SAFE-h^2 Manual

SAFE-h^2 is a python script designed for estimation of SNP heritability. It provides heritability profiles and thereby improves estimation accuracy of SNP heritability. SAFE-h^2 not only prevents the biologically meaningless negative contributions to heritability made by SNPs with large *p*-values (defined as "non-causals") but also avoids the false-positive SNP hits from the lower *p*-value tail. It can also extract intra-locus non-additive allelic effects and estimate the heritability by considering both the additive and non-additive allelic effects. This also makes it possible to perform GWAS using linear models while covering non-additive allelic effects.

Citation:

Behrooz Darbani, Mogens Nicolaisen. SNP Allocation For Estimating Heritability (SAFE-h^2): A tool to explore genomic origins of phenotypes for estimation of SNP heritability using additive-only allelic effects or additive and non-additive allelic effects. doi: https://doi.org/10.1101/2023.08.28.555092

Contact: behroozdarbani@gmail.com , bd@agro.au.dk

Department of Agroecology, Faculty of Technical Sciences, Aarhus University, Slagelse, Denmark

SAFE-h2 benefits from two major technical features:

I) A *p*-value guided SNP heritability profiling approach to include as many causal SNPs as possible in the heritability estimation models while expelling non-causal SNPs. To do this, SAFE-h^2 finds a *p*-value threshold differentiating causal and non-causal SNPs. This approach facilitates the estimation of phenotype-specific, *i.e.*, causal SNPs-specific, genetic variances. SAFE-h^2 also facilitates a false-positive free estimation of SNP heritability.

II) Exploiting an allele adjustment strategy which makes it possible to include the additive effects together with different non-additive allelic effects of dominance, overdominance, and heterosis-like in the heritability estimation models. This also makes it feasible to perform GWAS using linear models while taking both additive and non-additive allelic effects into account.

Installation requirements:

1) Install pip, gawk, and some python modules:
   - A) sudo apt update or sudo dnf update
   - B) sudo apt install python3-pip or sudo dnf install python3-pip
   - C) sudo apt install gawk or sudo dnf -y install gawk-devel
   - D) pip install pandas
   - E) pip install numpy
   - F) pip install rich
   - G) pip install matplot
   - H) pip install matplotlib
   - I) pip install pathlib

J) pip install pyfiglet

2) Download PLINK1.9 binary file from: https://www.cog-genomics.org/plink/
3) Download PLINK2 binary file from: https://www.cog-genomics.org/plink/2.0/
4) Download GCTA-1.94.1-linux-kernel-3 binary file from: https://yanglab.westlake.edu.cn/software/gcta/#Download
5) Download LDAK5.2.linux binary file from: https://dougspeed.com/downloads2/
6) Download EMMAX-beta-07Mar2010 binary files from: http://csg.sph.umich.edu//kang/emmax/download/index.html
7) Unpack the SAFE-h2 zip file.
8) Move the binary files into the SAFE-h2 folder (rename the binaries to plink, plink2, gcta, ldak, emmax, and emmax-kin).

Usage:

1) Prepare and put the bfiles and the GWAS *p*-value file into the SAFE-h2 folder. SAFE-h^2 uses these files as input. The bfiles should be named as "MAIN.fam, MAIN.bed, and MAIN.bim". The *p*-value file should be named MainPs. It is a tab-delimited text file with three columns as:

SNP1_ID	2.04562	9.5E-10
SNP2_ID	-1.87003	0.002
SNP3_ID	0.23419	0.092567
…	…	…

The first col. is for IDs, the second can be beta coefficient, and the third col. is for *p*-values. The covariates can also be included in the analyses. They should also be provided as tab-delimited text files.
For EMMAX model the covariate file should be named as "Covar_emmax".

Sample1	Sample1	1	2.657
Sample2	Sample2	1	3
Sample3	Sample3	1	9.342
…	…	…	…

For LDAK model the covariate file should be named as "Covar_ldak".

Sample1	Sample1	2.657
Sample2	Sample2	3
Sample3	Sample3	9.342
…	…	…

For GCTA-GREML model the quantitative and qualitative covariates should be provided by separate files. Please bear in mind that, for now, one of these should be provided. The quantitative covariate file should be named as "qCovar_greml" and the qualitative covariate as "Covar_greml".

Sample1	Sample1	2.657
Sample2	Sample2	3
Sample3	Sample3	9.342
…	…	…

Please bear in mind that the sample IDs should be identical to the sample IDs within your fam file.

2) Open the terminal from the SAFE-h2 folder and run:

python3 SAFE-h2.py or python3 SAFE-h2.py | tee output

This will perform the SNP heritability profiling by implementing the EMMAX model by default. It is possible to implement other estimation models by changing the configuration file in advance. For this:
Open the configuration file (Prog.config) and define the type of analyses you are going to use as:

Impl_status_Emmax, Impl_status_LdakGCTA, Impl_status_LdakThin, Impl_status_GCTA-GREML, Test_ADOH, PURE_Effects
X1 X2 X3 X4 X5 X6

X1-4 can get the values of 0 or 1 to ignore or implement the corresponding heritability estimation model within the analyses. **X4** can also get 2 when dealing with inbred lines or when the inbreeding coefficient indicates some levels of inbreeding.
The run will create and save the output folders of files and figures.

3) To consider the non-additive allelic effects in addition to the additive effects, define **X5** as "initial_files", "intermediate_files", or "plink_based_all" within the configura-tion file. The default is "plink_based_all". This allows SAFE-h2-preADOH to do the allele adjustments and association tests together. It uses the "glm" function of the PLINK for the association test.

To do this, put your bfiles ("MAIN.fam, MAIN.bed, and MAIN.bim") within the SAFE-h2 folder. Covariates can also be provided under the name of Covar_plink as a tab-delimited text file:

FID	IID	Cov001
Sample1	Sample1	3
Sample2	Sample2	9
…	…	…

Now, the user can open the terminal from the SAFE-h2 folder and run:

python3 SAFE-h2-preADOH.py or python3 SAFE-h2-preADOH.py | tee output1

The run will create and save the output folder of results for allele adjustments. To perform the SNP heritability profiling based on additive and non-additive effects (on the allele adjusted genotypes), run (again from the SAFE-h2 folder):

python3 SAFE-h2-ADOH.py or python3 SAFE-h2-ADOH.py | tee output2

(Before the run, **X1-4** within the configuration file can get the values of 0 or 1 to ignore or implement different heritability estimation models. **X4** can also get 2 when dealing with inbred lines or when the inbreeding coefficient indicates some levels of inbreeding)
The user can also consider providing covariates as Covar_emmax, Covar_ldak, and qCovar_greml or Covar_greml as explained at the beginning for SAFE-h^2.
The run will create and save the output folders of figures and files.

3-1) Instead of the "plink_based_all", user can set the option of "initial_files" within the configuration file to perform any other association test of interest on the 5 different adjusted genotypic datasets. Here, SAFE-h2-preADOH provides 5 different genotypic datasets (Additive, Dominance [over Ref allele], dominance [over Alt allele], Overdominance [over Ref allele], overdominance [over Alt allele], and Heterosis). For this run:
python3 SAFE-h2-preADOH.py or python3 SAFE-h2-preADOH.py | tee output1

3-2) The user will need to perform association tests and put the *p*-value files within the SAFE-h2 folder. The format of *p*-value files should be the same as explained for MainPs. They should be named as "**AdditivePs**" for Additive data, "**DominancePs**" for Dominance adjusted data, "**dominancePs**" for dominance adjusted data, "**OverdominancePs**" for Overdominance adjusted data, "**overdominancePs**" for overdominance adjusted data, and "**HeterPs**" for Heterosis adjusted data.

Furthermore, the user must set the option of "intermediate_files" instead of "initial_files" within the configuration file and run:
python3 SAFE-h2-preADOH.py or python3 SAFE-h2-preADOH.py | tee output2

This will create and save the output folder and combinatory files that can be used for SNP heritability profiling by SAFE-h2-ADOH. Before this step, the user needs to run another association test of interest and put all *p*-value files into the folder. They should be names as MainPs_Add, MainPs_AddDom, MainPs_AddDomOD, MainPs_AddDomODHet.

For SNP heritability profiling by SAFE-h2-ADOH, the user can also include covariates files. Finally, for performing the SNP heritability profiling based on additive and non-additive effects (on the allele adjusted genotypes), run:
python3 SAFE-h2-ADOH.py or python3 SAFE-h2-ADOH.py | tee output3
(Before the run, **X1-4** within the configuration file can get the values of 0 or 1 to ignore or implement different heritability estimation models. **X4** can also get 2 when dealing with inbred lines or when the inbreeding coefficient indicates some levels of inbreeding)
The run will create and save the output folders of figures and files.

4) To estimate the SNP heritability after minimizing the likelihood for contributions by false-positive SNP hits, **X6** in the configuration file (Prog.config) should be 1. The rest of the settings are already described in section 2. In addition to the MAIN.bim, MAIN.bed, MAIN.fam, and MainPs files for the original phenotype of interest, the user should also provide MAIN1.fam, Main1Ps, MAIN2.fam, Main2Ps, MAIN3.fam, Main3Ps, MAIN4.fam, Main4Ps, MAIN5.fam, Main5Ps, MAIN6.fam, and Main6Ps files related to the 6 independent random phenotypes simulated within the range of phenotype-of-interest. To perform the analysis, run:

   python3 SAFE-h2.py or python3 SAFE-h2.py | tee output

   The run will create and save the output folders of files and figures for the original phenotype (named as SAFE_...) and for the random phenotypes 1 to 6 (named as RP1_... to RP6_...). SAFE-h^2 also provides the user with aligned graphs and related datasets (named as RP1_6…).

Output files:

For heritability profiling by additive-only effects (see section 2):

The heritability profiling using Emmax model is saved as outfileE, outfileY_E, & Heritability_Bar_Graph_E

The heritability profiling using LDAK GCTA_model is saved as outfileL, outfileY_L, & Heritability_Bar_Graph_L

The heritability profiling using LDAK Thin_model is saved as outfileLT, outfileY_LT, & Heritability_Bar_Graph_LT

The heritability profiling using GCTA-GREML model is saved as outfile_G, outfileY_G, & Heritability_Bar_Graph_G

Clustered SNP hits are saved as Number_of_pvalues & Clustered_SNP_Hits_Bar_Graph

For heritability profiling by additive and non-additive effects (see section 3):

The heritability profiling using Emmax model is saved as outfileE_¤, outfileY_E_¤, & Heritability_Bar_Graph_E

The heritability profiling using LDAK GCTA_model is saved as outfileL_¤, outfileY_L_¤, & Heritability_Bar_Graph_L

The heritability profiling using LDAK Thin_model is saved as outfileLT_¤, outfileY_LT_¤, & Heritability_Bar_Graph_LT

The heritability profiling using GCTA-GREML model is saved as outfile_G_¤, outfileY_G_¤, & Heritability_Bar_Graph_G

Clustered SNP hits are saved as Number_of_pvalues_¤ & Clustered_SNP_Hits_Bar_Graph

¤ is Add for additive model.

¤ is AddDom for additive and dominance model.

¤ is AddDomOD for additive, dominance, and overdominance model.

¤ is AddDomODHet for additive, dominance, overdominance, and heterosis-like model.

The allele adjusted genotypic datasets also include:

MAIN_¤.bed, MAIN_¤.bim, MAIN_¤.fam, & MainPs_ ¤

For heritability profiling after minimizing the false-positive contributions (see section 4):

For each of the original and random phenotypes:

The heritability profiling using Emmax model is saved as outfileE, outfileY_E, & Heritability_Bar_Graph_E

The heritability profiling using LDAK GCTA_model is saved as outfileL, outfileY_L, & Heritability_Bar_Graph_L

The heritability profiling using LDAK Thin_model is saved as outfileLT, outfileY_LT, & Heritability_Bar_Graph_LT

The heritability profiling using GCTA-GREML model is saved as outfile_G, outfileY_G, & Heritability_Bar_Graph_G

Clustered SNP hits are saved as Number_of_pvalues & Clustered_SNP_Hits_Bar_Graph

And for all phenotypes together:

The heritability profiling using Emmax model is saved as outfileY_E, outfileY_E1, outfileY_E2, outfileY_E3, outfileY_E4, outfileY_E5, outfileY_E6, outfileY_E16, E16, E16_error, & Heritability_Bar_Graph_E-purifiedEffects

The heritability profiling using LDAK GCTA_model is saved as outfileY_L, outfileY_L1, outfileY_L2, outfileY_L3, outfileY_L4, outfileY_L5, outfileY_L6, outfileY_L16, L16, L16_error, & Heritability_Bar_Graph_L-purifiedEffects

The heritability profiling using LDAK Thin_model is saved as outfileY_LT, outfileY_LT1, outfileY_LT2, outfileY_LT3, outfileY_LT4, outfileY_LT5, outfileY_LT6, outfileY_LT16, LT16, LT6_error, & Heritability_Bar_Graph_LT-purifiedEffects

The heritability profiling using GCTA-GREML model is saved as outfileY_G, outfileY_G1, outfileY_G2, outfileY_G3, outfileY_G4, outfileY_G5, outfileY_G6, outfileY_G16, G16, G16_error, & Heritability_Bar_Graph_G-purifiedEffects