NetView

A high definition network-visualization approach to detect fine-scale population structures

Markus Neuditschko

Synbreed Winterschool, 2012

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

http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0048375

Overview

- The NetView idea
- Application on a simulated data set (5, populations, 2.5 Million SNPs)
- Comparison with common applied methods (PCA and Admixture)

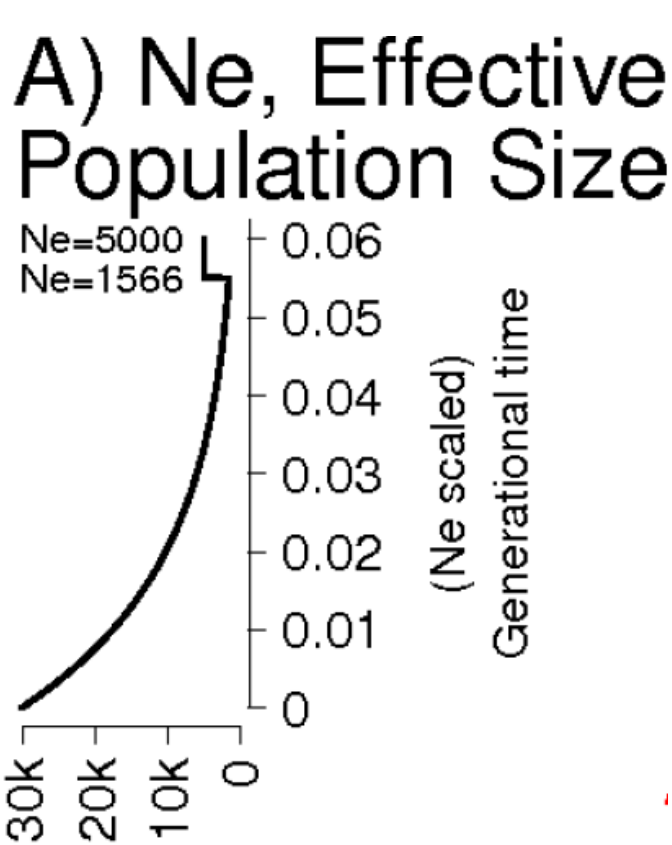
The NetView idea

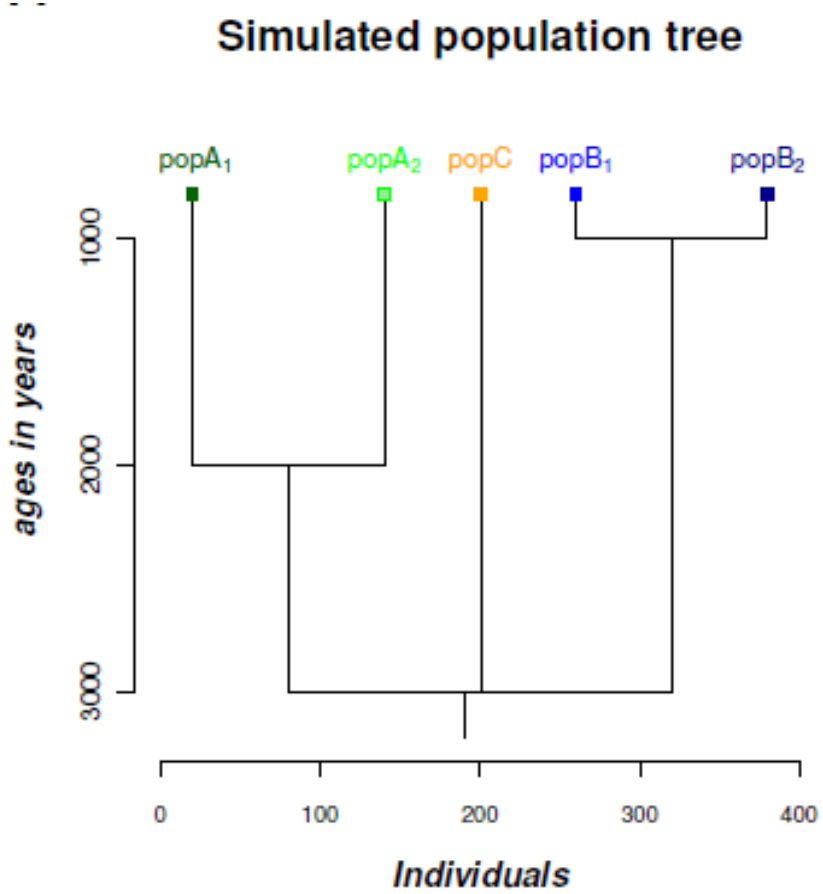
- The ***NetView*** approach is based on the so-called community structure, which arises by minimizing number of interactions in a symmetric distance matrix according to a mutual neighborhood criterion.

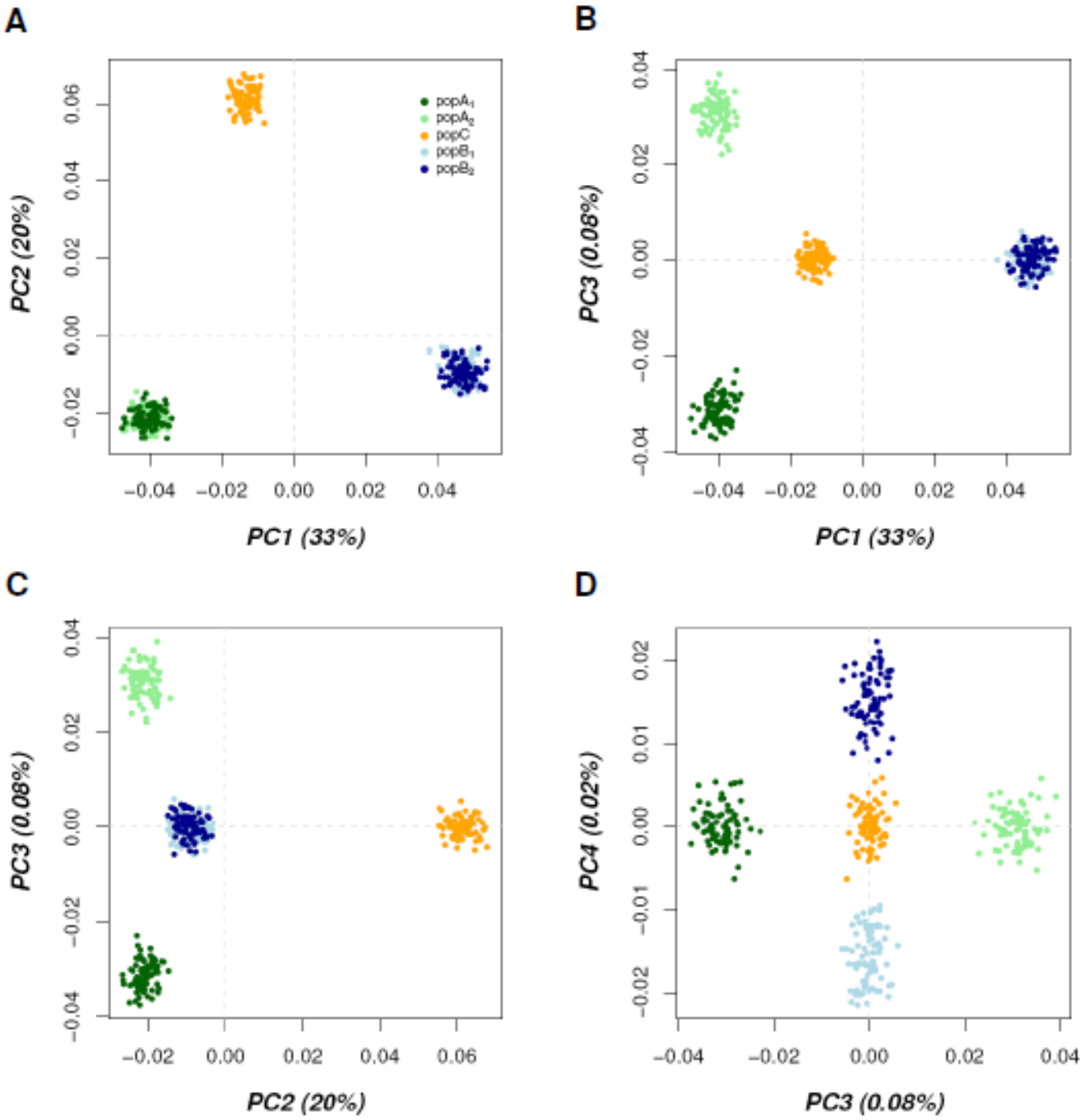
0	0.262797	0.249881	0.255202	0.256649	0.257468	0.264734	0.250277	0.257208
0.262797	0	0.265028	0.257817	0.261399	0.258552	0.260897	0.246905	0.266166
0.249881	0.265028	0	0.26651	0.250792	0.258734	0.262759	0.250613	0.255422
0.255202	0.257817	0.26651	0	0.268024	0.230432	0.244999	0.249137	0.265743
0.256649	0.261399	0.250792	0.268024	0	0.263960	0.264965	0.262013	0.256769
0.257468	0.258552	0.258734	0.230432	0.263960	0	0.239278	0.254612	0.267784
0.264734	0.260897	0.262759	0.244999	0.264965	0.239278	0	0.250467	0.260367
0.250277	0.246905	0.250613	0.249137	0.262013	0.254612	0.250467	0	0.26553
0.257208	0.266166	0.255422	0.265743	0.256769	0.267784	0.260367	0.26553	0

- Once, the community structures have been extracted different approaches can be applied to detect and characterize the community structures (e.g. the Potts Hamiltonian Model).

Simulated population structure



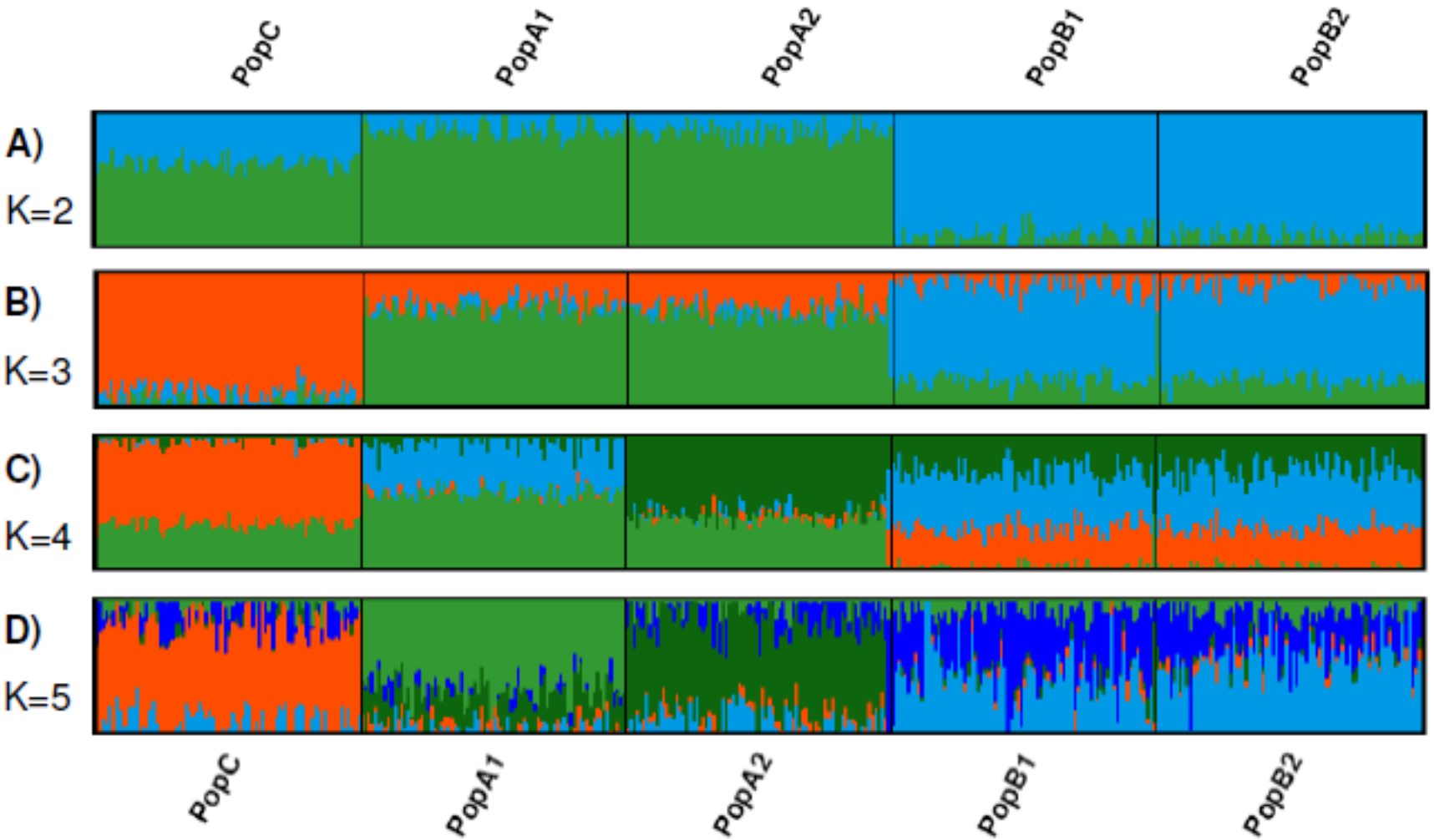



A
B
C
D
popA1
popA2
popC
popB1
popB2
PC1 (33%)
PC2 (20%)
PC3 (0.08%)
PC4 (0.02%)
−0.04
−0.02
0.00
0.02
0.04
0.06
−0.01
0.01

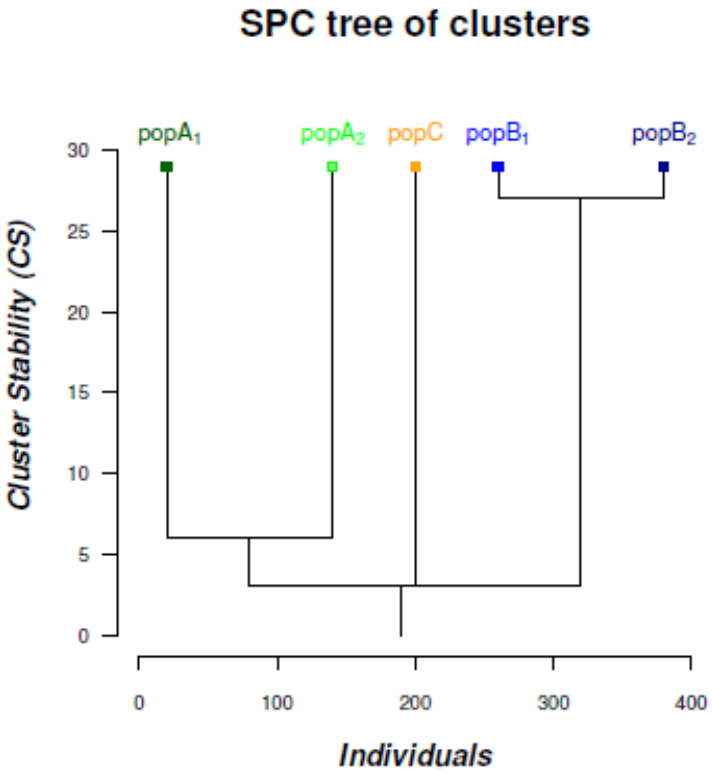
PCA and k-means clustering

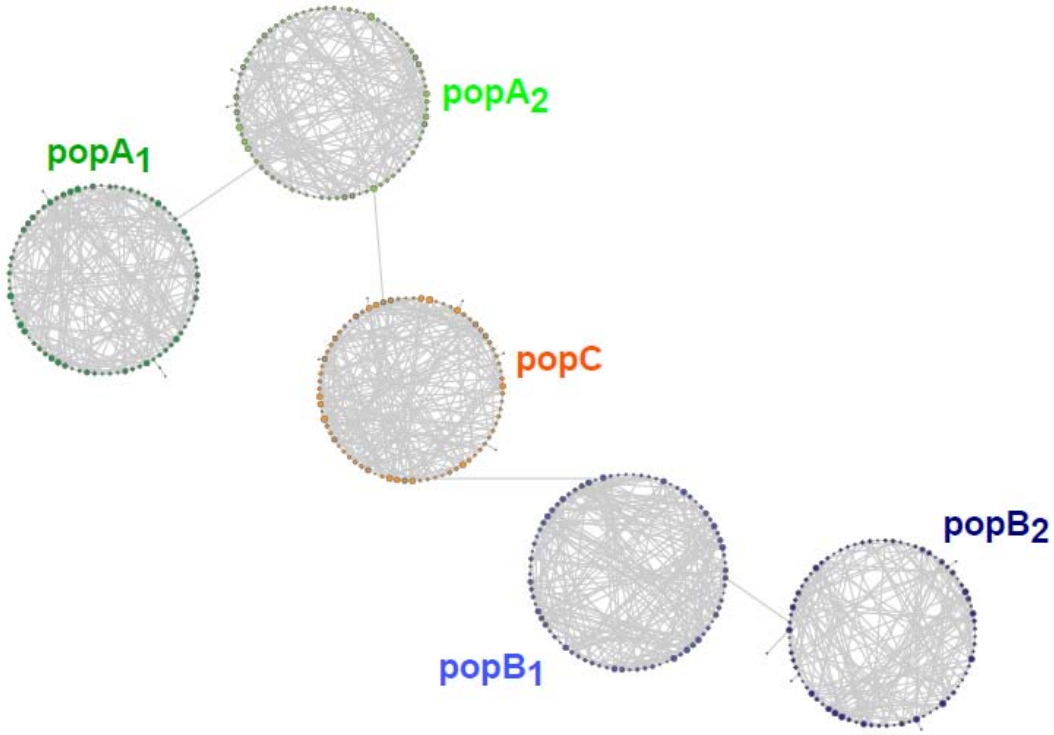
- Applied Horns Parallel analysis to determine the number of significant components. R package: paran (4 Components)
- Used the Calinksi Criterion to determine the optimal number of clusters: R package: vegan (5 clusters)
- Finally we used the k-means cluster alogrithm to assign individuals to respective clusters (Standard procedure in R)

Admixture


PopC
PopA1
PopA2
PopB1
PopB2
A)
K=2
B)
K=3
C)
K=4
D)
K=5
PopC
PopA1
PopA2
PopB1
PopB2

NetView


SPC tree of clusters
popA1
popA2
popC
popB1
popB2
Cluster Stability (CS)
30
25
20
15
10
5
0
0
100
200
300
400
Individuals


popA2
popA1
popC
popB2
popB1

Comparison of the applied methods

- **PCA** is an efficient method to determine the number of significant clusters, when it is applied in combination with various test statistics.
- **Admixture** provides useful information on the admixture between populations, but failed to seperate close related populations.
- **NetView** provides a high defintion network visualization and determines the significance of each individual.

Visit NetView hompage: http://sydney.edu.au/vetscience/reprogen/netview/index.shtml