Traditional and Expanded D-statistics

Durand, E.Y., Patterson, N., Reich, D. and Slatkin, M. (2011) Testing for Ancient Admixture between Closely Related Populations. *Molecular Biology and Evolution*, **28**, 2239-2252.

This paper came out shortly after Green et al. and formalized the D-statistic, solidifying it in the field as a valid test for introgression. The method used by Green et al. depends on having an archaic sample, so Durand et al. generalize it to depend only on extant populations. They point out their method does not depend on the time of admixture or ancestral population size, though it does assume random mating of ancestral populations. Their test does not directly tell you the direction of gene flow. A companion test tells you the proportion of genomic introgression between interbreeding taxa. They also introduce the f-statistic, which is the probability that the P2 population (one of your pair of interest) is derived from a lineage derived from P3 (your potential gene donor population). In other words, f is the fraction of the P2 genomes that are derived from P3.

Eaton, D.A.R. and Ree, R.H. (2013) Inferring Phylogeny and Introgression using RADseq Data: An Example from Flowering Plants (Pedicularis: Orobanchaceae). *Systematic Biology*, **62**, 689-706.

Introduces the 'partitioned D-statistic'. This was designed for datasets, like the one analyzed here, that feature a large number of closely related taxa. The authors point out that the classic ABBA-BABA test provides evidence for gene flow between populations P2 and P3, but if you have sequence from multiple close relatives of P3, it does not let you distinguish which P3 subpopulation was actually hybridizing—rather, it may appear that one of your P3 subpopulations was hybridizing simply because it shares the introgressed alleles with the actual donor population due to their common ancestry since diverging from the P2 branch. The partitioned D expands the D-statistic to five taxa and measures three D-statistics, each representing a comparison between two discordant trees. A drawback is that it assumes no gene flow between your subpopulations ($P3_1$ and $P3_2$). An even bigger drawback is revealed in Pease and Hahn 2015.

Fontaine, M.C., Pease, J.B., Steele, A., Waterhouse, R.M., Neafsey, D.E., Sharakhov, I.V., Jiang, X., Hall, A.B., Catteruccia, F., Kakani, E., Mitchell, S.N., Wu, Y.-C., Smith, H.A., Love, R.R., Lawniczak, M.K., Slotman, M.A., Emrich, S.J., Hahn, M.W. and Besansky, N.J. (2015) Extensive introgression in a malaria vector species complex revealed by phylogenomics. *Science*, **347**.

An interesting application of the five-taxon D-statistic (D_{FOIL}). They sequenced many individuals from the species complex of mosquitos that includes three vectors of malaria, and found disagreement between trees generated from the autosomes and X chromosome. Trees generated from autosomes grouped the three vector species together, while those from the X chromosome put them in separate clades. Like the Heliconius follow-up paper, they use the fact that introgressed alleles should have reduced divergence times: they compared divergence times among the possible discordant trees, and indeed found that one tree was an outlier with a very short

divergence time for the pair of species of interest. Turns out that the vast majority of the autosomes was actually introgressed, and the more introgression-resistant X chromosomes revealed the true branching pattern. D_{FOIL} tests with the correct branching pattern confirmed gene flow between one of the three vector species and the ancestor of the other two.

Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H.-Y., Hansen, N.F., Durand, E.Y., Malaspinas, A.-S., Jensen, J.D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano, H.A., Good, J.M., Schultz, R., Aximu-Petri, A., Butthof, A., Höber, B., Höffner, B., Siegemund, M., Weihmann, A., Nusbaum, C., Lander, E.S., Russ, C., Novod, N., Affourtit, J., Egholm, M., Verna, C., Rudan, P., Brajkovic, D., Kucan, Ž., Gušic, I., Doronichev, V.B., Golovanova, L.V., Lalueza-Fox, C., de la Rasilla, M., Fortea, J., Rosas, A., Schmitz, R.W., Johnson, P.L.F., Eichler, E.E., Falush, D., Birney, E., Mullikin, J.C., Slatkin, M., Nielsen, R., Kelso, J., Lachmann, M., Reich, D. and Pääbo, S. (2010) A Draft Sequence of the Neandertal Genome. *Science*, **328**, 710-722.

Widely cited as the first application of the D-statistic. They obtained ancient DNA from Neandertal fossils, and got whole genome sequences from a handful of humans from different regions around the world. They wanted to test whether the Neandertals are more closely related to some humans than others, which would be good evidence that there was admixture between Neandertals and the ancestors of those closely related human groups. They found Neandertals share significantly more derived alleles with non-Africans than with Africans, whereas they share equal amounts of derived alleles when compared either to individuals within Eurasia or to individuals within Africa. They inferred directionality of gene flow by comparing the magnitude of D for an African group that is more closely related to Europeans vs. D for an African group that is distantly related to Europeans.

Patterson, N., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T. and Reich, D. (2012) Ancient Admixture in Human History. *Genetics*, **192**, 1065-1093.

Expanded some of the functionality of Durand et al.'s f-statistics and incorporated into a package called ADMIXTOOLS, of which D statistics are only one part. Notably, they use their f-statistics to estimate the relative proportion of admixture between populations.

Pease, J.B. and Hahn, M.W. (2015) Detection and Polarization of Introgression in a Five-Taxon Phylogeny. *Systematic Biology*, **64**, 651-662.

This paper builds on the work of Eaton and Ree, improving their five-taxon test. This new five-taxon test, called D_{FOIL} is appropriate for biological scenarios where you have two pairs of populations and you want to know whether some combination of them have hybridized in the past. It works by calculating 4 different D-statistics for all 16 possible introgression scenarios. Unlike the classic D, but like the partitioned D, this test can infer the direction of gene flow between the pairs by analyzing the collective profile of all the D-statistics obtained. However, the partitioned D has a fatal flaw, which the authors name the "mirror effect." Essentially, certain patterns the

partitioned D aims to test could have been generated by either the candidate introgression, or by an introgression Eaton and Ree failed to account for. They demonstrate the power and accuracy of their method with a series of simulations.

Smith, J. and Kronforst, M.R. (2013) Do *Heliconius* butterfly species exchange mimicry alleles? *Biology Letters*, **9**.

This short paper argued that the *Heliconius* genome paper did not satisfactorily rule out the possibility that ancestral polymorphism in mimcry loci could have been passed down to the descendant species (trans-species polymorphism). They posit that introgression should result in recent splitting of alleles and low sequence divergence compared with the genomic background; meanwhile ancestral polymorphism should result in more ancient splitting of alleles and greater sequence divergence. They calculated nucleotide divergence between the species from the same RADseq data, and then looked at divergence values in the regions with the highest D values vs. surrounding regions. Found that for those regions with a high D, which includes the mimcry genes, sequence divergence was reduced relative to the whole genome. This provides further evidence for introgression.

The Heliconuis Genome Consortium (2012) Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. *Nature*, **487**, 94-98.

This publication of the *Heliconius* genome used the four-taxon D-statistic on RAD-reseq data to show evidence for introgression between certain species of butterfly. Not only did the entire genome show a positive D-statistic when testing a certain pair of species, but there was also a dramatically elevated D around mimicry loci. Furthermore, gene trees made from these loci group taxa by phenotype rather than by species, discordant with the whole-genome average. Taken together, these data suggest the mimicry loci may have been passed between the two species by hybridization.

Reduced Dimensionality Analysis of Treesets

Huang, W., G. Zhou, M. Marchand, J. R. Ash, D. Morris, P. Van Dooren, J. M. Brown, K. A. Gallivan, and J. C. Wilgenbusch. 2016. TreeScaper: Visualizing and extracting phylogenetic signal from sets of trees. *Mol. Biol. Evol.* 33:3314–3316.

Jombart, T., M. Kendall, J. Almagro-Garcia, and C. Colijn. 2017. Treespace: Statistical exploration of landscapes of phylogenetic trees. *Mol. Ecol. Resour.*

Larget, B. R., S. K. Kotha, C. N. Dewey, and C. Ané. 2010. BUCKy: Gene tree/species tree reconciliation with Bayesian concordance analysis. *Bioinformatics* 26:2910–2911.

Wilgenbusch, J. C., W. Huang, and K. A. Gallivan. 2017. Visualizing phylogenetic tree landscapes. *BMC Bioinformatics* 18:85.

Multispecies Network Coalescent

Pardi, F., and C. Scornavacca. 2015. Reconstructible phylogenetic networks: do not distinguish the indistinguishable. *PLoS Comput. Biol.* 11:1–23.

Identifies an issue of indistinguishability in phylogenetic network reconstruction in which the same constituent trees (e.g. gene trees) may imply multiple networks. The idea of canonical networks is introduced as a resolution in which the outflowing branch of all reticulation nodes are reduced to length zero. This provides a simple minimum identifiable member of a class of indistinguishable networks with different branch lengths. Also includes a review of previous attempts at reducing the analytic space via identifiable trees, including galled trees and regular networks.

Solís-Lemus, C., and C. Ané. 2016. Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *PLoS Genet.* 12:1–21.

Introduces SNaQ, a quartet method of inferring a species network form gene trees with either posterior probabilities from Bayesian tree inference or concordance factors from BUCKy. Uses maximum likelihood of all informative taxon quartets with reticulation to derive a maximum pseudolikelihood species network. A thorough investigation of informative and non-informative quartets is also included and allows the method to include reticulation edges with length. Implemented in PhyloNetworks (not PhyloNet).

Solís-Lemus, C., P. Bastide, and C. Ané. 2017. PhyloNetworks: A package for phylogenetic networks. *Mol. Biol. Evol.*:1–14.

Software (implemented in the Julia programming language) for the quartet-based method (SNaQ) in Solis-Lemus and Ane 2016. Also includes extensible tools for comparing and analyzing phylogenetic networks.

Solís-Lemus, C., M. Yang, and C. Ané. 2016. Inconsistency of species tree methods under gene flow. *Syst. Biol.* 65:843–851.

Demonstration via simulated data that non-reticulate gene tree-species tree methods (ASTRAL and NJst) systematically misidentify the underlying branching evolutionary pattern when gene flow is present. Even when gene flow is minor and a "traditional" tree is desired, the multispecies network coalescent method (in PhyloNet) is more reliable.

Than, C., D. Ruths, and L. Nakhleh. 2008. PhyloNet: A software package for analyzing and reconstructing reticulate evolutionary relationships. *BMC Bioinformatics* 9:322.

Introduces a set of tools that have become central in the field of phylogenetic networks, including a modified Newick format capable of displaying reticulation nodes and inheritance probabilities. Subsequent developments of the software have included full maximum likelihood and Bayesian inference of phylogenetic networks from gene trees or from sequence alignments.

Wen, D., and L. Nakhleh. 2017. Co-estimating reticulate phylogenies and gene trees from multi-locus sequence data. *bioRxiv* preprint.

Builds on the previously developed Bayesian method of phylogenetic network inference in Wen et al. 2016 by adding co-estimation of the gene trees from sequence alignment during reversible-jump MCMC. Integrates over gene tree histories during each MCMC step as a computational shortcut. Implemented in PhyloNet software.

Wen, D., Y. Yu, and L. Nakhleh. 2016. Bayesian inference of reticulate phylogenies under the multispecies network coalescent. *PLoS Genet.* 12:1–17.

Introduces a set of rules for proposing phylogenetic networks in a Bayesian framework with reversible-jump MCMC. Builds on multispecies network coalescent methods in Yu et al. 2012 and 2014 to compute the likelihood of a set of gene trees given a species network. Implemented in PhyloNet

Yu, Y., J. H. Degnan, and L. Nakhleh. 2012. The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. *PLoS Genet.* 8.

Introduces a method and likelihood function for calculating the probability of a set of gene trees given a phylogenetic network. The method involved unwinding each reticulate node to produce a multi-labeled dichotomous tree (MUL tree) and applying a standard coalescent analysis to that tree with the addition of a parameter (gamma) that estimates introgression on two equivalent branches of the MUL tree. The method is implemented in PhyloNet.

Yu, Y., and L. Nakhleh. 2015. A maximum pseudo-likelihood approach for phylogenetic networks. *BMC Genomics* 16:S10.

Introduces a method to infer a phylogenetic network using maximum likelihood of the coalescent on all rooted triplets (with no reticulation). It is shown that a maximum likelihood set of rooted triplets may imply more than maximum pseudolikelihood network, though. Suggestions for analysis of this set of MPL networks are given. The method is implemented in PhyloNet.

Yu, Y., C. Than, J. H. Degnan, and L. Nakhleh. 2011. Coalescent histories on phylogenetic networks and detection of hybridization despite incomplete lineage sorting. *Syst. Biol.* 60:138–149.

Demonstrates with simulations that both coalescent-based tree methods and non-coalescent based network methods bias the resulting species inferences, missing hybridization and incomplete lineage sorting, respectively. Formally introduces a multispecies network coalescent method , including a likelihood function for the probability of a set of tree branch-lengths (lambda) and reticulate introgressions (gamma) given a set of gene trees and population demographic values.

Zhang, C., H. A. Ogilvie, A. J. Drummond, and S. Tanja. 2017. Bayesian inference of species networks from multilocus sequence data. *bioRxiv* preprint.

Derives a full Bayesian method for the joint inferences of the species network and gene trees (along with population demographic parameters) from sequence alignments. The method is fully developed including a likelihood framework and a method of sampling the posterior distribution of gene trees and demographic parameters. As a result it is extremely computationally intensive. Implemented in BEAST 2 as the SpeciesNetwork add-on.