# Applying genotypic principal component scores as latent phenotypes in genome-wide and epistatic analyses of soybean agronomic traits.

Simon Lackey[1,2†], Siwar Haidar[1,2†], Martin Charette[1], Louise O'Donoughue[3], Istvan Rajcan[4], Francois Belzile[5], Ashkan Golshani[2], Elroy Cober[1*], Bahram Samanfar[1,2*]

Author affiliations:
[1]Agriculture and Agri-Food Canada, Ottawa Research and Development Centre, Ottawa, ON, Canada
[2]Department of Biology, Ottawa Institute of Systems Biology, Carleton University, Ottawa, ON, Canada
[3]CÉROM, Centre de recherche Sur Les Grains Inc., Saint-Mathieu de Beloeil, QC, Canada
[4]Department of Plant Agriculture, University of Guelph, Guelph, ON, Canada
[5]Department of Phytology, Institut de Biologie Intégrative et des Systèmes (IBIS), Université de Laval, Québec City, QC, Canada

*Contact information for corresponding authors:
Bahram Samanfar, bahram.samanfar@agr.gc.ca
Elroy Cober, elroy.cober@agr.gc.ca

† co-first authors

**1 Assemble starter files in correct format**

- Genotype file in .vcf format

- Phenotype file in .txt format with Family ID (FID), Individual ID (IID), and Phenotype Name with FID and IID matching the names in the .vcf file

| | A | B | C |
|---|---|---|---|
| 1 | FID | IID | oil |
| 2 | FC20160517_5_BC22_s1 | FC20160517_5_BC22_s1 | 19.8 |
| 3 | FC20160517_5_BC23_s1 | FC20160517_5_BC23_s1 | 21.3 |
| 4 | FC20160517_5_BC24_s1 | FC20160517_5_BC24_s1 | 20.2 |
| 5 | FC20160517_1_BC50_s1 | FC20160517_1_BC50_s1 | 21.8 |

- Note: Depending on the nature of your dataset FID and IID may not be the same, you may have multiple individuals belonging to the same family

- Set file four column .txt file with no headings: first column is chromosome number, second column is start in base pairs, third column is end in base pairs, fourth column is the trait

| | A | B | C | D |
|---|---|---|---|---|
| 1 | 14 | 191942 | 191942 | oil |
| 2 | 4 | 3466680 | 3466680 | oil |
| 3 | 15 | 5786875 | 5786875 | oil |
| 4 | 5 | 31424775 | 31424775 | oil |
| 5 | 4 | 46247981 | 46247981 | oil |
| 6 | | | | |

- Note: Depending on the epistatic analysis you are running you may set boundaries where column B and column C have different values. Having the same value forces the analysis to use only that SNP in the calculation of epistatic effects.

**2 Using PLINK v1.9 combine genotype and phenotype in .bed, .bim, and .fam format**

- plink --vcf GWAS_SNPs.vcf --double-id --allow-no-sex --make-bed --pheno oil.txt --pheno-name oil --out GWES_oil
- Note: see PLINK documentation for additional modifiers and descriptions of their functions

**3 Using PLINK v1.07 make a set for set by all epistasic analysis**

- plink --bfile GWES_oil --make-set oilset.txt --write-set
- Note: see PLINK documentation for additional modifiers and descriptions of their functions

**4 Using PLINK v1.07 to run set by all epistatic analysis**

- plink --bfile GWES_oil --epistasis --epi1 0.0000001 --set-test --set plink.set --set-by-all --allow-no-sex --out epi_results_oil
- Note: see PLINK documentation for additional modifiers and descriptions of their functions

**5 Read and interpret results in Microsoft Excel**

- Open Microsoft Excel first, then open the .qt or .qt.summary files generated in your PLINK directory
- .qt file contains all counted and reported epistatic interactions based on modifiers from Step 4
- .qt.summary displays the set SNPs with the best epistatic partner detected

**Supplementary Fig. S6:** Workflow used to identify epistatic interactions in a set by all epistatic analysis using SNPs identified by GWAS of gPCs of agronomic traits in soybean. All analysis was carried our using R Studio and PLINK. Additional information related to scripts and modifiers used in the analysis can be found at https://zzz.bwh.harvard.edu/plink/ (v1.07) and https://www.cog-genomics.org/plink/ (v1.9)