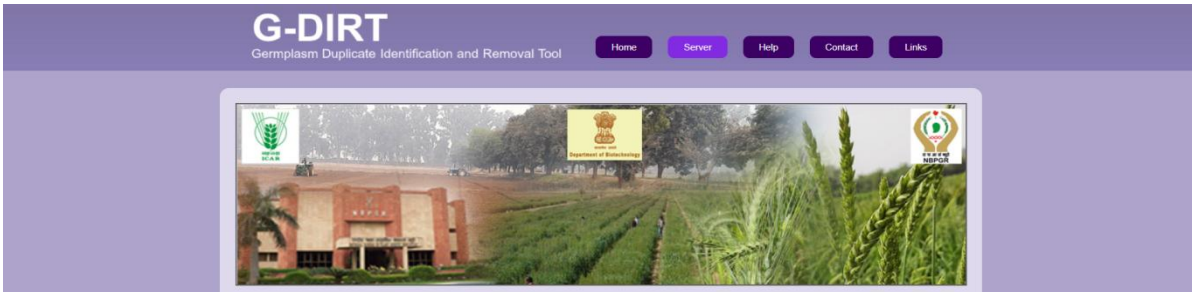


G-DIRT help for executing the tool and interpretation results

A. Input file preparation

The input file to GDIRT server has to be given in hapmap format, where markers are represented in rows and genotypes are represented in columns. Among the duplicates, the software keeps the first genotype and removes the second. The genotypes in the hapmap file should be arranged in order of preference. If the experimenter is biased towards a particular genotype, than he should keep the genotype prior to other genotypes of lesser importance.

Table 1. Column wise description of input file

Col. No.	Column Name	Column Description	Value	Note
1	rs	SNP identifier	Alpha-numeric	Mandatory
2	alleles	SNP alleles as per NCBI dbSNP	Alphabets ex. A/T	Mandatory Usually represented as reference / alternate
3	chrom	Chromosome on which SNP is present	Numeric	Mandatory Must be renumbered if not positive integer
4	pos	Position of SNP on the chromosome	Numeric	Mandatory
5	Strand	Orientation of the SNP in the DNA strand.	forward (+) or reverse (-)	Mandatory
6	assembly	Version of reference sequence assembly	Numeric	Put NA, if no data available
7	center	Name of genotyping center	Alphabet	Put NA, if no data available
8	protLSID	Identifier for HapMap protocol	Alpha-numeric	Put NA, if no data available
9	assayLSID	Identifier for HapMap assay	Alpha-numeric	Put NA, if no data available
10	panelLSID	Identifier for panel of individuals genotyped	Alpha-numeric	Put NA, if no data available
11	QCcode	Quality control code	Alpha-numeric	Put NA, if no data available
12 …	Sample accession/ name/ ID	The sample accession/ name/ ID that contain marker genotype in each row	Alphabet	Mandatory

Special points for input file preparation

1. The file should be a tab delimited text.
2. No hash (#) should append to the column names *rs* and *assembly*.
3. Missing data should be represented as NN.
4. Alleles should be capitalized with forward slash (/) under *alleles* column.
5. First four columns are mandatory to fill.

6. Make the labels of genotypes as small as possible for a better visual of clusters

rs	alleles	chrom	pos	strand	assembly	center	protLSID	assayLSID	panel	QCcode	EC313710	IC443766	IC252796A
AX-94422082	G/T	1	1145885	+	NA	NA	NA	NA	NA	NA	GG GG	GG GG TT	TG GG	GG GG	GG GG GG	GG GG TG	GG GG TT
AX-94598030	A/G	1	1159536	+	NA	NA	NA	NA	NA	NA	GG GG	GG GG AA	AG GG	GG GG	GG GG GG	GG GG AG	GG GG AA
AX-94669331	C/T	1	1159689	+	NA	NA	NA	NA	NA	NA	TT TT	TT TT CC	TC TT	TT TT	TT TT TT	TT TT NN	TT TT CC
AX-95183288	A/G	1	1159713	+	NA	NA	NA	NA	NA	NA	GG GG	GG GG AA	AG GG	GG GG	GG GG GG	GG GG AG	GG GG AA
AX-95217061	A/C	1	1161441	+	NA	NA	NA	NA	NA	NA	AC AA	AA AA AC	AA AA	AA AA	AC AC AA	AA AA AC	AA AA NN
AX-94493709	A/G	1	1174865	+	NA	NA	NA	NA	NA	NA	AG GG	AA AA AA	AA AG	GG GG	AG AG GG	AG GG AG	AG AG AA
AX-94449086	A/G	1	1190148	+	NA	NA	NA	NA	NA	NA	GG GG	AA AA GG	GG GG	GG GG	GG GG GG	GG GG GG	GG GG GG
AX-94889872	C/T	1	1211706	+	NA	NA	NA	NA	NA	NA	TT TT	TT TT CC	TT TT	TT TT	TT TT TT	TT TT TT	TT TT NN
AX-94583145	A/G	1	1211885	+	NA	NA	NA	NA	NA	NA	GG GG	GG GG AA	AG GG	GG AA	AA AA AA	GG GG AG	GG AA AA
AX-94745699	C/T	1	1235969	+	NA	NA	NA	NA	NA	NA	TT TT	TT CC TC	TC CC	CC CC	CC CC TT	CC CC CC	CC CC CC
AX-94772289	C/T	1	1236448	+	NA	NA	NA	NA	NA	NA	CC CC	CC CC TT	TC CC	CC CC	CC CC CC	CC CC NN	CC CC TT
AX-94974969	A/G	1	1338148	+	NA	NA	NA	NA	NA	NA	AG AG	AG AG GG	AG AG	AG GG	GG GG GG	AG AA AG	AG GG GG
AX-95211874	G/T	1	1340329	+	NA	NA	NA	NA	NA	NA	TG TG	TT TG TT	TG TT	TG TG	TG TG TG	TT GG TG	TG TG NN
AX-94778893	A/C	1	1645150	+	NA	NA	NA	NA	NA	NA	AA AA	AA AA NN	NN AA	AA AC	NN AA AA	AA AA NN	AA AA AA
AX-94496430	G/T	1	2377608	+	NA	NA	NA	NA	NA	NA	TG TG	TG TT GG	TG TG	TG TG	TT TT TG	TG TT TG	TT TT GG
AX-94733072	C/T	1	2378230	+	NA	NA	NA	NA	NA	NA	TC TC	TC TC TT	TC TC	TC CC	TC TC TC	TC TC TC	TC TC TT
AX-94700275	A/G	1	2441012	+	NA	NA	NA	NA	NA	NA	AA AA	AA AA AA	AA AA	AA AA	AA AA AA	AA AG AA	AA AA AA
AX-94581354	A/C	1	2920151	+	NA	NA	NA	NA	NA	NA	CC CC	CC CC CC	CC CC	AC CC	AC AC CC	CC AC CC	CC CC CC
AX-94492529	A/T	1	3118372	+	NA	NA	NA	NA	NA	NA	AA AA	AA AA TT	AT AA	AA AA	AA AA AA	AA AA NN	AA AA TT
AX-94550967	C/T	1	3381102	+	NA	NA	NA	NA	NA	NA	CC CC	CC CC CC	CC TT	CC CC	TT NN CC	TT TT CC	CC CC CC
AX-94936441	C/T	1	3388719	+	NA	NA	NA	NA	NA	NA	CC CC	CC CC TT	TC CC	CC CC	CC CC CC	CC CC CC	CC CC TT
AX-94386458	A/C	1	3392719	+	NA	NA	NA	NA	NA	NA	CC CC	CC CC AA	AC CC	CC CC	CC CC CC	CC NN NN	AC AC AA
AX-95162217	A/G	1	3761117	+	NA	NA	NA	NA	NA	NA	GG AA	GG GG AA	NN GG	AA GG	AG AG GG	GG AA AA	AA AA AA
AX-95158783	C/G	1	3845903	+	NA	NA	NA	NA	NA	NA	CG CG	CG NN CC	NN CG	GG CG	CG CG CG	CG CG GG	GG GG CC
AX-94664734	A/G	1	3847072	+	NA	NA	NA	NA	NA	NA	GG GG	GG GG NN	GG GG	GG GG	GG GG GG	GG GG GG	GG GG NN
AX-95082131	A/G	1	3847127	+	NA	NA	NA	NA	NA	NA	AA AA	AA AA GG	AG AA	AA AA	AA AA AA	AA AA AA	AA AA GG
AX-94438601	C/G	1	4033356	+	NA	NA	NA	NA	NA	NA	CC CC	CC CC CG	CG CC	CC CC	CC CC CC	CC CC CC	CC CC CG
AX-94608213	C/T	1	4048434	+	NA	NA	NA	NA	NA	NA	CC CC	CC CC CC	CC CC	CC CC	CC CC CC	CC CC CC	CC CC CC
AX-94979270	A/G	1	4053566	+	NA	NA	NA	NA	NA	NA	GG AA	GG GG AA	AG GG	GG GG	AA AA GG	GG AA GG	AA AA AA
AX-94945604	A/C	1	4502788	+	NA	NA	NA	NA	NA	NN	AC AC	CC AC AC	CC AC	AA CC	AC AC AC	AC NN AC	CC CC AC
AX-95219658	A/G	1	4812661	+	NA	NA	NA	NA	NA	NA	AA AA	AA AA AA	AA AA	GG GG	AA AA AA	AA NN AA	AA AA AA
AX-95139456	A/G	1	5025329	+	NA	NA	NA	NA	NA	NA	AA AA	AA AA GG	AG AA	GG GG	GG GG AA	GG AG GG	GG GG GG
AX-95230167	A/G	1	5028362	+	NA	NA	NA	NA	NA	NA	GG GG	GG GG AA	AG GG	GG GG	AA AA AA	AG AA AG	AA AA AA
AX-94562727	G/T	1	5646294	+	NA	NA	NA	NA	NA	NA	TG GG	GG GG GG	GG TG	GG TG	GG NN TG	TG NN TG	GG GG GG
AX-94661960	A/G	1	6067447	+	NA	NA	NA	NA	NA	NA	AG AA	AA AA AA	AA NN	AA AG	AA AA AG	AG AA NN	AA AA AA
AX-94741678	A/G	1	6150999	+	NA	NA	NA	NA	NA	NA	AA AG	AG AG AG	NN AA	AG AA	AG AG AA	AA AG AA	AG AG AG
AX-95106318	C/G	1	6153162	+	NA	NA	NA	NA	NA	NA	GG GG	GG GG GG	GG GG	GG GG	GG GG GG	CC GG GG	GG GG GG
AX-94482516	C/T	1	6153189	+	NA	NA	NA	NA	NA	NA	TT TC	TC TT TC	TT TT	TT TT	TT TC TT	TT TT CC	TT NN TT	TC

Figure 1. An example of a hapmap genotype file

Parameters

Minor allele frequency (MAF)

A common method of minimizing errors in large DNA sequence data sets is to drop variable sites with a MAF below some specified threshold. The rare genetic variants have a low MAF, which is usually less than 5 or 1 %. Therefore, the single nucleotide polymorphisms (SNPs) having MAF greater than 0.05 (5%) are considered in genome wide association studies. Thus, in this tool the default value for MAF is set as 0.05(5%). If the user does not want to use this parameter he can set the value to 0. The range for the MAF parameter varies from 0 to1.

Missing genotype data

The missing genotype data refers to the genotypes where one or more marker information is missing. The incidence of missing information in genotype data is due to unsuccessful assay of markers on genotyping platforms. Further, genotypes with missing information on high number of markers can often lead to a biased analysis. Thus, either the missing values are imputed computationally or genotypes/markers with missing values beyond a threshold (usually 5%) are removed from the analysis. Here, a provision in the tool has been made to filter the markers based on missing value. The default value for missing genotype data is 0.05 (5%) and user can set the value to 1 for not using this parameter. The range for the missing genotype data parameter varies from 0 to 1.

Linkage disequilibrium (LD) pruning

LD pruning is a method to select a subset of markers that are in approximate linkage equilibrium. LD pruning filters genetic markers by selecting only markers that are representatives of the genetic haplotype blocks. It avoids top ranked redundant SNP-SNP

interactions that are merely due to the high correlation between genetic markers. The default value for LD pruning is 0.75 and 1 can be set for not using this parameter.

Hardy Weinberg's equilibrium (HWE)

Violation of HWE law indicates that genotype frequencies are significantly different from expectations and the observed frequency should not be significantly different. In GWAS, it is generally assumed that the deviations from HWE are result of genotyping errors. The HWE thresholds in cases are often less stringent than those in controls, as the violation of the HWE law in cases can be indicative of true genetic association with disease risk. Hardy Weinberg's equilibrium (HWE) is estimated here based on p-value of Haldane's Exact test for HWE. A p-value < 0.05 is significant and rejects the null hypothesis *i.e.*, "The population is in equilibrium". Thus, the default HWE threshold is set to 0.05. If your data or analysis does not require HWE filtration, you may set the threshold value to 1.

Marker Heterozygosity

The marker loci having high heterozygosity indicate technical artifact or paralogous/repetitive regions that could not be distinguished through genotyping. Natural populations of self-pollinating crops and inbred lines are highly homozygous where, a marker loci with modest heterozygosity rate is also doubtful. Further, the extremely heterozygous markers could be due to the probes detecting homeologs and failing to distinguish between the two highly similar sub-genome sequences. So, highly heterozygous markers can be filtered out by using a suitable threshold. Here, a heterozygosity threshold of 0.1 (10%) is set as default. However, user has a flexibility to use a higher or lower threshold for filtering markers. The user can set the threshold to 1, if he doesn't want to use this parameter.

Homozygous difference

It refers to the difference between two germplasms based on only the homozygous markers. It is an ideal way of identifying and removing duplicates. It is not affected by heterozygous filtration. Based on literature, the default value is set to 0.05% which can be flexibly changed by the user.

Total genotypic difference

It refers to the difference between two germplasms based on both the homozygous and heterozygous markers. The default value is set to 2%. It is suggested that if you want to use *Total genotypic difference* than you should not filter markers based on *heterozygosity* (set threshold to 1) at the duplicate identification and removal step.

Monomorphic SNPs

Monomorphic SNPs represent the SNP in just one state, in contrast to polymorphic SNPs. In a monomorphic site all the individuals have the same genotype. As monomorphic SNPs give no information, it is a good idea to exclude them from analysis.

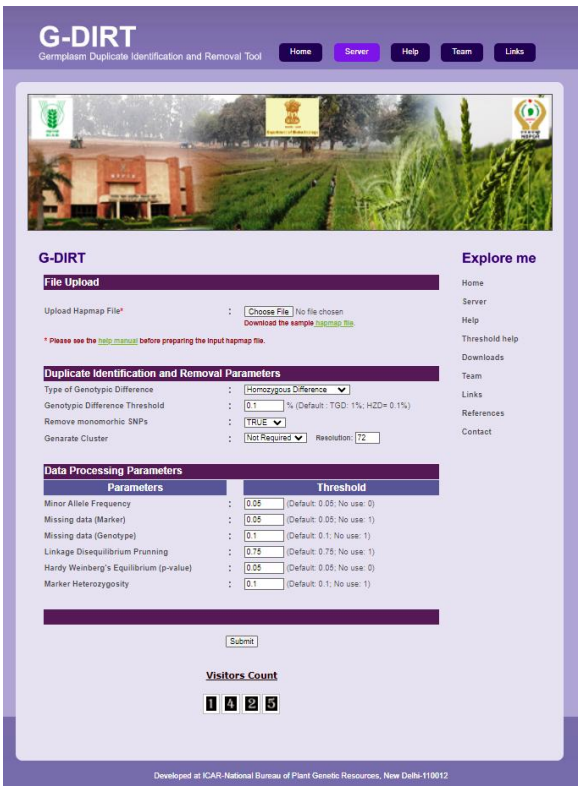


Figure 2. The server page showing all the parameters

Output

The web server gives information on the successful completion of the job or errors along with the assigned job id. Further, it provides duplicate removal summary, identified duplicates, cluster of genotypes based on selected parameters, K-density graph of genotypic difference and marker heterozygosity plot. The final filtered hapmap file is available for download at the bottom of the result page.

Parameters

In this section the information on user selected parameters is given.

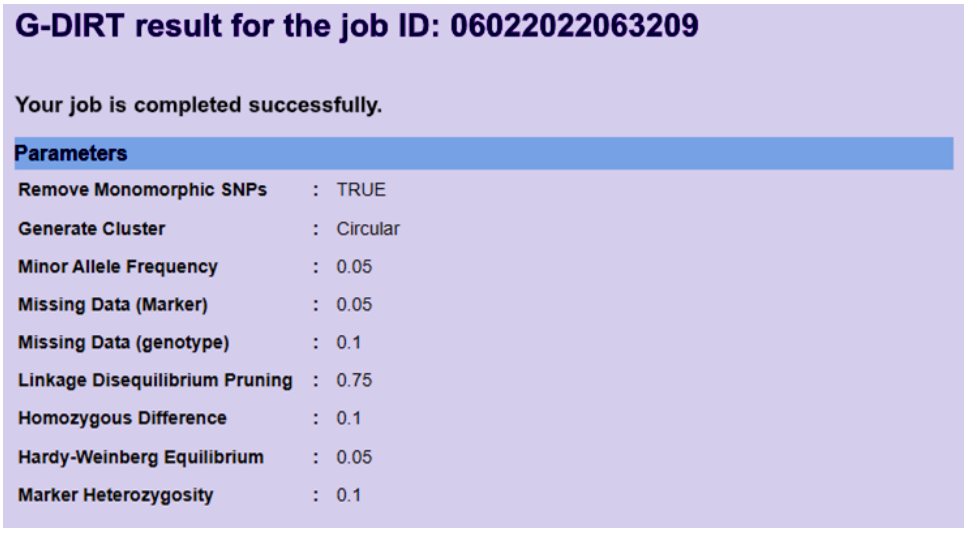
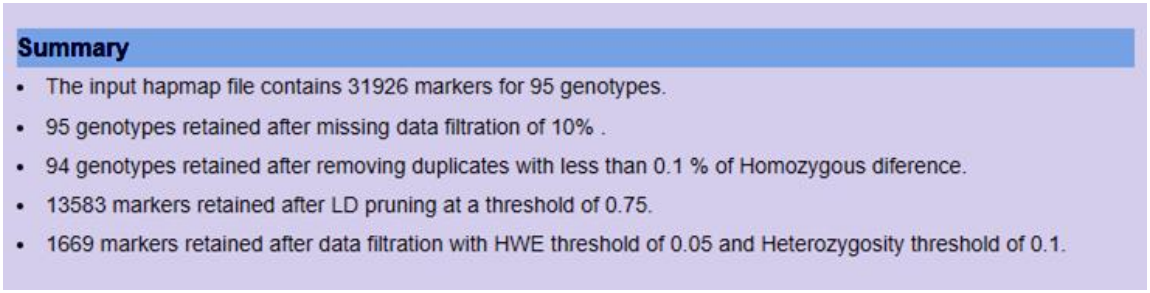


Figure 3. The result page showing input parameters

Duplicate removal summary

The number of genotype retained after duplicate removal, number of filtered markers after data pre-processing is provided.



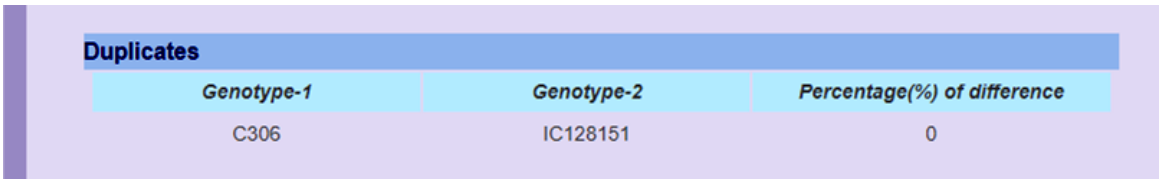
Summary

- The input hapmap file contains 31926 markers for 95 genotypes.
- 95 genotypes retained after missing data filtration of 10% .
- 94 genotypes retained after removing duplicates with less than 0.1 % of Homozygous diference.
- 13583 markers retained after LD pruning at a threshold of 0.75.
- 1669 markers retained after data filtration with HWE threshold of 0.05 and Heterozygosity threshold of 0.1.

Figure 4. The result page showing summary of results

Duplicates

A list of duplicates with the percentage difference is given in a tabular format.



Duplicates

Genotype-1	*Genotype-2*	*Percentage(%) of difference*
C306	IC128151	0

Figure 5. The result page list of duplicates with percentage difference

Genotype Clustering

In the genotype clustering, the genotypes are clustered based on the either total genotypic difference or homozygous difference. The clusters can be generated in rectangular, triangular and circular format. A dotted red line for threshold is marked on rectangular and triangular dendrograms below which the clusters represent duplicates.

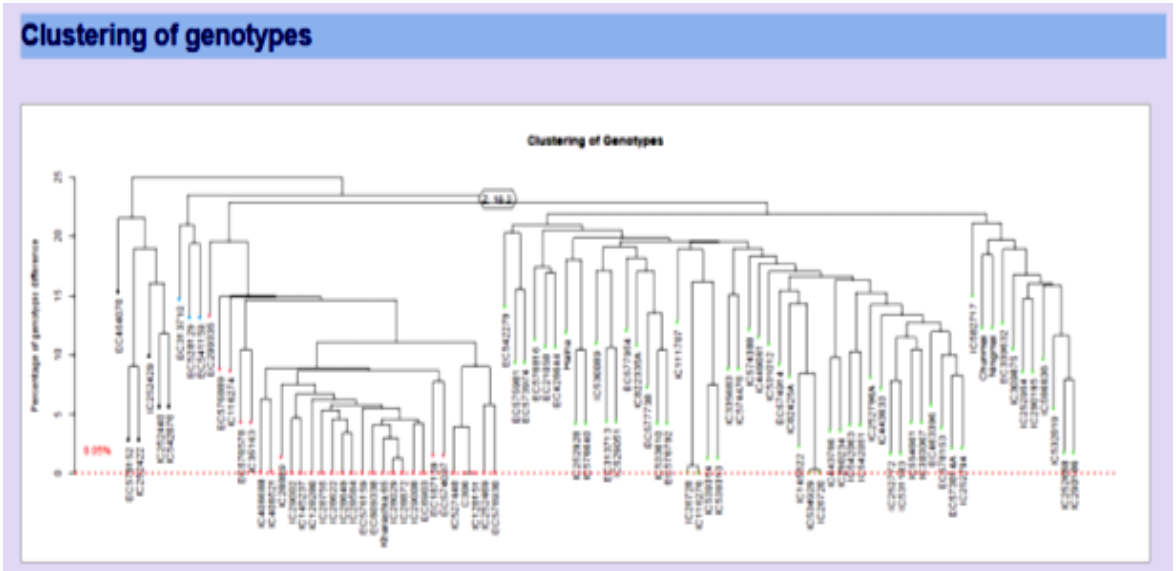


Figure 6. The rectangular cluster

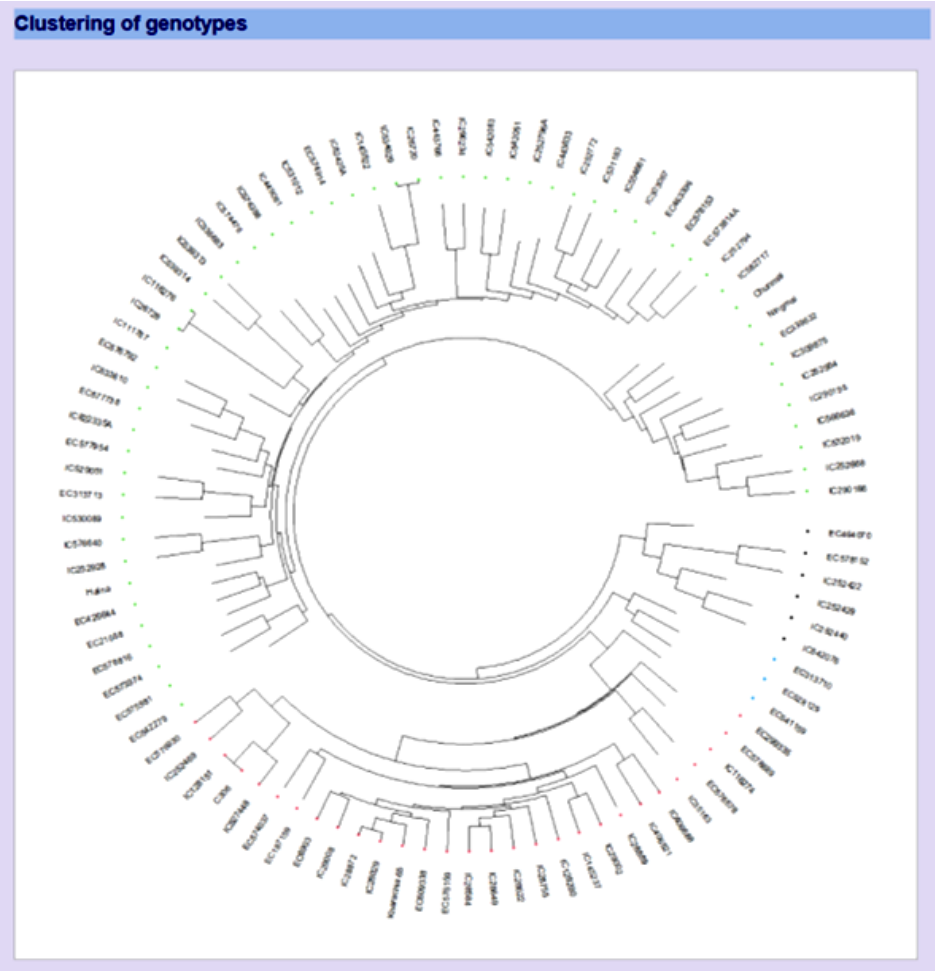


Figure 7. The circular cluster

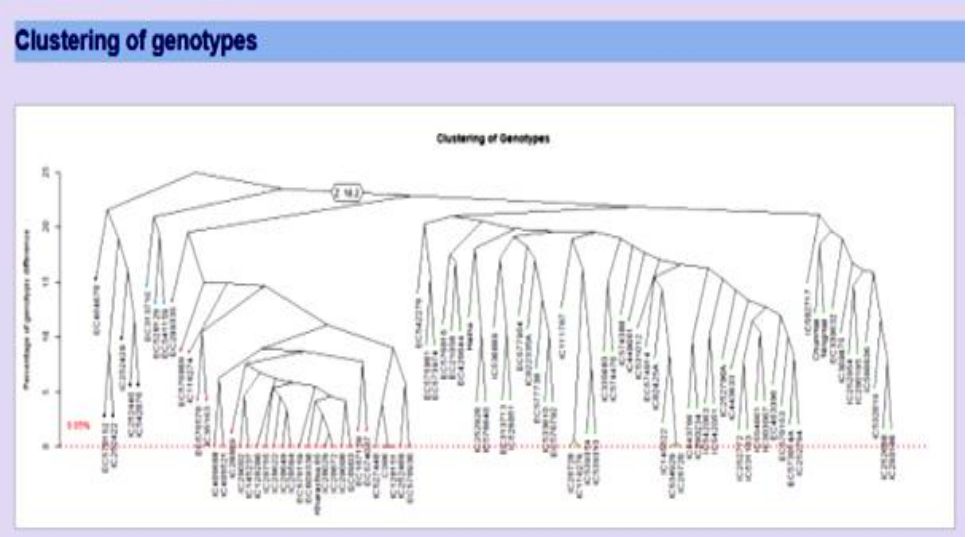


Figure 7. The triangular cluster

K-density graph

The K-density graph depicts the distribution of pairwise genotypic difference based on probability density function.

Genotype and marker heterozygosity plot

The marker heterozygosity plot gives an idea on the number of markers having the heterozygosity below the defined threshold.

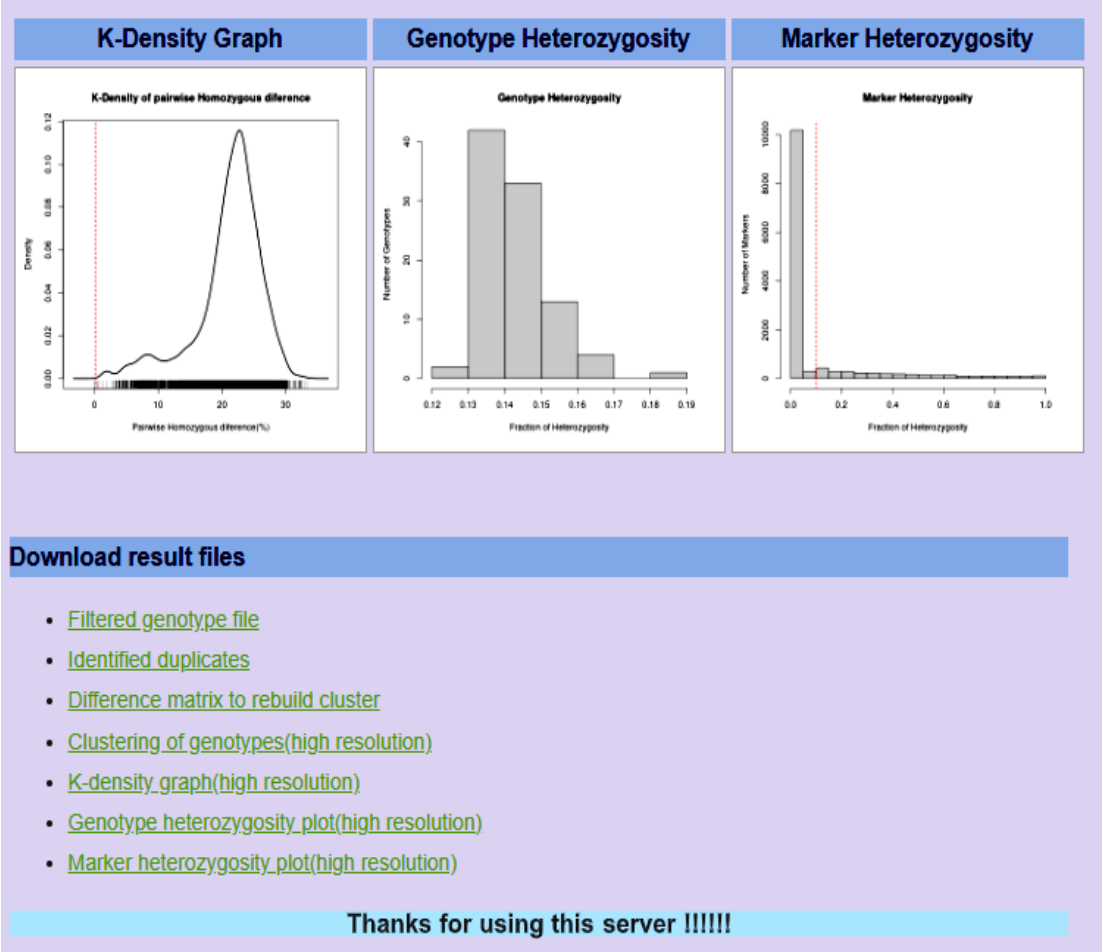


Figure 7. The K-density graph and marker heterozygosity plot

Downloading the result files

The links for downloading the original figures, files along with the final processed genotype file in hapmap format is provided at the bottom of the result page.

……Happy G-DIRTing…..