Discussion

Levels of Data

Under some circumstances, t-SNE is more useful than PCA when visualizing genetic and genomic data. Qualitatively, we find that t-SNE clusters individuals across multiple types of variance whilst PCA only depicts two or three dimensions of differentiation in a single plot. This conclusion has different implications for visualizing population structure or pedigree data.

When used with population-level data (Figure 1), PCA collapses fine-scale levels of variance in favor of maximizing the distances between populations, while t-SNE expands local population variance. These results are consistent with Platzer (2013), where local variation in human populations is expanded in t-SNE relative to PCA plots using SNP data. PCA is therefore effective 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. By contrast, because t-SNE uses a heavy-tailed distribution to eliminate undesirable attractive forces between moderately dissimilar points, family groups are often distinctly clustered (Figure 1; Van Der Maaten & Hinton 2008). This helps users visualize families, erroneously placed unrelated individuals, and half-sibling relationships (Figure 1). As such, while PCA and t-SNE are appropriate for population-level data when different questions are of interest, we believe that t-SNE is more appropriate for visualizing family-level data.

Limitations of t-SNE

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 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). Secondly, t-SNE can produce erroneous clusters under specific circumstances where PCA simply produces no clusters whatsoever. For example, high perplexity values cause both randomly generated and uniformly distributed datasets to erroneously cluster, while low perplexity values can erroneously separate points that do share relationships (Waatenberg et al. 2016; Figure 2 (bottom)). In our analyses, the perplexity parameter that had the greatest effect on resulting t-SNE plots. However, clustering was qualitatively different only at very low or high perplexities, which is consistent with how t-SNE was originally reported to be robust to large changes in perplexity (Figures 2, S2); Van Der Maaten & Hinton 2008). In our analysis, at least, changes in θ seemed to have little effect on the resulting plots (Figure S2).

t-SNE may also be prone to return spurious results is when analyzing small, limited data sets, particularly when that data is sub-structured, since minor background variance can create clustering artifacts. When visualizing 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 S2). 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 S4). 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 2). Note that at low perplexity, when points are not expected to be strongly clustered, this problem vanished.

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.

Because of t-SNE’s limitations, researchers must caution by testing different parameter values and running replicate t-SNE’s when doing 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 is a powerful exploratory tool for discovering patterns within datasets; and when creating plots for publication, t-SNE may be useful for presenting findings that are supported by other lines of evidence.

Flexibility of t-SNE

Like PCA, t-SNE can work with a wide range of input data types. For our analysis, we used allele presence-absence data with gene dosage when possible (i.e., 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, as did genotype posterior probabilities when distilled into the expected number of each allele at each loci (Figure S3). This is not surprising, given that other studies have used t-SNE for such varied tasks as visualizing RNA expression levels (Muraro et al. 2016), identifying tumor subpopulations from mass spectrometry data (Abdelmoula et al. 2016), visualizing complex researcher collaboration relationships (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. t-SNE requires exactly the same input data and is only slightly more complicated to run than PCA, since it only requires in addition only a handful of user-defined parameters. Furthermore, t-SNE should be easily accessible to most researchers, since it is 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.