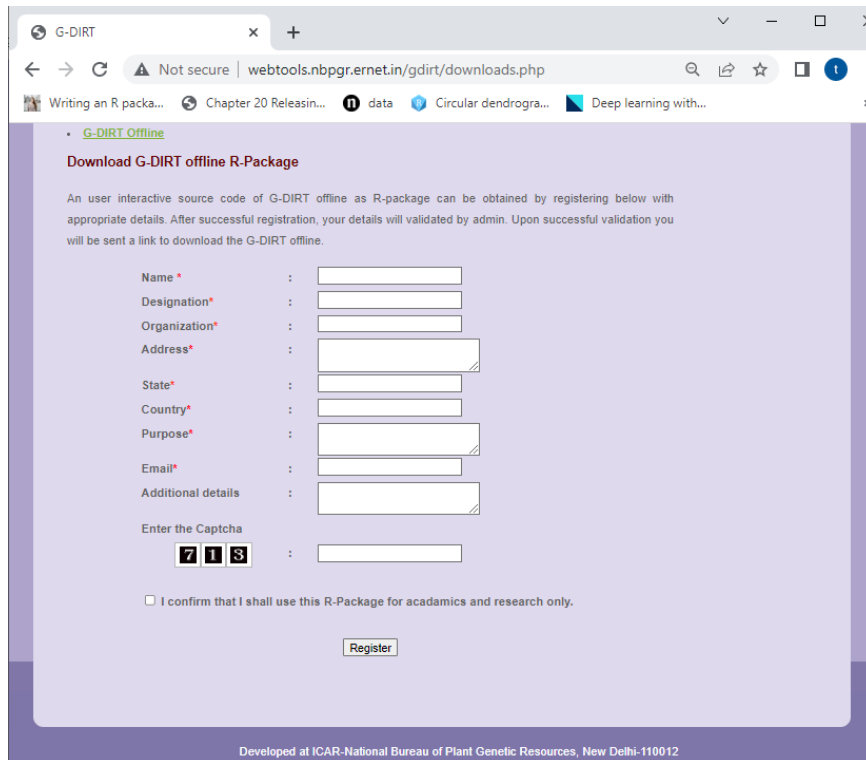


## G-DIRT Offline help for download, installation and execution of the tool

### A. Download and installation

The offline R-package can be obtained by filling up a form available at GDIRT web server downloads page. Upon submission of the form, the user request will be verified and a locally installable R-package will be sent to the user by email.



The screenshot shows a web browser window with the URL `webtools.nbpg.ernet.in/gdirt/downloads.php`. The page title is "G-DIRT Offline" and the main heading is "Download G-DIRT offline R-Package". The text below the heading states: "An user interactive source code of G-DIRT offline as R-package can be obtained by registering below with appropriate details. After successful registration, your details will be validated by admin. Upon successful validation you will be sent a link to download the G-DIRT offline." The form contains the following fields: Name\*, Designation\*, Organization\*, Address\*, State\*, Country\*, Purpose\*, Email\*, and Additional details. Below these fields is a "Enter the Captcha" section with a captcha image showing the numbers 7, 1, and 3, and an input field. At the bottom of the form is a checkbox labeled "I confirm that I shall use this R-Package for academics and research only." and a "Register" button. The footer of the page reads "Developed at ICAR-National Bureau of Plant Genetic Resources, New Delhi-110012".

Figure 1. Form for requesting offline G-DIRT

Installation can be done from the R console by using “**Install package(s) from local files...**” option. From the menu select **Packages** and then “**Install package(s) from local files...**” Give the path of the downloaded GDIRT R-package (`gdirt_0.1.tar.gz`). It will automatically install and/or load the dependent packages upon a function call to `gdirt()`.

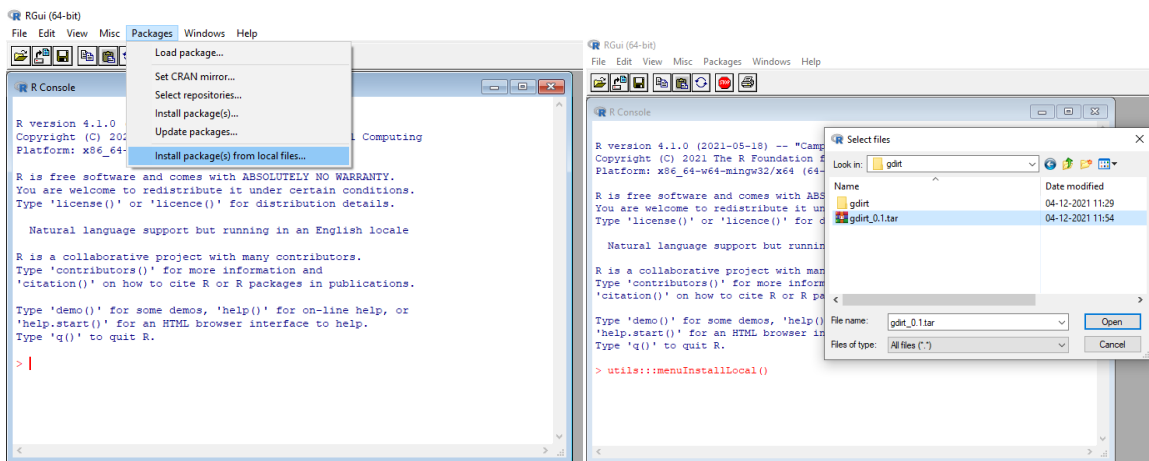


Figure 2. Installation of `gdirt_0.1.tar.gz` from local file

## B. Input file preparation

The input file to GDIRT 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 the order of preference. If the experimenter is biased towards a particular genotype, then 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

## C. 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. The range for the MAF parameter varies from 0 to 1.

rs	alleles	chrom	pos	strand	assembly	center	protLSID	assayLSID	panel	QCcode	EC313710	IC443766	IC252796A
AX-94422082	G/T	1	1145885	+	NA	NA	NA	GG	GG	GG	GG	GG	GG
AX-94598030	A/G	1	1159536	+	NA	NA	NA	GG	GG	GG	GG	GG	GG
AX-94669331	C/T	1	1159689	+	NA	NA	NA	TT	TT	TT	TT	TT	TT
AX-95183288	A/G	1	1159713	+	NA	NA	NA	GG	GG	GG	GG	GG	GG
AX-95217061	A/C	1	1161441	+	NA	NA	NA	AA	AA	AA	AA	AA	AA
AX-94493709	A/G	1	1174865	+	NA	NA	NA	AG	GG	AA	AA	AA	AA
AX-94449086	A/G	1	1190148	+	NA	NA	NA	GG	GG	AA	AA	AA	AA
AX-94898972	C/T	1	1211706	+	NA	NA	NA	TT	TT	TT	TT	TT	TT
AX-94583145	A/G	1	1211895	+	NA	NA	NA	GG	GG	GG	GG	GG	GG
AX-94745699	C/T	1	1235969	+	NA	NA	NA	TT	TT	TT	TT	TT	TT
AX-94772289	C/T	1	1236448	+	NA	NA	NA	CC	CC	CC	CC	CC	CC
AX-94974969	A/G	1	1338148	+	NA	NA	NA	AG	AG	AG	AG	AG	AG
AX-95211874	G/T	1	1340329	+	NA	NA	NA	TG	TG	TT	TG	TG	TG
AX-94778893	A/C	1	1645150	+	NA	NA	NA	AA	AA	AA	AA	AA	AA
AX-94496430	G/T	1	2377608	+	NA	NA	NA	TG	TG	TG	TG	TG	TG
AX-94733072	C/T	1	2378230	+	NA	NA	NA	TC	TC	TC	TC	TC	TC
AX-94700275	A/G	1	2441012	+	NA	NA	NA	AA	AA	AA	AA	AA	AA
AX-94581354	A/C	1	2920151	+	NA	NA	NA	CC	CC	CC	CC	CC	CC
AX-94492529	A/T	1	3118372	+	NA	NA	NA	AA	AA	AA	AA	AA	AA
AX-94550967	C/T	1	3381102	+	NA	NA	NA	CC	CC	CC	CC	CC	CC
AX-94936441	C/T	1	3388719	+	NA	NA	NA	CC	CC	CC	CC	CC	CC
AX-94386458	A/C	1	3392719	+	NA	NA	NA	CC	CC	CC	CC	CC	CC
AX-95162217	A/G	1	3761117	+	NA	NA	NA	GG	GG	AA	NN	GG	AA
AX-95158783	C/G	1	3845903	+	NA	NA	NA	CG	CG	CG	NN	CG	CG
AX-94664734	A/G	1	3847072	+	NA	NA	NA	GG	GG	GG	NN	GG	GG
AX-95082131	A/G	1	3847127	+	NA	NA	NA	AA	AA	AA	GG	AA	AA
AX-94438601	C/G	1	4033356	+	NA	NA	NA	CC	CC	CC	CC	CC	CC
AX-9468213	C/T	1	4048434	+	NA	NA	NA	CC	CC	CC	CC	CC	CC
AX-94979270	A/C	1	4053566	+	NA	NA	NA	GG	GG	AA	GG	GG	AA
AX-94945604	A/C	1	4502788	+	NA	NA	NA	NN	AC	AC	CC	AC	CC
AX-95219658	A/G	1	4812661	+	NA	NA	NA	AA	AA	AA	AA	AA	AA
AX-95139456	A/G	1	5025329	+	NA	NA	NA	AA	AA	AA	GG	AA	GG
AX-95230167	A/G	1	5028362	+	NA	NA	NA	GG	GG	GG	GG	AA	AA
AX-94562727	G/T	1	5646294	+	NA	NA	NA	TG	GG	GG	GG	TG	GG
AX-94661960	A/G	1	6067447	+	NA	NA	NA	AG	AA	AA	AA	AA	AA
AX-94741678	A/G	1	6150999	+	NA	NA	NA	AA	AG	AG	AG	AA	AG
AX-95106319	C/G	1	6153162	+	NA	NA	NA	GG	GG	GG	GG	GG	GG
AX-94482516	C/T	1	6153189	+	NA	NA	NA	TT	TC	TC	TT	TC	TT

Figure 3. An example of a hapmap genotype file

### 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 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. If your data or analysis does not LD pruning, you may set the threshold value to 1.

### 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.

## **E. Execution**

The GDIRT can be executed by a function call *i.e.*, `gdirt()` after loading the library using the command **`library(gdirt)`**. No argument is required to specify with the function call because the main **`gdirt()`** function interacts with the user by a set of statements requiring user input.

## **E. Output**

Upon successful execution the result summary will be displayed in the console and the result files will appear in the provided default directory. The default/output directory will contain many files such as (i) list of duplicates, (ii) difference matrix, (iii) final genotype file, (iv) IBS summary, (v) k-density graph, (vi) heterozygosity plot and (v) intermediate VCF and GDS files.

```

RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help

R version 4.1.0 (2021-05-18) -- "Camp Fontanezen"
Copyright (C) 2021 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.
> gdist()

Loading/installing required dependent packages.....

```

Figure 4. Calling gdist() function

```

Please press ENTER to Choose the default directory :
Your default directory is 'D:/TKSahu/IBS/tanmaya'

Please press ENTER to select the HAPMAP genotype file:
Remove Monomorphic SNPs(type 1 for TRUE and 0 for FALSE): 1
Please enter MAF threshold : 0.05
Please enter missing data(marker) threshold : 0.1
Please enter missing data(genotype) threshold : 0.1
Please enter LD pruning threshold : 0.75
Please enter 1 for 'Total genotypic difference' and 2 for 'Homozygous difference' based duplicates identification : 2
Please enter the threshold for homozygous difference(%): 0.05
Please enter Hardy Weinberg equilibrium(HWE) threshold : 0
Please enter the threshold for heterozygosity : 0.1
Generate Cluster(1-Not Required, 2-Rectangular, 3-Triangular, 4-Circular ):2

Parameters Entered
-----
Remove Monomorphic SNPs :TRUE
Minor Allele Frequency : 0.05
Missing rate(marker) : 0.1
Missing rate(genotype) : 0.1
LD pruning : 0.75
Homozygous difference(%) : 0.05
HWE : 0
Heterozygosity : 0.1
Generate Cluster : rectangle
-----

```

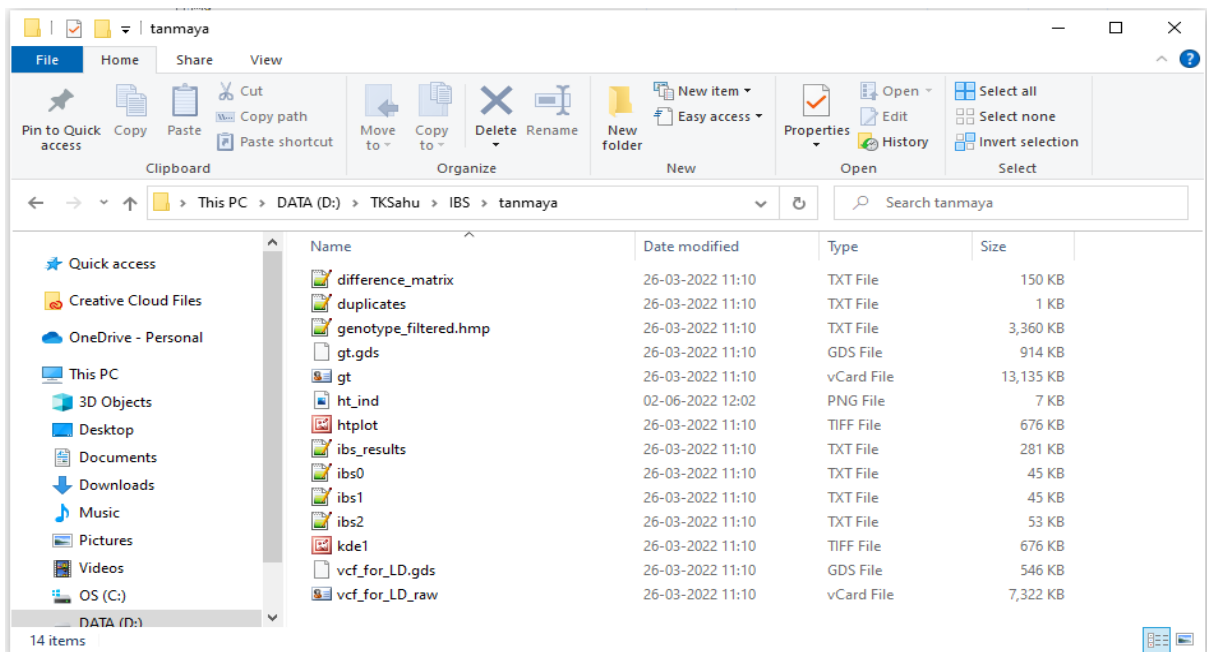
Figure 5. Interactive parameter input

```

The input hapmap file contains 31926 markers for 95 genotypes.
95 Genotypes retained after missing data filtration of 10%
Duplicate removal based on Identity By State analysis started at Homozygous difference of 0.05 %.
94 accessions retained after removing duplicates with less than 0.05 % of Homozygous difference.
Density graph generated....
Determine groups by permutation (Z threshold: 15, outlier threshold: 5):
Create 5 groups.
15194 markers retained after LD pruning at a threshold of 0.75.
HWE and Heterozygosity filtration of SNPs started...
11434 markers retained after data filtration with HWE threshold of 0 and Heterozygosity threshold of 0.1.
11434 markers and 94 genotypes retained after all filtration.

```

Figure 6. Result summary



*Figure 6. The output directory containing result files*

*Note: For interpretation of result user may refer online help manual.*

**.....Happy G-DIRting.....**