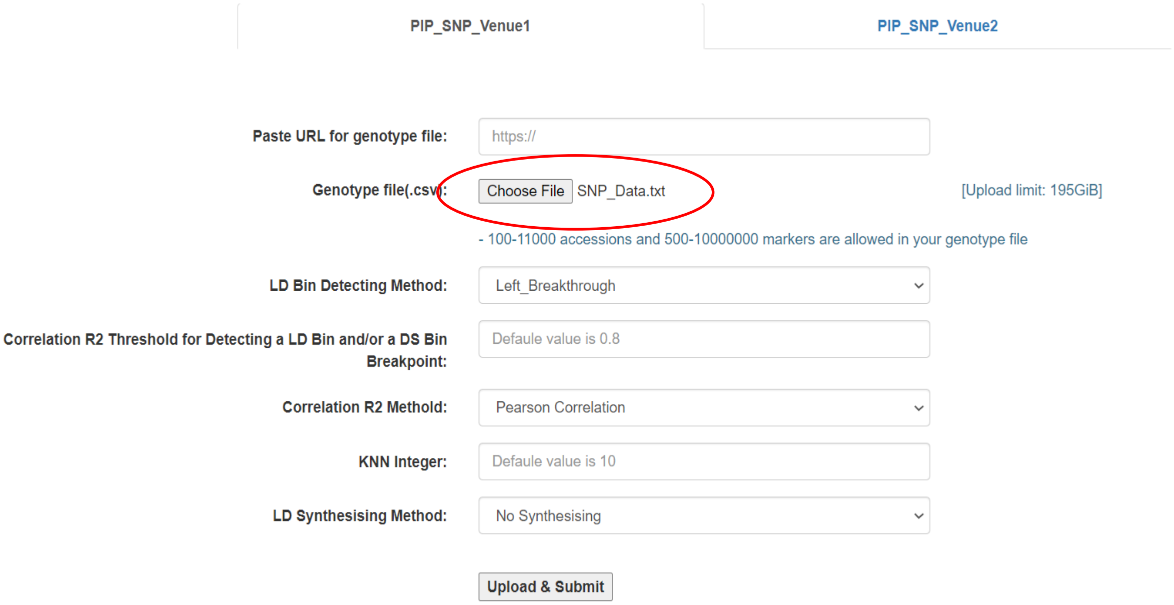
PIP-SNP Web Application

We are aiming to develop a well implemented pipeline to let user can easily preprocess their SNP data. To achieve this goal, we consider as much as more different application scenarios. Here, we prefer to suggest to use our user-friendly a web-interface PIP-SNP (https://bioinfo.noble.org/PIP_SNP/). This web application require no programming skills, no linux knowledge. The user only need to upload their data and configure the few options and parameters, which can avoid the tool's installation and updating, and is close zero pain for learning curve.

Two venues as two typical application

LD block mapping is the most crucial part of the process which will affect the whole implementation. Using the existing LD bin mapping information with higher accuracy usually is the top choice. Therefore, we considered fully the two practical application scenarios as PIP_SNP_Venue1 and PIP_SNP_Venue2, to directly detect the LD bins from the raw SNP data and to use the existing LD bin mapping information respectively.

PIP_SNP_Venue1 only require to upload the raw SNP data while the PIP_SNP_Venue2 require to upload raw SNP data and an existing LD bin mapping result file. The raw SNP data is .csv/.txt file which is an $M \times N$ matrix delimited with ‘,’ or ‘Tab’. The numerical values as -1, 0, 1, 2 represent the missing genotype, and three numerical biological genotypes as the homozygous major allele, heterozygous and homozygous minor allele, respectively. The LD bin map file is a .txt table and each row represent a LD bin recording the start and end of a bin. **Figure.10** and **Figure.11** illustrate two typical application tabs as PIP_SNP_Venue1 and PIP_SNP_Venue2 respectively. Additionally, the two files can be stored in **cloud** server and provided as **unique URL**.



PIP_SNP_Venue1 | PIP_SNP_Venue2

Paste URL for genotype file: https://

Genotype file(.csv): Choose File SNP_Data.txt [Upload limit: 195GiB]

- 100-11000 accessions and 500-10000000 markers are allowed in your genotype file

LD Bin Detecting Method: Left_Breakthrough

Correlation R2 Threshold for Detecting a LD Bin and/or a DS Bin Breakpoint: Defaule value is 0.8

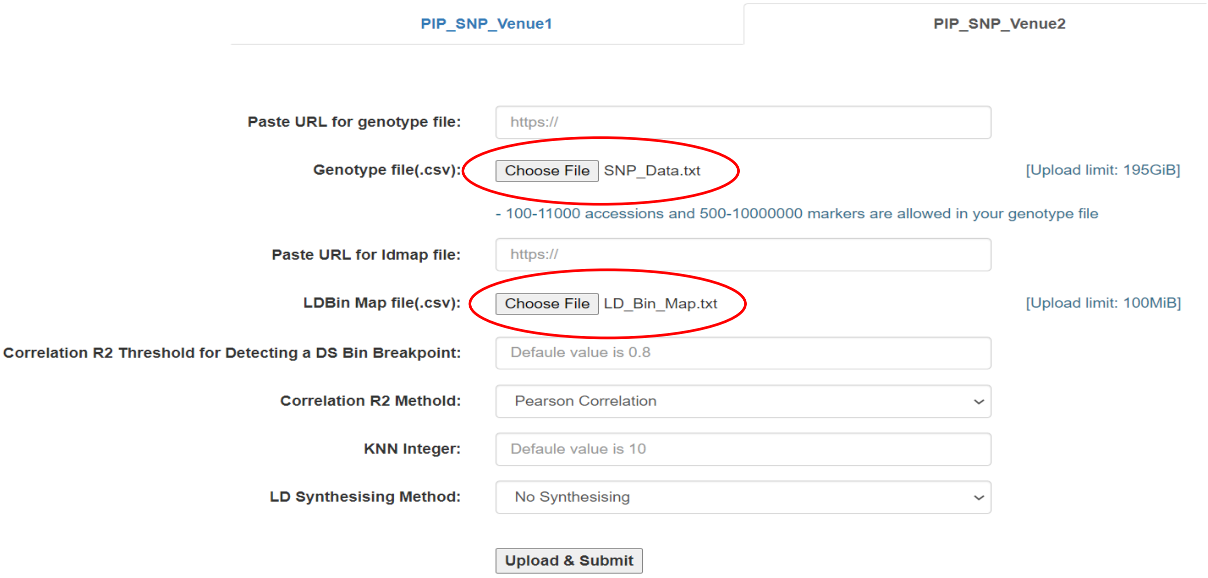
Correlation R2 Methold: Pearson Correlation

KNN Integer: Defaule value is 10

LD Synthesising Method: No Synthesising

Upload & Submit

Figure. 10 The Tab of PIP-SNP Venue1 to implement LD bin mapping, missing value imputing, and bin marker synthesizing.



PIP_SNP_Venue1 | PIP_SNP_Venue2

Paste URL for genotype file: https://

Genotype file(.csv): Choose File SNP_Data.txt [Upload limit: 195GiB]

- 100-11000 accessions and 500-10000000 markers are allowed in your genotype file

Paste URL for ldmap file: https://

LDBin Map file(.csv): Choose File LD_Bin_Map.txt [Upload limit: 100MiB]

Correlation R2 Threshold for Detecting a DS Bin Breakpoint: Defaule value is 0.8

Correlation R2 Methold: Pearson Correlation

KNN Integer: Defaule value is 10

LD Synthesising Method: No Synthesising

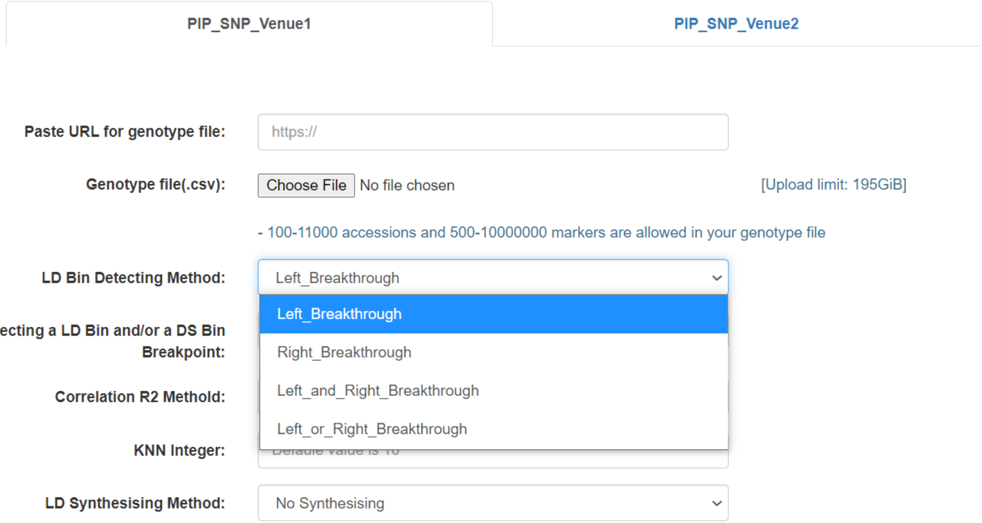
Upload & Submit

Figure. 11 The Tab of PIP-SNP Venue2 to use the existing LD bin mapping file to implement missing value imputing, and bin marker synthesizing.

Method selection and parameter Configuration

To select method for LD bin detection

To detect a LDbin, we need to select the method for LDbin detecting and the related correlation threshold to decide a breakthrough point as the right boundary of a bin. **Figure. 12** show us the four options in a dropdown menus and highlight the select option. The four options correspond to the four methods in **Figure.4**.



Paste URL for genotype file: https://

Genotype file(.csv): Choose File No file chosen [Upload limit: 195GiB]

- 100-11000 accessions and 500-10000000 markers are allowed in your genotype file

LD Bin Detecting Method: Left_Breakthrough

- Left_Breakthrough
- Right_Breakthrough
- Left_and_Right_Breakthrough
- Left_or_Right_Breakthrough

Correlation R2 Threshold for Detecting a LD Bin and/or a DS Bin Breakpoint:

Correlation R2 Methold:

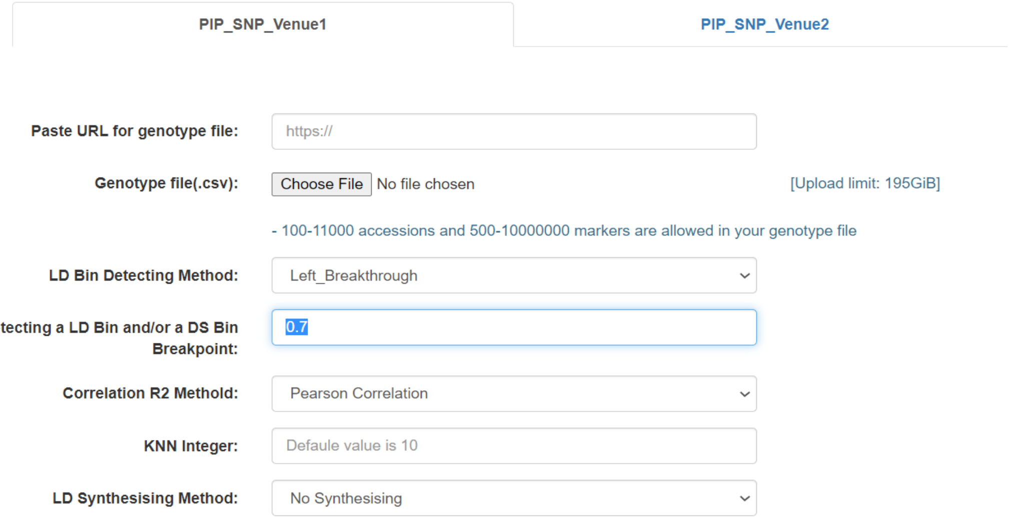
KNN Integer: Defaule value is 10

LD Synthesising Method: No Synthesising

Figure. 12 To select one of the four method options to detect LD bin.

To specific a correlation threshold for a boundary detection

User should specify the correlation threshold (**Figure.13**), which is a value at [0-1.0], and with the default as 0.8.



Paste URL for genotype file: https://

Genotype file(.csv): Choose File No file chosen [Upload limit: 195GiB]

- 100-11000 accessions and 500-10000000 markers are allowed in your genotype file

LD Bin Detecting Method: Left_Breakthrough

Correlation R2 Threshold for Detecting a LD Bin and/or a DS Bin Breakpoint: 0.7

Correlation R2 Methold: Pearson Correlation

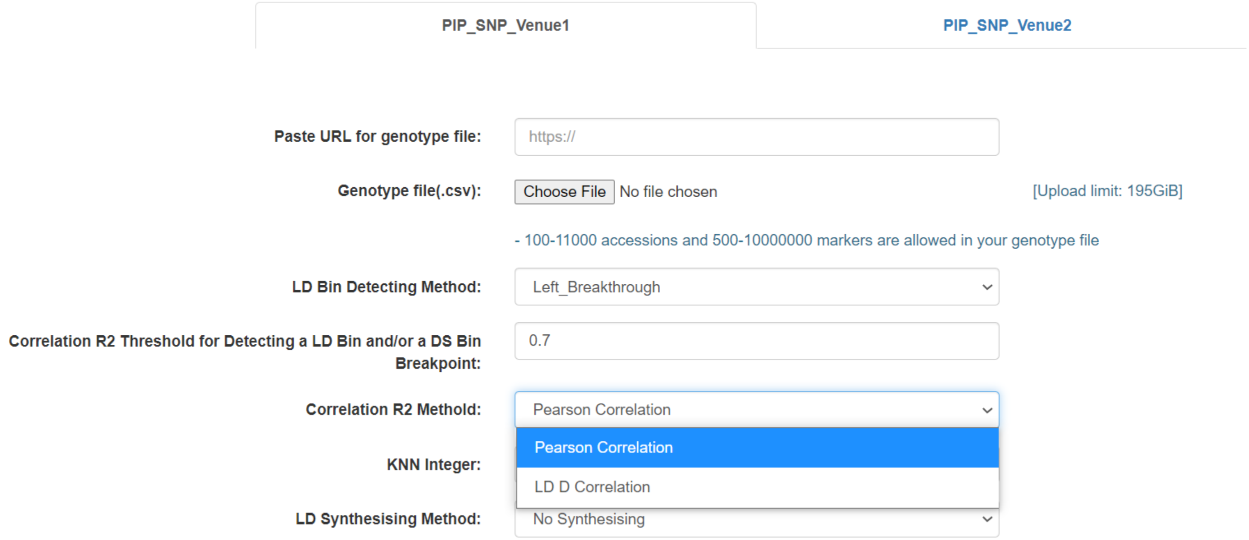
KNN Integer: Defaule value is 10

LD Synthesising Method: No Synthesising

Figure. 13 To configure a correlation value threshold R_th.

To select the correlation method

User should specify one of the two correlation methods: Pearson correlation or LD D correlation (**Figure.14**).



Paste URL for genotype file: https://

Genotype file(.csv): No file chosen [Upload limit: 195GiB]

- 100-11000 accessions and 500-10000000 markers are allowed in your genotype file

LD Bin Detecting Method: Left_Breakthrough

Correlation R2 Threshold for Detecting a LD Bin and/or a DS Bin Breakpoint: 0.7

Correlation R2 Methold: Pearson Correlation

Pearson Correlation

LD D Correlation

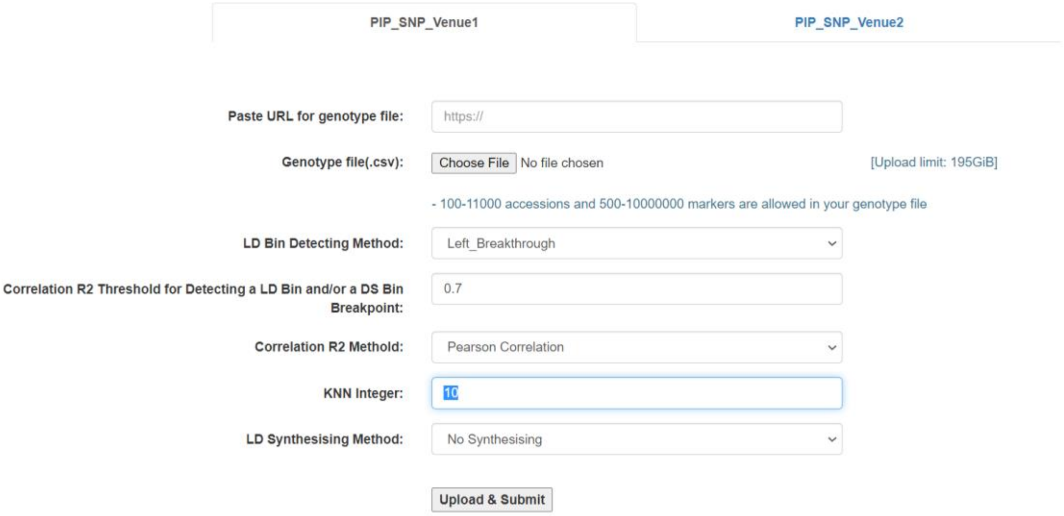
KNN Integer:

LD Synthesising Method: No Synthesising

Figure. 14 To configure a correlation method as Pearson correlation or LD D correlation.

To specific a kNN integer to define the sample number for imputing

PIP-SNP use LD-kNNi method to impute the missing genotype in each LD bin, which require a specific sample numbers (**Figure. 15**).



Paste URL for genotype file: https://

Genotype file(.csv): No file chosen [Upload limit: 195GiB]

- 100-11000 accessions and 500-10000000 markers are allowed in your genotype file

LD Bin Detecting Method: Left_Breakthrough

Correlation R2 Threshold for Detecting a LD Bin and/or a DS Bin Breakpoint: 0.7

Correlation R2 Methold: Pearson Correlation

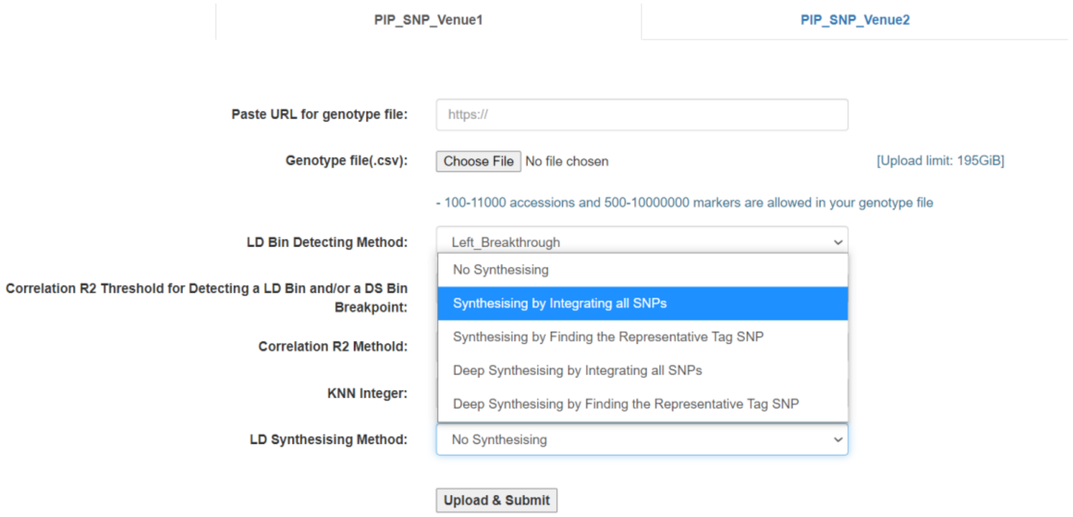
KNN Integer: 10

LD Synthesising Method: No Synthesising

Figure. 15 To configure an integer to define the sample numbers for LD-kNN Imputing method

To select method for synthesizing LD bin marker

For a LD bin, after the missing genotype values are imputed, we can choose to not synthesize and output the imputing result, or generate one synthesized marker by integrating all SNPs or finding one representative Tag SNP. Additionally, the user also can select the aggressive method- deep synthesizing at first and then generate a synthesizing marker. All together, there are five options for user to synthesize a marker of LD bin. **Figure.16** show us the five options in the dropdown menu and highlight the selected option.



PIP_SNP_Venue1 | PIP_SNP_Venue2

Paste URL for genotype file: https://

Genotype file(.csv): Choose File No file chosen [Upload limit: 195GiB]

- 100-11000 accessions and 500-10000000 markers are allowed in your genotype file

LD Bin Detecting Method: Left_Breakthrough

Correlation R2 Threshold for Detecting a LD Bin and/or a DS Bin Breakpoint:

Correlation R2 Methold:

KNN Integer:

LD Synthesising Method: No Synthesising

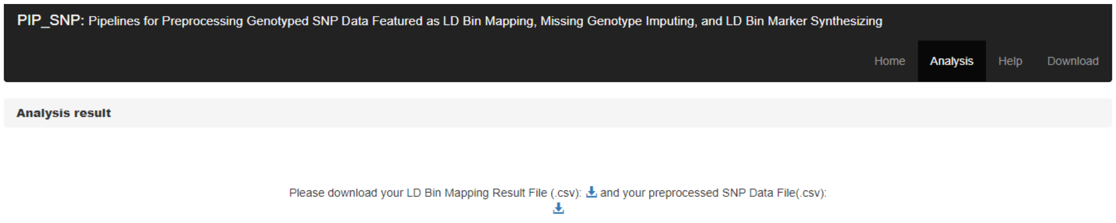
- No Synthesising
- Synthesising by Integrating all SNPs
- Synthesising by Finding the Representative Tag SNP
- Deep Synthesising by Integrating all SNPs
- Deep Synthesising by Finding the Representative Tag SNP

Upload & Submit

Figure. 16 To select one of the five options for a LD bin marker synthesizing.

User-interface to download the analyzing result

After submitting, PIP_SNP will proceed with all of the processing procedures, and return two files, including the LD bin mapping result and the final SNP data preprocessing result. **Figure.17** is the user-interface to download the analyzing result.



PIP_SNP: Pipelines for Preprocessing Genotyped SNP Data Featured as LD Bin Mapping, Missing Genotype Imputing, and LD Bin Marker Synthesizing

Home | Analysis | Help | Download

Analysis result

Please download your LD Bin Mapping Result File (.csv): ⤓ and your preprocessed SNP Data File(.csv): ⤓

Figure. 17 Interface of PIP-SNP for user to download the analysis result.

PIP-SNP source code and the command line application

Although we develop a web based pipeline PIP-SNP (https://bioinfo.noble.org/PIP_SNP/), which can provide user-friendly convenience by maximally avoiding the mistaken parameter configuration.

However, we are glad to open the main source codes which were developed in C++ in Linux IDE:: CodeBlocks. It mainly include three Linux CodeBlocks project files: Linux_CB_PIP_SNP_Venue1, and Linux_CB_PIP_SNP_Venue2, and Linux_CB_Deep_Synthesis, which correspond to the two application venues and deep synthesizing respectively. After complied, three command line executables as : PIP_SNP_V1, PIP_SNP_V2, DeepSynthesis will be generated respectively. The user only need to follow the step 1 to get the usage message, and step 2 to configure the corresponding parameters and option to analyze your data.

PIP_SNP_Venue1

Step 1.) ./PIP_SNP_V1 -u

```
Hello, Dear User! SNP Processing Pipeline for LD Bin Detecting, Missing Genotype Imputing,
and or Synthesizing is Starting!

Welcome to use this program to do LD Marker Bin Detecting

The usuage of input parameter arguments are listed as followings:

-u or -U: Output this help usage message

-g or -G: The full name of Genotype file

-l or -L: The full name of LD Bin Mapping Result file

-i or -I: The Individual number

-r or -R: The Threshold for the pairwise LD R2

-c or -C: the  Correlation Method for a Pairwise Genotypic Markers,0:Pearson_Correlation_R2;
1: LD_D_R2; Default(>2): Pearson_Correlation_R2

-d or -D: the  Detection Method for a Marker Group(LD Block), 0: Right_Breakthrough; 1:
Left_Breakthrough; 2: Left and Right Breakthrough; 3: Left or Right Breakthrough; Default(>3)
: Right_Breakthrough

-k or -K: the  K Nearest Neighbor individuals in The KNN method
```

-s or -S: the method how to generate the syntesized(binned) genotypic marker, 0: not to synthesize; 1: norm integration of the multiple markers' genotype values; 2: select the optimal one; Default(>2) :norm integration of the multiple markers' genotype values

-o or -O: the full name of Imputing results

Step 2.) ./PIP_SNP_V1 -g ./SNP_Data_Example.txt -l ldmap.txt -o PIP_SNP.txt -i 278 -c 0 -d 0 -r 0.8 -k 10 -s 2

PIP_SNP_Venue2

Step 1) ./PIP_SNP_V2 -U

Hello, Dear User! Welcome to use SNP Processing Pipeline for LD Bin Filling, Missing Genotype Imputing, and/or Marker Synthesizing!

Welcome to use this program to do LD Marker Bin Detecting

The usuage of input parameter arguments are listed as followings:

-u or -U: Output this help usuage message

-g or -G: The full name of Genotype file

-l or -L: The full name of LD Bin Mapping Result file

-i or -I: The Individual number

-c or -C: the Correlation Method for a Pairwise Genotypic Markers, 0: Pearson_Correlation_R2; 1: LD_D_R2; Default(>=2): Pearson_Correlation_R2

-k or -K: the K Nearest Neighbor individuals in The KNN method

-s or -S: the method how to generate the syntesized(binned) genotypic marker, 0: not to synthesize; 1: norm integration of the multiple markers' genotype values; 2: select the optimal one; default(>2): norm integration

-o or -O: the full name of Imputing results

-m or -M: the full name of LD Mapping results after synthesing

Step 2.) ./PIP_SNP_V2 -g ./SNP_Data_Example.txt -l ./LD_Bin_Map.txt -o PIP_SNP.txt -m ./Synthesis_Map.txt -i 278 -c 0 -k 10 -s 1

DeepSynthesis

Step 1.) ./DeepSynthesis -u

Hello, Dear User! To furrther reduce the SNP size, we develop such specific program to deep synthesis SNP Bins by considering the correlationship of a SNP and its neighbor jumped SNP Bins!

Welcome to use this program to do Deep Synthesising

The usuage of input parameter arguments are listed as followings:

-u or -U: Output this help usuage message

-g or -G: The full name of inputing Synthesised SNP file

-l or -L: The full name of inputting LD Bin file Mapping Synthesising SNPs

-i or -I: The Individual number

-c or -C: The Method for the pairwise R2 Correlation, 0:Pearson_Correlation_R2; 1: LD_D_R2; Default(>2): Pearson_Correlation_R2

-r or -R: The Threshold for the pairwise R2 Threshold, a float value (0~1.0)

-s or -S: The Method for the multiple SNP bin Synthesising, 1: norm integration of the multiple markers' genotype values; 2: select the optimal one; Default(0, 0r >2) :norm integration of the multiple markers' genotype values

-o or -O: The full name of Deep Synthesised SNP results

-m or -M: The full LD Bin mapping file name after deep synthesising

Step 2.) ./DeepSynthesis -g ./Propressed_SNP.csv -l ./LD_Map.csv -i 132 -c 0 -r 0.2 -s 2 -o ./DS_SNP.txt -m ./DS_Map.txt