PopCluster: A population genetics model-based toolset for simulating, inferring and visualizing individual admixture and population structure

Jinliang Wang

Institute of Zoology, Zoological Society of London, London NW1 4RY, United Kingdom

Left running head: J Wang

Right running head: Population admixture analysis

Key words: admixture, population structure, markers, hybridization, clustering analysis

Corresponding author:

Jinliang Wang

Institute of Zoology

Regent's Park

London NW1 4RY

United Kingdom

Tel: 0044 20 74496620

Fax: 0044 20 75862870

Email: jinliang.wang@ioz.ac.uk

Abstract

In this computer note I introduce software, PopCluster, that implements a new likelihood method for unsupervised population structure analysis from marker data. To infer a coarse population structure, it assumes the mixture model and adopts a simulated annealing algorithm to make a maximum likelihood clustering analysis, partitioning the sampled individuals into a predefined number of clusters. To deduce a fine population structure, it further assumes the admixture model and employs an expectation maximization algorithm to estimate individual admixture proportions. PopCluster has many features. First, it is one of just a couple of model-based methods that can handle both biallelic and multiallelic markers in the same framework. Second, it is the first population structure analysis method that uses both Message Passing Interface (MPI) and openMP to exploit multiple CPUs with both shared and distributed memories and has the capacity to handle genomic data with millions of individuals and millions of loci. Third, the algorithms for both mixture and admixture analyses are fast, rendering PopCluster favourably in computational efficiency over previous methods in analysing genomic data. Fourth, PopCluster is built for Windows, Linux and Mac platforms, and its Windows version has an integrated GUI that can conveniently visualize analysis results and facilitate data input. Fifth, its Windows version has a built-in simulation module designed to simulate genotype data under admixture, hybridization or migration models. PopCluster provides a valuable toolset for researchers to simulate, infer and visualize individual admixture and population genetic structure, hybridization and migration using marker data.

1. INTRODUCTION

Given a predefined subdivision of a population, it is straightforward to measure, by Wright's F_{ST} (Wright, 1931), the differentiation between subpopulations from the marker data of a sample of individuals drawn from each subpopulation (Nei, 1973; Weir & Cockerham, 1984). Frequently, however, we may have insufficient information to presume any population subdivision. Furthermore, individuals sampled from the same location (e.g. the same breeding or feeding ground) may come from multiple unknown but well differentiated source populations as is the case in mixed stock analysis (e.g. Smouse et al., 1990). In such circumstances, it becomes challenging to study population structure purely from the marker information of a sample of individuals. Since the seminal work by Pritchard et al. (2000), several Bayesian and likelihood methods (e.g. Dawson & Belkhir, 2001; Corander et al., 2003; Alexander et al., 2009; Frichot et al., 2014) have been developed to make unsupervised analysis of population structure from marker data. These methods can be used to learn the population structure solely from the sampled genotype data, inferring the most likely number of populations represented by the sampled individuals, partitioning the individuals into populations, and estimating the ancestry (admixture proportions) of an individual in the inferred source populations (Pritchard et al., 2000).

The Bayesian method of Pritchard et al. (2000), with improved models on linked loci and correlated allele frequencies (e.g. Falush et al. 2003, 2007) and on the use of sample group information (Hubisz et al., 2009), is proved to be highly popular because of its high accuracy, flexibility (e.g. the ability to handle recessive markers and null alleles), robustness (e.g. to unbalanced sampling, Wang, 2017), and user-friendly interface (e.g. with GUI for data input and results visualization). However, the method, implemented in software STRUCTURE, is only applicable to the analysis of small datasets with a small number of markers and individuals. It becomes computationally infeasible to apply to a large sample of genomic markers or individuals, especially considering the needs to make multiple replicate runs at the same assumed number of populations and at a range of different assumed numbers of populations. To handle large genomic SNP data, computationally efficient likelihood methods (e.g. Tang, 2005; Alexander et al., 2009; Frichot et al., 2014) and new methods using variational Bayesian inference (e.g. Raj et al., 2014; Gopalan et al., 2016) have been developed. The software ADMIXTURE (Alexander et al., 2009) that implements the likelihood method of Tang (2005) with a fast algorithm gains popularity in analysing genomic SNP data.

More recently, I proposed a new likelihood method that makes fast and accurate population structure analysis from a sample of individuals genotyped at a few multiallelic markers (such as microsatellites) or millions of biallelic markers (such as SNPs) (Wang, 2022). The method is shown to be more accurate than previous ones when many populations are assumed, when populations are little differentiated with intricate structures, or when samples drawn from the source populations are small or highly unbalanced in size (Wang, 2022). It is also more computationally efficient in handling large genomic data than previous methods because of its use of fast algorithms and the exploitation of multiple parallel computation protocols (e.g. openMP and MPI). In this computer note, I introduce the software, PopCluster, which implements the new method with a focus on describing its features. More details on how to use the software are described in the user manual included in the downloaded PopCluster package.

2. FEATURES OF THE SOFTWARE

2.1 Applicable to biallelic and multiallelic markers

Pritchard *et al*. (2000) developed the first Bayesian method for an unsupervised population structure analysis from markers with any number of alleles at a locus. Many new likelihood (e.g. Alexander et al., 2009) or Bayesian (e.g. Raj et al., 2014) methods have been developed since then to handle genomic marker data with much improved computational efficiency. Unfortunately, however, these new methods invariably assume biallelic codominant markers, and cannot be applied to microsatellites and other markers with 3 or more alleles at a locus.

Like STRUCTURE, PopCluster implemented a method that can handle markers with any number of codominant alleles, including biallelic SNPs and multiallelic microsatellites. The model and algorithms are invariable with the number of alleles per locus. However, to handle biallelic markers such as SNPs efficiently, a 2-bit encoding system is applied where each genotype is encoded by 2 bits such that a four-byte integer can be used to store 16 genotypes. This is important nowadays to save computer memory in dealing with genomic marker data where millions of SNPs can be genotyped for thousands or even millions of individuals. The human 1000 genomes phase I dataset (Abecasis et al., 2012), for example, has 1092 individuals sampled from 14 different populations across all continents, with each individual having 38 million SNP genotypes. The genotype data alone would take 83GB RAM when the two alleles of a genotype are coded by two four-byte integers. Using the 2-bit

encoding system, the same data take only 5.2GB RAM, small enough to be dealt with by a laptop computer.

2.2 Applicable to unbalanced sampling

All model-based population structure inference methods developed so far, except for those implemented in STRUCTURE and PopCluster, assume *a priori* that a sampled individual comes from each of K source populations at an equal probability of $1/K$. The assumption essentially dictates that an equal number of individuals are sampled from each source population. Realistically, however, the numbers of individuals drawn from different source populations are rarely equal and may differ substantially quite often, especially when sampling of individuals is made with little knowledge of the geographic distributions of the source populations. A good example is sampling individuals from a common breeding or feeding ground to identify the hidden structure of populations utilizing the ground (Smouse et al., 1990). In such a case, a big or adjacent population may have many individuals and a small or distant population may have few individuals included in a sample drawn randomly from the ground. Even in the most favourable situation of equal population sizes and equal distances from the ground, samples from different source populations may still differ substantially because of behaviour difference or uneven distributions of populations in the ground, or because of sampling bias.

An individual from an unbalanced sample (i.e. a sample containing many individuals from one population and few individuals from another population) comes from an over-represented (under-represented) population at a higher (lower) probability than the equal probability of $1/K$. This violates the assumption of equal prior ancestry adopted by the likelihood or Bayesian admixture analysis methods, causing an over-represented population being split and the under-represented populations being merged in the reconstruction of population structure as demonstrated using STRUCTURE with simulated data (Puechmaille, 2016; Wang, 2017). Unfortunately, this problem cannot be solved by increasing marker information, and it persists even when genomic data with millions of SNPs are used in the analysis. STRUCTURE has a built-in alternative ancestry model which assumes that a sampled individual comes from different source populations at different probabilities and is used to estimate these probabilities jointly with admixture proportions and population specific allele frequencies. The model was shown to provide accurate population structure inferences when sampling is unbalanced (Wang, 2017).

PopCluster tackles the unbalanced sampling problem by introducing a scaling scheme to minimize the merging of small populations and the split of large populations in clustering analysis (Wang, 2022). In brief, PopCluster partitions a sample of N individuals into K clusters. The original likelihood of cluster k (=1, 2, …, K), $L_k(\mathbf{\Omega}_k)$, is first calculated from the genotype data. It is then scaled by the cluster size (i.e. number of individuals belonging to cluster k), N_k, as $\mathcal{L}_k(\mathbf{\Omega}_k) = L_k(\mathbf{\Omega}_k)/(1 + e^{sN_k/(8N)})$, where s is the scaling factor taking values 1, 2, 3 and 4 for weak, medium, strong and very strong scaling, respectively. $\mathcal{L}_k(\mathbf{\Omega}_k)$ rather than $L_k(\mathbf{\Omega}_k)$ is used to assess the plausibility of a clustering configuration.

Applying this scaling, PopCluster can yield accurate admixture estimates of a highly unbalanced sample. An example is shown in Figure 1. As expected, STRUCTURE produced much better results with the alternative (unequal) ancestry prior than with the default (equal) ancestry prior. However, even the alternative ancestry prior does not preclude the merging of lightly-sampled populations and the splitting of heavily-sampled populations. PopCluster produced rather messy results when no scaling is applied, but highly accurate results when scaling is used. The other methods implemented in ADMIXTURE (Alexander et al., 2009), sNMF (Frichot et al., 2014) and SCOPE (Chiu et al., 2022) lack the sophisticated models of STRUCTURE and PopCluster to deal with unbalanced sampling. They all generated inaccurate admixture estimates for this example dataset, which is quite informative with a set of 10000 SNPs. Even when millions of SNPs are used, these methods may still fail to uncover the unbalanced population structure (Appendix 1).

The impact of unbalanced sampling on a population structure analysis depends on both the sizes and their relative differences (i.e. unbalance) of the samples drawn from different source populations, on the differentiation of the source populations, and on the marker information. Example 1 might provide an extreme yet realistic case where most samples are small (i.e. only 5 individuals), the samples are highly unbalanced (the largest sample has 50 individuals), and the F_{ST} is small (0.02) relative to the small and unbalanced samples. The performance differences among the methods become less striking with larger sample sizes, smaller differences in sample sizes from different source populations, larger F_{ST} among populations, and more markers.

In practice, most of the above-mentioned factors (e.g. the extent of unbalance in sampling intensity) are unknown. As a result, the strength of scaling (note, no scaling corresponds to s=0) most appropriate to a particular dataset is difficult to determine a priori.

When some non-genetic information such as sampling locations is available to judge whether the marker-based admixture estimates make sense or not, however, some experimentation with different strengths of scaling will be rather helpful. When the applied level of scaling (s=0,1,2,3,4) is too low, large populations tend to be split and small populations tend to be merged. When the applied level of scaling is too high, small populations tend to be merged among themselves or with a large population. With the help of some external information such as sampling locations, the appropriate scaling level can be determined from the admixture assignment analysis. Additionally, it is helpful to conduct and visualize a principal components analysis (PCA) of the data, as a means to assess a priori whether imbalance exists or not among the clusters and whether scaling should be applied to a PopCluster analysis or not.

2.3 High capacity to handle genomic data

The power of a population structure analysis using any method, either a non-model based method such as PCA or a model based method such as STRUCTURE, depends on the data size LN, where L is the number of SNPs and N is the number of sampled individuals (Patterson et al., 2006). The structure of a sample containing $N/2$ individuals from each of two populations can be reconstructed when the differentiation between the two populations, measured by F_{ST}, is larger than $1/\sqrt{LN}$ (Patterson et al., 2006). This rule means that more data (i.e. higher LN) permits the elucidation of more subtle structure (i.e. lower F_{ST}). The fine genetic structure of populations with weak differentiation (e.g. $F_{ST} < 0.01$) can still be detected and delineated by an analysis of marker data, given the data are sufficiently informative (i.e. LN large). For example, if N=L=10000, it is theoretically possible to reconstruct the structure of two populations differentiated as little as F_{ST} =0.0001, which is achieved by only one generation of separation (drift) of populations with an N_e as large as 10000 (Wright, 1965).

With the rapid development of DNA sequencing technology, larger and larger datasets are increasingly acquired, paving the way for studying fine-scale genetic structures at an unprecedented resolution. The UK's biobank data, for example, contain genotypes of about half a million British individuals across millions of SNPs (Bycroft et al., 2018). Datasets in this magnitude of size pose serious challenges for population structure analysis methods. One challenge is that the memory required to handle such a big dataset might be beyond the capacity of most computers. The genotype matrix of biobank data in PLINK's

most space-saving binary format (.bed file, Purcell et al., 2007) alone requires around 70 GB of storage. With some additional necessary working space, the total RAM requirement can easily go over 140GB which is well above the RAM capacity of most computers. The second challenge is computational time, which might be too long for datasets as large as the UK's biobank data.

PopCluster adopts an efficient encoding system in which two bits are used to store a biallelic SNP genotype. Therefore, a four-byte integer can be used to encode 16 genotypes. This data format is similar to that of PLINK's .bed file. Adopting this efficient data encoding system coupled with the exploitation of distributed memory across nodes of a computer cluster, PopCluster's capacity to handle large genomic data is only limited by the capacity of a computer.

PopCluster also adopts fast algorithms for maximum likelihood clustering analysis and admixture analysis. Assuming the mixture model (i.e. no admixture) of Pritchard et al., (2000) in clustering analysis, it updates only a small fraction of the variables (e.g. allele frequencies) because a cluster reconfiguration usually involves the membership changes of only one individual in an iteration. The computational efficiency advantage of PopCluster over other programs increases rapidly with an increasing *K* value (the assumed number of populations) in an analysis. For a simulated dataset consisting of 640 individuals genotyped at 10000 SNP loci, the running time (without parallelization) of PopCluster and ADMIXTURE is compared in Figure 2. While the running time of PopCluster increases linearly with *K*, the running time of ADMIXTURE increases log linearly with *K*. As a result, PopCluster runs many times faster than ADMIXTURE when *K* becomes large.

2.4 MPI and openMP parallel computation

PopCluster is the first admixture analysis program that employs MPI (Message Passing Interface) to use an arbitrary number of computer nodes for both storage of data and computation. When instructed to use *M* MPI processes, PopCluster will divide the RAM requirement for data storage and working space into *M* equal slices and each MPI process (running on a single node) stores and addresses data located in only one slice. This MPI capability essentially removes the memory constraint, as it can use the memory of all nodes in a computer cluster. PopCluster's MPI capability enables it to use as many nodes and cores as a computer has to speed up the computation for analysing large genomic data. Further speeding up is realized in analysing a large dataset by reducing cache misses, because only a

slice (a fraction of 1/M) of data and working space (such as the allele counts at each locus of each of K populations) needs to be addressed by any MPI process.

Like some other programs such as ADMIXTURE (Alexander et al., 2009), PopCluster can also employ openMP alone or in combination with MPI in parallel computation. OpenMP can exploit the hyperthreading technology to a physical core of some types of CPUs and different physical cores in a single node. It cannot use multiple nodes with distributed memory. Because all openMP processes share the same memory and address the same data and working space, cache misses can be a serious problem in slowing down the computation with a large dataset.

Armed with the 2-bit encoding system, fast algorithms, and multiple parallel computational methods (MPI and openMP), PopCluster runs faster than the popular model-based admixture analysis program, ADMIXTURE, in analysing large genomic data, especially when a parallel run is conducted with many threads. Table 1 compares PopCluster with ADMIXTURE in the computational time for analysing a large simulated dataset of 500 individuals and 1000000 SNPs, assuming K=10. For this dataset, PopCluster runs much faster than ADMIXTURE, especially when many parallel threads are used in the analyses. The computational time for PopCluster is from 45.1% to 3.5% of that for ADMIXTURE when the number of parallel threads used by both programs increases from 1 to 32. ADMIXTURE runs faster with the use of an increasing number of openMP parallel threads, n, until n reaches a small value of 4. It actually slows down the computation with an increasing n when $n > 4$. PopCluster always runs faster with an increasing value of n, no matter it uses MPI or openMP for parallel computation. Relatively, MPI is more efficient than openMP, although only a single node with shared memory is used in analysing this example dataset. Part of the reason is that MPI partitions the data and working space into equal-sized slices and each MPI process just needs to store and address one slice, reducing the risks of cache misses.

ADMIXTURE fails to run much larger datasets with many millions of SNPs. For example, I generated a simulated dataset with 100 individuals and 50 million SNPs. ADMIXTURE aborts after a few rounds of iterations conducted on the computer cluster (as used above for the smaller dataset) despite the allocation of a large RAM (50 GB). In contrast, PopCluster can analyse this dataset on the same cluster or on a PC. When many loci are available, we can trim the data by choosing a subset of the most ancestry informative markers (e.g. Wilkinson et al. 2011) for structure analysis. However, this is suitable only

when the original L markers provide more information than necessary for population structure and individual admixture analysis (i.e. when $F_{ST} \gg 1/\sqrt{LN}$ where the differentiation among populations is F_{ST} and the number of sampled individuals is N, Patterson et al., 2006). In the otherwise situations with weak structure or very small sample sizes of individuals, every marker contributes information to the analysis and we cannot afford giving up any marker data.

Like ADMIXTURE (Alexander et al., 2009), sNMF (Frichot et al., 2014) and other model-based methods proposed for population structure analysis, PopCluster does not model linkage disequilibrium (LD) and can be applied to markers no matter they are linked or not and no matter they are in linkage equilibrium or not. This means that the genomic data do not have to be LD-pruned before they can be analysed for population structure. LD-pruning removes some markers and thus reduces the information for use in individual admixture and population structure analysis. Wang (2022) applied PopCluster to the analysis of the human 1000 genomes phase I dataset (Abecasis et al., 2012) with 38 million SNPs without any data filtering (based on, for example, LD, data missingness and marker MAF), yielding meaningful results of the world-wide human population structure. Without data filtering, as much marker information as possible can be used to delineate population structure and the analysis is simplified with little arbitrariness in data selection and elimination.

2.5 Multi-platforms and GUI

PopCluster is written in Fortran and is compiled to produce executables runnable on Windows, Mac and Linux platforms. The Windows version has additionally a GUI written in VB.net, which facilitates the input of parameters and data and the visualization of population structure and individual admixture analysis results. For comparison purposes, the GUI can also visualize the admixture analysis results from other programs including STRUCTURE (Pritchard et al., 20000), ADMIXTURE (Alexander et al., 2009) and sNMF (Frichot et al., 2014). All bar charts in Figure 1 were plotted by PopCluster's GUI, where only the results from SCOPE (Chiu et al., 2022) were reformatted before being plotted by PopCluster's GUI.

Like STRUCTURE, PopCluster requires a data file and a parameter file containing the values of quite a few parameters such as the assumed K values, the number of replicate runs per K value, random number seeds, etc. A user-friendly GUI, shown by Figure 3, helps in preparing the two input files for a PopCluster analysis.

In some situations, it is more convenient to use multiple platforms for a PopCluster analysis. We can employ PopCluster's GUI to create the two input files on a PC running Windows, copy the two files to a Linux cluster to run PopCluster in parallel using both MPI and openMP, and then copy the analysis results back to a PC for visualization.

PopCluster's GUI provides additional functionalities such as data conversion. For example, it can convert SNP data in a VCF file to a given format accepted by PopCluster and save the reformatted data in a file. The numbers of loci and individuals of the data, the information required in setting up a PopCluster project, are also extracted from the VCF data and saved to a file.

2.6 Simulation module

PopCluster's Windows version has an integrated simulation module which allows a user to simulate genotype data under the admixture model, the hybridization model and the migration model. For each model, the module accepts a few parameter values (e.g. number of loci, number of simulated populations, number of sampled individuals per population), simulates the genotype data using these parameters by assuming unlinked markers in linkage equilibrium, and then outputs the simulated data in a user-defined format into a data file. Additionally, it outputs the true (simulated) population structure such as individual admixture proportions so that the inference from the data by any program can be assessed for accuracy against the truth. The simulated data can be analysed directly by PopCluster and STRUCTURE. With reformatting by Plink (Purcell et al., 2007), the data can also be analysed by programs such as ADMIXTURE and sNMF.

For the admixture model, the module can simulate a sample of individuals under the spatial admixture model (see Figure 4 as an example) or under the non-spatial admixture model. For the latter, the module can simulate populations with a hierarchical or non-hierarchical subdivision structure. For the hybridization model, the module simulates genotype data of individuals from various hybrid classes such as F1, F2, B1, B2 involving parents and grandparents from two, three or four source populations. For the migration model, the module makes a forward simulation of a set of populations with user-defined effective sizes and migration rates over generations until quasi-equilibrium is reached when a sample of individuals is taken from each population for population structure analysis, including estimating migration rates and effective population sizes.

The simulation module is particularly useful for generating replicated datasets to investigate factors affecting the power and accuracy of a marker-based population admixture analysis. Simulations are also valuable for optimizing the experimental design of a population structure study. It is helpful, for example, to determine the suitable sampling intensities (number of markers, number of individuals from each location) to yield accurate population structure inference. Before initiating an experiment, one can use simulations to generate replicated data in conditions similar to those of the conceived study, and to analyse the simulated data to get a feel of the estimation power and accuracy. For this same reason, simulations are also valuable for training and educational purposes.

3. CONCLUSION

PopCluster is a powerful software implementing an efficient population clustering method based on the mixture model and the simulated annealing algorithm, and an efficient individual admixture analysis method based on the admixture model and the expectation maximization algorithm. These fast algorithms coupled with the efficient encoding of genotype data (i.e. using 2 bits to store a SNP genotype) and the parallel computation capability using both openMP and MPI make PopCluster one of the fastest methods for inferring population structure, individual admixture, hybridization and migration from large genomic marker data. Its high capacity to handle a large volume of data (e.g. genomic SNPs of many individuals) makes it possible to analyse biobank data for understanding population structure at an unprecedented resolution. PopCluster is implemented for multiple platforms, and its Windows version has an integrated GUI to facilitate data and parameter input and to visualize the analysis results in publication-quality graphs. Furthermore, the GUI has a simulation module which allows a user to simulate replicated genotype data under several optional admixture models, the hybridization model and the migration model. It could become a valuable tool for the research and teaching in the fields of conservation genetics, evolutionary genetics, and molecular ecology.

PopCluster can be downloaded freely from https://www.zsl.org/about-zsl/resources/software/popcluster. The release has separate packages downloadable independently for the Windows, Mac and Linux platforms. Each package contains the executables, user's manual, and an example dataset.

CONFLICTS OF INTEREST

The author has no conflicts of interest to declare.

AUTHORS' CONTRIBUTIONS

J Wang conceived and conducted the study and wrote the manuscript.

DATA AVAILABILITY STATEMENT AND BENEFIT-SHARING

No new empirical data were generated in this study. The simulated data can be reproduced using the software package (Windows version) available on Zoological Society of London at https://www.zsl.org/about-zsl/resources/software/popcluster.

REFERENCES

Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T., & McVean, G. A. (2012). An integrated map of genetic variation from 1092 human genomes. *Nature, 491(7422)*, 56–65.

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

Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L. T., Sharp, K., Motyer, A., Vukcevic, D., Delaneau, O., O'Connell, J., Cortes, A., Welsh, S., Young, A., Effingham, M., McVean, G., Leslie, S., Allen, N., Donnelly, P., Marchini, J. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature*, *562(7726),* 203–209.

Chiu, A. M., Molloy, E. K., Tan, Z., Talwalkar, A., & Sankararaman, S. (2022). Inferring population structure in biobank-scale genomic data. *The American Journal of Human Genetics*, *109(4),* 727-737.

Corander, J., Waldmann, P., & Sillanpää, M. J. (2003). Bayesian analysis of genetic differentiation between populations. *Genetics*, *163(1),* 367–374.

Dawson, K., & Belkhir, K. (2001). A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genetics Research*, *78(1),* 59–77.

Falush, D., Stephens, M., & Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, *164(4),* 1567-1587.

Falush, D., Stephens, M., & Pritchard, J. K. (2007). Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular Ecology Notes*, *7(4),* 574-578.

Frichot, E., Mathieu, F., Trouillon, T., Bouchard, G., & François, O. (2014). Fast and efficient estimation of individual ancestry coefficients. *Genetics*, *196(4),* 973-983.

Gopalan, P., Hao, W., Blei, D. M., Storey, J. D. (2016). Scaling probabilistic models of genetic variation to millions of humans. *Nature Genetics*, *48(12),* 1587.

Hubisz, M.J., Falush, D., Stephens, M., & Pritchard, J. K. (2009). Inferring weak population structure with the assistance of sample group information. *Molecular Ecology Resources*, *9(5),* 1322-1332.

Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences of the United States of America*, *70(12),* 3321–3323.

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, *2(12),* e190.

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

Puechmaille, S. J. (2016). The program STRUCTURE does not reliably recover the correct population structure when sampling is uneven: sub-sampling and new estimators alleviate the problem. *Molecular Ecology Resources*, *16(3),* 608–627.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., & Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American journal of Human Genetics, 81(3),* 559-575.

Raj, A., Stephens, M., & Pritchard, J. K. (2014). fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics*, *197(2),* 573-589.

Smouse, P. E., Waples, R. S., & Tworek, J. A. (1990). A genetic mixture analysis for use with incomplete source population data. *Canadian Journal of Fisheries and Aquatic Sciences*, *47(3),* 620-34.

Tang, H., Peng, J., Wang, P., & Risch, N. J. (2005). Estimation of individual admixture: analytical and study design considerations. *Genetic Epidemiology*, *28(4),* 289-301.

Wang, J. (2017). The computer program structure for assigning individuals to populations: easy to use but easier to misuse. *Molecular Ecology Resources*, *17(5),* 981-990.

Wang, J. (2022). Fast and accurate population admixture inference from genotype data from a few microsatellites to millions of SNPs. *Heredity*, *129(2),* 79-92.

Weir, B. S., & Cockerham, C. (1984). Estimating *F*-statistics for the analysis of population structure. *Evolution*, *38(6),* 1358–1370.

Wilkinson, S., Wiener, P., Archibald, A. L., Law, A., Schnabel, R. D., McKay, S. D., Taylor, J. F., & Ogden, R. (2011). Evaluation of approaches for identifying population informative markers from high density SNP chips. *BMC Genetics*, *12,* 1-14.

Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, *16(2),* 97-159.

Wright, S. (1965). The interpretation of population structure by F-statistics with special regard to systems of mating. *Evolution*, *19(3)*, 395-420.

How to cite this article: Wang, J. (2024). PopCluster: A population genetics model-based toolset for simulating, inferring and visualizing individual admixture and population structure. Molecular Ecology Resources, ####. https://doi.org/####

TABLE 1 Computational time taken by PopCluster and ADMIXTURE for analysing a simulated dataset

Number of Parallel threads	**Running time (in minutes) by program (parallel protocol)**		
	ADMIXTURE	**PopCluster (MPI)**	**PopCluster (openMP)**
1	397	179	179
2	208	105	125
4	**162**	54	92
8	216	30	61
16	253	17	57
32	228	**8**	**46**

The simulated dataset consists of 500 individuals, with 50 from each of K=10 populations. Each individual is genotyped at 1 million SNP loci. The analyses (assuming K=10) were run on one Linux cluster node which has an Intel(R) Xeon(R) Gold 6140 CPU (@ 2.30GHz) with 36 cores. The RAM requested was 1GB while the maximum RAM actually used was 0.5GB in analysing this dataset. Running time is in minutes. For each program using each number of parallel threads, 3 independent replicate runs (with different random number seeds) were made and the average running time is reported.

FIGURE LEGENDS

FIGURE 1 The population structure of a simulated sample of 75 individuals. The true (simulated) structure is shown in the 1st row, with 50 individuals (1 to 50 on the *x* axis) coming from population 1, 5 individuals (51 to 55) from population 2, 5 individuals (56 to 60) from population 3, 5 individuals (61 to 65) from population 4, 5 individuals (66 to 70) from population 5, and 5 individuals (71 to 75) from population 6. The F_{ST} of each population is assumed 0.02, and each sampled individual is genotyped at 10000 SNPs. The structure was inferred from the genotype data by PopCluster with no scaling (scaling parameter $s = 0$, shown on the 2ed row) and with scaling (scaling parameter $s = 1$, shown on the 3rd row), by STRUCTURE with the default equal ancestry prior (shown on the 4th row) and with the alternative unequal ancestry prior (shown on the 5th row), by ADMIXTURE (shown on the 6th row), by sNMF (shown on the 7th row) and by SCOPE (shown on the 8th row).

FIGURE 2 A comparison of the running times of PopCluster and Admixture in analyzing a simulated dataset at different *K* values. The dataset consists of 640 individuals genotyped at 10000 SNP loci. For each program and each *K* value, 10 replicate runs with different random number seeds were conducted and the mean running time (in minutes) is presented. All runs were conducted on a Linux cluster (with an Intel(R) Xeon(R) Gold 6140 CPU @ 2.30GHz) using a single thread (i.e. no parallelization).

FIGURE 3 The PopCluster Windows GUI for setting up a new project

FIGURE 4 An example population structure generated by PopCluster's simulation module under the spatial admixture model. The sample contains N=500 individuals in the spatial admixture model with K=5 source populations whose ancestries are denoted in pink, blue, red, green and light blue colours.

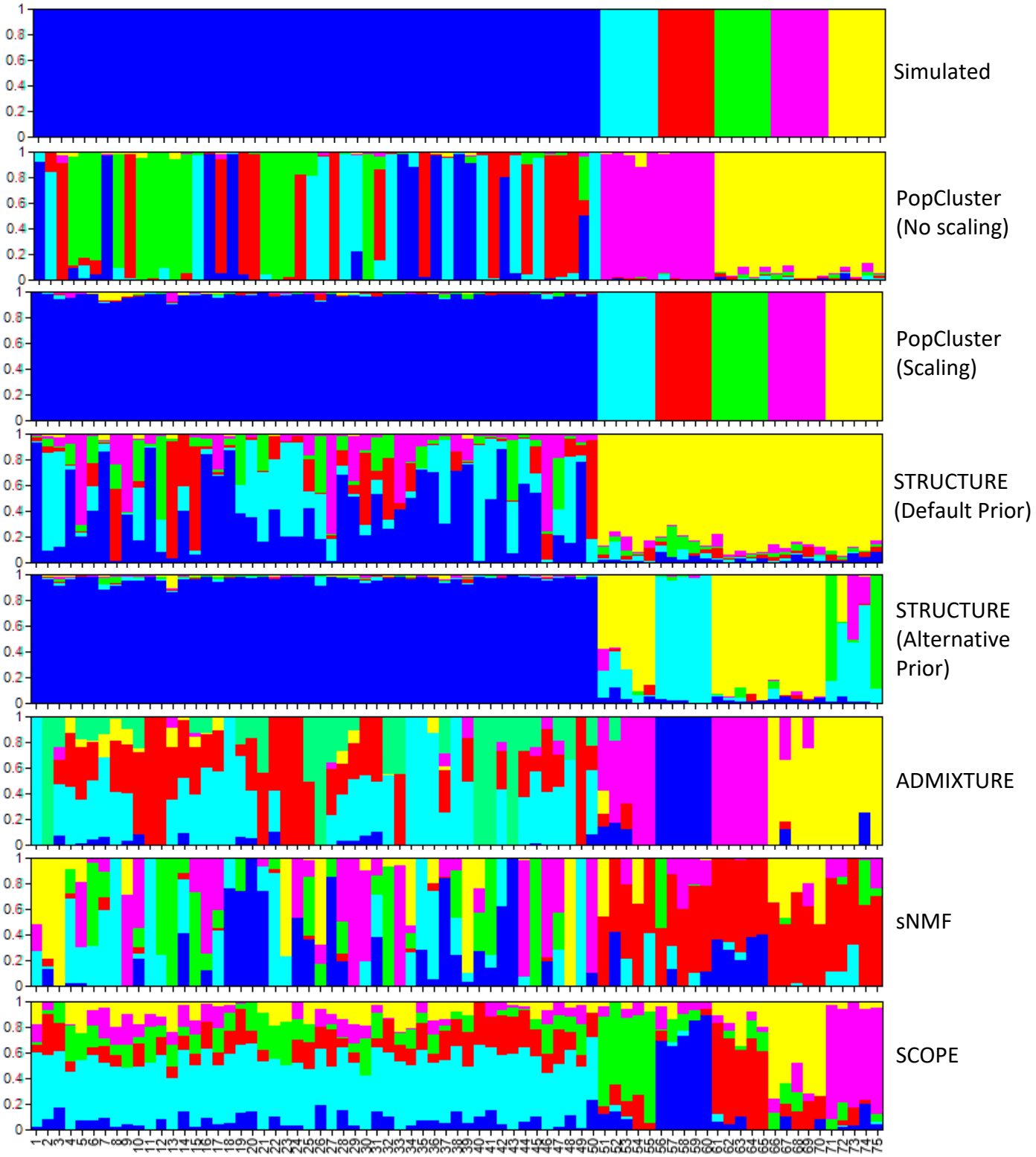
Simulated
PopCluster (No scaling)
PopCluster (Scaling)
STRUCTURE (Default Prior)
STRUCTURE (Alternative Prior)
ADMIXTURE
sNMF
SCOPE

FIGURE 1

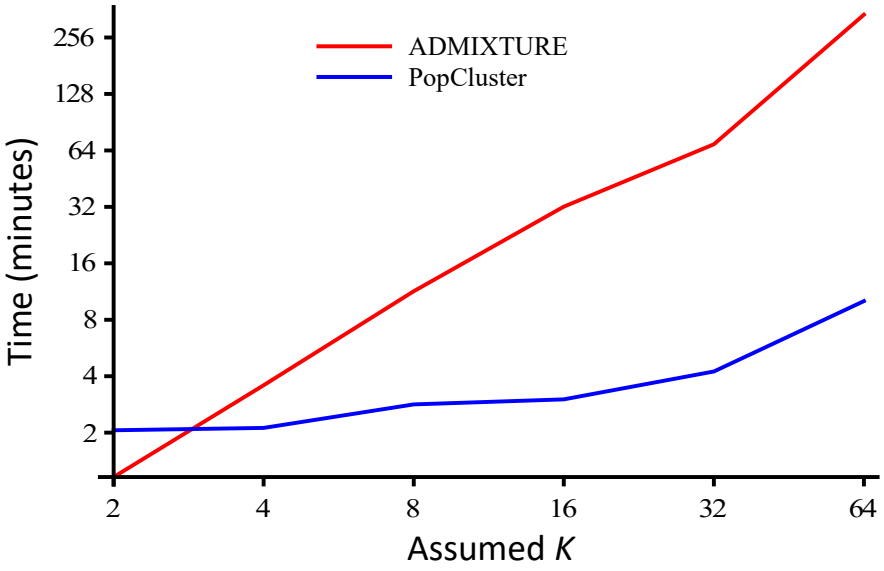
ADMIXTURE
PopCluster
Time (minutes)
256
128
64
32
16
8
4
2
2
4
8
16
32
64
Assumed K

FIGURE 2

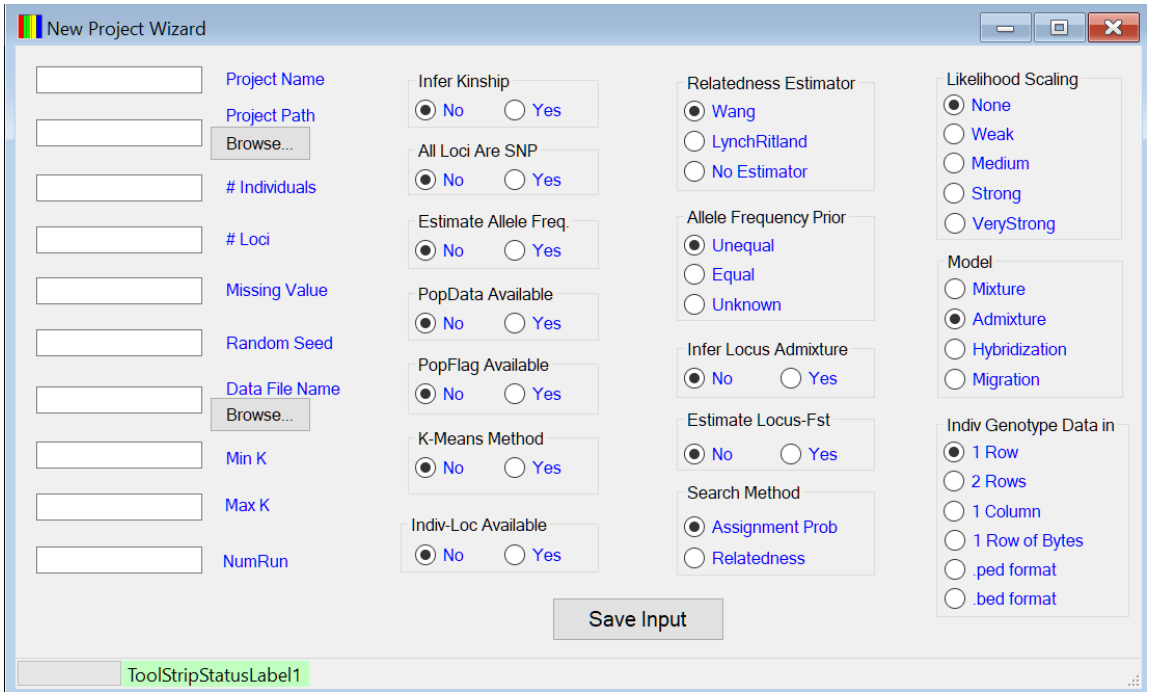
New Project Wizard
Project Name
Project Path
Browse...
Individuals
Loci
Missing Value
Random Seed
Data File Name
Browse...
Min K
Max K
NumRun
Infer Kinship
No
Yes
All Loci Are SNP
No
Yes
Estimate Allele Freq.
No
Yes
PopData Available
No
Yes
PopFlag Available
No
Yes
K-Means Method
No
Yes
Indiv-Loc Available
No
Yes
Relatedness Estimator
Wang
LynchRitland
No Estimator
Allele Frequency Prior
Unequal
Equal
Unknown
Infer Locus Admixture
No
Yes
Estimate Locus-Fst
No
Yes
Search Method
Assignment Prob
Relatedness
Likelihood Scaling
None
Weak
Medium
Strong
VeryStrong
Model
Mixture
Admixture
Hybridization
Migration
Indiv Genotype Data in
1 Row
2 Rows
1 Column
1 Row of Bytes
.ped format
.bed format
Save Input
ToolStripStatusLabel1

FIGURE 3

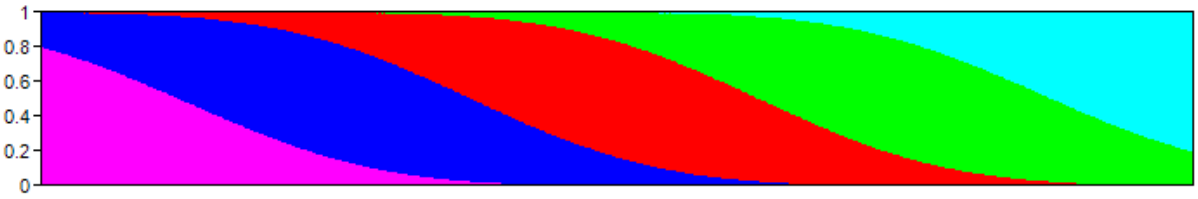

FIGURE 4

Appendix 1: An example showing the impact of unbalanced sampling on the structure analysis of a dataset with 2.5 million SNPs

A dataset was simulated using the same parameters as in Figure 1, except for the differentiation among populations (which was F_{ST} = 0.0015 instead of F_{ST} = 0.02) and the number of SNPs (which was L=2.5 million instead of L=10000). The data were not analyzed by STRUCTURE, because of its unrealistically long running time. They were analyzed by PopCluster (with scaling parameter s = 1), ADMIXTURE, sNMF and SCOPE, using default parameter settings with each method. The admixture proportions simulated and estimated by each method are shown below. Both PopCluster (with scaling parameter s = 1) and, to a much less extent, SCOPE recover the true (simulated) structure, while the other software, ADMIXTURE and sNMF, give rather poor estimates of the admixture proportions.