Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Federal Department of Economic Affairs,
Education and Research EAER

Agroscope

Breaking down big data:

A two-step method for visualising complex data structures

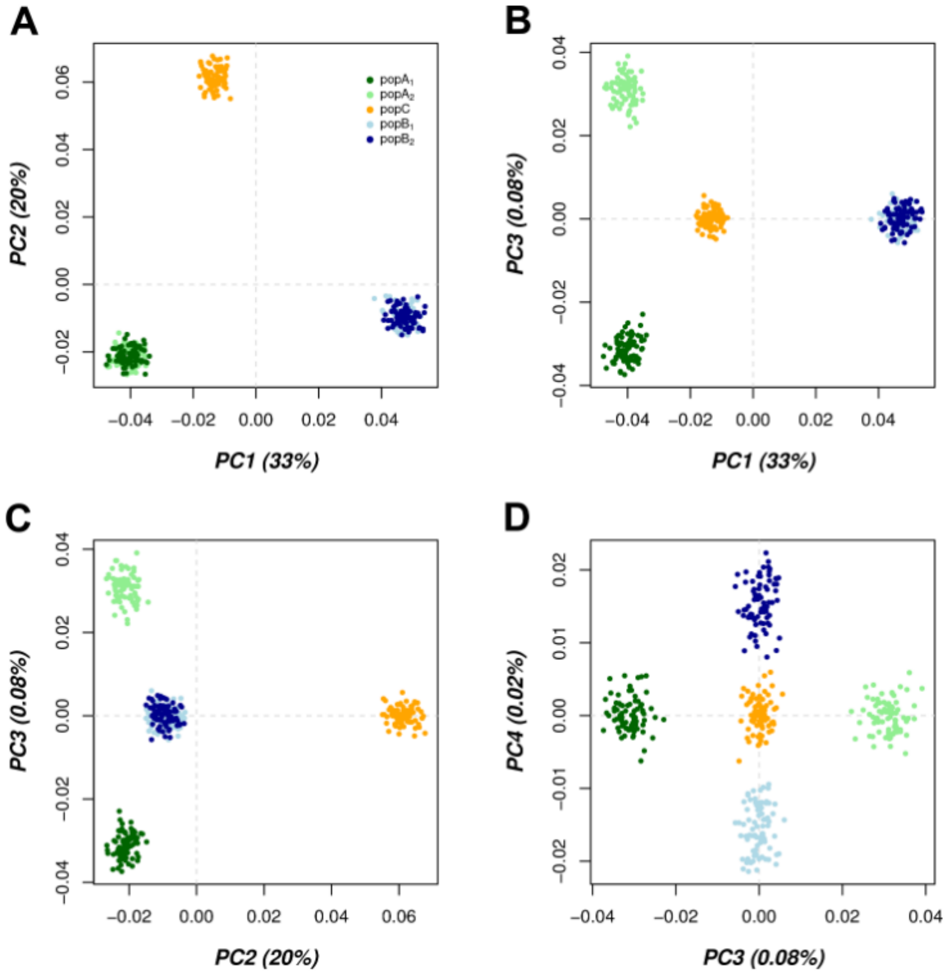
Markus Neuditschko

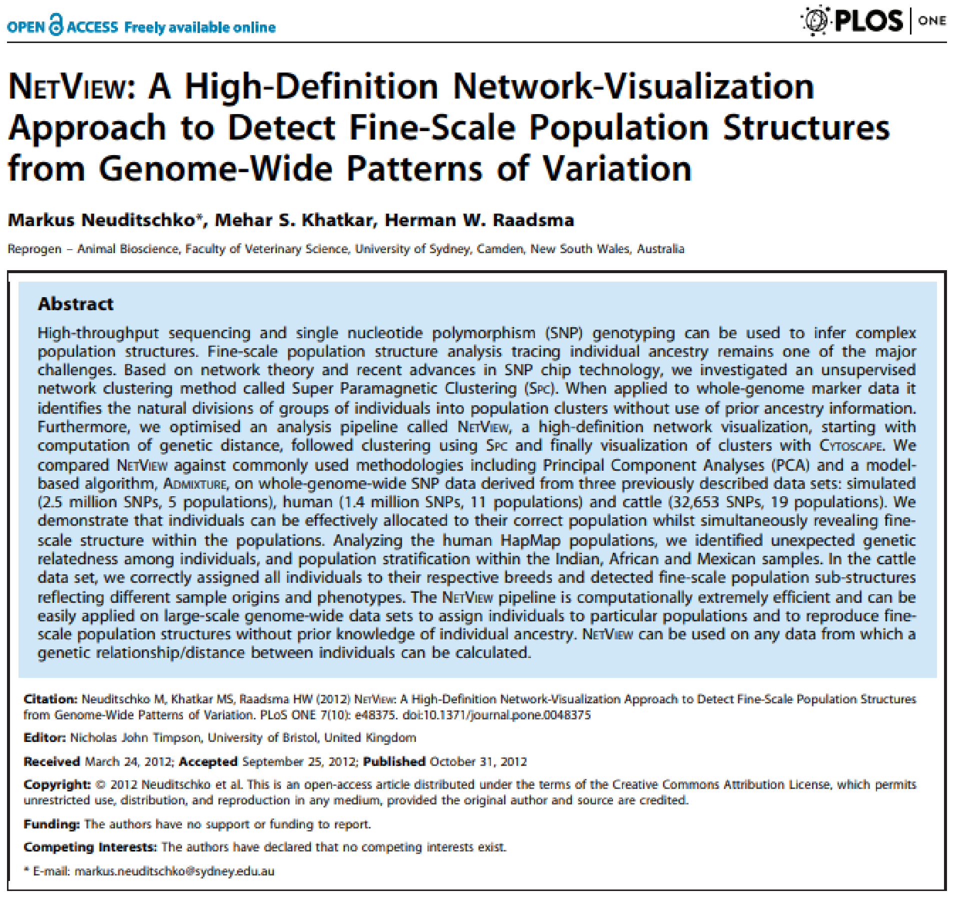
1^{st} EAAP conference on Artificial Intelligence 4 Animal Science

Background

- **Principal Component Analysis** (**PCA**) is a widely used method for uncovering patterns in complex data structures.
- If effectively simplifies complex data by reducing their complexity.
- This method is less-suited to explore big datasets including thousands of observations, as the visualization beyond three dimensions becomes ineffective.
- PCA is one of the prevailing methods to explore population structures using genotype information (SNP arrays or whole-genome sequencing).
- To assess **high-resolution** population structures we developed a two-step approach by visualizing the PCA result on a population network (identification of key contributors)

PCA results of a simulated population structure





OPEN ACCESS Freely available online

PLOS | ONE

NetView: A High-Definition Network-Visualization Approach to Detect Fine-Scale Population Structures from Genome-Wide Patterns of Variation

Markus Neuditschko*, Mehar S. Khatkar, Herman W. Raadsma

Reprogen – Animal Bioscience, Faculty of Veterinary Science, University of Sydney, Camden, New South Wales, Australia

Abstract

High-throughput sequencing and single nucleotide polymorphism (SNP) genotyping can be used to infer complex population structures. Fine-scale population structure analysis tracing individual ancestry remains one of the major challenges. Based on network theory and recent advances in SNP chip technology, we investigated an unsupervised network clustering method called Super Paramagnetic Clustering (Spc). When applied to whole-genome marker data it identifies the natural divisions of groups of individuals into population clusters without use of prior ancestry information. Furthermore, we optimised an analysis pipeline called NetView, a high-definition network visualization, starting with computation of genetic distance, followed clustering using Spc and finally visualization of clusters with Cytoscape. We compared NetView against commonly used methodologies including Principal Component Analyses (PCA) and a model-based algorithm, Admixture, on whole-genome-wide SNP data derived from three previously described data sets: simulated (2.5 million SNPs, 5 populations), human (1.4 million SNPs, 11 populations) and cattle (32,653 SNPs, 19 populations). We demonstrate that individuals can be effectively allocated to their correct population whilst simultaneously revealing fine-scale structure within the populations. Analyzing the human HapMap populations, we identified unexpected genetic relatedness among individuals, and population stratification within the Indian, African and Mexican samples. In the cattle data set, we correctly assigned all individuals to their respective breeds and detected fine-scale population sub-structures reflecting different sample origins and phenotypes. The NetView pipeline is computationally extremely efficient and can be easily applied on large-scale genome-wide data sets to assign individuals to particular populations and to reproduce fine-scale population structures without prior knowledge of individual ancestry. NetView can be used on any data from which a genetic relationship/distance between individuals can be calculated.

Citation: Neuditschko M, Khatkar MS, Raadsma HW (2012) NetView: A High-Definition Network-Visualization Approach to Detect Fine-Scale Population Structures from Genome-Wide Patterns of Variation. PLoS ONE 7(10): e48375. doi:10.1371/journal.pone.0048375

Editor: Nicholas John Timpson, University of Bristol, United Kingdom

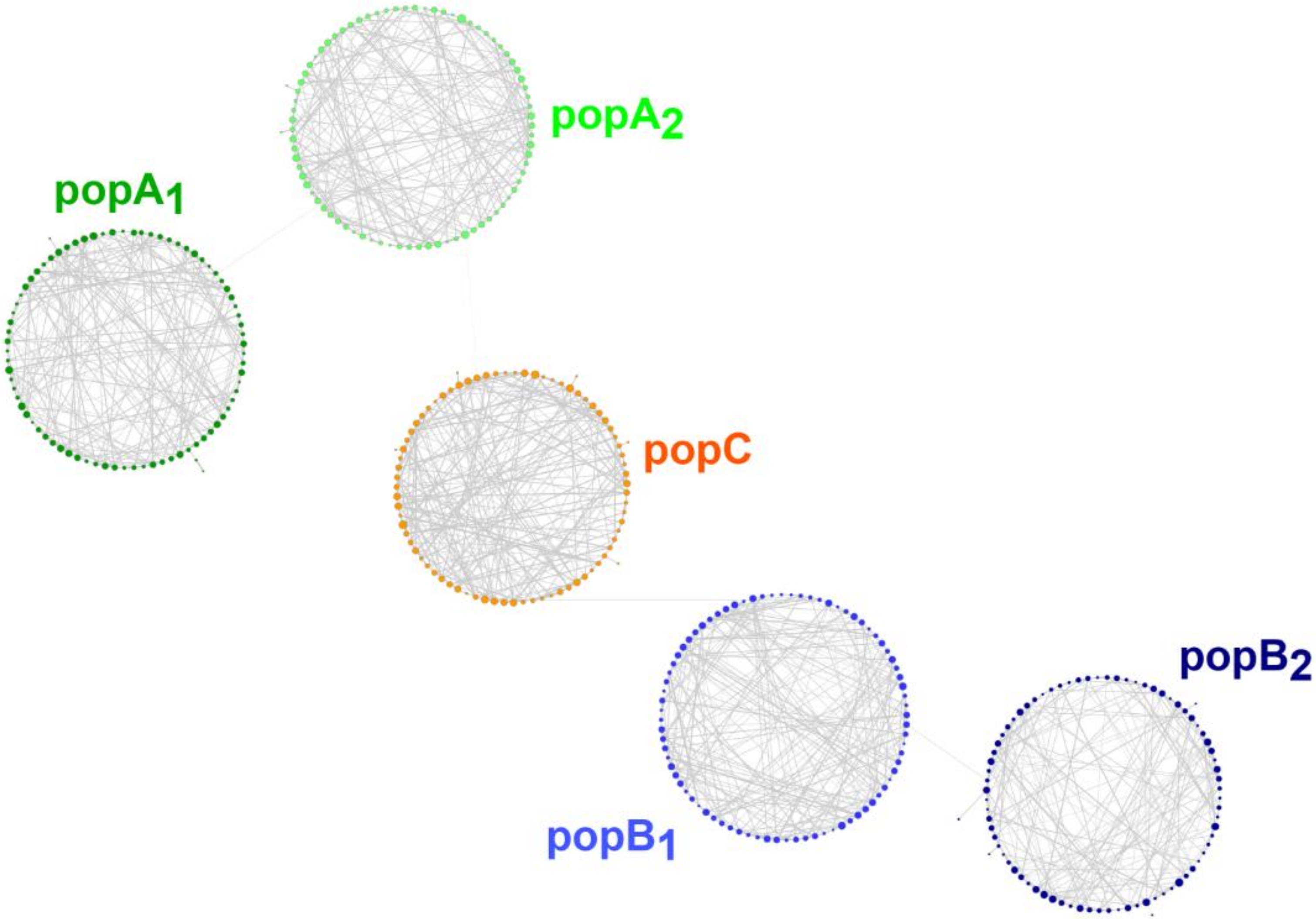
Received March 24, 2012; **Accepted** September 25, 2012; **Published** October 31, 2012

Funding: The authors have no support or funding to report.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: markus.neuditschko@sydney.edu.au

NetView of a simulated population structure



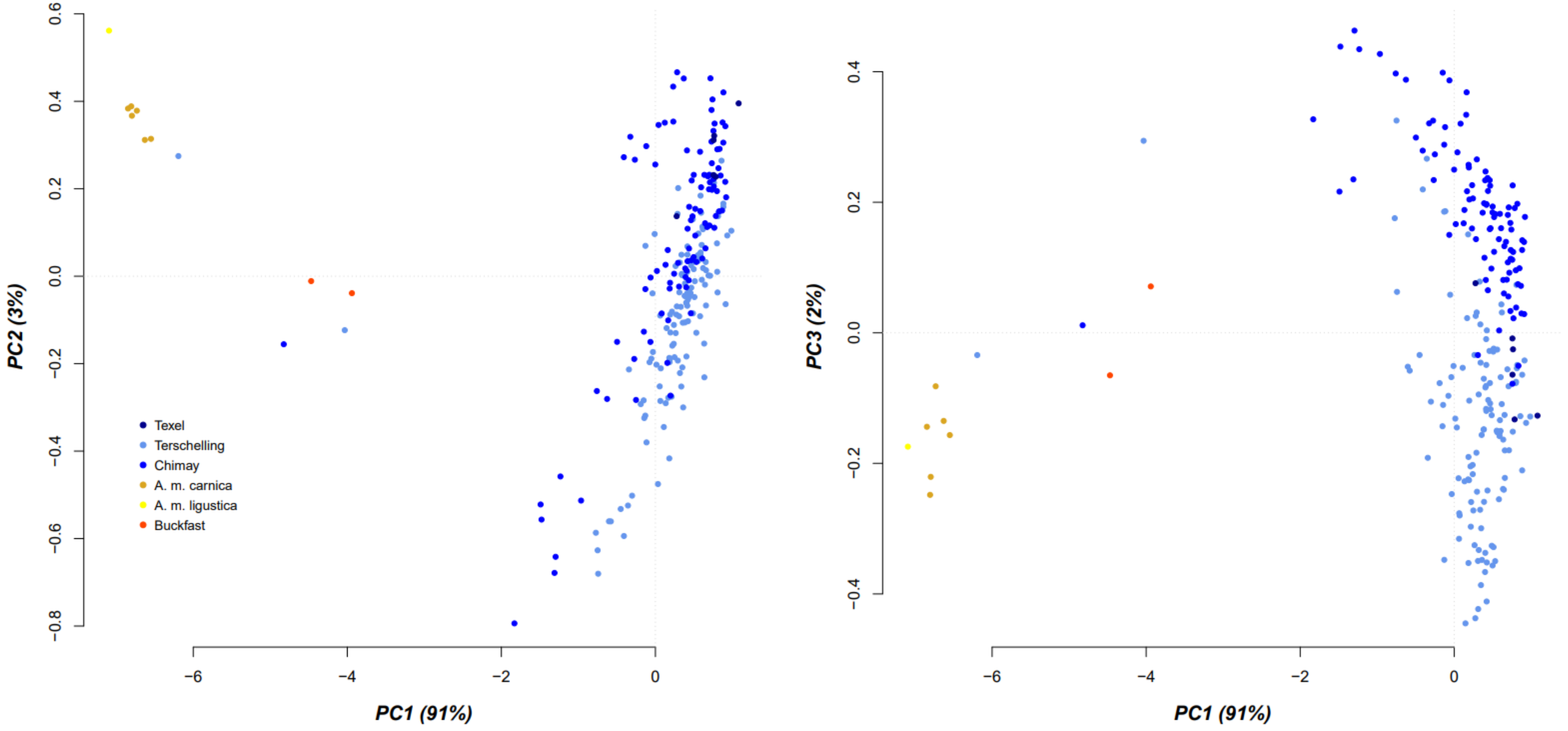
Identification of key contributors

- The method is based on the **Singular Value Decomposition** (SVD) and requires a symmetric relationship matrix between ***n*** individuals.

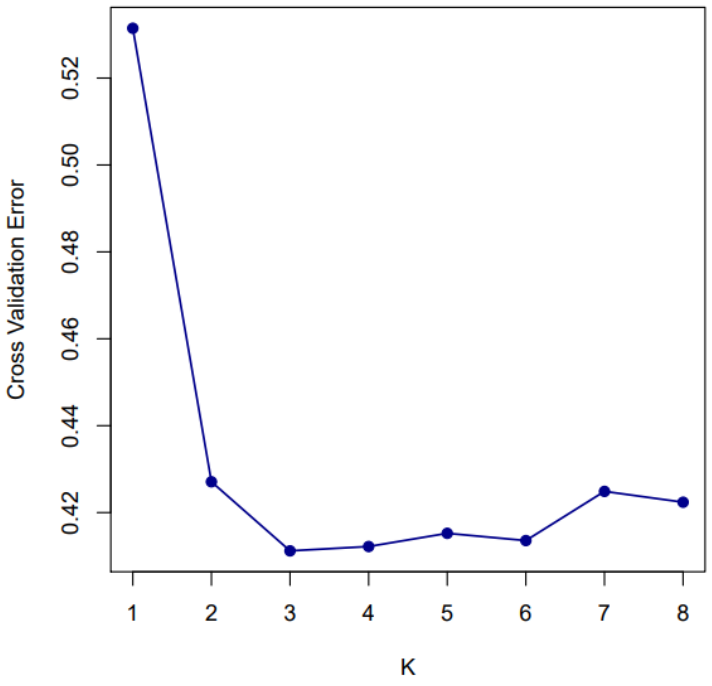
- **Key contributors** (individuals accountig for most of the genetic variance) are identified by calculating the correlation with the number of significant ***k*** principal components (PCs); so called **genetic contribution score.**

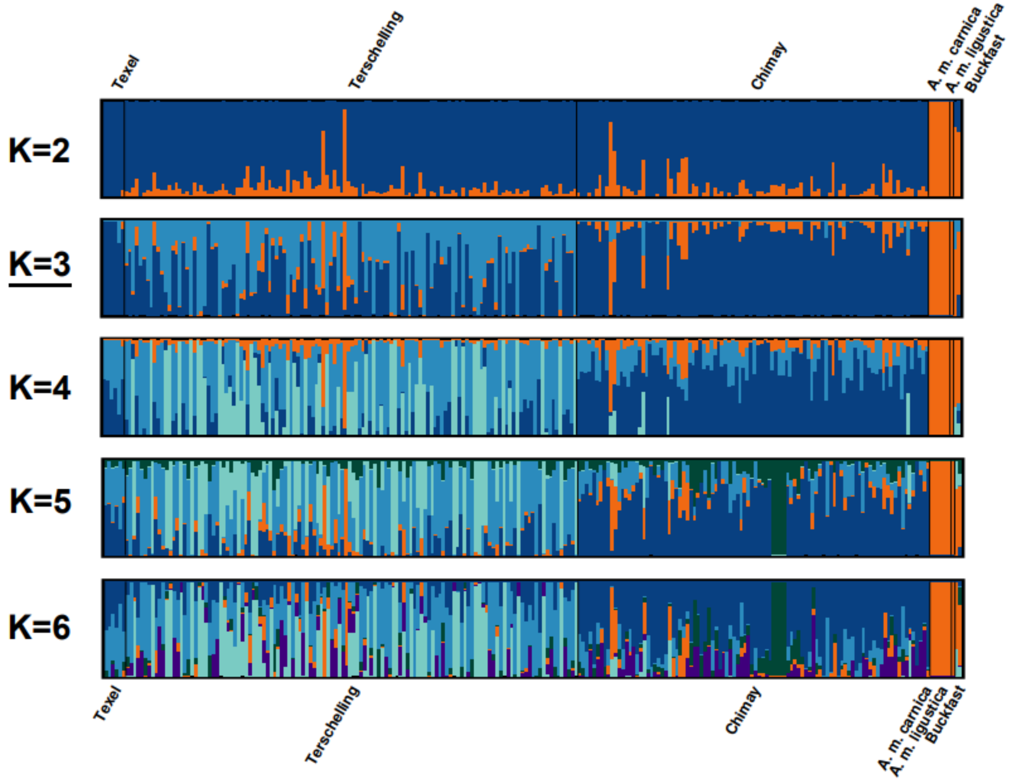
- In population genetics we applied the method on pedigree- and genome-derived relationship matrices and used the the empirical method **Horn's parallel analysis** to determine the number of significant PCs.

Results – PCA (Lowland Honeybees)

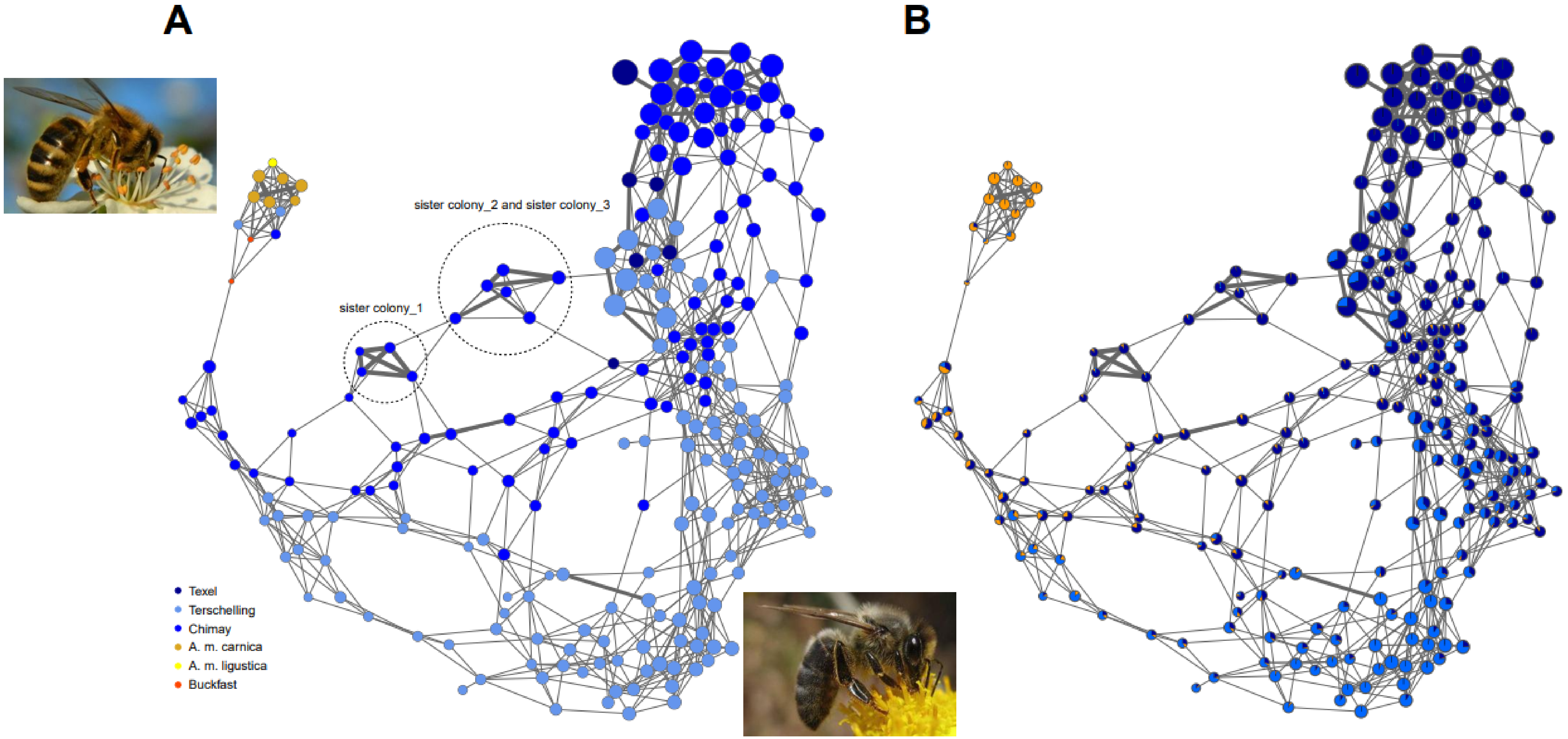


Results – Admixture (Lowland Honeybees)



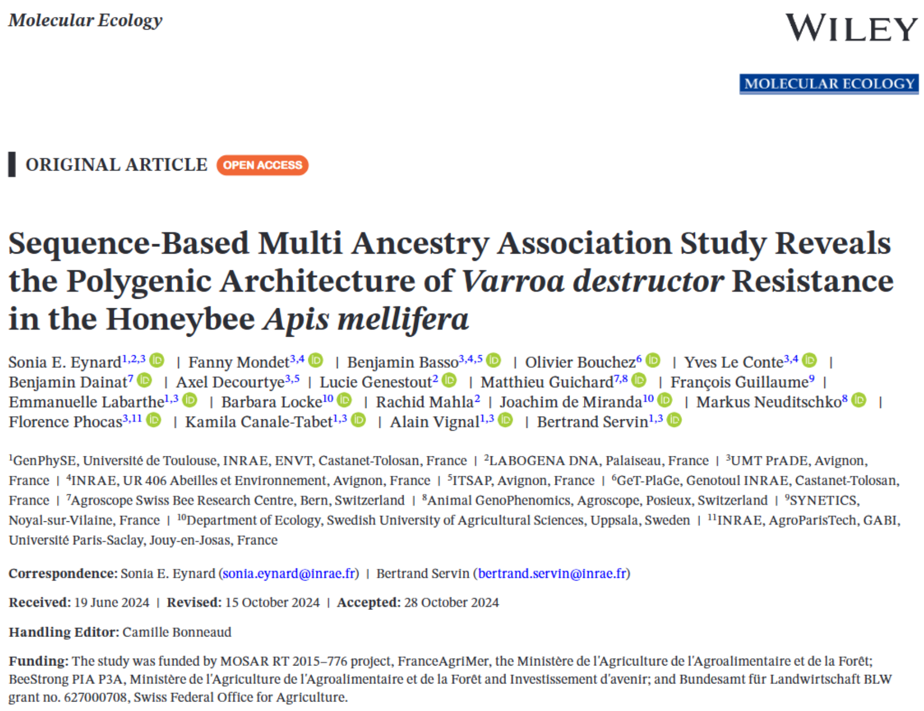


Results – High-resolution population structure



Results – Honeybees (symbiosphere data)

- Symbiosphere is defined by bacteria (1156 species), viruses (318 species), and fungi (139 species).



Molecular Ecology

WILEY

MOLECULAR ECOLOGY

ORIGINAL ARTICLE OPEN ACCESS

Sequence-Based Multi Ancestry Association Study Reveals the Polygenic Architecture of *Varroa destructor* Resistance in the Honeybee *Apis mellifera*

Sonia E. Eynard[1,2,3] | Fanny Mondet[3,4] | Benjamin Basso[3,4,5] | Olivier Bouchez[6] | Yves Le Conte[3,4] | Benjamin Dainat[7] | Axel Decourtye[3,5] | Lucie Genestout[2] | Matthieu Guichard[7,8] | François Guillaume[9] | Emmanuelle Labarthe[1,3] | Barbara Locke[10] | Rachid Mahla[2] | Joachim de Miranda[10] | Markus Neuditschko[8] | Florence Phocas[3,11] | Kamila Canale-Tabet[1,3] | Alain Vignal[1,3] | Bertrand Servin[1,3]

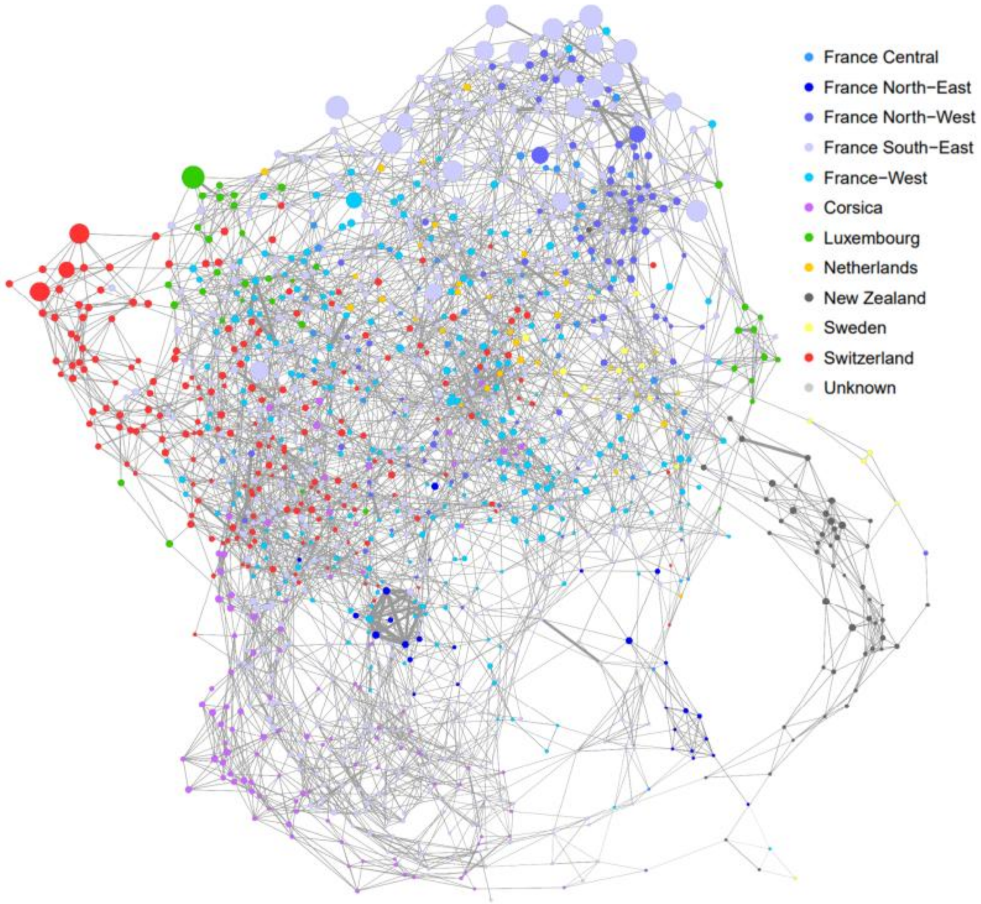
[1]GenPhySE, Université de Toulouse, INRAE, ENVT, Castanet-Tolosan, France | [2]LABOGENA DNA, Palaiseau, France | [3]UMT PrADE, Avignon, France | [4]INRAE, UR 406 Abeilles et Environnement, Avignon, France | [5]ITSAP, Avignon, France | [6]GeT-PlaGe, Genotoul INRAE, Castanet-Tolosan, France | [7]Agroscope Swiss Bee Research Centre, Bern, Switzerland | [8]Animal GenoPhenomics, Agroscope, Posieux, Switzerland | [9]SYNETICS, Noyal-sur-Vilaine, France | [10]Department of Ecology, Swedish University of Agricultural Sciences, Uppsala, Sweden | [11]INRAE, AgroParisTech, GABI, Université Paris-Saclay, Jouy-en-Josas, France

Correspondence: Sonia E. Eynard (sonia.eynard@inrae.fr) | Bertrand Servin (bertrand.servin@inrae.fr)

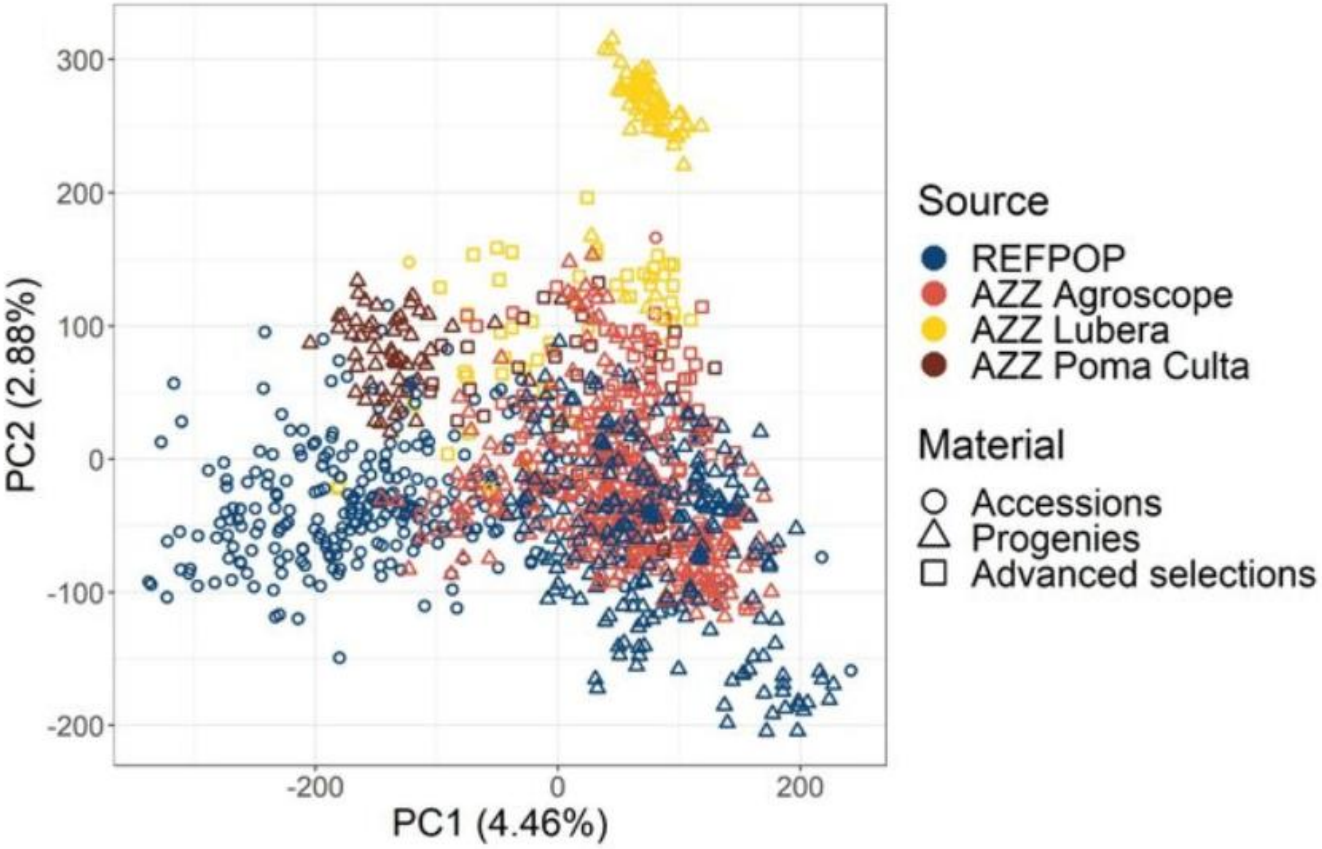
Received: 19 June 2024 | **Revised:** 15 October 2024 | **Accepted:** 28 October 2024

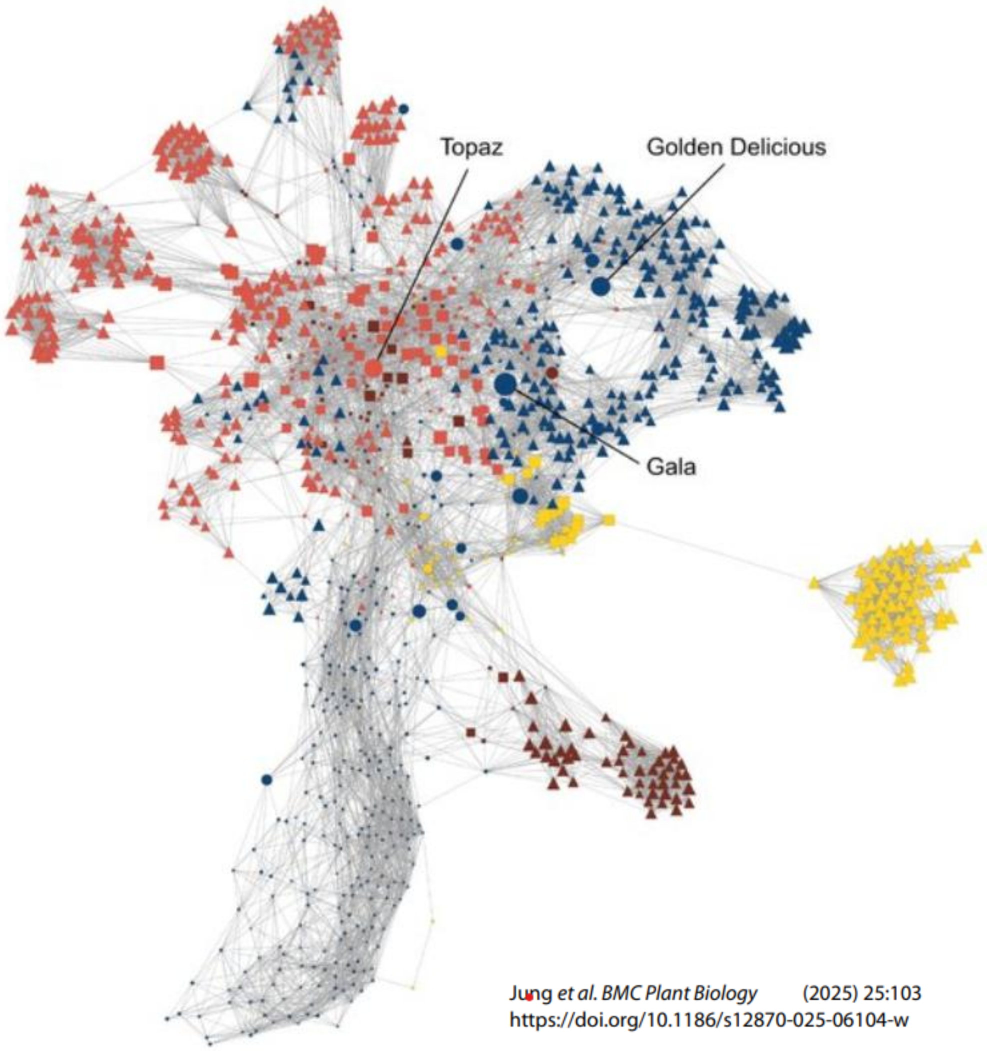
Handling Editor: Camille Bonneaud

Funding: The study was funded by MOSAR RT 2015–776 project, FranceAgriMer, the Ministère de l'Agriculture de l'Agroalimentaire et de la Forêt; BeeStrong PIA P3A, Ministère de l'Agriculture de l'Agroalimentaire et de la Forêt and Investissement d'avenir; and Bundesamt für Landwirtschaft BLW grant no. 627000708, Swiss Federal Office for Agriculture.



Results – Apple (genotype data)





Jung *et al. BMC Plant Biology* (2025) 25:103
https://doi.org/10.1186/s12870-025-06104-w

Conclusion

- In population genomics (genotype and microbiome data), we have demonstrated that combining the **identification of key contributors (PCA)** with **network visualization (NetView)** helps to uncover fine-scale population structures.

- Besides the assessment of high-resolution population structures, the selection of key contributors improved **imputation accuracy** and **genomic prediction** in target populations.

- We believe our **two-step method** offers substantial potential for visualizing complex data structures across various research disciplines, extending far beyond population genetics (e.g. by optimizing the training data in **machine learning**).

Get in touch with us

markus.neuditschko@agroscope.admin.ch

Agroscope good food, healthy environment
www.agroscope.admin.ch