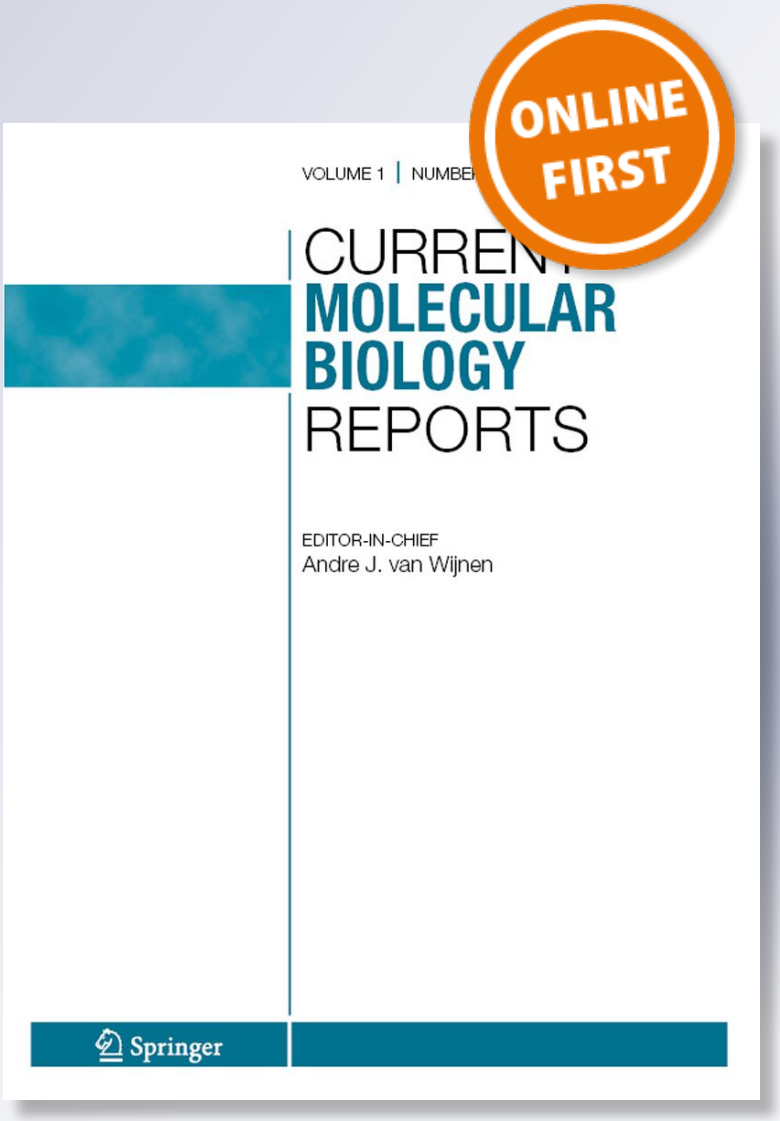
Recent Advances in the Inference of Gene Flow from Population Genomic Data

Richard H. Adams, Drew R. Schield & Todd A. Castoe

Current Molecular Biology Reports

e-ISSN 2198-6428

Curr Mol Bio Rep
DOI 10.1007/s40610-019-00120-0



Springer

Current Molecular Biology Reports
https://doi.org/10.1007/s40610-019-00120-0

POPULATION GENETICS (E LEWALLEN AND C BONIN, SECTION EDITORS)

Recent Advances in the Inference of Gene Flow from Population Genomic Data

Richard H. Adams[1] · Drew R. Schield[1] · Todd A. Castoe[1]

Abstract

Purpose of Review Detecting gene flow between populations or species is a fundamental goal of population genetics and speciation research and is also central for a thorough understanding of the demographic history of lineages. While population genomic data offer an unparalleled opportunity to study gene flow and other evolutionary processes at high resolution, extracting meaningful patterns from such large and complex datasets is rarely straightforward. Recent advances in both theory and methodology have led to a number of newly proposed analytical tools and frameworks for inferring genome-wide patterns of introgression and admixture that can more efficiently leverage population genomic data. Here, we provide an overview of several recent contributions to the problem of estimating gene flow, discuss advantages and potential pitfalls to these approaches, and provide an outlook for future developments.

Recent Findings Three prominent areas of recent research progress include (1) improving upon existing test statistics to detect and measure gene flow, (2) developing efficient frameworks for demographic model testing, and (3) applying supervised machine learning to identify introgressed loci across genomes. Over the past several years, contributions to these three areas have greatly enhanced our ability to study gene flow at various scales (i.e., species, populations, and individual genomes). Here, we highlight six relevant studies within these focal areas that represent particularly novel contributions to the goal of gene flow estimation from genome-scale data.

Summary The inference of gene flow is a notoriously challenging statistical problem that is an integral component of population genomic research. Our survey of the literature revealed a number of important recent contributions to this problem, from the improvement of admixture tests to demographic model testing and inference of specific regions of the genome likely to have crossed boundaries between populations and species. Although these studies represent only a sampling of the current literature, their contributions, along with those from numerous studies in the expanding field of population genomics, are markers of considerable progress in recent years toward addressing the issue of genomic inference of gene flow.

Keywords Migration introgression · Admixture · Hybridization · Next-generation sequencing

Introduction

Gene flow is an important evolutionary process that plays a fundamental role in shaping genetic variation both within and between populations and species. A number of recent genome-wide investigations have revealed that gene flow is pervasive in nature, and while expected to occur at some frequency between closely related species [1–5], studies have also found evidence of introgression between quite distantly related taxa [6–8]. For example, one such study found evidence of introgression and hybridization between two species of fern that diverged ~ 60 million years ago (Turissini and Matute 2017), which is slightly younger than the estimated age of the primate ancestor [9]. Genomic evidence for gene flow and introgression has been recovered in nearly all major lineages that have been surveyed, including primates [10], plants [11–13], reptiles [14–18], birds [19], insects [20], fungi [21], and protists (e.g., oomycetes) [22].

While gene flow is a fundamental evolutionary process that has shaped patterns of genetic diversity in many diverse taxa, confidently inferring various measures of gene flow (i.e.,

This article is part of the Topical Collection on *Population Genetics*

✉ Todd A. Castoe
todd.castoe@uta.edu

[1] Department of Biology, University of Texas at Arlington, 501 S. Nedderman Dr, Arlington, TX 76019, USA

Published online: 21 May 2019

migration rates and introgressed individuals and loci) from molecular data is often a challenging statistical problem. Next-generation sequencing technology has enabled genome-scale investigations of gene flow and its impact on genetic variation and the process of speciation in nature, yet many factors can prohibit or even mislead effective and efficient estimates of gene flow from complex genome-scale population datasets. In recent years, a wealth of new approaches have been proposed to address the factors that may confound studies of gene flow. Here, we conduct a survey of the current literature on the subject and highlight recent studies that have implemented particularly interesting and novel strategies that target three main aims: (1) improving the effectiveness of admixture tests for genomic scans of introgression, (2) establishing efficient methods for demographic model testing, and (3) using supervised machine learning to study introgression and admixture. In the following sections, we provide a background and current perspective on the state of each of these research areas and how recent advances in these areas are increasing the feasibility and precision of studying gene flow in natural populations using genomes.

Improving the Effectiveness of Admixture Tests

Genomic scans of population genetic statistics have been used extensively as a strategy for detecting signatures of gene flow between populations and species. First used to characterize patterns of gene flow between early humans and Neanderthals [23, 24], the *D* statistic [25] and its variants have risen in prominence over the past decade for studying introgression and admixture in a variety of contexts, including Old and New world Native Americans [26, 27], canids [28], butterflies [29], brown and polar bears [30], oak trees [31], and fungi [32].

The *D* statistic compares the observed and expected frequencies of two nucleotide patterns occurring across four individuals, each sampled from four different closely related populations: "ABBA" and "BABA." From left to right, the ABBA or BABA patterns are ordered based on the population sampled from one to four (Fig. 1), and the ancestral state of the nucleotide is determined based on the allele state of the individual sampled from the fourth outgroup population (i.e., the fourth individual always possesses the "A" state in both ABBA and BABA patterns). The ABBA pattern represents a site for which the individual sampled from the first population shares the same ancestral allelic state as the fourth population ("A"), while the derived allele "B" is found in the second and third populations (Fig. 1a, left). For the BABA pattern, the first and third individuals share the derived allele state "B," while the ancestral state "A" is found in the second individual (Fig. 1a, right). Due to random coalescence of gene lineages in ancestral populations, the frequency of ABBA and BABA site patterns is expected to be equal under the null

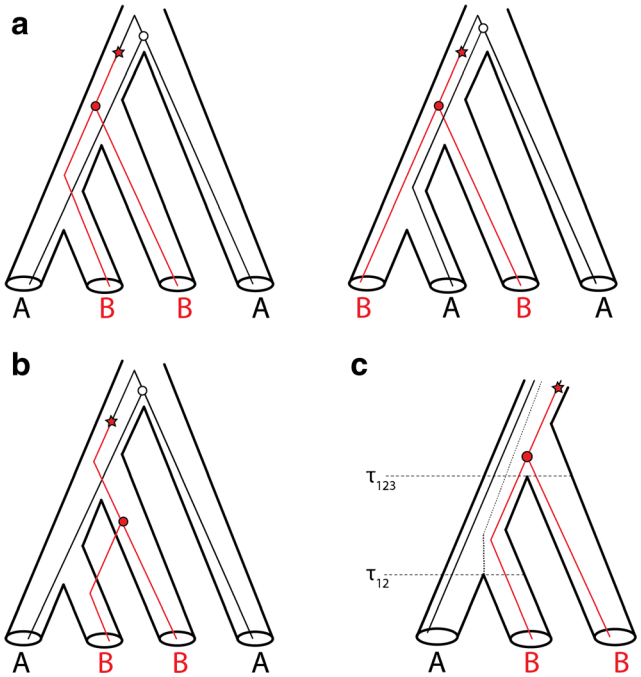


Fig. 1 Population tree models and genealogical examples used to demonstrate recent improvements to admixture tests based on the *D* statistic test. Circles represent coalescent events, and stars indicate mutations from the ancestral state "A" to the derived state "B." Null model (no gene flow) with example genealogies that can give rise to the ABBA and BABA patterns is shown in (**a**). Alternative migration model that can cause unequal ABBA vs. BABA frequencies is shown in (**b**), and an example of the ancestral structure model of [33] that can yield similar patterns is shown in (**c**)

hypothesis of no gene flow (Fig. 1a). Deviations from this expectation are often taken as evidence of introgression between one of the two ingroup populations (first or second population) and the outgroup taxa (Fig. 1b). This framework has also recently been extended to the analyses of five populations [34].

Although the *D* statistic can be a powerful method for detecting gene flow, recent analyses have revealed several limitations [e.g., 26, 33]. First, the *D* statistic is designed to detect signatures of gene flow, rather than to quantify it, and the nonlinear relationship between *D* and the fraction of introgressed individuals within a population suggests that *D* may be a biased estimator of migration (see Fig. 1 of Martin et al. [35]). Second, the *D* test statistic is also likely to be artificially inflated when it is used to infer introgression in genomic regions with low relative diversity, indicating that it may be unreliable in some contexts for identifying introgressed loci from individual loci affected by linked selection; for example, it is more appropriate to use *D* to measure genome-wide excess of shared derived alleles (see Fig. 2 of Martin et al. [35]). Lastly, most standard implementations of the *D* statistic are unable to differentiate between admixture and ancestral population structure, which can also yield unequal ABBA-BABA site pattern frequencies that can mimic

signatures of gene flow [25]. The *D* statistic is therefore likely to be ineffective at detecting gene flow using low-coverage genome sequencing and/or ancient DNA samples. To circumvent this issue, studies have typically down-sampled their data to include only a single base from one individual per population, thus reducing the information content from their data. Despite the apparent limitations and potential biases introduced by *D* statistics in certain contexts, they remain a prominent and useful family of metrics for inferring gene flow.

Distinguishing Between Ancestral Population Structure and Admixture

The power of the *D* statistic to reject the null model of no gene flow hinges on the assumption that, when present, admixture between one of the two populations with the outgroup is the sole cause of unequal ABBA-BABA site frequencies (Fig. 1a, b). However, other processes can also drive such patterns, including subdivision in ancestral populations, which has been shown to generate asymmetric gene trees using theory and empirical analyses [33, 36], which in turn can yield misleading evidence of gene flow [25]. Under these scenarios, ancestral structure can cause one of the two populations to exhibit greater than expected similarity with the outgroup population, even in the absence of admixture (Fig. 1c). Under simple models of ancestral structure, analyzing the doubly conditioned frequency spectrum (*dcfs*, i.e., conditioning on the allele state of an outgroup) can help distinguish admixture from ancestral subdivision [37]; however, more recent studies have shown that the *dcfs* could not distinguish between these two hypotheses under more complex scenarios likely to occur in nature [38]. In a recent study, Theunert and Slatkin (2017) [39] introduce two key improvements to these site-based tests that incorporate additional information from the unconditional site frequency spectrum (SFS) and linkage disequilibrium (LD) decay ratios to detect admixture.

In Theunert and Slatkin (2017), the authors first recapitulate the findings of previous studies [e.g., 37], demonstrating that a model of admixture (i.e., Fig. 1b here; Fig. 1A of Theunert and Slatkin 2017) yields essentially the same expected trends in the *dcfs* as a model of ancestral structure (i.e., Fig. 1c here; Fig. 1A of Theunert and Slatkin 2017). These results illustrate that the *dcfs* is unable to distinguish between these two models of population divergence (Fig. 2 of Theunert and Slatkin 2017). They also show that, unlike the *dcfs*, the unconditional site frequency spectrum (*ucfs*) will exhibit particular distortions under the ancestral structure model and is therefore useful for determining whether perturbations in site frequency patterns are due to admixture or ancestral structure (i.e., Fig. 3 of Theunert and Slatkin 2017). The authors then compare the *ucfs* under both models using simulations and empirical datasets collected from human populations in the South Pacific, which provide evidence of the power of the SFS for detecting admixture under complex demographic scenarios (Figs. 3 and SF5 of Theunert and Slatkin 2017).

These authors further investigate and compare the behavior of LD decay (i.e., DEFINE HERE) under the two competing models (i.e., admixture versus ancestral structure), finding that gene flow can generate a unique signal when measuring the ratio of LD between two sister populations (i.e., populations 1 and 2 in Fig. 1). Using simulations, the authors show that LD decays more slowly when derived alleles in one of the two sister populations have introgressed from the other through gene flow (Fig. 4B of Theunert and Slatkin 2017). This pattern is mirrored in their analyses of empirical human data from the 1000 Genomes Project [40], in which slower rates of LD decay were observed at sites in one population that shared the derived allele with Altai individuals (Fig. 6 of Theunert and Slatkin 2017). Taken together, these two approaches (*ucfs* and LD-based comparisons) mark major improvements to the power of SFS-based tests of admixture under complex models, and the authors recommend leveraging both approaches to reinforce inferences of admixture. These results also provide simulated and empirical evidence of the theoretical power of introgression to generate LD in the genome. One limitation to the broad usefulness of these approaches, however, is that LD-based methods often require phased genomic data, and the power of these approaches to distinguish admixture from ancestral structure is strongest when sites are closely linked to one another. Future investigations including comparisons of these methods under more complex models of gene flow, scenarios involving population size changes, and methods to account for evolutionary processes (e.g., genetic drift and natural selection) are required to thoroughly understand how these admixture tests behave under natural conditions.

Improving the Power of the *D* Statistic for Low-Coverage Analyses

Reconstructing the demographic history of both modern and ancient genomes has become a prominent approach for understanding the history of human populations [41], and the *D* statistic has been used extensively for this purpose [e.g., 23]. Relevant to the study of ancient genomes are issues inherent to the analyses of low-coverage genome data, which can be error-prone and subject to poor quality base calls. In a recent study, Soraggi et al. [42] address these concerns by increasing the robustness of the ABBA-BABA test to high sequencing error. In typical applications, the *D* statistic is computed by sampling a single base from a single individual, and Soraggi et al. [42] demonstrate that the power of the *D* statistic to detect admixture from low-coverage data can be improved by instead considering all reads from multiple individuals within each population (i.e., Fig. 4 of Soraggi et al. 2017). Furthermore, the authors apply type-error correction by

adjusting allele frequency estimates as a product of genotype probabilities (i.e., the likelihood of observing base *b* when the true state is *a*; Eq. 6 and Fig. 5 of Soraggi et al. 2017). Another valuable contribution of this study is a newly proposed approach to address admixture from "ghost populations" (i.e., external populations not directly considered in the tree), which can also confound inferences in other applications. The authors show that correcting for external admixture can be accomplished by using estimates of the external migration rate (α), although this requires reliable (or known) estimates of α (i.e., Fig. 6 of Soraggi et al. 2017).

Efficient Methods for Demographic Model Testing

Reconciling theoretical predictions and models based on these predictions with genetic sequence data has long been a core focus of population genetics research, and recent years have seen numerous studies that use model selection procedures to evaluate demographic histories across a wide range of organisms, including plants [e.g., 31, 42], arthropods [e.g., 43, 44], fish [e.g., 45, 46], amphibians [e.g., 47, 48], squamates [e.g., 14, 17, 18, 49, 50], birds [e.g., 51, 52], and mammals [e.g., 53, 54]. Estimating parameters and testing the fit of demographic models using traditional, full likelihood-based methods (including maximum likelihood estimation and Bayesian inference) are, however, often intractable for large genomic datasets, and a substantial amount of recent work in the field has been directed toward developing more efficient model-based frameworks that readily scale to genomic data.

Approximate likelihood and diffusion-based methods have both become popular approaches for testing demographic model fit from multilocus genetic data [e.g., 55–59]. Diffusion-based methods, such as *∂a∂i* [57], approximate the joint (or multi-population) allele frequency spectra using diffusion equations to test explicit hypotheses of population co-divergence, expansion and bottleneck events, presence and timing of gene flow, and modes of population divergence. Coalescent-based hidden Markov models have also seen widespread use for inferring demographic parameters, including migration rates, using one or more whole-genome sequences [e.g., 60, 61, 62]. By considering the genealogical history of each site as a "hidden" unknown state, these methods are capable of modeling molecular evolution through time and across the genome, providing estimates of coalescent parameters (e.g., population sizes, migration rates) as well as recombination rates. Given the large computational costs of inferring complex demographic models with many parameters using large datasets, a number of recent model-based approaches have been designed for more efficient analysis of genomic data using approximation methods [56, 63] or more simplistic models of molecular evolution [64, 65]. Computing full likelihoods for complex models can be intractable, and thus methods such as approximate Bayesian computation (ABC) and approximate likelihood methods use simulation techniques to approximate the likelihood, rather than compute it directly. ABC methods, in particular, have been leveraged extensively in recent years for testing the fit of complex demographic models involving migration, population size changes, and other complex evolutionary scenarios [63, 66–69]. Approximate likelihood approaches have been comparatively less popular but have nonetheless been useful for testing the fit of demographic models.

Approximate Likelihood Testing of Speciation Models with Gene Flow

In a recent study by Jackson et al. [70], the authors present a novel approximate likelihood-based method referred to as Phylogeographic Inference using Approximate Likelihoods (PHRAPL) to incorporate gene flow into demographic model tests. Although this study is primarily focused on the use of the method for species delimitation (i.e., the procedure of assigning individuals to discrete populations or species), the methods proposed by Jackson et al. [70] are fundamentally designed to test the fit of molecular datasets to models of population or species divergence that include or exclude gene flow. Due to the complexity of introducing migration rates in demographic models, PHRAPL compares the fit of empirically estimated genealogical trees for a given dataset to simulated genealogical distributions under different demographic models. The best fitting demographic model is determined by identifying a model that yields a simulated distribution of gene trees that most closely matches the empirical genealogies. The authors demonstrate the utility of this approach on both simulated and empirical datasets, and in nearly all scenarios explored in this study, PHRAPL appeared to perform better than other approaches that do not account for models of gene flow. For example, the authors show that another popular method, Bayesian Phylogenetics and Phylogeography (BPP) [71, 72], has a tendency to lump human samples into a single-population model, while PHRAPL correctly delimited distinct human populations (Fig. 5 of Jackson et al. 2017).

Given that nearly all other species delimitation methods ignore gene flow, the approach of Jackson et al. [70] represents an important improvement in model-based approaches for the delimitation of populations and species, although there are several limitations inherent to this approach. Specifically, PHRAPL uses gene tree estimates to compare the fit of models, which can be highly unreliable when populations are closely related, for example, and in cases such as these, the authors suggest that PHRAPL may not be as powerful as BPP for detecting genetically isolated populations. A common issue with each of these methods is that they operate under the assumptions of strictly neutral evolution, free recombination among loci, and a lack of recombination within loci, which can influence parameter inference to various degrees [73, 74].

A subsequent study [75] also highlights some disadvantages of this approach and shows using simulations that BPP (which assumes no gene flow) yields more reliable inferences when applying the same heuristic thresholds used by PHRAPL, in some case. Collectively, these studies highlight the importance of developing robust methods that consider the realistic complexities of the divergence process of populations and species and have provided new constructive inroads toward addressing such complexities.

Machine Learning Algorithms for Studying Gene Flow

Another area of recent interest in the context of population genomic inference of gene flow is the use of machine learning techniques to detect regions of the genome likely to have undergone introgression between populations or species. Machine learning (ML) methods have revolutionized an enormous diversity of research fields and applications—from computer engineering and data mining [76] to language processing [77] and agriculture [78]—and yet, ML algorithms are comparatively underused in population genetics and evolutionary research in general, except for handful of recent examples [79–85]. Perhaps one reason for this is that, as a field, ML represents a clear departure from more classical population genetic approaches that have traditionally focused on using theory and models to study genetic variation and to estimate parameters of interest. By contrast, ML is primarily concerned with optimizing predictive accuracy of an algorithm without the use of parametric models and assumed probability distributions. ML has also been applied recently to other related areas, including the inference of deep-time reticulation of hybridization network models [86].

The accuracy of ML algorithms is improved by "learning" (i.e., improving with experience), rather than fitting models to data. Free from the constraints of models and their assumptions, ML algorithms can be informative while also efficiently leveraging high-dimensional input data that would otherwise be intractable to analyze with typical model-based approaches. This is one outstanding advantage of ML methods for analyses of population genomic datasets, which can comprise thousands to millions of loci or SNPs sampled across multiple individuals and diverse genomic regions, and for which likelihood computations of complex, parameter-rich models would be otherwise infeasible using traditional approaches. ML methods are primarily classified into unsupervised and supervised learning algorithms, as well as semi-supervised ML which combines aspects of both [87]. The primary aim of unsupervised ML is to detect underlying structure within a dataset in the absence of knowledge about that structure (i.e., clustering algorithms), while supervised algorithms are typically optimized using training data designed to predict response variables or to classify input. Applications of unsupervised ML methods in population genetics and bioinformatics include a number of related *k*-means clustering algorithms [88] and the familiar principle component analysis (PCA; [89, 90]), which has been used extensively to uncover population genetic structure and haplotype clusters [e.g., 87]. Coalescent-based hidden Markov models are sometimes considered to be ML methods as well. Supervised learning algorithms, while commonly used in other bioinformatics analyses [87, 91], such as predictions of genes [92], gene expression [93], cancer classification [94], haplotype assembly [95], and chromatin structure [96], are relatively new in the context of population genomics.

Using Supervised Learning to Detect Introgressed Regions of the Genome

Recent applications of ML have addressed a particularly challenging area of gene flow inference: identifying specific genomic regions that appear to have moved between populations or species, as well as the directionality of gene flow. Most genome-wide approaches have traditionally focused more on determining whether gene flow has occurred between populations, migration rate estimation, and demographic model testing. These approaches are therefore not informative about which loci have likely crossed population boundaries and how their specific movement has shaped the genomic landscape of genetic variation and speciation. A recent and creative use of ML was employed specifically for this purpose in a study by Schrider et al. [81], in which the authors present a new supervised ML algorithm applied to study introgressed loci between two closely related species of *Drosophila*. The method, called Finding Introgressed Loci using Extra Tree Classifiers (FILET), combines information across a large number of summary statistics to classify the genealogical history of a genomic region. FILET implements an Extra-Trees algorithm that uses an ensemble of learning technique to "vote" among randomly generated decision trees and assign specific windows of the genome to three different classes of gene flow: two directional classes with gene flow from one population to another and a third "no gene flow" class. This Extra-Trees algorithm is an extension of the random forest algorithms [97] and has been used previously to distinguish between hard and soft sweeps in the human genome [83, 98].

One obvious advantage of the ML-based approaches of FILET is the ability to combine genealogical information encoded by a suite of summary statistics, including 18 single-population statistics (e.g., π, Tajima's D) and at least 14 two-population statistics (e.g., F_{ST}, d_{xy}) and including five newly proposed statistics shown to be powerful for identifying introgressed loci. The ability to efficiently incorporate such a large and diverse amount of information from population genomic-scale datasets (e.g., 157 whole genomes) enables FILET to detect introgression between two populations with high sensitivity—even for particularly challenging scenarios of

very recent divergence, when other approaches might be intractable or imprecise, such as methods that use only a single or few statistics. Comparisons with another popular approach, ChromoPainter [99], reveal the greater statistical power of FILET to identify introgressed loci and the directionality of gene flow at loci under challenging demographic scenarios (e.g., Fig. 2 of Schrider et al. 2018) and the relative robustness of FILET to potentially confounding effects of introgression from ghost populations (e.g., Fig. S4 of Schrider et al. 2018). Another interesting advantage of using ML for this purpose is the ability to rank population statistics in order of their effectiveness for detecting introgressed loci (Table S2 of Schrider et al. 2018)—such information can be useful for many aspects of population genetics research, including the development of new approaches (both ML-based and otherwise), and for strategically selecting particular statistics to measure based on their usefulness.

In a follow-up review article [80], the authors argue for a paradigm shift in the field of population genetics occurring with the use of supervised ML-based methods. Indeed, given the favorable performance of FILET and its easily extendable framework, it does seem likely that supervised learning techniques hold exciting promise for furthering our understanding of introgression and admixture in nature. Some areas of future improvements include the incorporation of additional processes (such as natural selection) and more complex demographic models (e.g., additional populations and ghost populations), both of which should be relatively straightforward to incorporate into the simulated datasets used to train the algorithm (another advantage of ML-based methods). Currently, FILET only considers three different classifications (no gene flow or two directional gene flow classes), and it seems likely that other classes could be easily considered (e.g., reciprocal gene glow), as well as variation in the specific dynamics of gene flow. For example, FILET currently models gene flow as a discrete "pulse" rather than a continuous flow, and FILET does not currently consider adaptive introgression, which has been inferred in a variety of empirical systems [100–102]. Nonetheless, the ML-based approach of Schrider et al. [81] represents an excellent foundation for future ML-based approaches and highlights the utility of supervised learning to target important yet often difficult questions in genomic studies of natural populations. The authors of these two papers note, as we do here, that other ML-based methods, such as vector machines [103] and deep learning [79, 104], may also prove useful for the purpose of studying gene flow using population genomic data.

Conclusions

An enormous amount of progress has been made over the past decade to advance our ability to effectively study gene flow using population genomic data. In this review, we have discussed how emerging approaches, such machine learning-based methods and approximate likelihood computation, hold particular promise for studying complex models of demographic and targeting specific regions of the genome that may be introgressed. We have also explored how new improvements on widely used test statistics may provide more accurate and precise methods for detecting gene flow in nature. Despite these strides forward, key challenges still remain, such as the incorporation of more complex and realistic models of population divergence and structure. Perhaps most notably, there is still a general lack of efficient, full model-based frameworks (i.e., without approximate computation) that hinder the broad utility of powerful full likelihood-based approaches for the analyses of genome-scale data under models of gene flow and population divergence. This may be an especially useful avenue for further investigation, in order to strategically leverage full likelihood methods in the context of whole-genome data from populations.

Acknowledgements Support was provided from an NSF grant to TAC (DEB-1655571) and Phi Sigma Support to RHA. Additionally, both the Lonestar and Stampede compute systems of the Texas Advanced Computing Center (TACC) were utilized for these analyses.

Compliance with Ethical Standards

Conflict of Interest All authors have no conflict of interest to disclose.

Human and Animal Rights and Informed Consent This article does not contain any studies with human or animal subjects performed by any of the authors.

References

1. Brandvain Y, Kenney AM, Flagel L, Coop G, Sweigart AL. Speciation and introgression between *Mimulus nasutus* and *Mimulus guttatus*. PLoS Genet. 2014;10:e1004410.
2. Begun DJ, Holloway AK, Stevens K, Hillier LDW, Poh YP, Hahn MW, et al. Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. PLoS Biol. 2007;5:e310.
3. Kulathinal RJ, Stevison LS, Noor MAF. The genomics of speciation in Drosophila: diversity, divergence, and introgression estimated using low-coverage genome sequencing. PLoS Genet. 2009;5:e1000550.
4. Martin SH, Dasmahapatra KK, Nadeau NJ, Salazar C, Walters JR, Simpson F, et al. Genome-wide evidence for speciation with gene flow in Heliconius butterflies. Genome Res. 2013;23:1817–28.
5. Fontaine MC, Pease JB, Steele A, Waterhouse RM, Neafsey DE, Sharakhov IV, et al. Extensive introgression in a malaria vector species complex revealed by phylogenomics. Science. 2015;347(80):1258524.
6. Nadeau NJ, Ruiz M, Salazar P, Counterman B, Medina JA, Ortiz-Zuazaga H, et al. Population genomics of parallel hybrid zones in the mimetic butterflies, *H. melpomene* and *H. erato*. Genome Res. 2014;24:1316–33.
7. Rothfels CJ, Johnson AK, Hovenkamp PH, Swofford DL, Roskam HC, Fraser-Jenkins CR, et al. Natural hybridization

between genera that diverged from each other approximately 60 million years ago. Am Nat. 2015;185:433–42.
8. Nürnberger B, Lohse K, Fijarczyk A, Szymura JM, Blaxter ML. Para-allopatry in hybridizing fire-bellied toads (*Bombina bombina* and *B. variegata*): inference from transcriptome-wide coalescence analyses. Evolution. 2016;70:1803–18.
9. Foley NM, Springer MS, Teeling EC. Mammal madness: is the mammal tree of life not yet resolved? Philos Trans R Soc Lond B Biol Sci. 2016;371:20150140.
10. Tung J, Barreiro LB. The contribution of admixture to primate evolution. Curr Opin Genet Dev. 2017;47:61–8.
11. Goulet BE, Roda F, Hopkins R. Hybridization in plants: old ideas, new techniques. Plant Physiol. Am Soc Plant Biol. 2017;173:65–78.
12. Baack EJ, Rieseberg LH. A genomic view of introgression and hybrid speciation. Curr Opin Genet Dev. 2007;17(6):513–8.
13. Whitney KD, Ahern JR, Campbell LG, Albert LP, King MS. Patterns of hybridization in plants. Perspect Plant Ecol Evol Syst. 2010;12:175–82.
14. Leaché AD, Harris RB, Maliska ME, Linkem CW. Comparative species divergence across eight triplets of spiny lizards (Sceloporus) using genomic sequence data. Genome Biol Evol. 2013;5:2410–9.
15. Burbrink FT, Guiher TJ. Considering gene flow when using coalescent methods to delimit lineages of North American pitvipers of the genus Agkistrodon. Zool J Linnean Soc. 2015;173:505–26.
16. Schield DR, Card DC, Adams RH, Jezkova T, Reyes-Velasco J, Proctor FN, et al. Incipient speciation with biased gene flow between two lineages of the Western Diamondback Rattlesnake (*Crotalus atrox*). Mol Phylogenet Evol. 2015;83:213–23.
17. Schield DR, Adams RH, Card DC, Perry BW, Pasquesi GM, Jezkova T, et al. Insight into the roles of selection in speciation from genomic patterns of divergence and introgression in secondary contact in venomous rattlesnakes. Ecol Evol. 2017;7:3951–66.
18. Harrington SM, Hollingsworth BD, Higham TE, Reeder TW. Pleistocene climatic fluctuations drive isolation and secondary contact in the red diamond rattlesnake (*Crotalus ruber*) in Baja California. J Biogeogr. 2018;45:64–75.
19. Rheindt FE, Edwards SV. Genetic introgression: an integral but neglected component of speciation in birds. Auk. 2011;128:620–32.
20. Clarkson CS, Weetman D, Essandoh J, Yawson AE, Maslen G, Manske M, et al. Adaptive introgression between Anopheles sibling species eliminates a major genomic island but not reproductive isolation. Nat Commun. 2014;5:4248.
21. Gladieux P, Ropars J, Badouin H, Branca A, Aguileta G, De Vienne DM, et al. Fungal evolutionary genomics provides insight into the mechanisms of adaptive divergence in eukaryotes. Mol Ecol. 2014;23:753–73.
22. Schardl CL, Craven KD. Interspecific hybridization in plant-associated fungi and oomycetes: a review. Mol Ecol. 2003;12:2861–73.
23. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A draft sequence of the Neandertal genome. Science. 2010;328(5979):710–22.
24. Wall JD, Yang MA, Jay F, Kim SK, Durand EY, Stevison LS, et al. Higher levels of Neanderthal ancestry in east Asians than in Europeans. Genetics. 2013;194:199–209.
25. Durand EY, Patterson N, Reich D, Slatkin M. Testing for ancient admixture between closely related populations. Mol Biol Evol. 2011;28:2239–52.
26. Raghavan M, Steinrücken M, Harris K, Schiffels S, Rasmussen S, DeGiorgio M, et al. Genomic evidence for the Pleistocene and recent population history of native Americans. Science. 2015;349:aab3884.
27. Skoglund P, Mallick S, Bortolini MC, Chennagiri N, Hünemeier T, Petzl-Erler ML, et al. Genetic evidence for two founding populations of the Americas. Nature. 2015.
28. Gopalakrishnan S, Sinding MHS, Ramos-Madrigal J, Niemann J, Samaniego Castruita JA, Vieira FG, et al. Interspecific gene flow shaped the evolution of the genus Canis. Curr Biol. 2018;28:3441–3449.e5.
29. Nadeau NJ, Martin SH, Kozak KM, Salazar C, Dasmahapatra KK, Davey JW, et al. Genome-wide patterns of divergence and gene flow across a butterfly radiation. Mol Ecol. 2013;22:814–26.
30. Cahill JA, Green RE, Fulton TL, Stiller M, Jay F, Ovsyanikov N, et al. Genomic evidence for island population conversion resolves conflicting theories of polar bear evolution. PLoS Genet. 2013;9:e1003345.
31. Eaton DAR, Hipp AL, González-Rodríguez A, Cavender-Bares J. Historical introgression among the American live oaks and the comparative nature of tests for introgression. Evolution. 2015;69(10):2587–601.
32. Gladieux P, Condon B, Ravel S, Soanes D, Maciel JLN, Nhani A, et al. Gene flow between divergent cereal- and grass-specific lineages of the rice blast fungus *Magnaporthe oryzae*. MBio. 2018;9.
33. Slatkin M, Pollack JL. Subdivision in an ancestral species creates asymmetry in gene trees. Mol Biol Evol. 2008;25:2241–6.
34. Pease JB, Hahn MW. Detection and polarization of introgression in a five-taxon phylogeny. Syst Biol. 2015;64:651–62.
35. Martin SH, Davey JW, Jiggins CD. Evaluating the use of ABBA-BABA statistics to locate introgressed loci. Mol Biol Evol. 2015;32:244–57.
36. DeGiorgio M, Rosenberg NA. Consistency and inconsistency of consensus methods for inferring species trees from gene trees in the presence of ancestral population structure. Theor Popul Biol. 2016.
37. Yang MA, Malaspinas AS, Durand EY, Slatkin M. Ancient structure in Africa unlikely to explain Neanderthal and non-African genetic similarity. Mol Biol Evol. 2012;29:2987–95.
38. Eriksson A, Manica A. Effect of ancient population structure on the degree of polymorphism shared between modern human populations and ancient hominins. Proc Natl Acad Sci. 2012;109:13956–60.
39. Theunert C, Slatkin M. Distinguishing recent admixture from ancestral population structure. Genome Biol Evol. 2017;9:427–37.
40. Siva N. 1000 Genomes project. London: Nature Publishing Group; 2008.
41. Stoneking M, Krause J. Learning about human population history from ancient and modern genomes. Nat Rev Genet. 2011;12:603–14.
42. Soraggi S, Wiuf C, Albrechtsen A. Powerful inference with the D-statistic on low-coverage whole-genome data. G3 Genes, Genomes, Genet. G3: Genes,Genomes, Genetics. 2018;8:551–66.
43. Satler JD, Carstens BC. Phylogeographic concordance factors quantify phylogeographic congruence among co-distributed species in the *Sarracenia alata* pitcher plant system. Evolution. 2016;70(5):1105–19.
44. Krehenwinkel H, Rödder D, Tautz D. Eco-genomic analysis of the poleward range expansion of the wasp spider *Argiope bruennichi* shows rapid adaptation and genomic admixture. Glob Chang Biol. 2015;21:4320–32.
45. Anna P, Lacey KL. Genomic tests of the species-pump hypothesis: recent island connectivity cycles drive population divergence but not speciation in Caribbean crickets across the Virgin Islands. Evolution(N Y). 2015;69:1501–17.
46. Roesti M, Kueng B, Moser D, Berner D. The genomics of ecological vicariance in threespine stickleback fish. Nat Commun. 2015;6:8767.
47. Meier JI, Sousa VC, Marques DA, Selz OM, Wagner CE, Excoffier L, et al. Demographic modelling with whole-genome

data reveals parallel origin of similar Pundamilia cichlid species after hybridization. Mol Ecol. 2017;26:123–41.
48. Thomé MTC, Carstens BC. Phylogeographic model selection leads to insight into the evolutionary history of four-eyed frogs. Proc Natl Acad Sci. 2016;113:8010–7.
49. Portik DM, Leaché AD, Rivera D, Barej MF, Burger M, Hirschfeld M, et al. Evaluating mechanisms of diversification in a Guineo-Congolian tropical forest frog using demographic model selection. Mol Ecol. 2017;26:5245–63.
50. Barley AJ, Monnahan PJ, Thomson RC, Grismer LL, Brown RM. Sun skink landscape genomics: assessing the roles of micro-evolutionary processes in shaping genetic and phenotypic diversity across a heterogeneous and fragmented landscape. Mol Ecol. 2015;24:1696–712.
51. Laurent S, Pfeifer SP, Settles ML, Hunter SS, Hardwick KM, Ormond L, et al. The population genomics of rapid adaptation: disentangling signatures of selection and demography in white sands lizards. Mol Ecol. 2016;25:306–23.
52. Nater A, Burri R, Kawakami T, Smeds L, Ellegren H. Resolving evolutionary relationships in closely related species with whole-genome sequencing data. Syst Biol. 2015;64:1000–17.
53. Provost KL, Mauck WM, Smith BT. Genomic divergence in allopatric Northern Cardinals of the North American warm deserts is linked to behavioral differentiation. Ecol Evol. 2018;8(24): 12456–78.
54. Jónsson H, Schubert M, Seguin-Orlando A, Ginolhac A, Petersen L, Fumagalli M, et al. Speciation with gene flow in equids despite extensive chromosomal plasticity. Proc Natl Acad Sci. 2014;111: 18655–60.
55. De Manuel M, Kuhlwilm M, Frandsen P, Sousa VC, Desai T, Prado-Martinez J, et al. Chimpanzee genomic diversity reveals ancient admixture with bonobos. Science. 2016;354(6311):477–81.
56. Hickerson M, Stahl E, Takebayashi N. msBayes: pipeline for testing comparative phylogeographic histories using hierarchical approximate Bayesian computation. BMC Bioinformatics. 2007;8: 268.
57. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. 2009;5:e1000695.
58. Excoffier L, Foll M. fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. Bioinformatics. 2011;27:1332–4.
59. Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M. Robust demographic inference from genomic and SNP data. PLoS Genet. 2013;9:e1003905.
60. Li H, Durbin R. Inference of human population history from individual whole-genome sequences. Nature. 2011;475:493–6.
61. Sethuraman A, Hey J. IMa2p—parallel MCMC and inference of ancient demography under the isolation with migration (IM) model. Mol Ecol Resour. 2016;16:206–15.
62. Hobolth A, Christensen OF, Mailund T, Schierup MH. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. PLoS Genet. 2007;3:e7.
63. Hickerson MJ, Stahl EA, Lessios HA. Test for simultaneous divergence using approximate Bayesian computation. Evolution (N Y). Wiley Online Library. 2006;60:2435–53.
64. Adams RH, Schield DR, Card DC, Blackmon H, Castoe TA. GppFst: genomic posterior predictive simulations of FST and dxy for identifying outlier loci from population genomic data. Bioinformatics. 2017;33(9):1414–5.
65. Adams RH, Schield DR, Card DC, Corbin A, Castoe TA. ThetaMater: Bayesian estimation of population size parameter from genomic data. Bioinformatics. 2018;34:1072–3.
66. Beaumont MA, Zhang W, Balding DJ. Approximate Bayesian computation in population genetics. Genetics. 2002;162(4): 2025–35.
67. Beaumont MA. Approximate Bayesian computation in evolution and ecology. Annu Rev Ecol Evol Syst. 2010;41:379–406.
68. Hickerson MJ, Carstens BC, Cavender-Bares J, Crandall KA, Graham CH, Johnson JB, et al. Phylogeography's past, present, and future: 10 years after Avise, 2000. Mol Phylogenet Evol. 2010;54:291–301.
69. Hickerson MJ, Meyer CP. Testing comparative phylogeographic models of marine vicariance and dispersal using a hierarchical Bayesian approach. BMC Evol Biol. 2008;8:322.
70. Jackson ND, Carstens BC, Morales AE, O'Meara BC. Species delimitation with gene flow. Syst Biol. 2017;66(5):799–812.
71. Yang Z, Rannala B. Unguided species delimitation using DNA sequence data from multiple loci. Mol Biol Evol. 2014;31:3125–35.
72. Yang Z, Rannala B. Bayesian species delimitation using multilocus sequence data. Proc Natl Acad Sci. 2010;107(20): 9264–9.
73. Adams RH, Schield DR, Card DC, Castoe TA. Assessing the impacts of positive selection on coalescent-based species tree estimation and species delimitation. Syst Biol. 2018;67:1076–90.
74. Edwards SV, Xi Z, Janke A, Faircloth BC, McCormack JE, Glenn TC, et al. Implementing and testing the multispecies coalescent model: a valuable paradigm for phylogenomics. Mol Phylogenet Evol. 2016;94:447–62.
75. Leaché AD, Zhu T, Rannala B, Yang Z. The spectre of too many species. Syst Biol. 2019;68:168–81.
76. Witten, Frank, Hall. Data mining: practical machine learning tools and techniques (Google eBook). Complement. Lit. None. 2011.
77. McCallum A. MALLET: a machine learning for language toolkit. http://mallet.cs.umass.edu. 2002.
78. McQueen RJ, Garner SR, Nevill-Manning CG, Witten IH. Applying machine learning to agricultural data. Comput Electron Agric. 1995;12:275–93.
79. Sheehan S, Song YS. Deep learning for population genetic inference. PLoS Comput Biol. 2016;12:e1004845.
80. Schrider DR, Kern AD. Supervised machine learning for population genetics: a new paradigm. Trends Genet. 2018;34:301–12.
81. Schrider DR, Ayroles J, Matute DR, Kern AD. Supervised machine learning reveals introgressed loci in the genomes of *Drosophila simulans* and *D. sechellia*. PLoS Genet. 2018;14: e1007341.
82. Pybus M, Luisi P, Dall'Olio GM, Uzkudun M, Laayouni H, Bertranpetit J, et al. Hierarchical boosting: a machine-learning framework to detect and classify hard selective sweeps in human populations. Bioinformatics. 2015;31(24):3946–52.
83. Schrider DR, Kern AD. S/HIC: robust identification of soft and hard sweeps using machine learning. PLoS Genet. 2016;12: e1005928.
84. Ronen R, Udpa N, Halperin E, Bafna V. Learning natural selection from the site frequency spectrum. Genetics. 2013;195:181–93.
85. Lin K, Li H, Schlötterer C, Futschik A. Distinguishing positive selection from neutral evolution: boosting the performance of summary statistics. Genetics. 2011;187:229–44.
86. Burbrink FT, Gehara M. The biogeography of deep time phylogenetic reticulation. Syst Biol. 2018;67:743–55.
87. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat. Rev. Genet. 2015;16:321–32.
88. Oyelade J, Isewon I, Oladipupo F, Aromolaran O, Uwoghiren E, Aameh F, et al. Clustering algorithms: their application to gene expression data. Bioinform Biol Insights. 2016;10:BBI.S38316.
89. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, et al. Genes mirror geography within Europe. Nature. 2008;456(7218):98–101.

90. Ma S, Dai Y. Principal component analysis based methods in bioinformatics studies. Brief Bioinform. 2011;12:714–22.
91. Tan AC, Gilbert D. An empirical comparison of supervised machine learning techniques in bioinformatics. Proc First Asia-Pacific Bioinforma Conf Bioinforma 2003.
92. Hoff KJ, Tech M, Lingner T, Daniel R, Morgenstern B, Meinicke P. Gene prediction in metagenomic fragments: a large scale machine learning approach. BMC Bioinformatics. 2008;9:217.
93. Brown MPS, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci U S A. 2000;97:262–7.
94. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn. 2002;46(1):389–422.
95. Xu X-S, Li Y-X. Semi-supervised clustering algorithm for haplotype assembly problem based on MEC model. Int J Data Min Bioinform Inderscience Publishers. 2012;6:429–46.
96. Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nat Methods. 2012;9:473–6.
97. Breiman L. Random Forrest. Mach Learn. 2001;45:5–32.
98. Schrider DR, Kern AD. Soft sweeps are the dominant mode of adaptation in the human genome. Mol Biol Evol. 2017;34:1863–77.
99. Lawson DJ, Hellenthal G, Myers S, Falush D. Inference of population structure using dense haplotype data. PLoS Genet. 2012;8:e1002453.
100. Racimo F, Sankararaman S, Nielsen R, Huerta-Sánchez E. Evidence for archaic adaptive introgression in humans. Nat. Rev. Genet. 2015;16:359–71.
101. Hedrick PW. Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. Mol Ecol. 2013;22:4606–18.
102. Pardo-Diaz C, Salazar C, Baxter SW, Merot C, Figueiredo-Ready W, Joron M, et al. Adaptive introgression across species boundaries in Heliconius butterflies. PLoS Genet. 2012;8:e1002752.
103. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20:273–97.
104. Lecun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44.