File S1

Supporting Information for PSIKO

September 26, 2014

This is the supplement for (POPESCU *et al.*, 2014). We begin by complementing our simulation results for PSIKO for $K = 3$ founders presented in (POPESCU *et al.*, 2014) with those for larger values of K i. e. $K = 4 \ldots 10$. We then present a pseudo-code version of PSIKO (Algorithm 1). Subsequent to this, we present mathematical details on how equations underpinning PSIKO are solved. We conclude with details on the `msms` commands used to generate our simulated datasets. Unless stated otherwise, our notation follows that of (POPESCU *et al.*, 2014).

LARGER VALUES FOR K

For various Dirichlet distribution parameter settings in a symmetric simulation scenario (see POPESCU *et al.* (2014) for details), we present in Table **S**1 the average Root Mean Squared Error (RMSE) between inferred and true Q-matrices for values of $K = 4, \ldots, 10$. As can be readily observed, the average RMSE over all 100 datasets for each Dirichlet distribution parameter choice and each value for K is below 1.6% which suggests that PSIKO performs very well for larger values of K.

Dirichlet parameters	1	0.5	0.1
$K = 4$	0.013	0.009	0.007
$K = 5$	0.013	0.01	0.01
$K = 6$	0.015	0.01	0.01
$K = 7$	0.015	0.011	0.011
$K = 8$	0.015	0.012	0.012
$K = 9$	0.016	0.013	0.013
$K = 10$	0.016	0.013	0.01

Table S1: Denoting the Dirichlet distribution parameter settings of all 1s, all 0.5s and all 0.1s, by 1, 0.5 and 0.1 respectively and using the latter as column labels, we present the average RMSE between the true and the estimated Q-matrix for PSIKO for the values $K = 4, \ldots, 10$.

PSIKO PSEUDO-CODE

A representation of PSIKO in pseudocode is given in Algorithm 1. Let *nComp* denote the number of found significant components.

Algorithm 1 PSIKO

Input: A dataset in the form of a SNP matrix $\mathbf{X}$ with accession loci encoded as 2's, 1's and 0's.

Output: The number K of founders, the significant principal components (PCs) and a Q-matrix $Q = (q_{\mathbf{cx}})$ for $\mathbf{X}$, where $\mathbf{c}$ is a founder of $\mathbf{X}$ and $\mathbf{x}$ is a an accession of $\mathbf{X}$.

STEP I (Dimensionality Reduction):
1 : first apply linear kernel-PCA to $\mathbf{X}$ to reduce dimensionality of the dataset and then the Tracy-Widom test for non-zero eigenvalues to infer the number *nComp* of significant principal components. Finally use those components to compute a *nComp* dimensional dataset X'
2 : normalize $\mathbf{X}'$ to have zero mean and unit variance
STEP II (Population Structure Inference):
3 : find the vertices (and thus the number K of founders) of the $(nComp-1)$-simplex representing $\mathbf{X}'$ by minimising Equation (1) below
4 : return K and the matrix Q found in Step 3 and the significant PCs found in line 1

OPTIMISING EQUATIONS UNDERPINNING PSIKO

In this section, we give details on the algorithm used to minimise Equation (2) of (POPESCU *et al.*, 2014), that is:

$$\mathscr{L}(\mathbf{A}, Q) = ||\mathbf{X} - \mathbf{A}Q||_F^2, \tag{1}$$

where $\mathbf{A} = (\mathbf{a}_i)_{1 \leq i \leq K}$ and $\mathbf{a}_1, \mathbf{a}_2, \ldots \mathbf{a}_K$ are the founders of $\mathbf{X}$ represented as column vectors, $Q = (q_{\mathbf{x}i})_{1 \leq i \leq K}$ is the matrix of ancestry coefficients for each accession $\mathbf{x} \in \mathbf{X}$, and K is the number of putative founders.

We start by making some observations that are specific to optimising Equation (1), and then present in Algorithm 2 an efficient algorithm for minimising it. The notation used follows that of (POPESCU *et al.*, 2014).

Suppose we are given a matrix Q. Finding a matrix $\mathbf{A}$ which minimises Equation (1) can easily be achieved via linear least-squares optimisation. More precisely, we have that

$$\mathbf{x} = \sum_{i=1}^{K} q_{\mathbf{x}i} \mathbf{a}_i, \tag{2}$$

holds for any accession $\mathbf{x}$ in our data set. In the context of optimising Equation (1) we are interested in finding values for $\mathbf{a}_i$, $1 \leq i \leq K$ such that a given accession $\mathbf{x}$ in $\mathbf{X}$ is approximated as closely as possible by Equation (2). This can be achieved by using:

Observation 0.1.

$$\mathbf{A} = (Q^T Q + \Gamma\Gamma^T)^{-1} Q^T \mathbf{X}^T, \tag{3}$$

where Q and $\mathbf{X}$ are as before and $\Gamma = \mathbf{I}$ is a Tikhonov regularisation matrix.

Now consider the converse problem, i. e. that the matrix $\mathbf{A}$ is known, and that we are interested in finding the matrix Q. For this we once again use Equation (2) above. More precisely, utilising the fact that $\sum_{i=1}^{K} q_{\mathbf{x}i} = 1$ holds for all $\mathbf{x} \in \mathbf{X}$, we obtain:

Observation 0.2. *Let* $\mathbf{B} := \begin{pmatrix} 1 & 1 & \dots & 1 \\ \mathbf{a}_1 & \mathbf{a}_2 & \dots & \mathbf{a}_k \end{pmatrix}$, $\mathbf{q_x} := \begin{pmatrix} q_{\mathbf{x}1} \\ q_{\mathbf{x}2} \\ \dots \\ q_{\mathbf{x}K} \end{pmatrix}$ *and* $\mathbf{x}' := \begin{pmatrix} 1 \\ \mathbf{x} \end{pmatrix}$.

Then $\mathbf{Bq_x} = \mathbf{x}'$ *or, equivalently,*

$$\mathbf{q_x} = \mathbf{B}^{-1}\mathbf{x}' \tag{4}$$

We note that the above solution for $\mathbf{q_x}$ can produce entries which are outside the interval $[0,1]$. To address this we followed the strategy employed in sNMF (FRICHO *et al.*, 2014), and first set for all $\mathbf{x} \in \mathbf{X}$ all entries of $\mathbf{q_x}$ that are negative to zero. We then divide each entry of $\mathbf{q_x}$ by the sum of entries of $\mathbf{q_x}$. This ensures that the values of $\mathbf{q_x}$ lie in the interval $[0,1]$ and that they also sum to one.

Using Observations 0.1 and 0.2, we can optimise Equation (1) iteratively (see Algorithm 2, with ε set to 10^{-5}). We found that that algorithm returns accurate estimates of the Q matrix across all simulation scenarios as well as for the two biological datasets under investigation in (POPESCU *et al.*, 2014).

Algorithm 2 Algorithm used to optimise Equation (1)

Input: A data matrix $\mathbf{X}$ as returned by Step I of PSIKO
Output: A matrix $\mathbf{A}$ of founders for $\mathbf{X}$ as well as Q-matrix Q for $\mathbf{X}$, minimising Equation (1).

Initialise $\mathbf{A}$ and Q randomly.
$prev = 0$
$cur = \mathscr{L}(\mathbf{A}, Q)$
set ε to a small number, say 10^{-5}
while $|prev - cur| < \varepsilon$ **do**
  estimate Q given $\mathbf{A}$ using Equation (4)
  estimate $\mathbf{A}$ given Q using Equation (3)
  $prev = cur$
  $cur = \mathscr{L}(\mathbf{A}, Q)$
end while
return $\mathbf{A}, Q$

MSMS COMMANDS USED

In order to simulate $K > 1$ independent, randomly mating populations, we used the `msms` coalescent simulator (EWING and HERMISSON, 2010). We simulate K independent demes (populations) with no migration between them over a period of 10,000 generations. Each deme is represented by 100 simulated individuals. After 10,000 generations, all K demes are merged and the coalescent process is allowed to terminate. We simulate a fixed number of segregating sites (SNPs in our case) in each case. Specifically, for $K = 3$ and $13,626$ segregating sites, we used the following `msms` command:

```
msms.jar 300 1 -s 13626 -N 1000 -I 3 100 100 100 -ej 2.5 1 2 -ej
2.5 2 3
```

By modifying the -I flag and adding more -ej flags, this command can be used to simulate an arbitrary number of independent populations. The user is referred to the `msms` manual for more details.

LITERATURE CITED

EWING, G., and J. HERMISSON, 2010 MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics **26**: 2064–2065.

FRICHO, E., F. MATHIEU, T. TROUILLON, G. BOUCHARD, and O. FRANOIS, 2014 Fast inference of admixture coefficients using sparse non-negative matrix factorization algorithms. Genetics Early online access, 10.1534/genetics.113.160572.

POPESCU, A.-A., L. A. HARPER, M. TRICK, I. BANCROFT, and T. K. HUBER, 2014 A Novel and Fast Approach for Population Structure Inference using Kernel-PCA and optimisation (PSIKO) .