Working title SoyAdapt

Josephine Estelle Ananda Connelly[1,†,*], Guillaume Paul Ramstein[2], Wolf L Eiserhardt[3], Torben Asp[4] and Josephine Estelle Ananda Connelly[5]

[1]Centre for Quantitative Genetics and Genomics, Faculty of Technical Sciences, Aarhus University (DK).
[2]Department of Biology, Faculty of Natural Sciences, Aarhus University, (DK).
[3]Royal Botanic Gardens, Kew, (UK).

*Corresponding author: AU, Connelly.

Abstract

Soybean (Glycine max L. Merrill) is the worlds leading oilseed crop used as a primary source of vegetable oil for human consumption and in proteinmeal for animal feed.

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

Abstract

Soybean, belonging to the Fabaceae family and scientifically known as *Glycine max* (L. Merrill), is globally recognized 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 utilization 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: ...

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: This argument but then for spread of growth area! "Modern U.S. soybean breeding has led to a yield increase of 29 kg ha1 yr1 (Rincker et al., 2014)" ""Better understanding the genomic basis behind this improvement may provide indicators for further soybean improvement and adaptation. " from: Sequencing the USDA core soybean collection reveals gene loss during domestication and breeding

the swedish soybeans (goals and scope)

mention approaches

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

Materials and methods

what is the data: 153 soybean accessions of Swedish origin were obtained from the Nordgen genebank. These accessions 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 north western Europe Holmberg (1973).

The Core collection used here is a subset of the from a 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 phenoptpic 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 .

For a detiailed account of programs and commands used see supplimentry material (wgs What was done: The Swedish accessions cosist of 160 accessions and the Core collection selected for Nordic regions were whole genome sequenced. where 4 of the CCA didnt germinate in time for the sequencing resulting in 409 accessions sequenced. . GATK: The reads were aligned to the Williams 82 2a refrence genome ((Schmutz et al. 2010).) MQ30 and biallelic sites and applied a MAF filter of 0.01 in order to remove rare variants from the data, and that reduced the number of SNPs from 17,648,123 to 10,000,122.

From when i recived it: further filtering: A further three accessions were removed due to missing metadata. and a single accessions with missing data > 5%. resluting in 156 nordgen accessions and 406 CCA accessions.

SNP 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 35486

H1 Q

Method pca

IBS Matrix: The package SnpRelate in R (A Parallel Computing Toolset for Relatedness and Principal Component Analysis of SNP Data) 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.

snpgdsHCluster: Hierarchical cluster analysis In SNPRelate: Parallel Computing Toolset for Relatedness and Principal Component Analysis of SNP Data

realisation that 2 of the swedish were likely not a part of the swedish breeding program and excluded see Supplimentary material for accession data. Fst of the swedish vs the cca

H2 Q diversity

Method LD pi and

H3 Q

Method

Statistical analysis

Indicate what statistical analysis has been performed.

Results and Discussion

population structure and genetic differentiation

H1: 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 compared to a core collection of 409 soybean from the USDA genebank selected from the soybean (Glycine max) germplasm for Central European breeding. The Core collection used here is a subset of the from a large soybean germplasm collection, the USDA genebank accessions, that are likely preadapted to cultivation in Central Europe, while simultaneously conserving a high level of genetic diversity (CCA paper). Additionally accessions were grouped into categories based on known breeding history, So the suspected Founders of the Swedish breeding program are grouped together. Here an 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.

To understand the population differentiation/ diversity, a Principal Component Analysis (PCA) and the genome-wide Identity By State (IBS) pairwise distance matrix were applied. The resulting PCA shows the Swedish accessions as a sub group seen in green in (Figure 4) . A closer look at the Nordgen Swedish accessions reviled two accessions that were not part of the particular breeding program as seen in (Figure X showing the SBPA PCA with circles around the two accessions not from the breeding program) and also thereafter confirmed by the genebank passport data. (Or do i simply remove these from the analysis and only mention that they are excluded up in the methods when describing the data)

and the IBS (Figure 2) shows . IBS numbers in a table showing ibs within sbpa and out?

The genetic differentiation measured by FST

Genetic diversity

MAF? LD pi theta

Genome wide selection signatures

gws.

Discussion

The Swedish soybeans are 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

Additional guidelines

Numbers

In the text, write out numbers nine or less except as part of a date, a fraction or decimal, a percentage, or a unit of measurement. Use Arabic numbers for those larger than nine, except as the first word of a sentence; however, try to avoid starting a sentence with such a number.

Units

Use abbreviations of the customary units of measurement only when they are preceded by a number: "3 min" but "several minutes". Write "percent" as one word, except when used with a number: "several percent" but "75%." To indicate temperature in centigrade, use ° (for example, 37°); include a letter after the degree symbol only when some other scale is intended (for example, 45°K).

Nomenclature and italicization

Italicize names of organisms even when when the species is not indicated. Italicize the first three letters of the names of restriction enzyme cleavage sites, as in HindIII. Write the names of strains in roman except when incorporating specific genotypic designations. Italicize genotype names and symbols, including all components of alleles, but not when the name of a gene is the same as the name of an enzyme. Do not use "+" to indicate wild type. Carefully distinguish between genotype (italicized) and phenotype (not italicized) in both the writing and the symbolism.

Cross references

Use the `\nameref` command with the `\label` command to insert cross-references to section headings. For example, a `\label` has been defined in the section Materials and methods.

In-text citations

Add citations using the `\citep{}` command, for example (**?**) or for multiple citations, (**???**)

Examples of article components

The sections below show examples of different header levels, which you can use in the primary sections of the manuscript (Results, Discussion, etc.) to organize your content.

First level section header

Use this level to group two or more closely related headings in a long article.

Second level section header

Second level section text.

Third level section header: Third level section text. These headings may be numbered, but only when the numbers must be cited in the text.

Figures and tables

Figures and Tables should be labelled and referenced in the standard way using the `\label{}` and `\ref{}` commands.

PCA figure

Figure 4 shows an example figure.

Figures and Tables should be labelled and referenced in the standard way using the `\label{}` and `\ref{}` commands.

Dendogram figure

Figure 2 shows an example figure.

Figures and Tables should be labelled and referenced in the standard way using the `\label{}` and `\ref{}` commands.

Sample figure

Figure 5 shows an example figure.

Sample table

Table 1 shows an example table. Avoid shading, color type, line drawings, graphics, or other illustrations within tables. Use tables for data only; present drawings, graphics, and illustrations as separate figures. Histograms should not be used to present data that can be captured easily in text or small tables, as they take up much more space.

Tables numbers are given in Arabic numerals. Tables should not be numbered 1A, 1B, etc., but if necessary, interior parts of the table can be labeled A, B, etc. for easy reference in the text.

Sample equation

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables with $\mathrm{E}[X_i] = \mu$ and $\mathrm{Var}[X_i] = \sigma^2 < \infty$, and let

$$S_n = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i}^{n} X_i \tag{1}$$

denote their mean. Then as n approaches infinity, the random variables $\sqrt{n}(S_n - \mu)$ converge in distribution to a normal $\mathcal{N}(0, \sigma^2)$.

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.

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.

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.

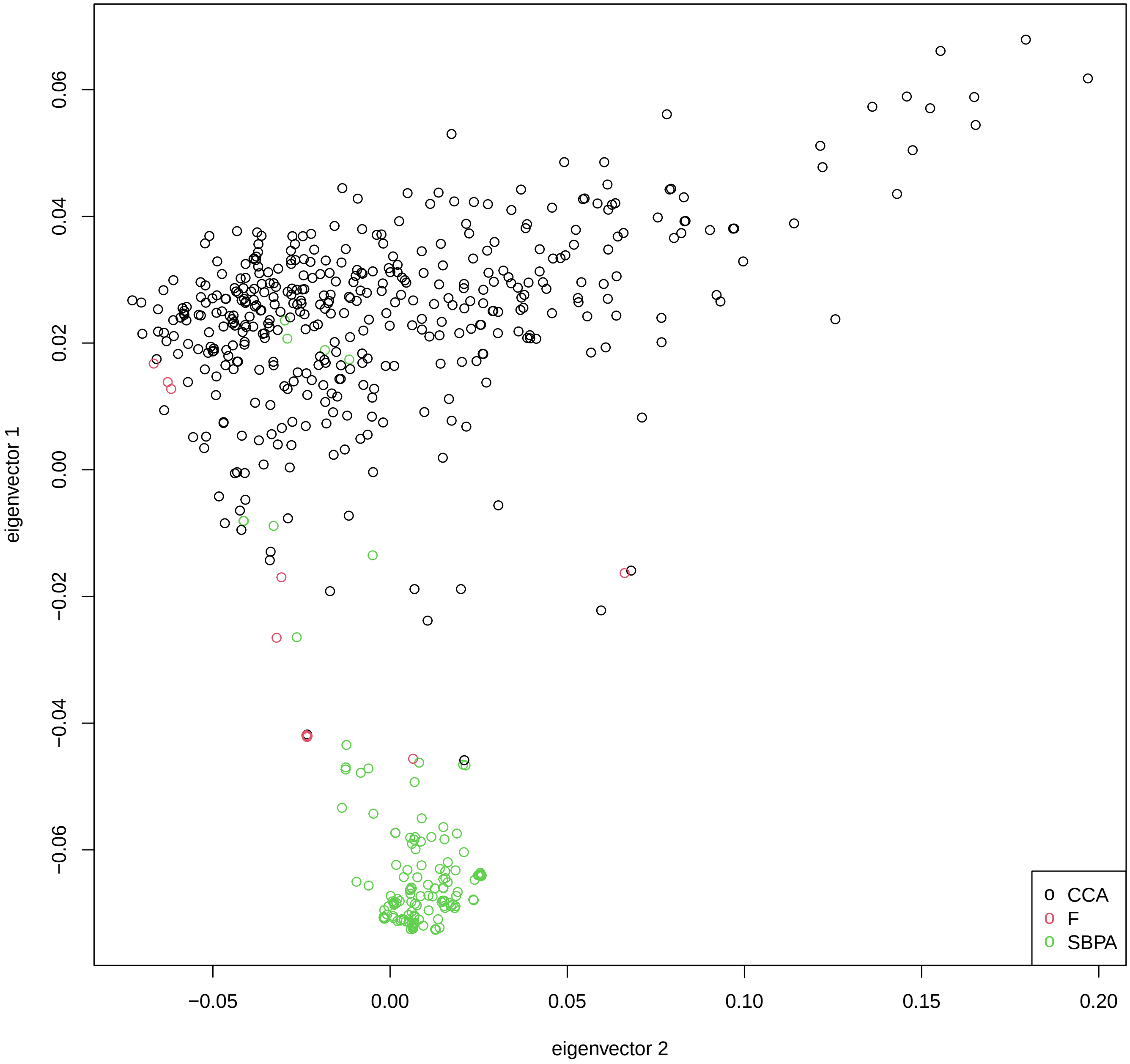

Figure 1 This PCA shows the CCA in black, the SBPA in green and the thought to be founders in red.

Table 1 Students and their grades

Student	Grade[a]	Rank	Notes
Alice	82%	1	Performed very well.
Bob	65%	3	Not up to his usual standard.
Charlie	73%	2	A good attempt.

[a] This is an example of a footnote in a table. Lowercase, superscript italic letters (a, b, c, etc.) are used by default. You can also use *, **, and *** to indicate conventional levels of statistical significance, explained below the table.

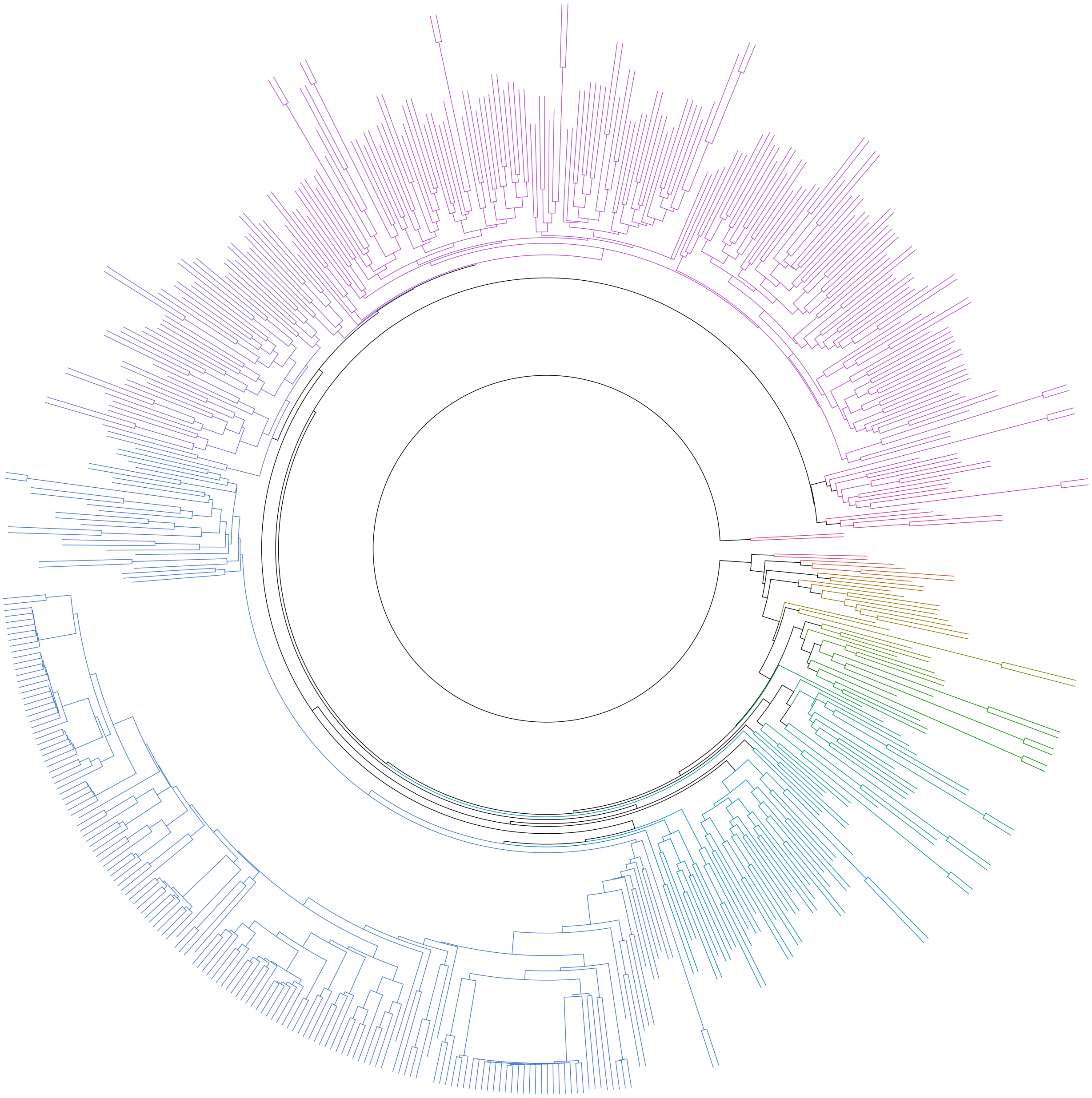

Figure 2 Hierarchical cluster dendrogram based on pairwise identity-by-state (IBS) values from WGS data for all samples.

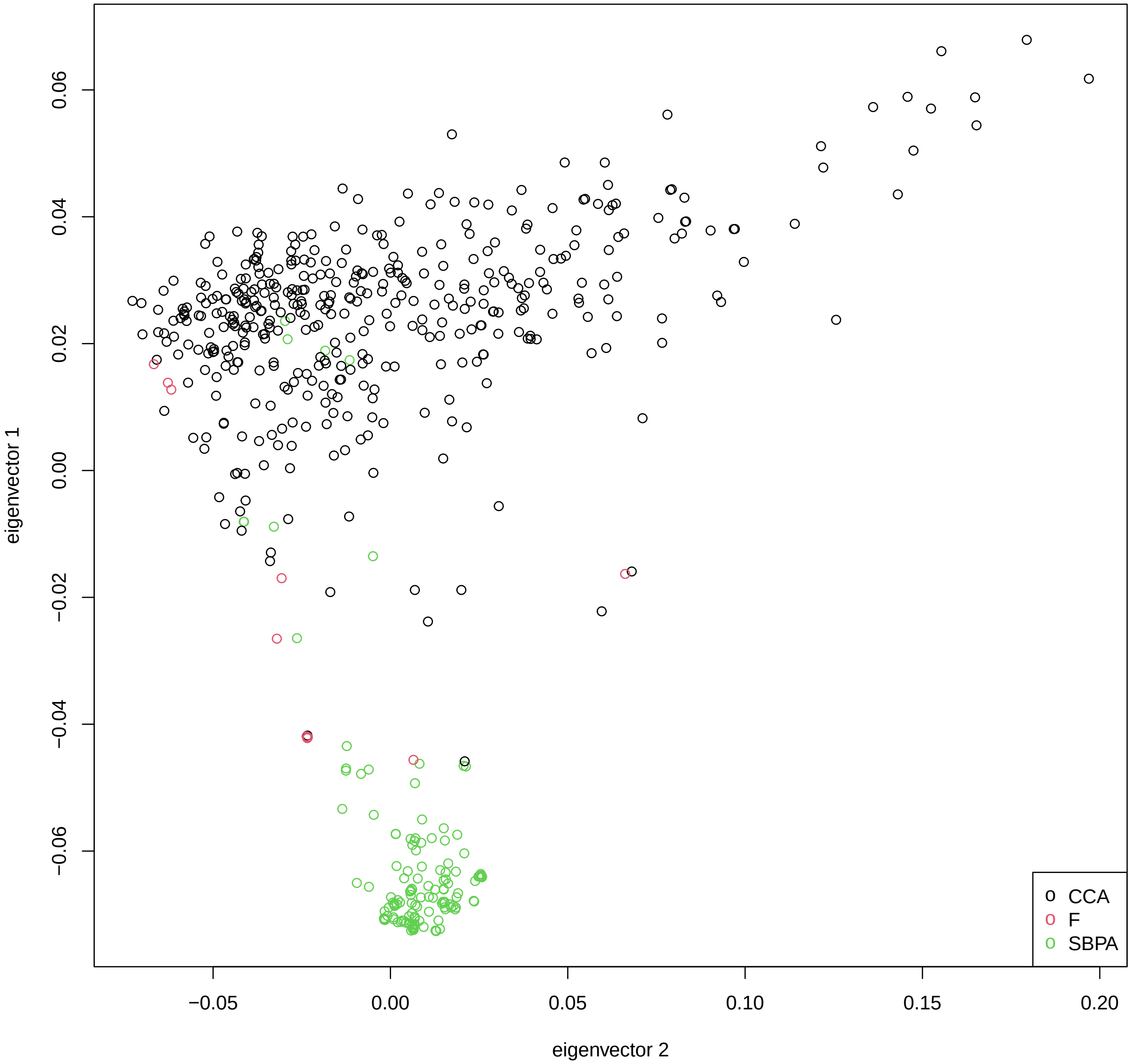

Figure 3 This PCA shows the CCA in black, the SBPA in green and the thought to be founders in red.

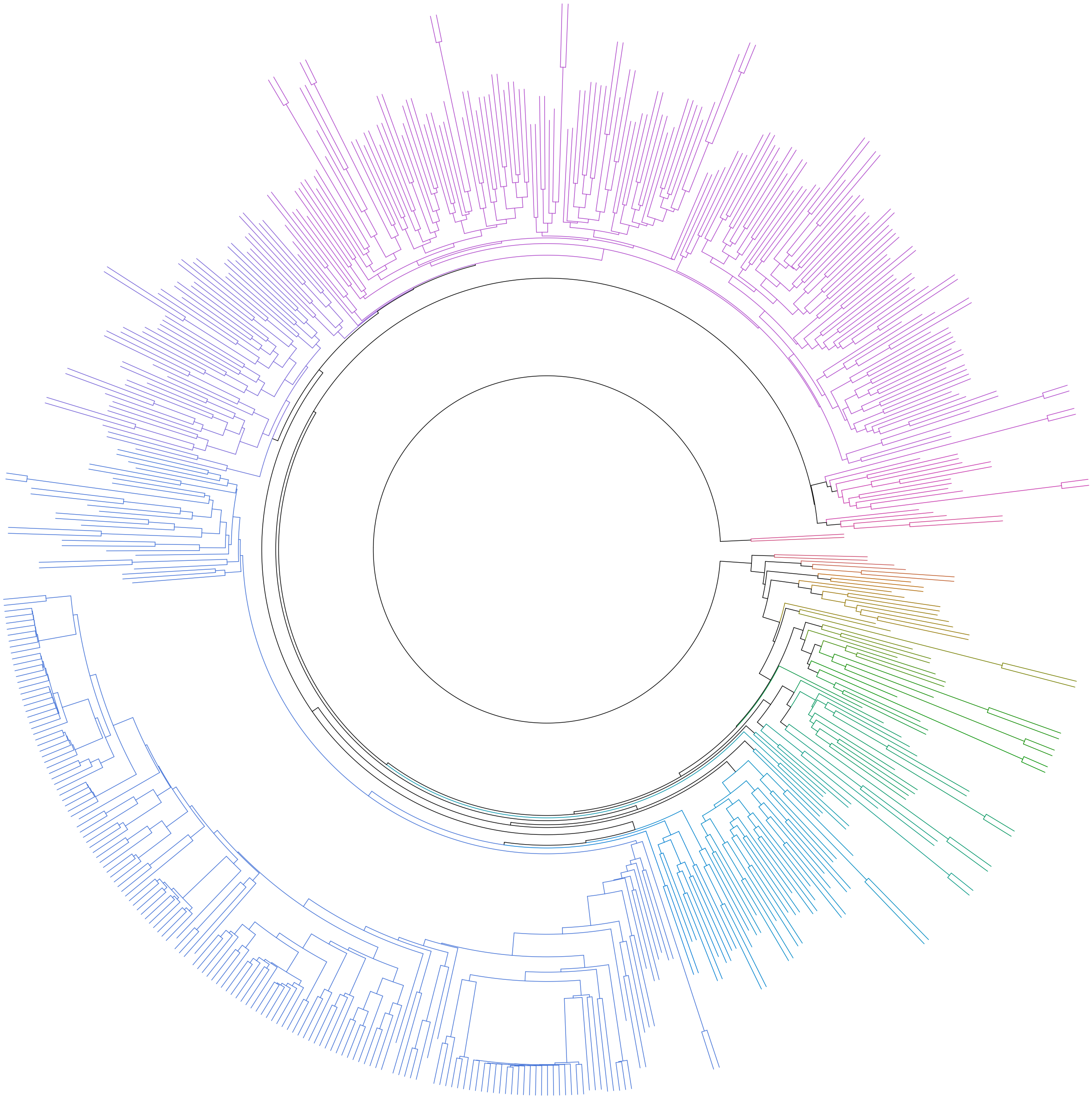

Figure 4 Hierarchical cluster dendrogram based on pairwise identity-by-state (IBS) values from SNP data for all samples. describe dendogram

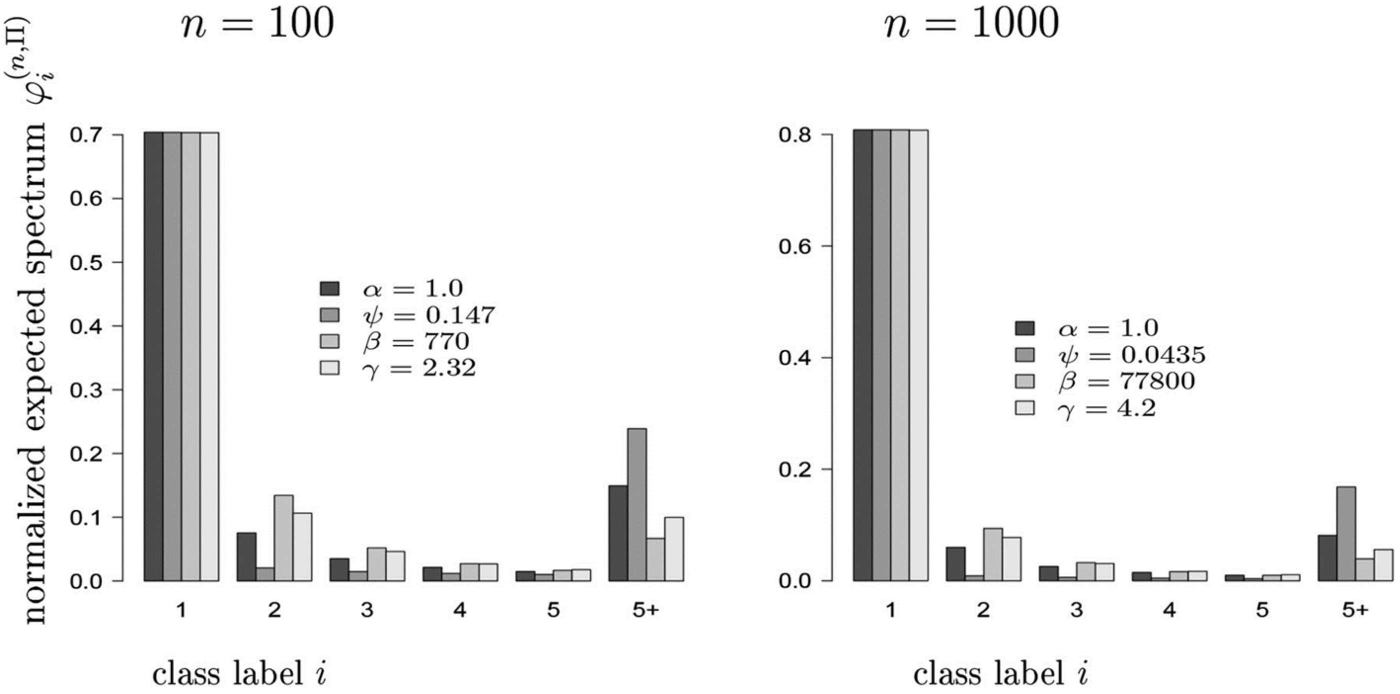

Figure 5 Example figure from 10.1534/genetics.114.173807. Please include your figures in the manuscript for the review process. You can upload figures to Overleaf via the Project menu. Images of photographs or paintings can be provided as raster images. Common examples of raster images are .tif/.tiff, .raw, .gif, and .bmp file types. The resolution of raster files is measured by the number of dots or pixels in a given area, referred to as "dpi" or "ppi."

- minimum resolution required for printed images or pictures: 350dpi
- minimum resolution for printed line art: 600dpi (complex or finely drawn line art should be 1200dpi)
- minimum resolution for electronic images (i.e., for on-screen viewing): 72dpi

Images of maps, charts, graphs, and diagrams are best rendered digitally as geometric forms called vector graphics. Common file types are .eps, .ai, and .pdf. Vector images use mathematical relationships between points and the lines connecting them to describe an image. These file types do not use pixels; therefore resolution does not apply to vector images. Label multiple figure parts with A, B, etc. in bolded. Legends should start with a brief title and should be a self-contained description of the content of the figure that provides enough detail to fully understand the data presented. All conventional symbols used to indicate figure data points are available for typesetting; unconventional symbols should not be used. Italicize all mathematical variables (both in the figure legend and figure) , genotypes, and additional symbols that are normally italicized.