Discussion

Levels of Data

Under some circumstances, t-SNE performs better than PCA when visualizing genetic and genomic data. Primarily, t-SNE clusters individuals across multiple dimensions of differentiation whilst PCA only depicts two or three dimensions. When visualizing population structuring or pedigree data, this has different implications.

When used with population-level data (figures X, X), PCA collapses fine-scale levels of variance in favor of maximizing the distances between populations, while t-SNE expands local population structure. PCA is therefore excellent at showing differences between populations when these are the only matters of interest, such as when providing evidence for population structure. However, when one wishes to visualize hierarchical population structure, subpopulation structure, or within-population differences between individuals in one figure, the multidimensionality of t-SNE may be more useful.

For family-based data where pedigrees or progeny arrays are of interest, we observed that PCA collapses individuals into population level clusters without regard to family group. As such, visualizing family structure with PCA is challenging and can be misleading. In contrast, because t-SNE displays data across multiple dimensions of variance, family clusters are often distinctly clustered (Figure X). This helps users visualize families, erroneously placed unrelated individuals, and half-sibling relationships (Figure X). As such, while arguments can be made for using both the PCA and t-SNE on population-level data, we believe that t-SNE is more appropriate for visualizing family-level data. However, like PCA, t-SNE can erroneously cluster individuals with few called genotypes (Figure X), so caution must be taken when inferring family groups for such individuals.

STICKLEBACK STUFF

Limitations of t-SNE

As a dimension reduction algorithm, t-SNE attempts to visualize multi-dimensional data on a 2 or 3-dimensional plane. We caution that t-SNE should be used for data analysis with care. Most importantly, since t-SNE is a parameterized model, researcher latitude can potentially influence results. Researcher-set parameters of perplexity, theta (or gravity under the Barnes-Hut implementation of t-SNE; van der Maaten 2013), initial dimensions, and number of iterations can have large effects on final visualizations (Waatenberg et al. 2016; Figure X). Secondly, t-SNE can produce erroneous clusters under specific circumstances where PCA simply produces no clusters whatsoever. For example, high perplexity values cause even random datasets to cluster (Waatenberg et al. 2016) while low values can erroneously separate points that do share relationships (CITATION? Figure X?).

While researcher latitude can lead to false conclusions, the parameterized nature of t-SNE allows users to test different configurations and investigate how robust possible patterns are to different parameter values. For example, if subpopulation structure is consistent across a wide range of perplexities in a t-SNE plot, then a researcher may feel more safely inclined to believe that it is a true pattern. The same logic holds for running replicate t-SNE analyses, since weak or non-existent patterns often disappear in replicate runs of the analysis with identical parameters. In contrast, PCA, as a deterministic method, returns the same results every time. With patterns possibly spread across hundreds of loadings from a PCA (Citation?), visualizing such data can be impossible in a single plot.

t-SNE may be prone to create spurious results when presented with small data sets of limited or sub-structured data, since minor background variance can create clustering artifacts. For instance, we visualized the chromosomal inversion in threespine sticklebacks, t-SNE accurately separates individuals which carry the inversion and clusters the remaining individuals into several different groups (Figure X). While these clusters persist across multiple t-SNE runs, they only differ in allele frequencies at a handful of snps (Table S1) and do not persist when the data is bootstrapped (Figure SX). On the other hand, PCA accurately clustered the individuals with the inversion while separating out the individuals without it, and even separates homozygous from heterozygous individuals (Figure X).

Because of t-SNE’s limitations, care must be taken to test different parameter values and run replicate t-SNE’s when using this method to do a first-pass on new data. If care is taken to show that patterns persist across multiple parameter sets and permutations of the data, t-SNE may is a powerful exploratory tool for discovering patterns within datasets; and, when creating plots for publication, t-SNE may be valuable for presenting findings that are supported by other lines of evidence.

Flexibility of t-SNE

Like PCA, t-SNE can work with a massive range of input data types. For our analysis, we used allele presence-absence data with gene dosage when possible (ie, for diploid but not polyploid organisms). With this data, both the PCA and t-SNE can be thought of as delineating allele frequency differences between individuals across multiple dimensions in order to represent those similarities in visually. We found that IBS calls worked similarly well as input data for t-SNE (Figure X), as did genotype posterior probabilities when distilled into the expected number of each allele at each loci. This is not surprising, given that other studies have used t-SNE for such varied tasks as visualizing RNA expression levels (Citation), identifying tumor subpopulations from mass spectrometry data (Abdelmoula et al. 2016), clustering researcher collaborations (van der Maaten and Hinton 2012), visualizing and classifying epileptic seizure events (Birjandtalab et al. 2016), and to provide online karaoke song suggestions based on user vocal competence (Guan et al. 2017). Therefore, analyzing allelic presence-absence data is a natural extension of t-SNE for use in molecular ecology, and we suspect there are many more uses than we describe here.

Importantly, using t-SNE is easy to implement in existing work flows. Fundamentally, t-SNE requires exactly the same input data as PCA, and is only slightly more complicated to run than the later, since it only requires a handful of user-defined parameters. Furthermore, t-SNE should be easily accessible to most researchers, since implemented in many different software packages across many programming languages, such as R, Python and Matlab. While it is slightly more computationally intense than a PCA, none of the data presented in this paper took more than 10 minutes to run on a laptop computer, and most took less than a minute.

Finally, we promote the use of t-SNE not only for data visualization, but also because extensions of the method exist that can explore non-metric relationships and be used to classify data. For instance, multiple maps t-SNE weights points differently in different t-SNE plots, revealing similarities between points that would be impossible to show in single plots (van der Maaten et al. 2012). Other researchers have built classifier algorithms to be used in tandem with t-SNE for classifying data points under various conditions of interest (Abdelmoula et al. 2016; Birjandtalab et al. 2016; Guan et al. 2017). Some of these, such as a *k*-nearest-neighbor function tailored to population-level data, could be easily tailored to provide statistical support for population structure results visualized by t-SNE.