PSIKO MANUAL

1. Introduction

Inferring individual admixture coefficients of an admixed population is a problem encountered in many whole genome association studies, as it allows one to, for example, correct for population stratification effects in an association analysis. Traditionally this problem has been solved using computationally intensive model based approaches such as ADMIXTURE (Alexander *et al.*, 2009) or STRUCTURE (Pritchard *et al.*, 2000), but recently non-model based approaches such as sNMF (Fricho *et al.*, 2014) have made solving this problem feasible even on large NGS data. PSIKO is such an approach, building on noted properties of PCA to quickly and accurately estimate the sought admixture coefficients coming from a number K of founders, forming what is known as a Q-matrix. PSIKO2 is a fully compatible extension of PSIKO that allows for local ancestry inference and also usage of PSIKO within a Mac environment. PSIKO is released under GPL and comes with NO WARRANTY. If you use PSIKO, please cite (Popescu *et al.*, 2014). If you use the extension to carry out local ancestry inference, please cite (Popescu and Huber, subm).

2. Obtaining PSIKO and PSIKO2

PSIKO2 is freely available at https://www.uea.ac.uk/computing/psiko in source code or compiled binary form. A user can compile this code to obtain an executable which can be called from the command line. PSIKO requires the `armadillo` (Sanderson, 2010), `blas` (Lawson *et al.*, 1979) and `lapack` (Anderson *et al.*, 1999) libraries. To compile PSIKO, the user can run the following command after obtaining all dependencies.

```
g++ -o PSIKO main.cpp -O3 -larmadillo -llapack -lblas -lgfortran -lpthread
-Xlinker -zmuldefs -mpopcnt
```

For the convenience of the user, a precompiled statically linked binary is also provided at the aforementioned URL. It can be run straight from the command line without having to install any dependencies. It has been linked with the most up-to-date versions of the dependencies, and uses `open-blas` (Wang *et al.*, 2013) instead of `blas` for gains in speed. This binary file should work on linux kernel versions 2.6.32 or later. In the form of PSIKO2 there is also a pre-compiled version of PSIKO for Mac, which should run on systems having the Accelerate framework installed.

3. Program options

3.1. **Population Structure.** PSIKO takes as input a file in .geno (Price *et al.*, 2006) format and produces the following output files.

- a file named reduced_ data.csv, which contains the PCA reduced data.
- a file means.csv which contains the inferred positions of the putative founders.

Date: May 19, 2015.

- a file which contains the inferred admixture coefficients. By default this file is the input file name with the extension .PSI.Q (see Section 4 for an example).

PSIKO takes the following command line options:

- `-i`: required, the name of the input file, in .geno format
- `-K`: optional, the value of K. The default setting is the Tracy-Widom statistic as described in (Patterson *et al.*, 2006).
- `-q`: optional, the name of the file where the ancestry coefficients are stored. See above for default value

PSIKO2 additionally takes the following command line options:

- `-a`: the name of the file where local ancestry estimates are to be output
- `-f`: optional, the name of the file containing founder information (see below)

3.2. PSIKO2**: Local Ancestry Inference.** PSIKO2 allows one to estimate the founder of each locus (i. e. SNPs in our case) using a sliding window approach (Popescu and Huber, subm). To do this, the user is required to provide, in addition to a .geno input file, an ancestry output file with the -a option.

Estimates of local ancestry are output to the file specified by the -a option, with one row per individual. Each row has, for each SNP a number $0 \leq k < K$, (either provided as input or obtained via the Tracy-Widom statistic) indicating the founder of that SNPs.

When using the -a option, the user can also optionally provide founder information. For this the -f option is used. This will tell PSIKO2 to look in the file specified by this option for information on founder individuals. The format of the founder file is as follows. Each line contains two numbers, a and b. This means that the a-th individual in the input .geno file is a non-admixed individual with founder b. In the absence of the -f option, the Q-matrix is used to find proxies for founders.

4. Example Run

4.1. **Population Stratification.** PSIKO come with an example file included. This is an oilseed rape dataset from (Harper *et al.*, 2012). To run PSIKO with the included example file, run the following:

```
./ PSIKO -i OSRMatrix_ Complete.txt.geno -K 2
```

This will produce the reduced_ data.csv and the means.csv files as above as well as the OSRMatrix_ Complete.txt.geno.PSI.Q ancestry coefficients file in the current working directory. It also displays some run-time information, such as the part of the algorithm currently being executed.

4.2. **Local ancestry inference.** To obtain local ancestry estimates from PSIKO2, use the -a option. If information about non-admixed individuals is available, the user may provided it using the -f option (see above).

```
./ PSIKO -i OSRMatrix_ Complete.txt.geno -K 2 -a ancestryFileName.txt
-f founderFileName.txt
```

5. Contact

For help with PSIKO or to report issues, please send an email to Andrei-Alin.Popescu@uea.ac.uk or K.Huber@uea.ac.uk.

REFERENCES

Alexander, D. H., Novembre, J., and Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, **19**(9), 1655–1664.

Anderson, E., Bai, Z., Bischof, C., Blackford, S., Demmel, J., Dongarra, J., Du Croz, J., Greenbaum, A., Hammarling, S., McKenney, A., and Sorensen, D. (1999). *LAPACK Users' Guide*. Society for Industrial and Applied Mathematics, Philadelphia, PA, third edition.

Fricho, E., Mathieu, F., Trouillon, T., Bouchard, G., and Franois, O. (2014). Fast inference of admixture coefficients using sparse non-negative matrix factorization algorithms. *Genetics*, **196**(4), 973–983.

Harper, A. L., Trick, M., Higgins, J., Fraser, F., Clissold, L., Wells, R., Hattori, C., Werner, P., and Bancroft, I. (2012). Associative transcriptomics of traits in the polyploid crop species Brassica napus. *Nature Biotechnology*, **30**(8), 798–802.

Lawson, C. L., Hanson, R. J., Kincaid, D. R., and Krogh, F. T. (1979). Basic linear algebra subprograms for fortran usage. *ACM Trans. Math. Softw.*, **5**(3), 308–323.

Patterson, N., Price, A. L., and Reich, D. (2006). Population structure and Eigenanalysis. *PLoS Genet*, **2**(12).

Popescu, A.-A. and Huber, K. T. (subm). Psiko2: a fast and versatile tool to infer population stratification on various levels in gwas. *Bioinformatics*.

Popescu, A.-A., Harper, A. L., Trick, M., Bancroft, I., and Huber, K. T. (2014). A novel and fast approach for population structure inference using kernel-pca and optimisation (PSIKO). *Genetics*, **198**(4), 1421–1431.

Price, A. L., Patterson, N. J., Plenge, R. M., Michael, E. W., Shadick, N. A., and Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, **38**(8), 904–909.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, **155**(2), 945–959.

Sanderson, C. (2010). Armadillo: An Open Source C++ Linear Algebra Library for Fast Prototyping and Computationally Intensive Experiments. Technical report, NICTA.

Wang, Q., Zhang, X., Zhang, Y., and Yi, Q. (2013). Augem: Automatically generate high performance dense linear algebra kernels on x86 cpus. In *Proceedings of SC13: International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '13, pages 25:1–25:12, New York, NY, USA. ACM.