**SoyAdapt (Working title)**

Josephine Estelle Ananda Connelly1,†,∗, Guillaume Paul Ramstein2, Wolf L Eiserhardt3 and Torben Asp4

1,2,4Centre for Quantitative Genetics and Genomics, Faculty of Technical Sciences, Aarhus University (DK).

1,3Department of Biology, Faculty of Natural Sciences, Aarhus University, (DK).

3Royal Botanic Gardens, Kew, (UK).

∗Corresponding author: AU, JosephineConnelly@qgg.uni.dk.

# Abstract

Soybean, belonging to the Fabaceae family and scientifically known as *Glycine max* (L. Merrill), is globally recognised as the primary legume crop due to its high protein and oil content. However, its genetic diversity has decreased over time due to domestication and intensive artificial selection, driving the need to look elsewhere to increase germplasm utilisation for abiotic adaptation (Hyten *et al.* 2006; Gizlice *et al.* 1996). Relevant accessions with the potential for abiotic adaptation to cold tolerance have been identified [@haupt20]. Additionally 160 Swedish accessions, originating from a Swedish breed for the adaptation to the cool climate of north western Europe and stored in the NordGen gene bank have been unearthed [@holmberg1973]. We will interpret the plant genetic resources available, so as to utilise the genetic potential. To facilitate targeted selection of promising accessions from soybeans vast germplasm collections, and unveiling the genetic data stored in the NordGen seed banks. findings are: ...

**Keywords:** Soy; population genetics; Diversity; Selection; more keywords

# Introduction

*what we already know:* Several severe genetic bottlenecks occurred in soybean. Compared to the wild species, the genetic diversity

was halved, resulting in a loss of 81% of rare alleles (Hyten et al., 2006). Additionally, two major bottlenecks occurred during the development of North American modern cultivars, where only a few landraces were used. And as a result of intensive soybean breeding over the past 75 years. Elite cultivars have emerged, but they are derived from only about 19 landraces. Consequently, the North American breeding pools now retain only 72% of genome diversity and have lost 79% of rare alleles found in diverse landraces (Gizlice et al., 1996).

*soybean needs to be grown in the north*: Modern soybean breeding has led to yield increases of 229 kg *ha*−1 yr−1 (Rincker et al., 2014),

gaining a deeper understanding of the genomic basis of this progress could offer valuable insights for advancing and adapting soybean cultivation even further.

Expanding the production of soy to the northern climates by leveraging the advancements in soybean breeding and focusing on the genes relevant to adaption, can diversify the global supply chain, reduce dependency, and potentially alleviate the environmental pressures associated with soy production in high biodiversity, sensitive areas/ecologically vulnerable regions. With the global demand for soy and the need for sustainable agricultural practices, investing in breeding programs to develop soybean varieties suitable for northern climates can present a promising solution. The swedish breedingprogram by Holmberg *the swedish soybeans (goals and scope)*

*mention approaches*

*In this study we* The goal is to help precipitate the utilisation of these soybean genetic resources.

**Materials and methods**

what is the data: 160 soybean accessions were obtained from the Nordgen genebank, and 155 of these accessions are of Swedish origin. These accessions are thought to originate from a Swedish breeding program running from the 1840s to the 1970s which used material consisting of a mixture of elite cultivars of the time and Japanese landraces, breeding for the adaptation to the cool climate of northwestern Europe Holmberg (1973).

The Core collection used here is a subset of the large soybean germplasm collection, the USDA genebank accessions consisting of 415 of the USDA soy germplasm accessions. This Core collection has been selected with a focus on adaptation to high-latitude cold regions using environmental data from phenotypic trials in Germany, and comparing Donor opulation of Environments (DPE) in Asia and the Target Population of Environments (TPE) in Central Europe. From the 3663 accessions, two diverse core collections of 183 and 366 accessions were created. These 514 diversity panels are used here due to that they are likely preadapted to cultivation in Central Europe, while simultaneously conserving a high level of genetic diversity.

USDA germplasm of soybean consists of ... soybase.org. Founders 10.

# Genotyping and filtering of whole genome sequence SNP

*What was done:* The accessions from the Nordgen genebank and the Core collection accessions were whole genome sequenced. Four of the Core collection seeds didn’t germinate in time for the sequencing, resulting in accessions sequenced 574. Of the Nordgen accessions, Four were of Latvian origin and therefore excluded.

*GATK:* The reads were aligned to the Williams 82 2a reference genome ( (Schmutz et al. 2010).) adapted from GATK Best Practices, keeping MQ30 and biallelic sites and applying a MAF filter of 0.01 in order to remove rare variants from the data, reducing the number of SNPs from 17,648,123 to 10,000,122. For a detailed account of programs and commands used see supplementary material:

Supplementary Material

*From when i received it:* A further three accessions were removed due to missing metadata and a single accession with missing data > 5% using BCFtools, resulting in 153 Nordgen accessions and 409 CCA accessions and a SNP count of 8,533,444.

# SNP filtering

The SNP data is a high-density genotyping array for soybean with 47,337 single nucleotide polymorphisms (SNPs) was develop from soybean (Glycine max L. Merr.). The SoySNP50K iSelect BeadChip has been used to genotype the USDA Soybean Germplasm

Collection and is available for use from Soybase An intersect was made of the whole genome sequenced 35486

For a detailed account of programs and commands used see supplementary material, Supplementary Material and Data availability for data.

# H1

To infer population stratification of the Swedish accessions from the NordGen gene bank and the Core collection and of the 12 accessions assumed to be founders of the Swedish breeding program.

Principal Component Analysis (PCA) on SNP genotype data were applied using the package SnpRelate in R (A Parallel

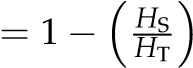
Computing Toolset for Relatedness and Principal Component Analysis of SNP Data). The functions in SNPRelate for PCA used

(snpgdsPCA) calculate the genetic covariance matrix from genotypes, computing the correlation coefficients between sample loadings and genotypes for each SNP. SNP eigenvectors were calculated. For the PCA the whole genome sequence was used for some of the PCA or the intersect of the whole genome sequences the Core Collection accessions, the Swedish Nordgen accessions and the 50kSNP data available of the ten accessions identified as possible genetic Founders to the Swedish breeding program is used as indicated in the figure 5

The package SNPRelate in R was used to compute an identity-by-state (IBS) matrix that calculates the proportion that two randomly selected reads that contain a certain SNP locus are the same or different between two individuals. The resulting pairwise IBS matrix

was used to generate a cluster dendrogram using the function *snpgdsHCluster: Hierarchical cluster analysis In SNPRelate* in R.

Relatedness measures F-statistics (also known as fixation indices) for given populations were calculated using the function *snpgdsFst: F-statistics (fixation indices) in SNPRelate* in R. The whole genome data was used to calculte the pairwise population differentiation measured in the populations the Swedish accessions as a subpopulation with the Core Collection as a representation of the total population Fst is comparing the expected heterozygosity in the subpopulation (HS) to that expected in the total population (HT) :

*F*ST 

**H2**

diversity Method LD Nucleotide diversity values (*π* ) were calculated using x using all sites?.

**H3**

**Statistical analysis**

Indicate what statistical analysis has been performed.

# Results and Discussion

# population structure and genetic differentiation

To better understand the population relationships of the accessions received from the Nordic gene bank and how they are related to each other, and the rest of the soybean germplasm, they are here compared to a Core Collection of 409 soybeans from the USDA genebank chosen/curated from the soybean germplasm for Central European breeding. The Core collection used here is a subset of the soybean germplasm collection, the USDA genebank accessions, which are assessed to be preadapted to cultivation in Central Europe, while conserving a high genetic diversity Haupt and Schmid (2020). Additionally, accessions were grouped into categories based on the suspected breeding history. There are the ten suspected Founders of the Swedish breeding program and the Swedish accessions of the Nordgen genebank.

A Principal Component Analysis (PCA) and the genome-wide Identity By State (IBS) pairwise distance matrix were applied to visualise the overall variance of the data. The resulting PCA (Figure 5) shows the clustering of the Swedish accessions as a sub-group based on similarity. The variance proportions explained.

A closer look at the PCA revealed two accessions of the Nordgen Swedish that were not part of the particular breeding program of interest as seen in (Figure 3 showing the PCA indicating the two accessions not from the breeding program in the cluster with the CCA in orange and black) and also thereafter confirmed by the genebank passport data that they originated from other Swedish soybean breeding programs (mention more here when certain. Altonagård and Ugra soja/ saja, malmåhus). This figure also shows the accessions of the Nordgen and core collection that are Swedish of origin and with a name consisting of letters (excluding numbers or non-alphanumeric characters of the accessions’ names) as included in the database. The IBS (Figure **??**) clearly shows the Swedish accessions of the breeding program of interest clustering together. The IBS numbers in a table showing ibs within sbpa and out?

Heatmap based on the sbpa shows high similarity of the SBPA and shows that the 2 accessions Altonagård/Altonagard and Ugra

Soja(Saja) are dissimilar.

The population differentiation measured by *F*ST . Population subdivision analysis will quantify the SPBA will be a distinct subpopulation with a moderate to distinct allele frequency difference.

**Table 1** *F*ST estimation on genotypes. Measuring the population differentiation of the 155 Accessions of the Nordgen genebank (Do again with only the 153 that were part of the breeding program)(SBPA) to the Core Collection of 409 accessions (CCA).

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Samples: | 564 |  |  |  |  |  |
| SNPs: | 8,533,444 |  |  |  |  |  |
| Method: Weir & Cockerham, 1984 |  |  |  |  |  |  |
| Populations: | CCA | 409 | SBPA | 155 |  |  |
| Weighted Fst estimate: | 0.2231884 |  |  |  |  |  |
| Mean Fst: | 0.1108533 |  |  |  |  |  |
| summary: | Min. | 1st Qu | Median | Mean | 3rd Qu. | Max. |
|  | -0.002515 | 0.011669 | 0.041730 | 0.110853 | 0.145697 | 0.975485 |

What do the numbers say?

How would showing the

Does the population differ primarily in their profiles of rare alleles?

**Genetic diversity**

H2 LD Drift

# Genome-wide selection signatures H3

## Discussion

The Swedish soybeans are mainly a snapshot of a breeding program stopped around 1978 and all lines or recent crosses from this program were then set into the

**Figures and tables**

# PCA1 figure

Figure 5 shows an example figure.

**PCA2 figure** Figure 2 .

**PCA3 figure** Figure 3 .

**PCA figure4** Figure 4 .

# Dendogram figure

Figure **??** shows an example figure.

**Heatmap figure** Figure 6 .

# Table 1

Table 1 shows table 1.

**Supplementary Material**

Supplementary Material Methods (S Methods)

## Data availability

For example: Strains and plasmids are available upon request. File S1 contains detailed descriptions of all supplemental files. File S2 contains SNP ID numbers and locations. File S3 contains genotypes for each individual. Sequence data are available at GenBank and the accession numbers are listed in File S3. Gene expression data are available at GEO with the accession number: GDS1234. Code used to generate the simulated data can be found at [https://github.com/JosephineConnelly/soyadapt\_data\_analysis.](https://github.com/JosephineConnelly/soyadapt_data_analysis)

**Acknowledgments**

Acknowledgments should be included here.

**Funding**

Funding, including Funder Names and Grant numbers should be included here.

**Conflicts of interest**

There are no known conflicts of interest.

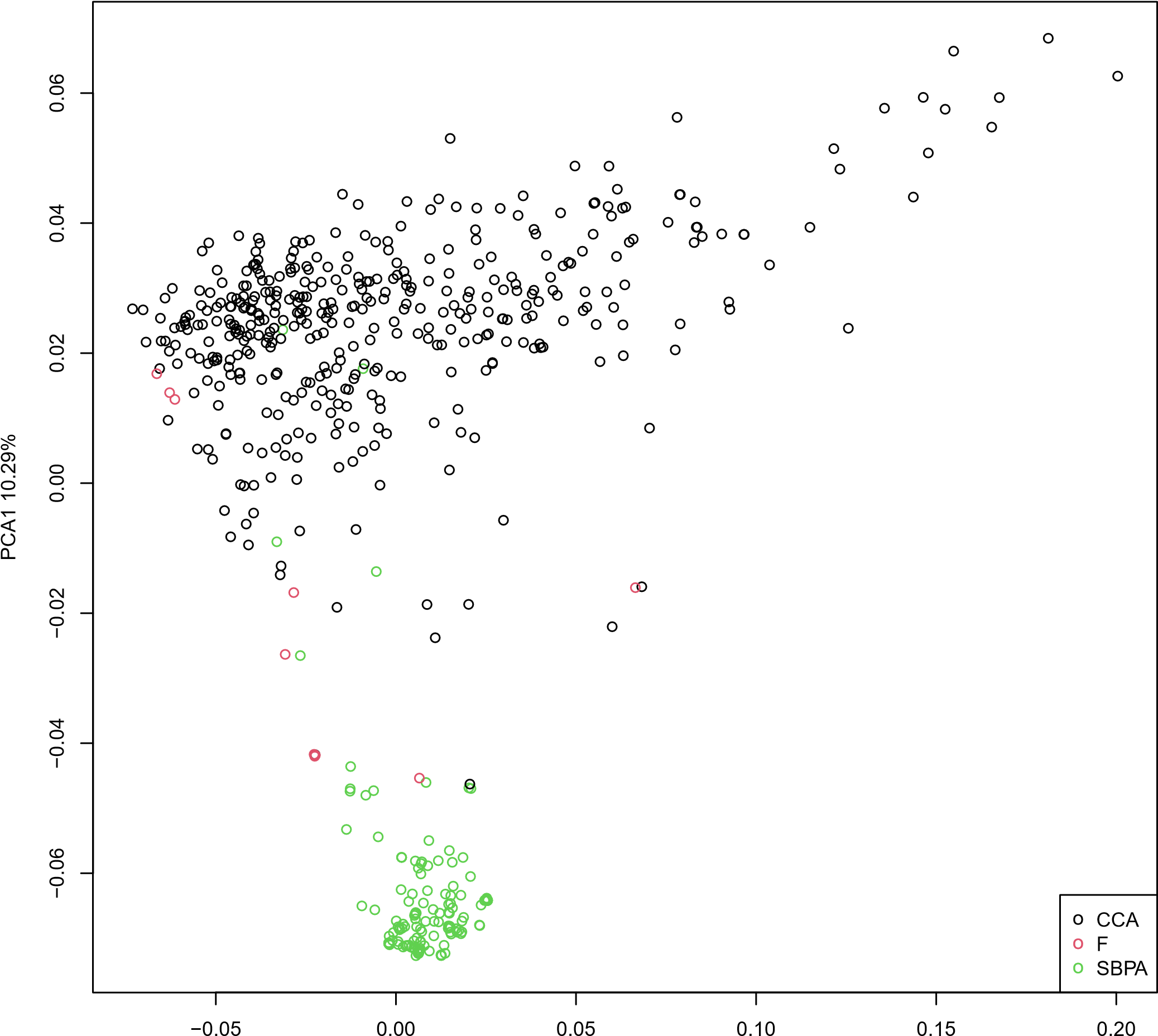
## Literature cited

Gizlice Z, Carter Jr. TE, Gerig TM, Burton JW. 1996. Genetic Diversity Patterns in North American Public Soybean Cultivars based on Coefficient of Parentage. Crop Science. 36:cropsci1996.0011183X003600030038x.

Haupt M, Schmid K. 2020. Combining focused identification of germplasm and core collection strategies to identify genebank accessions for central European soybean breeding. Plant, Cell & Environment. 43:1421–1436.

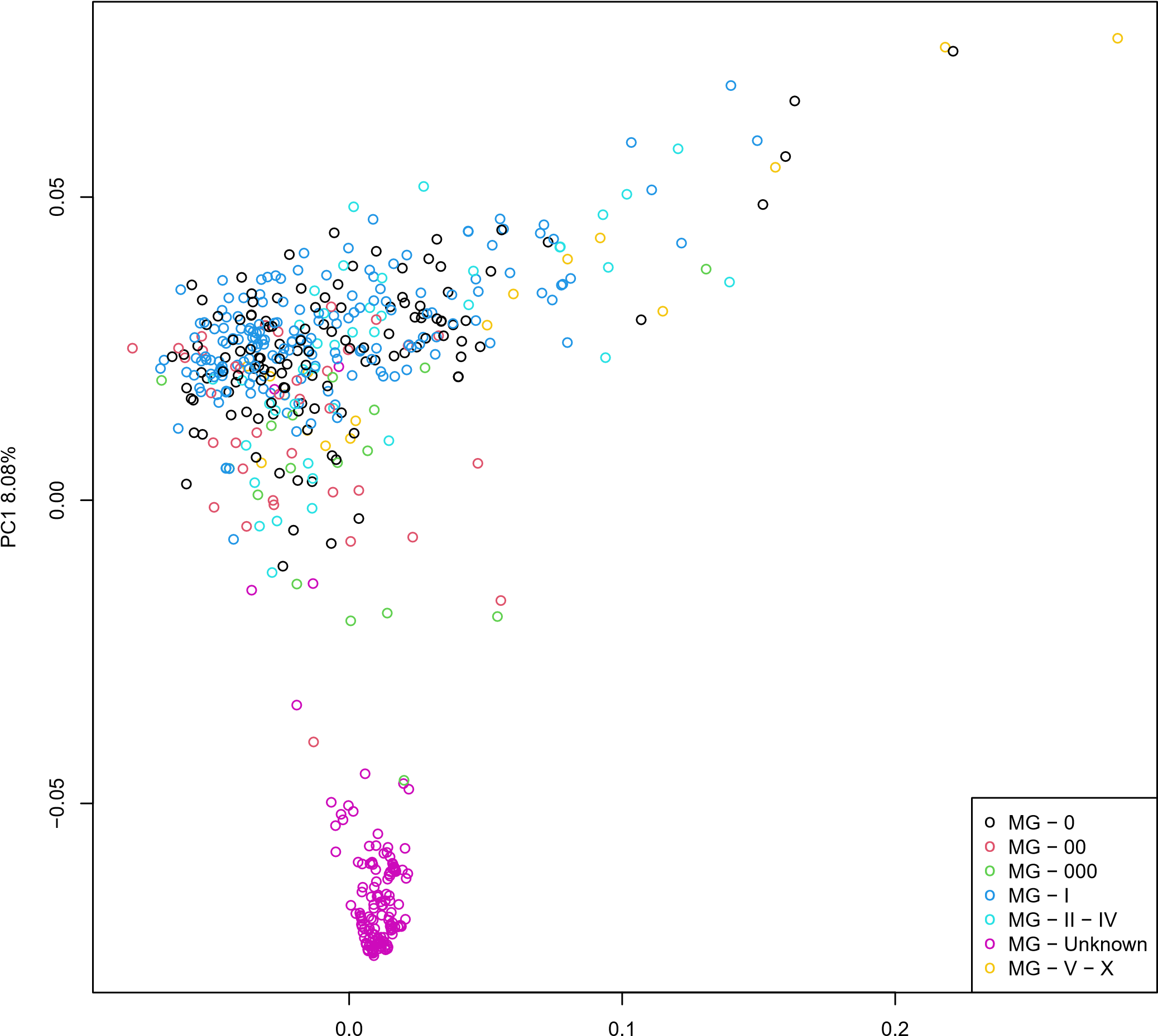
Holmberg SA. 1973. Soybeans for cool temperate climates. AGR HORTIQUE GENET. 31:1–20.

Hyten DL, Song Q, Zhu Y, Choi IY, Nelson RL, Costa JM, Specht JE, Shoemaker RC, Cregan PB. 2006. Impacts of genetic bottlenecks on soybean genome diversity. Proceedings of the National Academy of Sciences. 103:16666–16671.



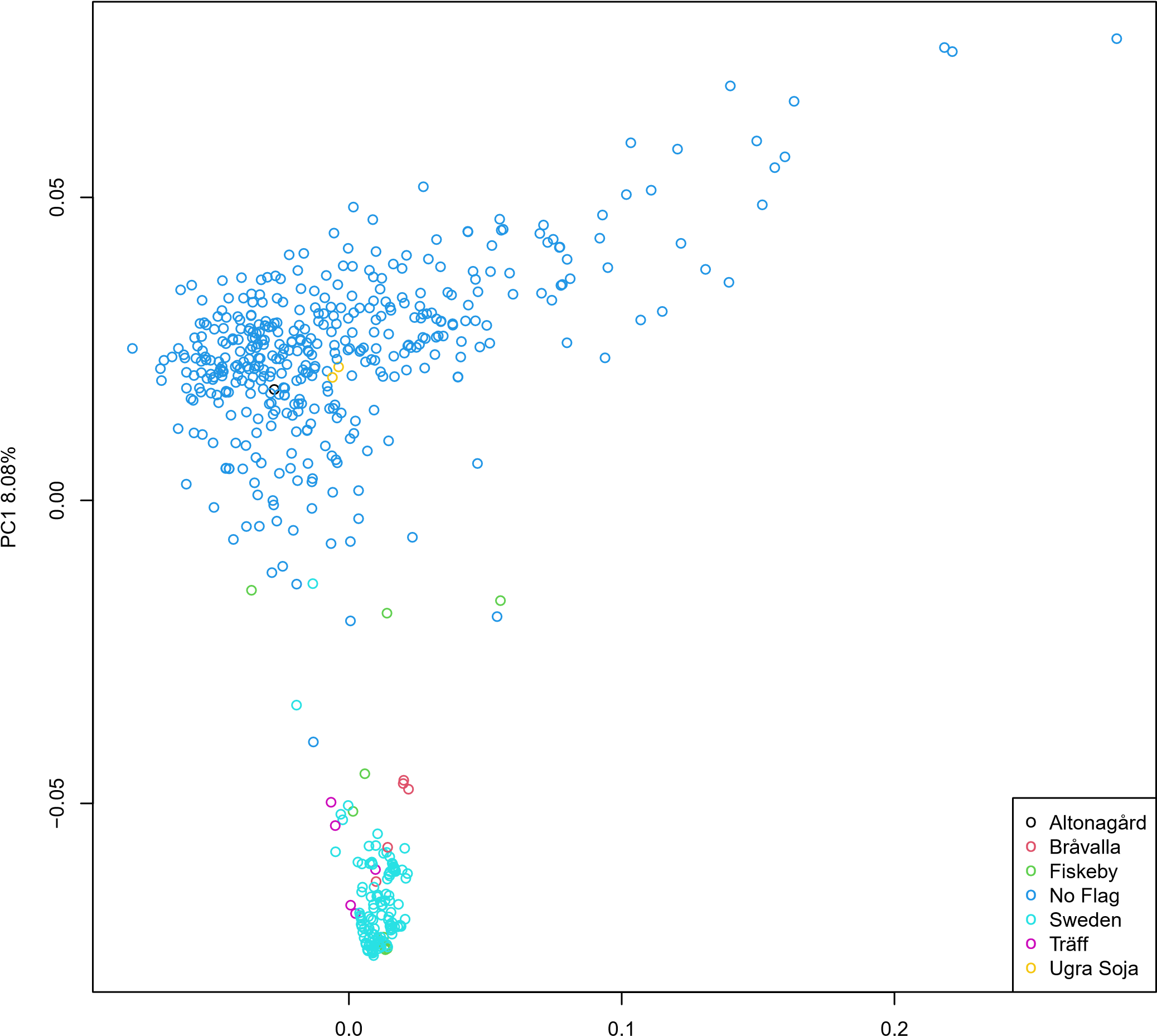
### PCA2 3.43%

**Figure 1** Principal component analysis for soybean samples from defined groups, Core collection Accessions (CCA) in black, the Swedish Breeding Program Accessions (SBPA) in green and the Founders of the breeding program (F) in red.



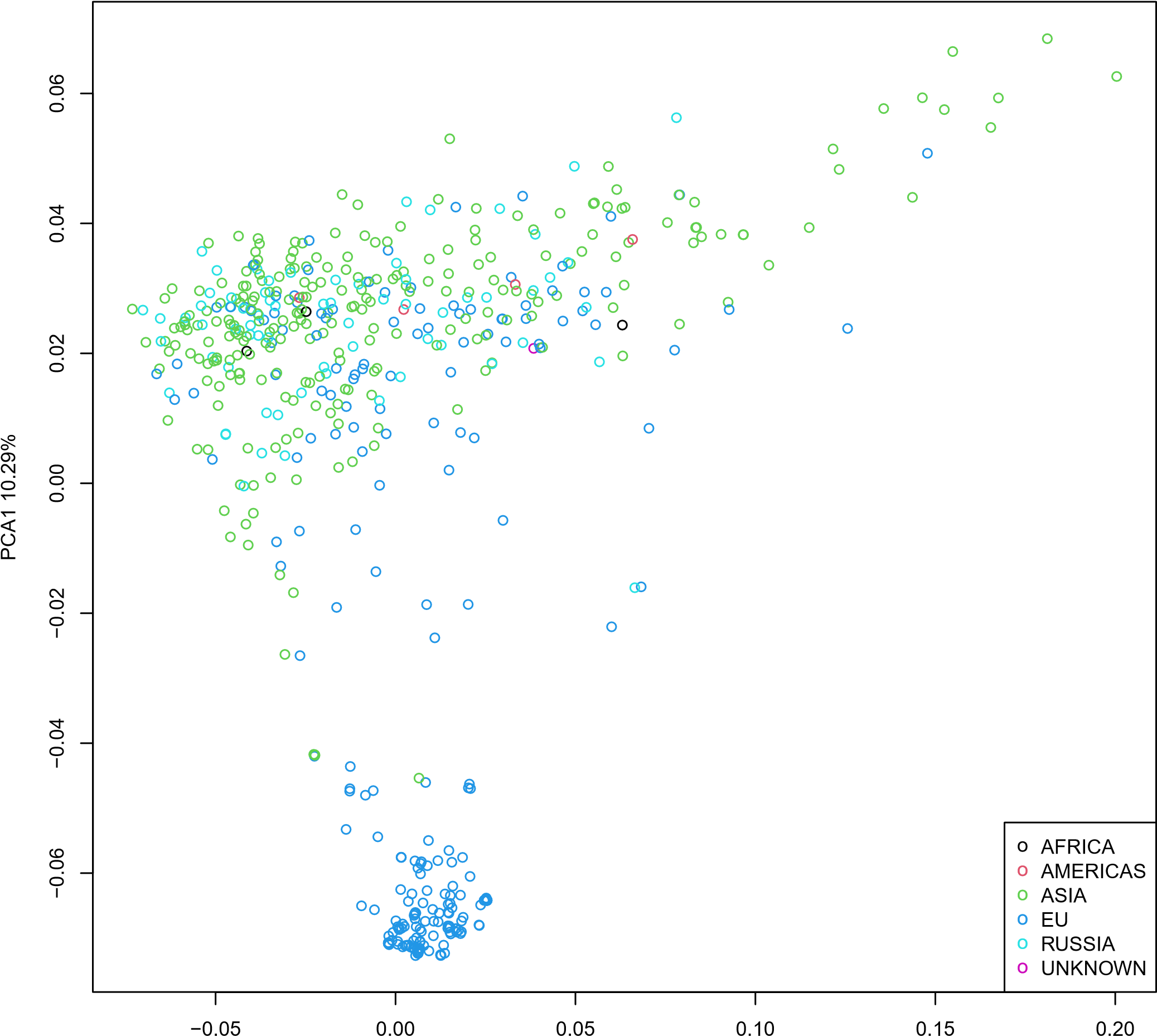
### PC2 2.91%

**Figure 2** Principal component analysis for soybean samples from the whole genome sequence data, including the 155 Nordgen accessions and the 409 Core collection accessions. Indicated are the Maturity groups assigned to the accessions.



PC2 2.91%

**Figure 3** PCA flag



PCA2 3.43%

**Figure 4** PCA regions

Soy scientific manuscript

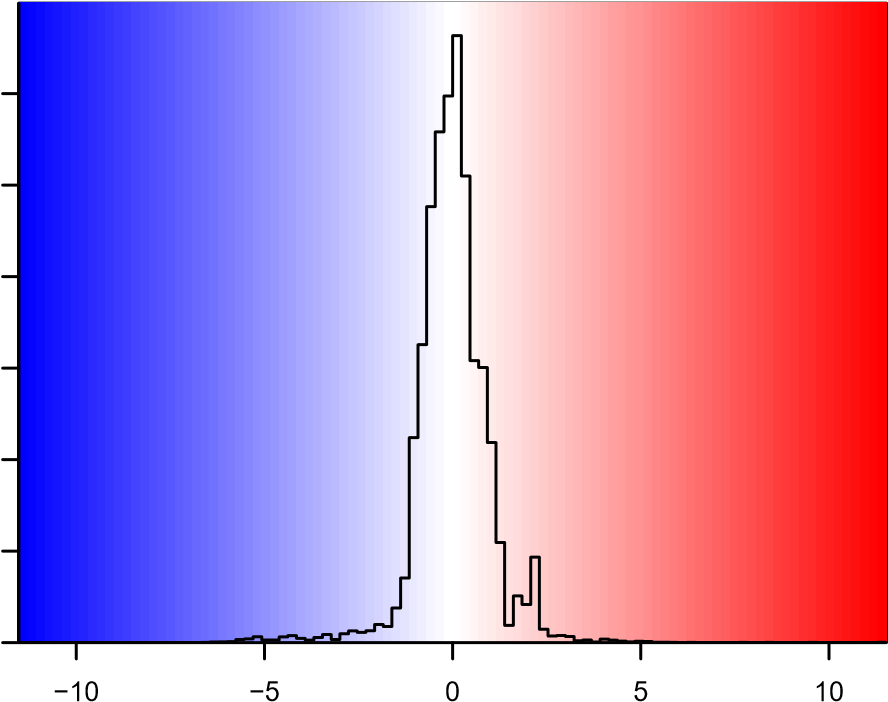


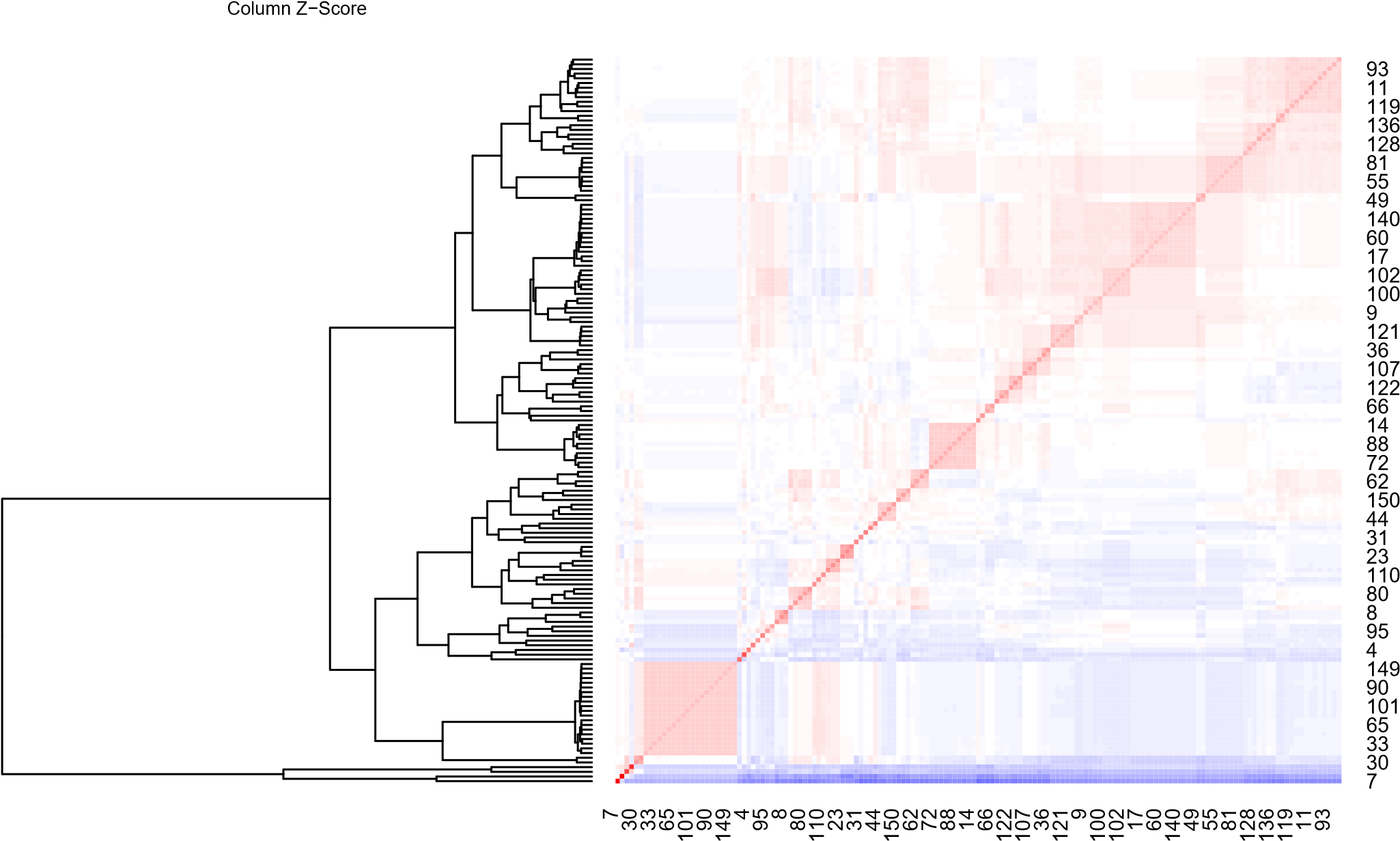
**Figure 5** Hierarchical cluster dendrogram based on pairwise identity-by-state (IBS) values from SNP data for all samples. describe dendogram

Connelly *et al.*

**Color Key**

**and Histogram Heatmap**





**Figure 6** PCA regions