

LD Score Regression

Author: Golden Helix, Inc.

Overview

This document describes the LD Score Regression methodology capability in SVS.

The implementation leverages some of the functionality from the Python implementation of the LDSC paper, found at [this github location](#). The input and output of these functions are managed by SVS, along with a couple of utility features to prepare your data.

The overall workflow consists of the following steps:

1. Acquire precomputed LD Scores or compute your own LD Scores for a population represented by the samples in your GWAS study.
2. Join the LD Scores with your GWAS association test results.
3. Run LD Score Regression on your merged spreadsheet, choosing to run in one of two modes:
 - a. Heritability estimate for a single study
 - b. Genetic Correlation between multiple studies

More background on these workflows can be found at the tutorial page for LDSC on [Heritability and Genetic Correlation](#).

Installing the Script

First, save the python script file to the following directory:

***..\Application Data\Golden Helix SVS\UserScripts\Spreadsheet\Genotype**

Note: The **Application Data** folder is a hidden folder on Windows operating systems and its location varies between XP and Vista. The easiest way to locate this directory on your computer is to open SVS and go to **Tools > Open Folder > UserScripts Folder**. If saved to the proper folder, this script will be accessible from the spreadsheet **Genotype** menu.

Second, download the ldscore.zip file and extract the folder named ldscore.

Copy or move the folder to the SVS Install directory under the **Lib\stats** sub-folder.

Note: You can navigate to the SVS install location by going to **Tools > Open Folder > SVS Install Folder**.

Obtaining LD Scores

The first step needed to perform LD score regression is to obtain LD scores. The following methods are available:

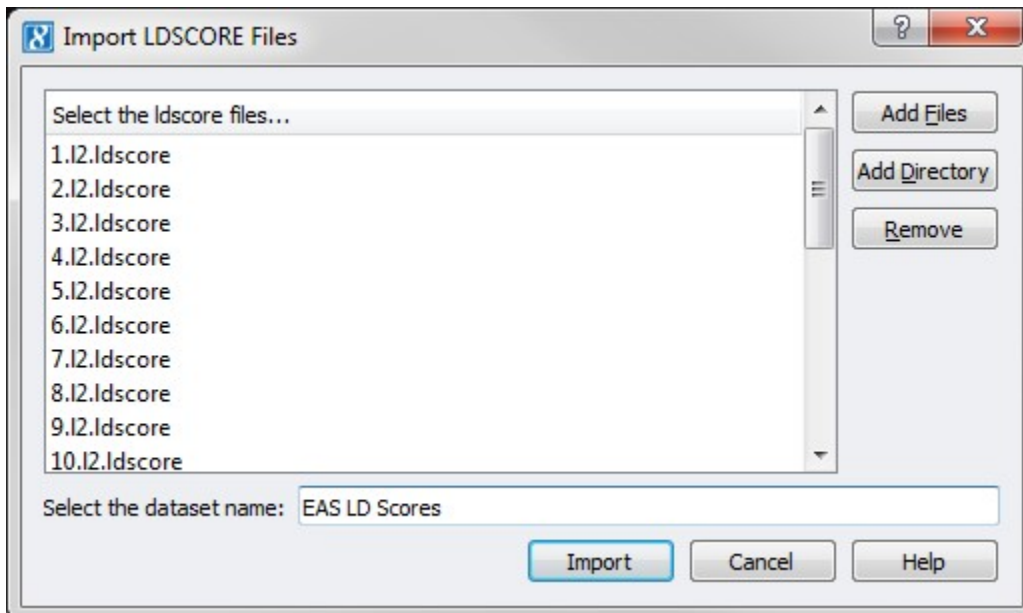
1. Using your own test data, perform **Genotype > Quality Assurance and Utilities > LD Score Computation and Binning**.
2. Use the LDSC utility described in the LDSC paper to create LDscore files on your own population, then import these files.
3. Use the DSF files Golden Helix created by importing the precomputed LDSC files provided by the LDSC paper, for:
 - East Asian populations
 - European populations

Importing LD Scores When Using Method 2 or Method 3

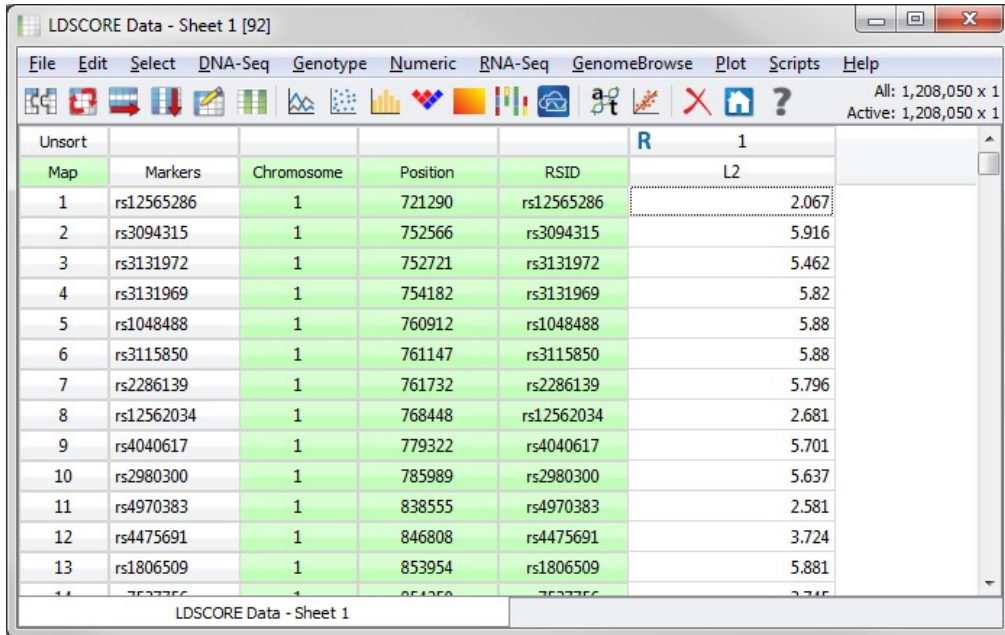
To import LD scores from LDSC go to **Import > Import LDSCORE Files**.

Note: The LDscore files provided by LDSC are broken up by chromosome and each individually compressed with gzip. You will need to uncompress them first to get to “ldscore” files.

On the dialog screen, select **Add Files** and select *.ldscore files. The spreadsheet name can be edited in the **Select the dataset name** field.



The *.ldscore files will be compiled into one spreadsheet.



Map	Markers	Chromosome	Position	RSID	L2
1	rs12565286	1	721290	rs12565286	2.067
2	rs3094315	1	752566	rs3094315	5.916
3	rs3131972	1	752721	rs3131972	5.462
4	rs3131969	1	754182	rs3131969	5.82
5	rs1048488	1	760912	rs1048488	5.88
6	rs3115850	1	761147	rs3115850	5.88
7	rs2286139	1	761732	rs2286139	5.796
8	rs12562034	1	768448	rs12562034	2.681
9	rs4040617	1	779322	rs4040617	5.701
10	rs2980300	1	785989	rs2980300	5.637
11	rs4970383	1	838555	rs4970383	2.581
12	rs4475691	1	846808	rs4475691	3.724
13	rs1806509	1	853954	rs1806509	5.881

The raw precomputed LD Scores can be downloaded at the sites below:

- East Asian 1kG LD Scores: https://data.broadinstitute.org/alkesgroup/LDSCORE/eas_ldscore.s.tar.bz2
- European 1kG LD Scores: https://data.broadinstitute.org/alkesgroup/LDSCORE/eur_w_ld_c/hr.tar.bz2

The pre-processed and ready to import into SVS versions are at:

- East Asian LD Scores: http://data.goldenhelix.com/d/ldsc/EAS_LD_Scores.dsf
- European LD Scores: http://data.goldenhelix.com/d/ldsc/EUR_LD_Scores.dsf

Using the Script

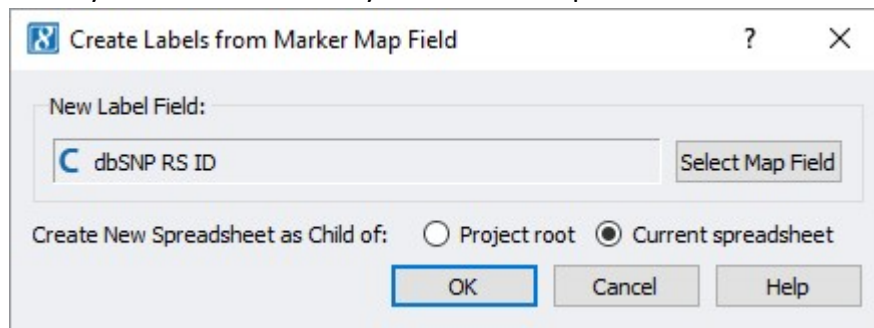
Join the LD Scores with the GWAS Association Test Results

Next, join the LD Scores and the results of your GWAS Association Test Results spreadsheets. The best way to do this, if you are using Method 2 or Method 3, is to match on the RSID field as row labels in both spreadsheets.

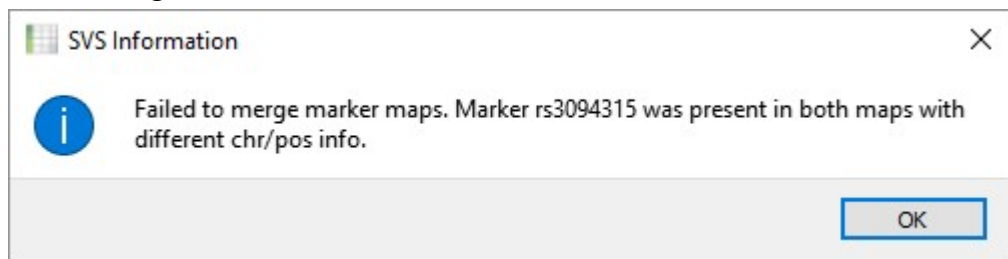
Creating RSID Row Labels for Method 2 or Method 3

If your row labels of your association test results spreadsheet are not RSID values, you will need to recode them to be RSID values before joining. If you have RSID values in your marker map, you can follow these steps to do this:

- From your association test spreadsheet, go to **Edit -> Create Labels from Marker Map**.
- Select your RS ID field from your marker map:



- Join the output spreadsheet with the LDSCORE imported spreadsheet.
- Note that if your data is not using the GRCh37 human reference, you may receive this message:



In this case, from the LDSCORE spreadsheet, go to **File -> Drop Genetic Marker Map** and then do the join again with the un-mapped spreadsheet.

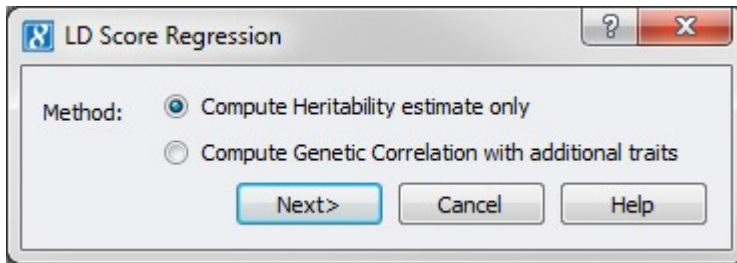
Assumptions About Your Data

It is suggested to perform quality assurance by filtering out low quality markers based on their imputation quality and/or minor allele frequency (MAF). This should be done before running **LD Score Regression**.

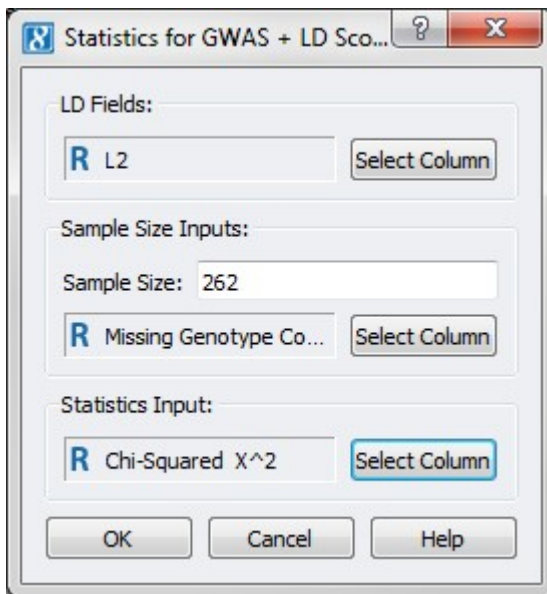
For example, using the recommendations of the [Heritability and Genetic Correlation Tutorial](#), filter out poorly-imputed SNPs with a R^2 or "INFO" value of less than 0.9.

Computing LD Score Heritability Estimates

From the merged spreadsheet, go to **Genotype > LD Score Regression** and select **Compute Heritability estimate only**.



Fill in the fields in the next window. The **Missing Genotype Column** is not required, but if included, it will decrease the sample size count for a given marker. The sample size should be the number of samples included for each SNP.



Click **OK** to start the analysis.

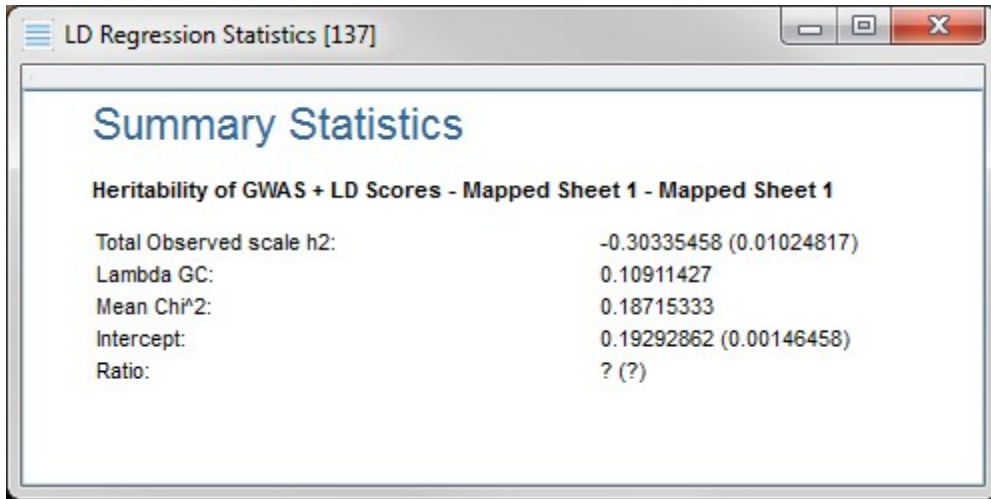
The remaining options are as follows:

- **LD Score Field:** This is the joined “L2” field from the LDSC Scores field you previously joined to your association results.
- **Sample Size:** The number of samples in your genotype spreadsheet used to compute the association results.
- **Missing Genotype Column:** When running Genotype Association Tests, select “Genotype counts” in the “Overall Marker Statistics” tab to get a “Missing Genotype Count” output to your output spreadsheet. That field will be detected automatically by LD Score Regression if present. It is used to adjust the sample

size to a marker-specific value of the samples used for computing that marker's test statistic.

- **Summary Statistic Column:** This is the P-value or other test statistic you want to use for the computation. If it is a signed test statistic, that sign indicates the direction of the effect for the minor allele.

When the script is done, a result viewer will appear with the estimated heritability.



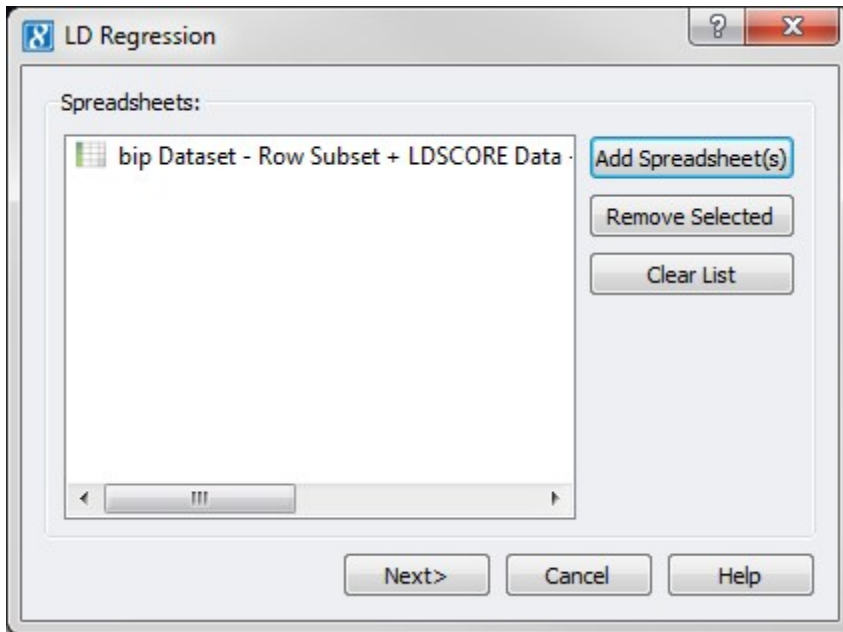
The screenshot shows a window titled "LD Regression Statistics [137]" with a "Summary Statistics" section. The data is as follows:

Heritability of GWAS + LD Scores - Mapped Sheet 1 - Mapped Sheet 1	
Total Observed scale h2:	-0.30335458 (0.01024817)
Lambda GC:	0.10911427
Mean Chi ² :	0.18715333
Intercept:	0.19292862 (0.00146458)
Ratio:	? (?)

Computing LD Score Genetic Correlation with Additional Traits Estimates

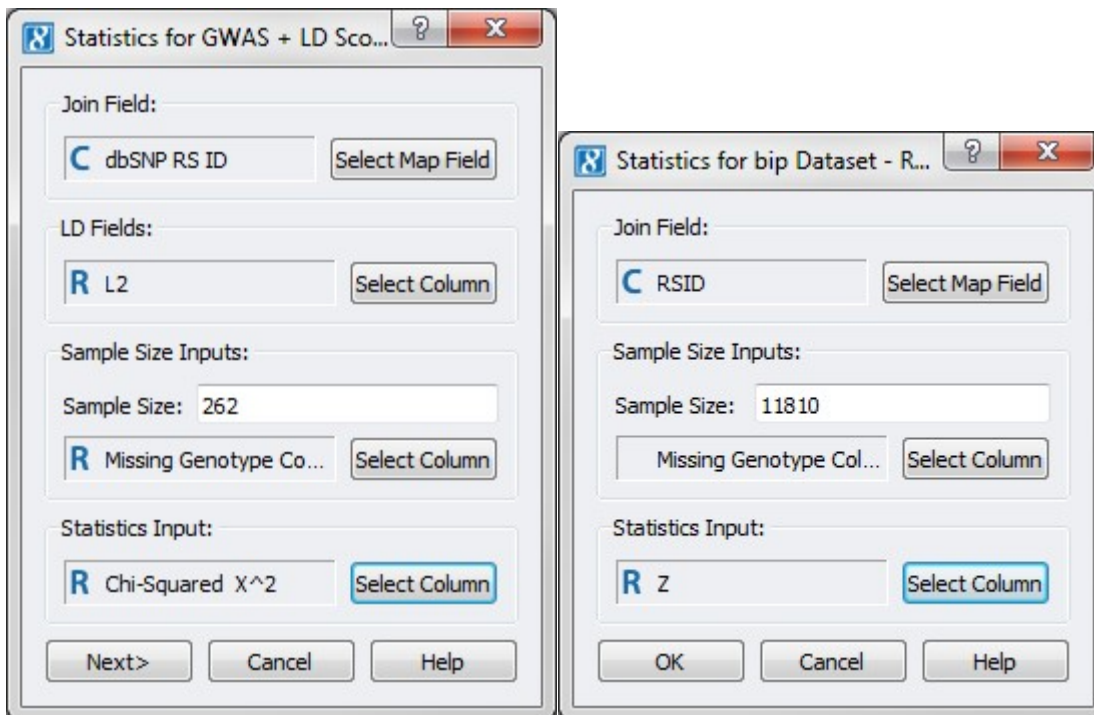
To get the genetic covariance and correlation in addition to the heritability select **Compute Genetic Correlation with additional traits** on the first dialog.

In the second spreadsheet, select additional spreadsheets. Pairwise comparisons will be made between the first spreadsheet and each of the selected spreadsheets.



The computation will be done on the intersection of all the shared markers between traits.

The first dialog will allow you to enter information from the first spreadsheet and the subsequent dialogs will pertain to the additional trait spreadsheets.



The “Join Field” is the row labels or the marker map field that can be used to match the markers in your current spreadsheet with the ones in the additional trait spreadsheets. The algorithm will create a Results Viewer with the output of the computation. In the results viewer, there will be heritability estimates for each spreadsheet and the genetic covariance and correlation between the first spreadsheet and each subsequent spreadsheet. Standard errors are included in parentheses.

