Here we describe a typical analysis using SparSNP. We assume that we have two datasets in PLINK BED/BIM/FAM format, named discovery and validation, on the same SNP microarray platform. The phenotype in the FAM file can be discrete (1 for controls and 2 for cases) or continuous. In the following, we assume the user is working with a Unix-like command line shell such as Bash.
1. Quality control
SparSNP implements simple imputation of genotypes such that missing genotypes are randomly assigned the value \{0,1,2\} with probability of 1/3 each, which does not introduce substantial bias to the model when the proportion of missingness is small and the genotypes are missing at random. We recommend removing samples and SNPs with high missingness, and testing for differential missingness between cases and controls.

- Remove markers with missingness \(\geq 0.01\), MAF \(\leq 0.01\), and Hardy-Weinberg test for equilibrium in controls of \(p \leq 10^{-4}\), and samples with missingness \(\geq 0.01\):
  
  $\text{plink --bfile discovery --geno 0.01 --mind 0.01 \ 
  --maf 0.01 --hwe 0.0001 \ 
  --make-bed --out discovery\_filtered}$

- Test for differential missingness:
  $\text{plink --file discovery\_filtered --missing}$
  (remove SNPs as indicated by output)

- Test for sample relatedness with a threshold of \(\hat{\pi} = 0.05\):
  $\text{plink --file discovery\_filtered --Z-genome --min 0.05}$
  (remove related samples as indicated by output)

- Check for stratification using PCA, for example, using smartpca in Eigensoft (Price et al., 2006).

- Two locus test for detecting batch effects (Lee et al., 2010).
  (remove SNPs as indicated by output)

2. SparSNP on discovery data

- Run cross-validation. By default, 3 x 3-fold cross-validation will be performed, allowing up to 1024 SNPs in the model, using a case/control phenotype:
  for classification (case/control)
  $\text{./cv3.sh discovery\_filtered sqrhinge}$
  and for continuous outputs (linear regression)
  $\text{./cv3.sh discovery\_filtered linear}$
  The cv3.sh script can be modified to perform more cross-validation folds or to tune the model parameters.

- Plot AUC and explained phenotypic and genetic variance in the 3 cross-validation replications, with the optional population prevalence \(K = 1\%\) and heritability \(h^2_L = 50\%:\)
  $\text{./eval.sh discovery\_filtered prev=0.01 h2l=0.5}$
  The plots discovery\_filtered\_AUC.pdf, discovery\_filtered\_VarExp.pdf, and discovery\_filtered\_GenVarExp.pdf will be produced in the directory discovery. The raw AUC and explained variance data is stored in the R-data file named discovery\_filtered.RData. The prevalence and heritability are optional. Prevalence is needed for computing explained genetic and phenotypic variance, and heritability is needed for computing explained genetic variance.

- The set of models with best predictive ability will automatically be chosen from the results, based on smoothing of the AUC or \(R^2\). The SNPs appearing in these models will be tabulated according to how many time they were included over all
cross-validation folds. To inspect the SNPs selected by models that maximise the predictive ability:

```bash
cat discovery/topsnps.txt
```

| RS Counts | Proportion | Replications |
|-----------|------------|--------------|
| rs2050189 | 60         | 1            |
| rs2187668 | 60         | 1            |
| rs9357152 | 60         | 1            |
| rs7774954 | 60         | 1            |
| rs3129763 | 58         | 0.966666666666667 |

The SNPs are ordered by the number of times they were included in a model with non-zero weight (Counts) out of the total number of cross-validation folds (Replications), also shown as a Proportion. SNPs at the top of the list are more stably selected by the lasso and are potentially more robust markers than SNPs at the bottom of the list.

3. Optional: Apply models to validation data

SparSNP models trained on the discovery dataset can be tested on an independent dataset, if one is available. It is crucial for the validation dataset to contain the same SNPs in the same order and use the same allele as the minor allele.

- Use the same SNP list as discovery dataset:
  ```bash
  awk '{print $2}' discovery_filtered.bim > discovery_filtered.snps
  plink --bfile validation --extract discovery_filtered.snps --make-bed --out validation_filtered
  ```

- Use same reference allele as discovery dataset:
  ```bash
  awk '{print $2,$5}' discovery_filtered.bim > discovery_filtered.ref
  plink --bfile validation_filtered --reference-allele discovery_filtered.ref --make-bed --out validation_filtered_ref
  ```

- Apply SparSNP models to validation dataset:
  ```bash
  ./validation.sh validation_filtered_ref
  ```
  Each model trained in cross-validation on the discovery dataset will be applied separately to the validation dataset.

- Plot AUC and optionally, phenotypic and genetic variance explained, on validation dataset:
  ```bash
  ./eval.R mode=validation prev=0.01 h2l=0.5
  ```
  The plots `validation_filtered_AUC.pdf`, `validation_filtered_VarExp.pdf`, and `validation_filtered_GenVarExp.pdf`, will be produced in the directory `validation`.

4. Other post processing

- The model weights are stored in each cross-validation directory `discovery/crossvalXX/beta.csv.XX.XX` using a sparse text format `<index:weight>`, where index is the zero-based index.
of the SNP in the data (0 is the intercept), and weight is the model weight (a real number). The weights can be read into R (using \texttt{read.table} with \texttt{sep=":"}, \texttt{header=FALSE}) or any other tool for visualisation or for validating the model on other datasets.

Some options that can be set by editing the file \texttt{cv3.sh} are:

- \textbf{NFOLDS}: number of cross-validation folds (default=3)
- \textbf{NREPS}: number of cross-validation replications (default=3)
- \textbf{NZMAX}: maximum number of SNPs to allow in model (default=1024)
- \textbf{NLAMBDA1}: number of \( \lambda \) penalties on the grid (default=20)
- \textbf{L1MIN}: multiplier on smallest \( \lambda \) used; it should be a some positive fraction such as 0.01. Setting it lower will increase the number of SNPs in the models, but will also increase computational time.

\section*{References}

A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, et al. Principal components analysis corrects for stratification in genome-wide association studies. \textit{Nat. Genet.}, 38:904–909, 2006.

S. Lee, D. Nyholt, S. Macgregor, A. Henders, K. Zondervan, G. Montgomery, and P. Visscher. A simple and fast two-locus quality control test to detect false positives due to batch effects in genome-wide association studies. \textit{Genet. Epidemiol}, 34: 854–862, 2010.