Visualizing Pedigree and Population-Level Data in Molecular Ecology: A Tale of Two Methods

Matt Thorstensen\*, William Hemstrom\*, and Andrea Schreier

Introduction

There is a pressing need for better methods to visualize multidimensional data in molecular ecology. Visualizations can be helpful tools for first-pass investigations on a data set and can produce informative figures to support numerical results for publication, and so are almost always part of analysis pipelines across the discipline. Existing methods typically work well with low dimensional data, but struggle to depict more complex relationships, such as pedigrees or within and between-population diversity. For example, plots produced by the program STRUCTURE clearly show information about population structure and hybridization (Pritchard et al., 2000), but often fail to show hierarchical or fine-scale population structure (Hubisz et al., 2009; Kalinowski, 2011). Principle Component Analysis (PCA) is likewise excellent at showing large scale clustering patterns, but, since the first several principle components (those which are typically plotted) delineate only the largest dimensions of variance in the data, they often depict only the largest, between population differences while neglecting differing degrees of variance within-population. Similarly, PCA’s struggle to display pedigree information in one plot, since each individual family groups may constitute a single dimension of difference. Multiple family groups are therefore usually impossible to show in a 2 or 3-dimensional PCA plot.

Other disciplines have long struggled with similar problems, and several solutions have been proposed. One of these, t-Distributed Stochastic Neighbor Embedding (t-SNE) has been used in machine learning and biology for dimension reduction and data visualization to great effect (Maaten & Hinton, 2008). For instance, t-SNE has been used to support research on deep convolutional networks being trained in visual recognition (Donahue et al. 2014), applied to cancer and epileptic seizure research to help identify tumor subpopulations that affect patient outcomes and for detecting epileptic seizures (Abdelmoula et al. 2016; Birjandtalab et al. 2016), and has been useful for differential gene expression, such as in islets of Langerhans within the human pancreas (Muraro et al. 2016).Interestingly, t-SNE applied to SNP data has also been shown to be more effective than PCA at resolving human population structure (Platzer 2013).

Briefly, t-SNE works by considering its input data set a high-dimensional matrix that it attempts to display in less (typically 2 or 3) dimensions. In the process of reducing dimensions, it considers the distance between data points conditional probabilities, and the similarity between data points is calculated as the conditional probability that one point would pick another as its neighbor randomly following a Gaussian distribution (Maaten & Hinton, 2008). Conditional probabilities are also calculated between points in the low-dimensional space for plotting using a Student-t distribution with one degree of freedom. This distribution has a heavy tail that leads to the compromise of within-cluster and between-cluster distance displayed in a t-SNE plot. t-SNE then attempts to minimize the difference between its low and high-dimensional plots by minimizing the sum of Kullback-Leibler divergences over all data points, or in other words attempts to minimize how much the high-dimensionality probability distribution differs from low-dimensionality distribution. A more complete explanation of how the t-SNE works and important parameters to consider when running the algorithm can be found in van der Maaten and Hinton (2008).

In this paper, we demonstrate the ability of t-SNE to accurately depict multi-dimensional genetic data. Critically, t-SNE displays population genetic data in 2 or 3 dimensions while making compromises between within- and between-cluster distances so that fine-scale as well as large-scale patterns in data may be visualized. While this method has been proposed for use in population genetics before (THAT ONE PAPER), we show here that t-SNE can be applied to microsatellite, single nucleotide polymorphism (SNP), posterior genotype probability, and other data sets. In particular, we demonstrate the ability of t-SNE to depict complex pedigree data accurately and quickly, despite lacking any kind of explicit parentage model. Our results show conclusively the utility and applicability of t-SNE in visualizing challenging, multidimensional genetic data.

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. doi:10.1111/j.1755-0998.2009.02591.x

Kalinowski, S. T. (2011). The computer program STRUCTURE does not reliably identify the main genetic clusters within species: simulations and implications for human population structure. *Heredity, 106*(4), 625-632.

Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research, 9*(Nov), 2579-2605.

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