Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Eidgenössisches Departement für
Wirtschaft, Bildung und Forschung WBF
Agroscope

harasnational.ch

Identification of PCA informative Individuals (PCA-IIs) within populations

Markus Neuditschko

64th Annual Meeting of the European Federation of Animal Science, August 26th – 30th , 2013
Nantes France

The two questions we have addressed in this study

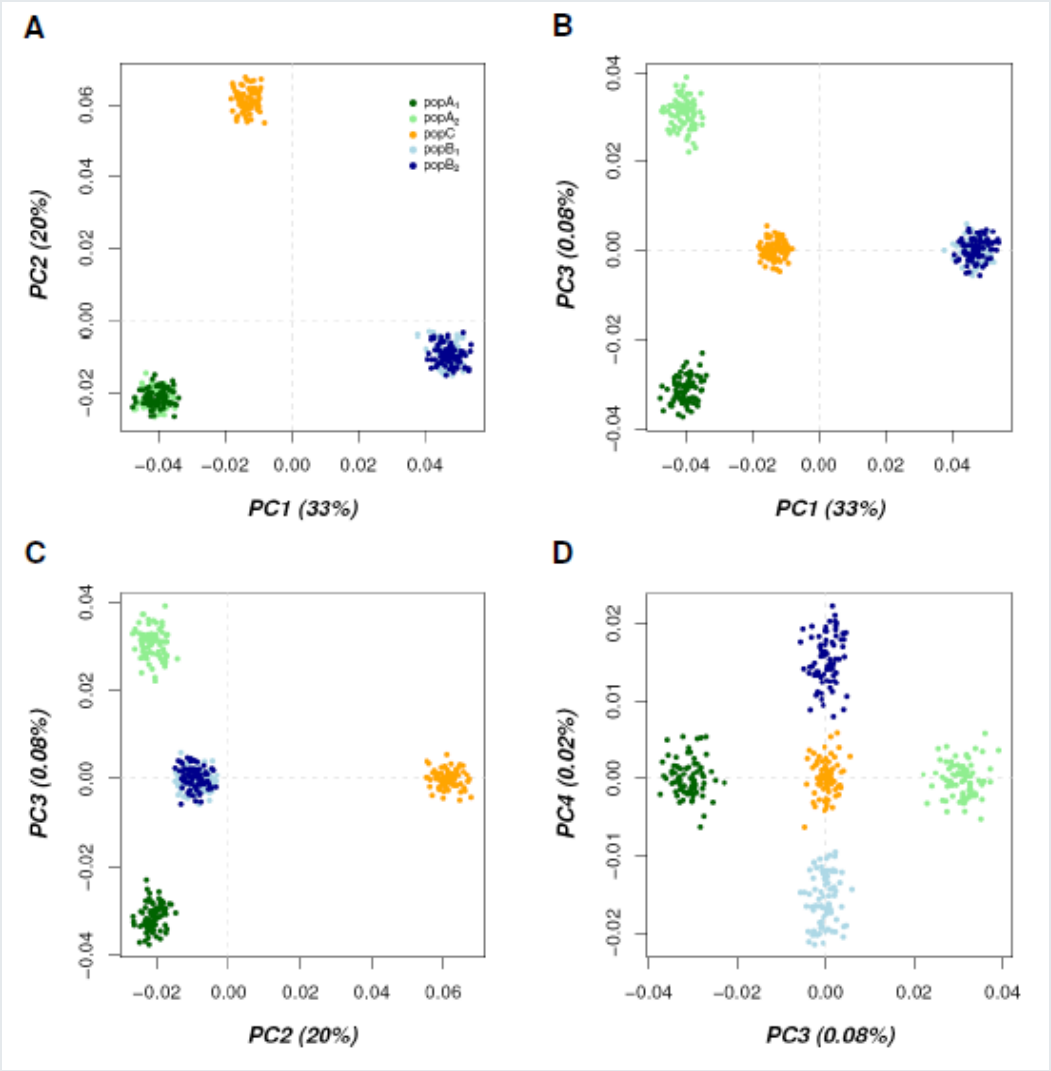
- How to can influential/informative individuals be identified within populations using genome-wide single nucleotide polymorphism (SNPs)?

- How can influential/informative (key ancestors) be visualized within populations?

Principal Component Analysis (PCA)

- **PCA** is one of the prevailing methods to identify population structures and to correct for population stratification.

- It has been already applied in numerous studies to study fine-scale population structures and to effectively assign individuals to population clusters.



Neuditschko et al. 2012

- Based on this principle Paschou *et al. 2007* developed a method to extract small sets of SNPs that correlate well with the population structure.

OPEN ACCESS Freely available online PLOS GENETICS

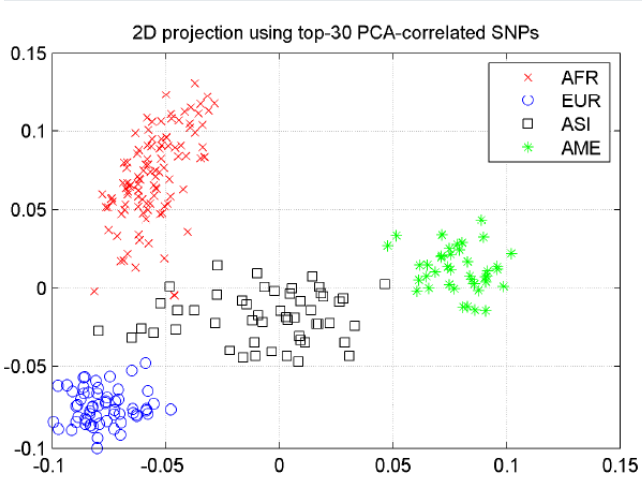
PCA-Correlated SNPs for Structure Identification in Worldwide Human Populations

Peristera Paschou[1*], Elad Ziv[2,3,4], Esteban G. Burchard[5,6], Shweta Choudhry[7], William Rodriguez-Cintron[8], Michael W. Mahoney[9], Petros Drineas[10]

1 Department of Molecular Biology and Genetics, Democritus University of Thrace, Alexandroupoli, Greece, 2 Division of General Internal Medicine, University of California San Francisco, San Francisco, California, United States of America, 3 Institute for Human Genetics, University of California San Francisco, San Francisco, California, United States of America, 4 Comprehensive Cancer Center, University of California San Francisco, San Francisco, California, United States of America, 5 Department of Biopharmaceutical Sciences, University of California San Francisco, San Francisco, California, United States of America, 6 Department of Medicine, University of California San Francisco, San Francisco, California, United States of America, 7 Lung Biology Center, Department of Medicine, University of California San Francisco, San Francisco, California, United States of America, 8 Pulmonary/CCM Veterans Caribbean Healthcare System, University of Puerto Rico School of Medicine, San Juan, Puerto Rico, United States of America, 9 Yahoo Research, Sunnyvale, California, United States of America, 10 Department of Computer Science, Rensselaer Polytechnic Institute, Troy, New York, United States of America,

Existing methods to ascertain small sets of markers for the identification of human population structure require prior knowledge of individual ancestry. Based on Principal Components Analysis (PCA), and recent results in theoretical computer science, we present a novel algorithm that, applied on genomewide data, selects small subsets of SNPs (PCA-correlated SNPs) to reproduce the structure found by PCA on the complete dataset, without use of ancestry information. Evaluating our method on a previously described dataset (10,805 SNPs, 11 populations), we demonstrate that a very small set of PCA-correlated SNPs can be effectively employed to assign individuals to particular continents or populations, using a simple clustering algorithm. We validate our methods on the HapMap populations and achieve perfect intercontinental differentiation with 14 PCA-correlated SNPs. The Chinese and Japanese populations can be easily differentiated using less than 100 PCA-correlated SNPs ascertained after evaluating 1.7 million SNPs from HapMap. We show that, in general, structure informative SNPs are not portable across geographic regions. However, we manage to identify a general set of 50 PCA-correlated SNPs that effectively assigns individuals to one of nine different populations. Compared to analysis with the measure of informativeness, our methods, although unsupervised, achieved similar results. We proceed to demonstrate that our algorithm can be effectively used for the analysis of admixed populations without having to trace the origin of individuals. Analyzing a Puerto Rican dataset (192 individuals, 7,257 SNPs), we show that PCA-correlated SNPs can be used to successfully predict structure and ancestry proportions. We subsequently validate these SNPs for structure identification in an independent Puerto Rican dataset. The algorithm that we introduce runs in seconds and can be easily applied on large genome-wide datasets, facilitating the identification of population substructure, stratification assessment in multi-stage whole-genome association studies, and the study of demographic history in human populations.

Citation: Paschou P, Ziv E, Burchard EG, Choudhry S, Rodriguez-Cintron W, et al. (2007) PCA-correlated SNPs for structure identification in worldwide human populations. PLoS Genet 3(9): e160. doi:10.1371/journal.pgen.0030160



Selecting PCA-Correlated SNPs

- The method presented in this study is based on the Singular Value Decomposition (SVD) of a SNP-Data matrix.
- PCA-Correlated SNPs are identified by calculating the correlation of all SNPs with the number of significant *k* principal components (PCs); so called PCA score.
- Inferring eigenvalues or PCs is a common strategy in population genetics to visualize population structures based on a small number of (PCs) (e.g. two or three).

Identify PCA Informative Individuals

- In order to identify PCA informative individuals (PCA-IIs) within populations, we applied the same principle as described by Paschou *et al. 2007*.

- The PCA-IIs selection procedure we present requires as input a symmetric relationship matrix and the number of significant principal components.

- Here, we have used IBD and pedigree derived relationship matrices and the empirical method Horn's parallel analysis to determine the number of significant PCs.

Applied datasets

- We exemplified our approach on two designed but disparate livestock populations namely sheep and horse.

- The sheep dataset we have applied describes a designed 2 breed intercross sheep resource folk, where the formation of population has been staged in three phases:

(1) Heterozygous F1 males and femals were created by crossing 4 Awassi founder sires with 30 wool Merino ewes.

(2) F1 sires representing each of the founder (F0) were selected and back-crossed to Merino ewes. In total 400, 150, 150 and 150 progeny were born for each for each of the F1 sires.

(3) Back cross ewes were being mated to F1 sires and F1 ewes and sires were inter-crossed to produce F2 progeny.

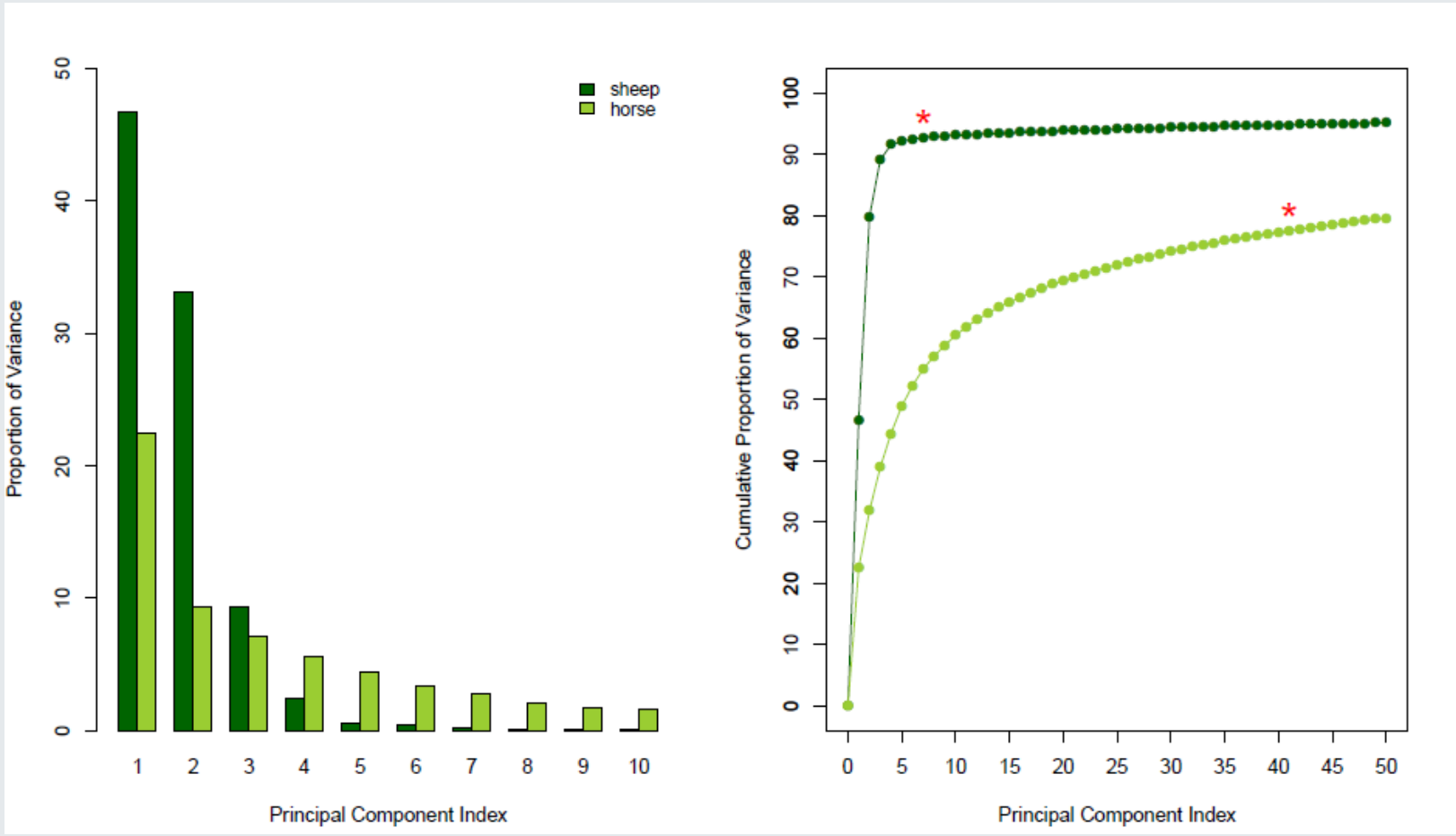
Applied datasets

- The horse dataset we have used represents an active breeding population including stallions with many offspring, younger stallions and breeding mares of the Swiss **Franches-Montagnes** breed.

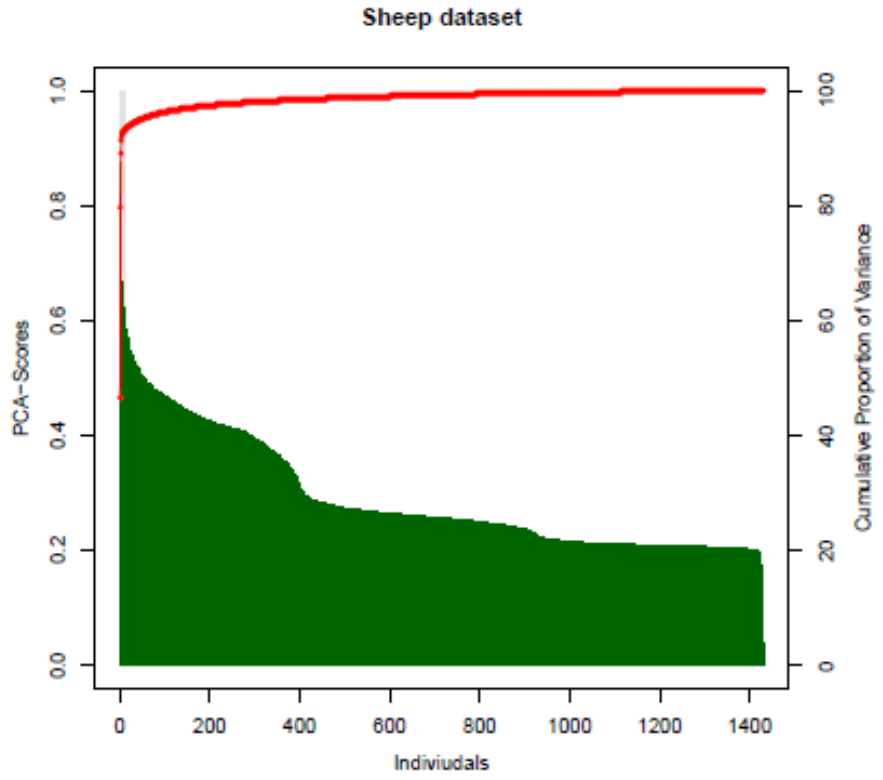
- The current population structure of this breed is based upon the formation of nine stallion linages.

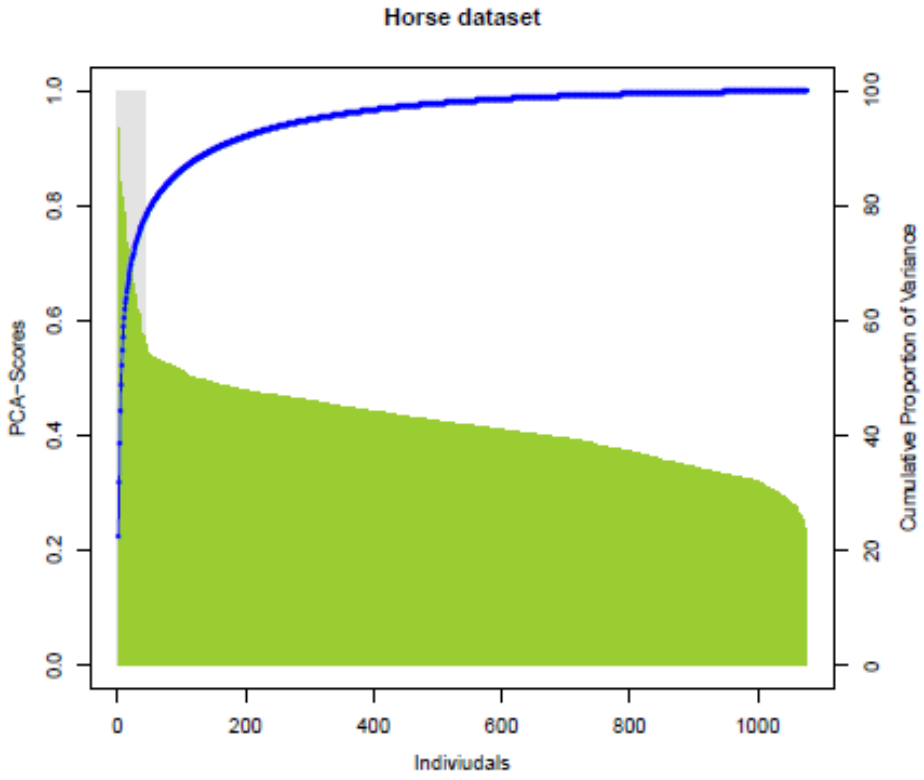
- For the sheep dataset we analyzed a total of 1’430 sheep genotyped for 44,693 autosomal SNPs, whilst for horse we have investigated 1’077 horses genotyped for 38,124 autosomal SNPs.

Results (significant PCs)



Results (PCA-Scores)





Results (Screening PCA-IIs)

- Within sheep the most informative individual is an inbred ram descending from the F1 foundation sire with the most progeny in the population (400) followed by the 4 F1 foundation sires.

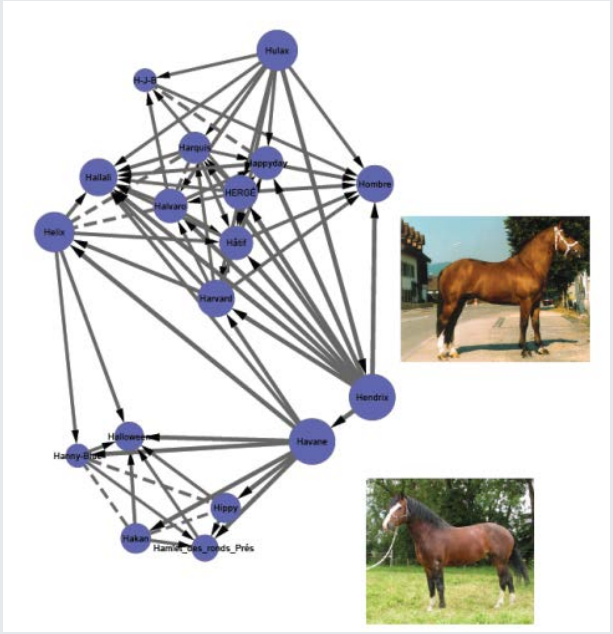
- Screening the top 20 PCA-IIs within the horse population we noticed that:

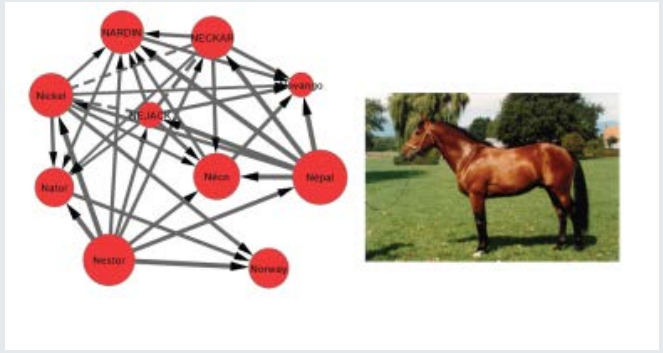
  (1) Only stallions are in the top ranking.
  (2) Top PCA-IIs descending from 4 different linages (H,E,L,N).
  (3) No stallions from three linages (Don, Q and V).
  (4) Many father son relationships, especially between stallions descending from lineages (H and E).

- Considering the top 41 PCA-IIs stallions descending from all linages are in the top ranking as well as important breeding mares.

Visualization of PCA-IIs within populations

- To visualize PCA-IIs within populations we have the applied recently published approach NETVIEW (Neuditschko *et al. 2012*).
- Using NETVIEW population structures are presented in terms of nodes, edges between nodes and thickness of edges.
- In the final network presentation the node size is associated with the information score and the direction of edges with ancestry information.





Conclusions

- The method allows a successfull identification of influential individuals using any kind of relationship matrices.
- The combination of PCA-IIs with high definition network analysis allows the accurate identification of key ancestors without the need of individual ancestry.
- Especially useful to investigate population structures in indigenous breeds and wild species, where ancestry information is often lacking or not available.
- Useful for assembling resource populations to facilitate accurate genotype imputation across and within populations.


Thank you for your attention!