GAMA v1.0 User Manual

Shilei Zhao, Guo-Dong Wang, Yanhu Liu, Ya-Ping Zhang, Hua Chen

August 10, 2023

1 Introduction

GAMA (Genomic Analysis of Multiple Admixture) is designed for decoding the fine-scale ancestral origin of chromosome segments in the genome of admixed animal breeds or plant lines. The genome of an admixed individual resembles a mosaic of ancestry blocks that are inherited from different ancestral populations, with changes in ancestry occurring at recombination points. The mosaic pattern depends on time (in generations) since admixture, recombination rate, and ancestry proportions. We proposed a novel method, Genomic Analysis of Multiple Admixture (GAMA), to infer local ancestry segments from multi-way admixture along the genome. The model contains three free parameters, the number of generations since admixture λ, ancestry proportions π of each contributor, and error rates for the haplotypes α. GAMA performs well in inferring local ancestries (especially for long-term admixture and multiple-way admixture) and parameters. The running time is acceptable for whole-genome and large-scale admixture populations. Thus, GAMA is suitable for studying domestic animals and plants with complex breeding histories. The GAMA software is written in Matlab language. We plan to provide the C version of GAMA for faster running speed.

If you have any issues or suggestions with the software, please contact Shilei Zhao at zhaoshilei2018d@big.ac.cn.

If you use GAMA and publish your analysis, please cite the publication:

Shilei Zhao, Guo-Dong Wang, Yanhu Liu, Ya-Ping Zhang, Hua Chen. Decoding genetic architecture of dog complex traits by constructing fine-scale genomic ancestry of admixture (not published yet).

2 Inputs and Outputs

2.1 Inputs

GAMA requires phased haplotypes from unrelated individuals for ancestral populations and admixed individuals and genetic distances between adjacent loci. The genetic distances can be

obtained from the genetic map or approximated using the global recombination rate. The phased haplotypes are subsequently divided into L successive local fragments that contain the same number of *k* SNPs.

2.1.1 Ancestral haplotype

The phased haplotypes of ancestral populations are provided in a cell array. Each cell contains the matrix of haplotypes of an ancestral population. The rows denote haplotypes, and the columns denote SNPs. The SNPs should be organized in ascending order of physical location. The SNPs for all ancestral populations should be identical.

2.1.2 Admixed population haplotype

Admixed population haplotypes are provided in a matrix. The rows denote haplotypes, and the columns denote SNPs. The SNPs should be organized in ascending order of physical location and consistent with SNPs of Ancestral populations.

2.1.3 Physical position

Physical positions for SNPs are organized by a vector with the same order as that of ancestral populations and admixed populations.

2.1.4 Genetic position

Users can provide a vector of genetic positions or a scalar of the overall recombination rate. The genetic positions are with a unit of Morgan, and the recombination rate is in units per bp per generation.

2.1.5 Bin size

For the current version of GAMA, users can specify the bin size in a range of 1-12. The running speed will be too slow for a bin size larger than 12. We recommend using a bin size of 10.

2.2 Output

The output of GAMA is a structure with fields hidstate, par, iterpar, postp_raw, postp_norm, observeProb. Hidstate is a matrix with rows denote haplotype and columns denote bins. The elements in the matrix represent the inferred local ancestral segments. The first row is the physical position. Par is a vector containing the inferred parameters of GAMA. The first N elements denote the ancestral proportions. The N+1-th element denotes admixture time. The N+1-th element denotes the error rate. Iterpar is a matrix with the i-th row denoting the temporary states of parameters in the EM step. Postp_raw is a three-dimensional matrix with a size of $H \times L \times N$, where H denotes the number of haplotypes of the admixed population, L denotes the number of

bins, and N denotes the number of ancestral populations. Postp_raw(i,j,k) denotes the probability of the j-th bin of haplotype i inherited from the ancestral population k. Postp_norm is the normalization of Postp_raw, with $\sum_{k=1}^{N} \text{Postp\_norm}(\text{i},\text{j},\text{k}) = 1$. ObserveProb is also a three-dimensional matrix with a size of $H \times L \times N$, with ObserveProb(i,j,k) denoting the emission probability of i-th bin of haplotype i to the ancestral population k.

3 How to use GAMA

In environment of Matlab, users can run GAMA using the following command:

Output=GAMA(Ancestral_haplotype,Admixed_population_haplotype,Physical_position,Genetic_position,bin_size);

The inputs and output are described in the section 2.