WiDGeT PopGen Summer Notes - Updated 26 July 2019

Population Genetic Summary Statistics

Watterson's θ, **Tajima's *D***, and π are measures used to assess (nucleotide) diversity. Assessing differences in genetic diversity can provide insights into, for example, demographic history and selective processes. Calculating these summary statistics requires sequence data (i.e., SNPs or sequence) and one should be aware of, or take into account via modeling, the historical demography of the focal population.

Nucleotide diversity (π) is the average proportion of nucleotides that differ between any randomly sampled pair of sequences. The calculation requires p_i that is the frequency of sequence i in the sample, p_j that is the frequency of sequence j in the sample, and π_{ij} that is the proportion of sites that differ between i and j. The formula is as follows:

$$\pi = \sum_{i=1}^{n} \sum_{j=1}^{n} p_i p_j \pi_{ij}$$

It can be seen with that formula that the p parameter or the frequency of alleles are paramount to the calculation of π. If the ancestral state is known of autosomes, π can also be derived as follows:

$$\pi = \frac{2p(1-p)n}{n-1}$$

where p is the ancestral allele frequency that is segregating in the population and n is the sample size.

Watterson's estimator of θ relies on the the number of segregating sites (S_n) of n alleles and is formulated as follows:

$$\hat{\theta} = \frac{S_n}{1+\frac{1}{2}+\frac{1}{3}\cdots\frac{1}{n}}$$

The $\hat{\theta}$ means that this formula should be a good estimator for $\theta = 4N\mu$ that is expected under neutrality. We also see that the numerator S_n is a count of segregating sites, and thus rare and common alleles have the same weight. The number of alleles (n) in this equation are those that differ by *origin*, which simply refers to a sample of n (different) alleles from a population. We also assume the infinite sites model and that $n << N_e$. Of special importance is under neutrality the sum of site heterozygosities (i.e. $2pq = 2p(1\text{-}p)$ is:

$$\pi = E\left\{ \sum_{i=1}^{\infty} 2p_i(1-p_i) \right\} = \theta$$

Lastly, Tajima's D takes the difference between the Watterson's θ and π:

$$D = \frac{\hat{\theta} - \hat{\pi}}{C}$$

where C is a normalizing constant - see full equation for C in Gillesipie (2004). We immediatly see that when θ is equal to π, D is zero: this is expected under neutrality. When D is > 0, this is suggestive of a bottleneck or balancing selection; when a population expands or is under positive selection D is < 0. Tajima's D capitalizes on the fact that π encompasses more of the intermediate frequency alleles, where as θ reflects more the rare (low frequency) alleles; deviations due to demography and selection cause these two values to change and produce non-zero values of D.

The **Effective population size** or N_e is the size of an "ideal" population that would have the same rate of inbreeding or decrease in genetic diversity due to genetic drift as the real population of interest. Strictly speaking, N_e formalizes the non-adaptive loss of genetic variation via genetic drift. Microsatellite data, sequence, and SNP data can be used to determine the value of N_e, and linkage will bias estimates downwardly. Most estimates of N_e encompass one to a few generations of a focal population; in contrast the coalescent *Ne* - see Wakely & Sargsyan 2009 - that encompassess genetic variation of a larger taxonomic unit (subspecies or species) over a much longer time frame (i.e. $4N_e$ generations).

The program NeEstimator is ubiquitously used in literature and the linkage disequilibrium method (Waples & Do 2004) is the most accurate among the single-sample methods. The equation below calculates a correlation coefficient for each pair of alleles at each pair of loci and computes an average of all of the correlation coefficients:

$$\hat{r}_\Delta = \frac{\hat{\Delta}}{\sqrt{[\hat{p}(1-\hat{p}) + (h_i - \hat{p}^2)][\hat{q}(1-\hat{q}) + (h_i - \hat{q}^2)]}'}$$

Here h_{ij} are the homozygote frequenecies, and $\hat{p}$ and $\hat{q}$ are the allele frequencies for i and j respectively. Weir's unbiased estimator $\hat{\Delta}$ is equal to $\hat{\Delta}S/(S\text{-}1)$ where S is the number of individuals sampled. This value is then input into the below formula for a randomly mating populations when S is > 30:

$$N_e = \frac{1/3 + \sqrt{1/9 - 2.76\hat{r}^{2\prime}}}{2\hat{r}^{2\prime}}$$

and less than 30:

$$N_e = \frac{0.308 + \sqrt{0.308 - 2.08\hat{r}^{2\prime}}}{2\hat{r}^{2\prime}}$$

The coalescent *Ne* can be derived from π, as we know that under neutrality:

$$\theta_W = \theta_\pi = 4N\mu$$

Rearranging this allows the coalescent N_e to be solved from θ_π / 2cμ, where c is the ploidy level and μ is the mutation rate, per site, per generation. Thus calculating π allows for the extrapolation of the coalescent N_e.

F_{ST}, **Nei's D** and **Jost's D** all measure population differentiation on a scale from 0 (no differentiation) to 1 (completely differentiated). Metrics rely on allele frequencies, so SNP and microsatellite data are suitable. F_{ST} is arguably the most famous metric and was introduced by Sewall Wright in 1949. The fixation index (F) measures differentiation due to drift and can be derived when Hardy-Weinberg proportions have been violated, presumably due to non random mating. F is integrated into genotype frequency estimates as follows:

Genotype	A_1A_1	A_1A_2	A_2A_2
random mating	p^2	$2pq$	q^2
non-random mating	$p^2(1-F)+pF$	$2pq(1-F)$	$q^2(1-F)+qF$

When $F > 0$ there is an excess of homozygotes, where < 0 occurs from an excess of heterozygotes. When two populations are combined (the "species" average below) the expected (H-W) proportions differ from what is observed.

Genotype	A_1	A_1A_1	A_1A_2	A_2A_2
Population 1	0.25	0.0625	0.375	0.5625
Population 2	0.75	0.5625	0.375	0.0625
Species (combined)	0.5	0.3125	0.375	0.3125
Species (H-W)	0.5	0.25	0.50	0.25

Combining the two above tables we can see that $2pq(1-F) = 0.375$, allowing us to solve for $F(_{ST})$, which is 0.25 in this example. Other derivations of F_{ST} have emerged, notably:

$$F_{ST} = \frac{G_S - G_T}{1 - G_T} = \frac{2Var\{p_i\}}{1 - G_T}$$

where $G_T = p^2 + q^2$ and is the probability that two alleles drawn at random with replacement from the entire species are identical by state. G_S is the probability that two alleles drawn at random with replacement from a randomly chosen subpopulation are identical by state. Intuitively this equation is nice as it means if you always grab the same allele (A_1) in both G_S and G_T, the populations are identical and F_{ST} is zero; if the the alleles are always different (A_1 and A_2), F_{ST} is one. One last derivation worth highlighting is based on within and between population diversity:

$$F_{ST} = \frac{\pi_{Between} - \pi_{Within}}{\pi_{Between}}$$

Nei's D was introduced in 1972 and assumes genetic differences are caused by mutation and genetic drift. D is derived first by calculationg I as follows:

$$I = \frac{\sum_{i=1}^{L} \sum_{j=1}^{l_i} p_{ij,x} p_{ij,y}}{\sqrt{\sum_{i=1}^{L} (\sum_{j=1}^{l_i} p_{ij,x}^2) \sum_{i=1}^{L} (\sum_{j=1}^{l_i} p_{ij,y}^2)}}$$

where $p_{ij,x}$ frequency of j^{th} allele at i^{th} locus in population X. Once solved, D_{Nei} is equal to $-ln(I)$. When populations enter into a mutation-drift equilibrium, assuming all mutations result in new alleles in accordance with the the infinite alleles model, the expected value of D increases in proportion to the time after divergence between two populations.

Jost's D is a relatively new metric that arose out of concerns related to the derivation of F_{ST}. Jost's D is independent of average within-subpopulation heterozygosity (i.e. π_{Within}. The equation is strikingly similar to variants of F_{ST} but takes into account the number of subpopulations (n):

$$D = \frac{H_T - H_S}{[1 - H_S][\frac{n}{n-1}]}$$

To date we have observed no major discrepancies - or alteration of population genetic interpretations - using either F_ST, Nei's D or Jost's D.

Linkage disequilibrium (LD) is the non-random association or non-independence of alleles at different loci in a population. Understanding important for limiting pseudoreplication, association mapping, understanding demographic events and selection. Alleles that are proximal typically experience higher LD, while mutation, recombination, and population structure result in decreased LD over time. Two common metrics - r^2 and D' - are used to quantify LD. Let D represent the amound of LD between alleles p_1 and p_2, and thus:

$$D = p_{11}p_{22} - p_{12}p_{21}$$

To estimate r^2 we apply the following equation:

$$r^2 = \frac{D^2}{p_1(1-p_1)q_1(1-q_1)}$$

Because D can be negative, it is standardized to a relative measure that ranges between 0 and 1 as follows:

$$D' = \frac{D}{D_{max}}\text{; -D } D_{max} = \min[p_1q_2 \text{ or } p_2q_1], \text{ +D, } D_{max} = \min[p_1q_1 or p_2q_2]$$

Signficnace of LD can then be assessed with a χ^2 test, with the expected being random associations between alleles.

Relevent Readings

Ardlie KG, Leonid K & Seielstad M (2002) Patterns of linkage disequilibrium in the human genome. *Nature Reviews Genetics* **3**: 299–309

Gaut SB & Long AD (2003) The lowdown on linkage disequilibrium. *Plant Cell* **15**: 1502-1506

Gillespie JG (2004) Population Genetics: A Concise Guide. The John Hopkins University Press

Jost L (2008) G(ST) and its relatives do not measure differentiation. *Molecular Ecology* **17**:4015-4026

Wright S (1949) The Genetical Structure of Populations. *Annals of Eugenics* **15**:323-354

Nei M. (1972) Genetic Distance between Populations. *American Naturalist* **106**:283-292

Peart CR, et al. (In review) Determinants of genetic variation across eco-evolutionary scales and their implications for the Anthropocene

Rogers AR & Huff C (2009) Linkage disequilibrium between loci with unknown phase. *Genetics* **182**: 839-834.

Wakely J & Sargsyan O (2009) Extensions of the coalescent effective population size. *Genetics* **181**: 341-345

Waples RS & Do C (2008) LDNE: a program for estimating effective population size from data on linkage disequilibrium. *Molecular Ecology* **8**: 753-756

Population Genetic Analyses

Pedigree reconstructions depict ancestral relationships and for example are useful for inferring patterns of inheritance, understanding life-history traits, partitioning phenotypic variance between the environment and genetics, and estimating inbreeding coefficients. Reconstructing a pedigree relies on generating genotypic data (SNPs or microsatellites) and calculating allele frequencies.

In many human and wildlife cases maternity is known so the focus is assigning a father. This can be done by identifying mismatches (i.e. father is A_1A_1 and offspring is A_3A_3) and setting an exclusion criteria. In practice, good field data, meaning knowing possible parents, combined with mismatch analysis can readily identify the father. In many cases, however, likelihood ratios are used to determine paternity and this analysis relies on: i) frequency of offspring alleles that came from candidate parents; ii) genotypic state (heterozygous or homozygous) of the parents. As will be shown with the likelihood formula, paternal homozygous and rare genotypes will result in higher likelihoods.

In the case of a known mother, Marshall et al. (2007) defined the following likelihood for a father being tested at a single locus:

$$L(H_1|g_M, g_P, g_O) = T(G_O|G_M, G_P)P(G_M)P(G_P)$$

Similarly, the likelihood that the mother and random father are the parents is as follows:

$$L(H_2|g_M, g_O) = T(G_O|G_M)P(G_M)P(G_{P-random})$$

$P(G_i)$ and $P(G_j)$ are the frequencies of the mother's and alleged father's genotypes in the population; $T(G_ij)$ is the mendelian segregation or transmission probability from each parent and *should* be 1 if homozygous and 0.5 if heterozygous. The likelihood ratio (LR) at the locus is then calculated, which is simply the first equation divided by the second:

$$\frac{L(H_1|g_M,g_P,g_O)}{L(H_2|g_M,g_O)}$$

The LR for each locus are multiplied to produce a final LOD score; true fathers should have a positive LOD score and you can see in the first equation how a homozygous ($T(G_P)$=1) rare genotype $P(G_P<0.05)$ would result in a higher likelihood value.

Simulations are run using the population allele frequency and age/sex data of the population to determine a significant delta value, meaning a LOD score that results in a 95% probability of correct paternal assignment. Marshall et al. (2007) provides additional likelihood derivations for additional scenarios such as when the mother is unknown.

Outlier tests are used to detect loci affected by selection. Outlier loci are of interest because they can reflect local adaptation, meaning the genes underlying a trait that confers a fitness advantage in one environment but not another. It should be noted that differentiation selection from other processes (i.e. drift) is not trivial (see Kardos & Shafer 2018). Outlier tests typically rely on an estimate of population genetic differentiation, with F_{ST} being the default, and screen for loci with significantly higher (or lower) F_{ST} than expected under neutrality. Outlier loci should be viewed as candidate genes for further investigation.

Outlier tests based on F_{ST} generally assume: i)No pair of populations is more closely related than any other pair; ii) Independent divergence of each population from mean allele frequencies; iii) local and global allele frequencies are known without error; and iv) the distribution of allele frequencies approximates a normal distribution. Standard methods include:

***F*$_{ST}$ histograms** – used in high-density genome scans;

FDIST / FDIST2 – incorporates heterozygosity and simulates null model distribution;

BayeScan – incorporates uncertainty of allele frequencies into the model;

FLK – uses a population tree to build a null model;

OutFLANK – fits a trimmed χ^2 distribution to the null model.

Of the above approaches, **Bayescan** consistently had the highest power to detect outlier loci under increasingly more complex demographic scenarios; however, it was also prone to false positives. Comparing heterozygosity to to F_{ST} (as is done in FDIST with simulations) can be used to detect the upper level of

genetic divergence expected under neutral divergence due to genetic drift in the pair of populations. Here, loci with higher F_{ST} relative to the null model given heterozygosity are candidates for positive selection, where loci with lower F_{ST} relative to the null model given heterozygosity are candidates for balancing selection.

Additional **tests of selection** are listed in Vitti et al. (2013), but we will focus on the use of dN/dS or ω that is measured between species. Neutral Theory assumes that most non-synonymous mutations are neutral; the nonsynonymous to synonymous differences (ω) reflect non-neutral changes relative to neutral change. In this context, dN = number of non-synonymous changes per non–synonymous site and dS = number of synonymous change per synonymous site. An $\omega = 1$ is expected under neutrality; $\omega > 1$ is indicative of positive selection, with <1 reflecting negative selection. For the latter ($\omega < 1$) we would expect this for critical genes (e.g. Hox) that are conserved across taxa.

Maximum-Likelihood (ML) approaches have been adopted for teh detection of positive selection. ML methods evaluate the probability (i.e., likelihood) of obtaining a set of DNA sequences given a specific phylogenetic tree and an explicit model of nucleotide substitution. **PAML** is the standard softward and uses a Markov process to describe substitutions between sense codons. Parameters include: transition/transversion ratio (k), codon frequencies (π) adn branch lengths scaled for time (t). While a variety of models exist:

Model M7 assumes ω ratios follow a beta distribution (i.e., constrained in the interval 0-1);

Model M8 adds a second class of sites to M7 at which ω ratios can exceed unity (i.e., positive selection).

Next, we obtain log likelihood scores for M7 (L_7) and M8 (L_8), with signficance estimated as follows to select the best model for the given data:

$$\chi^2 = 2(L_7 - L_8), \text{ with 1 d.f.}$$

A **population bottleneck** is a pronounced reduction in population size that can result in the loss of genetic diversity, compromise a species' ability to adapt to environmental change, and increase the likelihood of extinction via a host of genetic and demographic processes. Detecting a population bottleneck is often used in evaluating the need for conservation efforts, especially when dealing with threatened or endangered species. Geneoypic data from microsatellites and SNPs can be used to test whether a bottleneck has occurred.

There are some simple tests to evaluate if a bottleneck has occurred (M-ratio test, heterozygote excess using BOTTLENECK) that rely on the number of alleles, allelic range, and heterozygosity. These are baed on the expectation that alleles are lost during a bottleneck, while heterozygosity is miminally impacted (i.e. gets restored to Hardy-Weinberg proportions).

Among the more accurate bottleneck detection tests are those that use Markov Chain Monte Carlo (MCMC) coalescent simulations of the population history to assess the possibility of a population decline or expansion. The following formula utilizes current population size (N_0), ancestral population size (N_1), and the ratio of time between time of sampling (t) and time of ancestral population size (Ta). Going backwards in time, the population size *N(t)* changes deterministically (either linearly or exponentially) to an ancestral size N_1 at time $t = T_a$ and then remains constant at N_1 for $t > T_a$.

$$N(t) = N_0 \frac{N_1}{N_0}^{t/T_a}$$

As the time seperating the population change from sampling (t) increases (i.e. t/T_a) the ratio becomes smaller, thereby weakening the influence on *N(t)*. Likewise, the ratio of N_1 to N_0 clearly influences *N(t)*, with a value one equating to no change, >1 a bottleneck, and <1 and expansion. To determine if an expansion or decline has occurred, likelihoods are calculated from the genealogical history of the sample represented as a sequence of events (coalescences and mutations) and Bayes factors are computed (e.g., a population decline is the ratio of the posterior probability of a population decline divided by the posterior probability of a population expansion).

The Pairwise Sequentially Markovian Coalescent model or **PSMC** estimates the effective population size (N_e) over time from a single diploid individual. The model creates a time of the most recent common ancestor (TMRCA) distribution for each allele across the entire genome, with the rate of coalescent events is inversely proportional to N_e:

$$N_e = \frac{1}{\text{Coalescent rate}}$$

In the coalescent the probability of a coalescent event occurring per generation is $\frac{1}{N}$, and the TMRCA is exponentially distributed with a mean *N*. This means that most coalescent events occur quickly (or recently), and likewise, the strongest PSMC signal is most recent - thus the approximate time range is is 10k to 1M years ago. Key considerations are i) >18X but <30X genome coverage; ii) <25% missing data; iii) identifying a biologically relevant mutation rate and generation time.

PSMC has been extended to include multiple phased genomes known as **MSMC**. Recent work has shown that harmonic mean of the PSMC N_e estimate is very similar to that of the coalescent N_e derived from π and μ.

Principal component analysis or PCA is a statistical method for datasets with multiple measurements (or dimensions); this is exactly what SNP and genotype data are. PCA is a data summary tool that can identify patterns (structure) within your data without applying a modeling framework. PCA reduces the dimensions into uncorrelated principal components (PCs), with PC1

for example being a linear combination of measurements that encompasses the most variation in the dataset. Assumbing biallelic SNPs we can let $Z_{si}\epsilon\{0,1\}$ be the allelic state for individual i at locus s, with 0 reflecting the ancestral allele.

This will create an L by n binary matrix, with L reflecting the number of SNPs. The data are typically zero-centred as follows to create a new matrix:

$$X_{si} = Z_{si} - \frac{1}{n}\sum_{j=1}^{n} Z_{sj}$$

Using these centred data we then compute a covariance matrix (call it A) for the whole dataset. For loci X and Y as follows:

$$cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{x})(Y_i - \bar{y})$$

With the completed matrix A, $\boldsymbol{v}$ is a vector and λ a scalar such that $Av = \lambda v$; here λ is the eigenvalue association with the eigenvector v of covariance matrix A. Here we note that the total variance is equal to $\sum_{i=1}^{N}\lambda_i$, with the retained variance of **D** eigenvectors equal to $\sum_{i=1}^{D}\lambda_i$; the ultimately goal is to retain the a large amount of the variance, while using a fraction of the eigenvectors. Calculating the eignevalues and vectors requires more complex maths, but after solving for both we can then sort by decreasing eigenvalues and select eigenvectors (generally, low eigenvalues equate to less information about the data). Here we select the top **M** pairs and then plot the two eigenvectors with the highest eigenvalues (PC1 vs PC2).

Relevent Readings

Dubley A (2018) The mathematics behind principal components analysis. Towards Data Science blog (https://towardsdatascience.com/the-mathematics-behind-principal-component-analysis-fff2d7f4b643), accessed 25 July 2019

Girod C, Vitalis R, Leblois R & Freville H (2011) Inferring population decline and expansion from microsatellite data: A simulation-based evaluation of the Msvar method. *Genetics* **188**: 165-179

Kardos M & Shafer ABA (2018) The peril of gene-targeted conservation. *Trends in Ecology & Evolution* **33**: 827-839

Li H & Durbin R (2011) Inference of human population history from individual whole genome sequences. *Nature* **475**:493-496

Luikart G, Allendorf FW, Cornuet JM & Sherwin WB (1988) Distortion of allele frequency distributions provides a test for recent population bottlenecks. *Journal of Heredity* **89**: 238-247

Luikart G, England PR, Tallmon D, et al. (2003) The power and promise of population genomics: From genotyping to genome typing. *Nature Reviews Genetics* **4**:981–994

Marshall TC, Slate J, Kruuk LEB & Pemberton JM (1998) Statistical confidence for likelihood-based paternity inference in natural populations. *Molecular Ecology* **7**: 639-655

McVean G (2009) A genealogical interpreation of principal components analysis. *PLOS Genetics* **5**:e1000686

Nadachowska Brzyska K, Burri R, Smeds L, Ellegren H (2016) PSMC analysis of effective population sizes in molecular ecology and its application to black and white Ficedula flycatchers. *Molecular Ecology* **25**:1058-1072

Narum SR, Hess JE (2011) Comparison of F_{ST} outlier tests for SNP loci under selection. *Molecular Ecology Resources* **11**:184–194 Peart CR, et al. (In review) Determinants of genetic variation across eco-evolutionary scales and their implications for the Anthropocene

Peery ZM, Kirby R, Reid BN, Stoelting R, Doucet-Beer E, Robinson S, Vasquez-Carrillo C, Pauli JN & Palsboll PJ (2012) Reliability of genetic bottleneck tests for detecting recent population declines. *Molecular Ecology* **21**: 3403-3418

Storz JF & Beaumont MA (2002) Testing for genetic evidence of population expansion and contraction: an empirical analysis of microsatellite DNA variation using a hierarchical Bayesian model. *Evolution* **56**: 154-166

Vitti JJ, Grossman SR & Sabeti PC (2013) Detecting natural selection in genomic data. *Annual Reviews in Genetics* **47**: 97–120

Whitlock MC, Lotterhos KE (2015) Reliable detection of loci responsible for local adaptation: inference of a null model through trimming the distribution of F_{ST}. *American Naturalist* **186**:S24–S36