

Missing genotype imputation in non-model species using Self-Organizing Maps

Fernando Mora-Márquez¹, Juan Carlos Nuño¹, Álvaro Soto¹, and Unai López de Heredia¹

¹Universidad Politécnica de Madrid

June 5, 2023

Abstract

Current methodologies of genome-wide Single Nucleotide Polymorphism (SNP) genotyping produce large amounts of missing data that may affect statistical inference and bias the outcome of experiments. Genotype imputation is routinely used in well-studied species to buffer the impact in downstream analysis and several algorithms are available to fill in missing genotypes. The lack of reference haplotype panels precludes the use of these methods in genomic studies on non-model organisms. As an alternative, machine learning algorithms are employed to explore the genotype data and to estimate the missing genotypes. Here, we propose an imputation method based on Self-Organizing Maps (SOM), a widely used neural networks formed by spatially distributed neurons that cluster similar inputs into close neurons. We follow a classical approach that explores genotype datasets to select SNP loci for each query missing SNP genotype to build training sets, and that initializes and trains the neural networks to finally use the SOM-derived clustering to impute the best genotype. To automate the imputation process, we have implemented GTIMPUTATION, an open source application programmed in Python3 and with a user-friendly GUI to facilitate the whole process. The method performance was validated by comparing its accuracy, precision and sensitivity on several benchmark genotype datasets with other available imputation algorithms. Our approach produced highly accurate and precise genotype imputations and outperformed other algorithms, especially for datasets from mixed populations with unrelated individuals.

Hosted file

Mora-Marquezetal-SOMbasedImputation.docx available at <https://authorea.com/users/347523/articles/647482-missing-genotype-imputation-in-non-model-species-using-self-organizing-maps>