**Methods:**

*Datasets:* We collected datasets from several other in prep and previously published studies. To explore the capacity of t-SNE to depict within and between population differences, as well as family structure based on different types of genomic data, we both PCA and t-SNE to plot 1) microsatellite pedigree, 2) SNP pedigree, 3) microsatellite population structure, and 4) SNP population structure datasets. These datasets consist of: sibling assignments (from SOLOMON, Christie et al., 2013) and genotypes at 11 microsatellite for 2,077 juvenile Chinook salmon (*Oncorhynchus tshawytscha*) collected in 2011 by Sard et al. (2015); 11,588 SNP genotypes and sibling assignments (from breeding records) of 1,302 Chinook salmon collected by Pepping et al. (*in prep*); genotypes at 13 microsatellite loci from 1,574 steelhead salmon (*Oncorhynchus mykiss*) from Hemstrom et al. (2017); and 238,915 SNP genotypes from 289 monarch butterflies collected by Freedman et al (*in prep*), respectively. In addition, we used 8923 SNP genotypes for 420 three-spined stickleback (*Gasterosteus aculeatus*) collected by Hemstrom et al. (*in prep*) in order to evaluate the efficacy of t-SNE and PCA at multiple levels of population sub-structure. Lastly, we used 13 microsatellite genotypes for 250 white sturgeon (*Acipenser transmontanus*) collected by X and X to show the efficacy of t-SNE to visualize population structuring and family relationships in a polyploid organism, respectively.

To gauge the viability of t-SNE and PCA to visualize datasets other than simple called genotypes, we generated several other datasets from the raw data for the SNP pedigree dataset (Pepping et al. *in prep*). We first used the ANGSD software package to estimate the posterior probabilities that each individual had each possible genotype at every putative SNP (SNP *p*-value <= 1x10-12) with a minor allele frequency above 0.05 (Korneliussen et al., 2014). We used mapping and base quality filters of 20, the SAMtools genotype likelihood model, and a uniform allele prior. We also used the ANGSD package to sample single bases at each putative SNP (SNP *p*-value <= 1x10-12) in each sample using the same filters (the Identity By State, or IBS, approach). Since this does not require calling genotypes or even estimating genotype likelihoods, this method tends to create larger datasets with less missing data.

*Data filtering and formatting:*

In order to have a consistent methodology across all of our datasets, we re-formatted all of the datasets into a “presence-absence” format, in which each possible allele at each locus is noted as a 1 if present or a 0 if absent. We used this approach rather than translating each genotype to a 0, 1, or 2 (for the minor homozygote, heterozygote, or major homozygote, respectively) since this format does not work for poly-allelic markers such as microsatellites. For the call genotype posterior probabilities, however, we instead calculated the expected count of each allele at each site according to the formula , where *E­i* is the expected count of allele *i*, *p­ii* is the posterior probability of the homozygote *i* genotype, and *pij* is the posterior probability of the *ij* heterozgote genotype. For all basic SNP and microsatellite datasets we used the snpR package to reformat genotypes (Hemstrom et al. *in prep*). We used custom R scripts to reformat the polyploid, IBS, and posterior genotype probability datasets.

Prior to further analysis, we filtered each of these datasets to remove poorly sequenced individuals and loci and interpolated any missing genotypes which remained. Since neither PCA nor t-SNE allow for missing data, it is necessary to interpolate missing genotypes prior to analysis. To reduce the need for interpolation, we first removed all loci sequenced at less than 50% of individuals in each of our data sets, then all individuals sequenced at less than 50% of the remaining loci. For the posterior genotype probability dataset, individuals were determined to have missing data at a site if all three genotype probabilities were equal to the specified prior genotype distribution (1/3 for each genotype). For each data set other than the genotype posterior dataset, any remaining missing allele presence-absence calls were interpolated with the mean value of the presence-absence call for that allele. We used the snpR package to filter the SNP datasets and custom R scripts for everything else.

*PCA and t-SNE*: We created both PCA and t-SNE plots from each of the above datasets. PCA plots were created using the “prcomp” function in R, t-SNE plots were created using the Rtsne R package (Krijthe, 2015). Briefly, t-SNE minimizes the divergence between multi-dimensional input data and points embedded stochastically in either fewer (typically 2 or 3) dimensions for visualization (Maaten & Hinton, 2008). During this process, the expected number of data points adjacent to each other data point is given by the user defined perplexity parameter. More exactly, perplexity is equal to , where *H(Pi)* is the Shannon entropy of probability distribution *Pi* over all *i* data points based on the variance of the Gaussian in the multi-dimensional input data (Maaten & Hinton, 2008). We estimated the perplexity value to use during t-SNE as described by van der Maaten and Hinton (2012) via the “hbeta” function as implemented in the mmtsne R package unless otherwise stated. Additionally, Rtsne uses the Barnes-Hut implementation of t-SNE to minimize the number of pairwise differences the t-SNE algorithm must explore by approximating to a user defined degree (θ) the influence of groups of nearby points on distant points rather than calculating each individual pairwise difference (Van Der Maaten, 2013, 2014). Since our datasets were not prohibitively large, we used a θ of 0 (for a full, non-summarized search) in most cases. We used the default values for all other parameters.

We tested the sensitivity of t-SNE results to user-defined parameters in using the stickleback SNP and chinook microsatellite pedigree data. We ran PCA as well as t-SNE at perplexities of 2, 15, and 50 on full stickleback dataset, without SNPs on the sex chromosome, and only on SNPs diagnostic for the inversion on linkage group IX. For the chinook dataset, we conducted t-SNE runs with each combination of perplexity = 2, 18, 34, and 50 and θ = 0, 0.33, 0.66, and 1.

Christie, M. R., Tennessen, J. A., & Blouin, M. S. (2013). Bayesian parentage analysis with systematic accountability of genotyping error, missing data and false matching. *Bioinformatics, 29*(6), 725-732. doi:10.1093/bioinformatics/btt039

Hemstrom, W., van de Wetering, S., & Banks, M. A. (2017). Fish ladder installation across a historic barrier asymmetrically increased conspecific introgressive hybridization between wild winter and summer run steelhead salmon in the Siletz River, Oregon. *Canadian Journal of Fisheries and Aquatic Sciences*. doi:10.1139/cjfas-2016-0411

Korneliussen, T. S., Albrechtsen, A., & Nielsen, R. (2014). ANGSD: analysis of next generation sequencing data. *BMC bioinformatics, 15*(1), 356.

Krijthe, J. (2015). Rtsne: T-distributed stochastic neighbor embedding using Barnes-Hut implementation. *R package version 0.13, URL* [*https://github*](https://github)*.com/jkrijthe/Rtsne*.

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

Sard, N. M., O’Malley, K. G., Jacobson, D. P., Hogansen, M. J., Johnson, M. A., & Banks, M. A. (2015). Factors influencing spawner success in a spring Chinook salmon (Oncorhynchus tshawytscha) reintroduction program. *Canadian Journal of Fisheries and Aquatic Sciences, 72*(9), 1390-1397. doi:10.1139/cjfas-2015-0007

Van Der Maaten, L. (2013). Barnes-hut-sne. *arXiv preprint arXiv:1301.3342*.

Van Der Maaten, L. (2014). Accelerating t-SNE using tree-based algorithms. *Journal of machine learning research, 15*(1), 3221-3245.

van der Maaten, L., & Hinton, G. (2012). Visualizing non-metric similarities in multiple maps. *Machine Learning, 87*(1), 33-55. doi:10.1007/s10994-011-5273-4