Review

Decoding Genomic Landscapes of Introgression

Xin Huang (黄欣)[1, 2, *], Josef Hackl[1], Martin Kuhlwilm[1, 2, *]

1 Department of Evolutionary Anthropology, University of Vienna, Vienna, Austria

2 Human Evolution and Archaeological Sciences (HEAS), University of Vienna, Vienna, Austria

* Correspondence: xin.huang@univie.ac.at (X. Huang) and martin.kuhlwilm@univie.ac.at (M. Kuhlwilm)

Genomic landscapes of introgression provide valuable information for how different evolutionary processes interact and leave signatures in genomes. The recent expansion of genomic datasets across diverse taxa, together with advances in methodological development, has created new opportunities to investigate the impact of introgression along individual genomes in various clades, making the precise identification of introgressed loci a rapidly evolving area of research. In this review, we summarize recent methodological progress within three major categories: summary statistics, probabilistic modeling, and supervised learning. We examine how these approaches have been applied to data beyond humans and discuss the challenges associated with their application. Finally, we outline future directions for each category, including accessible implementation, transparent analysis, and systematic benchmarking.

Highlights

- Recent advances in methods and tools have enabled the study of genomic landscapes of introgression across diverse and complex evolutionary scenarios, including adaptive and ghost introgression.
- Despite their long history, summary statistics-based methods continue to evolve, with new implementations broadening their applicability across taxa.

- Probabilistic modeling is a major approach that provides a powerful framework to explicitly incorporate evolutionary processes and has yielded fine-scale insights across diverse species.
- Supervised learning is an emerging approach with great potential, particularly when the detection of introgressed loci is framed as a semantic segmentation task.
- Various methods have been applied across clades, revealing introgressed loci linked to immunity, reproduction, and environmental adaptation, especially in cases of adaptive and ghost introgression.

Genomic Landscapes of Introgression

Introgression (see Glossary) plays an important role in evolution. Beyond merely studying introgression events through phylogenetic approaches [1], understanding their genomic footprints—how introgressed loci are retained, eliminated, or distributed within genomes—is essential (Figure 1A), because these patterns are shaped by demographic histories, selective pressures, and genomic architectures. Although many methods for detecting introgressed loci have provided crucial insights into past gene flow from archaic hominins to contemporary human populations, they have largely been developed with modern and archaic human genomes [2].

With the increasing availability of genomic data from diverse taxa, such as hickories, peafowls, and corn earworms [3–5], studies on the detection of introgressed loci are becoming more prevalent in the field. Consequently, characterizing genomic landscapes of introgression offers deeper insights into the evolutionary forces driving hybridization outcomes and the functional roles of introgressed loci in different species (Figure 1B). Although developed in the pre-genomic era, summary statistics-based approaches remain widely used. In recent years, substantial methodological advances have emerged from probabilistic modeling and supervised learning (Figure 2). These developments motivate a critical assessment of existing approaches and a comprehensive review of emerging strategies for decoding genomic landscapes of introgression.

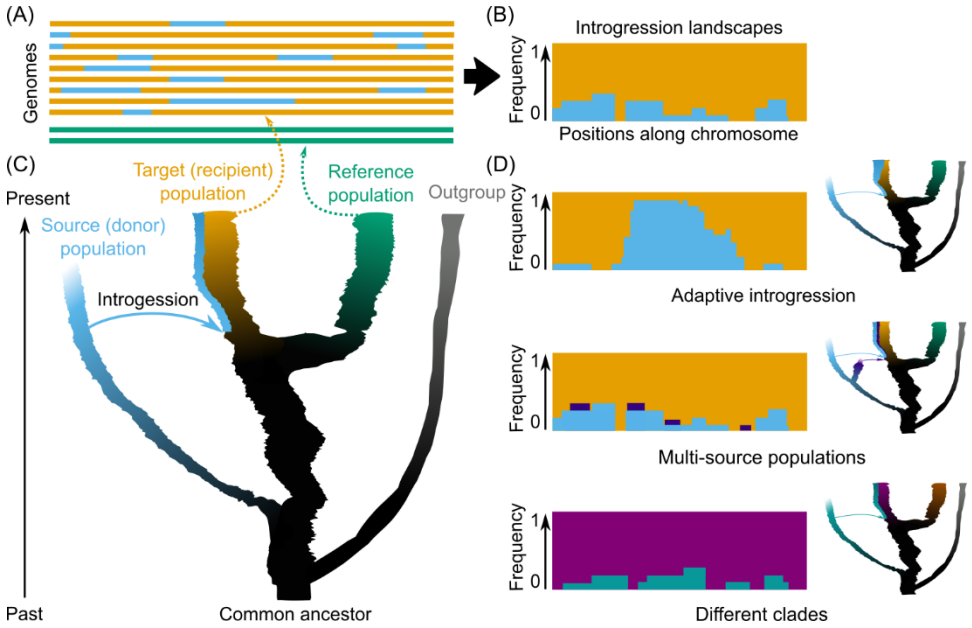

Figure 1. Introgression and its genomic landscapes. (A) Individual genomes in the target population carry short introgressed segments because recombination breaks down long haplotypes. (B) Due to genetic drift, the frequency of these segments in the target population varies along chromosomes. (C) Population setting in the context of introgression: deriving from a common ancestor, lineage-specific genetic variation arises over time in the diverging populations (color gradients). Introgression transfers this variation from the source population (blue) into the target population (orange). A reference population (green), more closely related to the target than the source population, is often used to determine non-introgressed variation in a lineage. Some methods use an outgroup (grey) to infer whether an allele is ancestral or derived within this topology. If no data is available from the source population, the scenario is referred to as ghost introgression. (D) Adaptive introgression represents a special case where introgressed ancestry surrounding an adaptive locus rises in frequency beyond the expectation under neutrality. When multiple source populations are involved, fragments from divergent origins may co-occur in the same genome, potentially confounding the detection of introgressed loci. Genomic landscapes of introgression might have different distribution patterns and dynamics in different clades (illustrated by different color gradients).

Summary Statistics-based Methods

Summary statistics are simple yet effective approaches for exploratory data analysis, commonly used to distill complex genomic data into simple numeric representations, including frequency-

based, linkage-based, or topology-based measures (Figure 2A). The *D* statistic is a widely used method for detecting genome-wide evidence of introgression [6], based on asymmetries in derived allele sharing between the target and source populations relative to a reference population and an outgroup (Figure 1C). However, it is not well suited for pinpointing introgressed loci, as it can be biased in regions of low genetic diversity [7–9].

To address this, alternative statistics that are calculated in windows along the genome have been developed. The dynamic estimator of the proportion of introgression (f_d) and the distance fraction (d_f) both scale the observed excess of shared derived alleles but differ in their normalization approaches: f_d uses the maximum possible level of derived allele sharing due to introgression, whereas d_f normalizes against the expected derived allele sharing under the species tree topology [7,8]. Both reduce the bias of *D* by avoiding direct dependence on the genetic distance between the reference and target populations [8]. The D^+ statistic further extends *D* by incorporating both shared derived and ancestral alleles, thereby increasing the number of informative sites and reducing variance [9]. Additionally, f_d has a bounded variant, f_{dM}, which ranges from −1 to 1, and is symmetrically distributed around zero under no introgression [10].

To detect loci under **adaptive introgression** (Figure 1D), additional methods have been introduced that also leverage allele sharing patterns between the target and source populations. These include the number of uniquely shared sites (*U*) and the quantile of derived allele frequency distributions (*Q*) [11]. Such methods retain variants shared between target and source populations that are rare or absent in the reference population, thereby enriching for candidates likely introduced via introgression. While adaptive variants in the target population often reach high frequency due to **positive selection**, these methods are primarily sensitive to such cases and may miss beneficial alleles that are at low or intermediate frequencies.

In cases of **ghost introgression**, where source samples are unavailable, inference relies on alleles present in the target but absent in the reference population. *S** was initially developed to detect archaic introgression in human populations without a source genome, by identifying clusters of private mutations in strong linkage within the target population [12]. However, *S** does not account for local mutation or recombination rate variation. To handle this, one approach involves simulating data under varying local rates, fitting a **generalized additive model** to the resulting *S** scores, and using this model to estimate expected values and assess significance in real data

[13,14]. Another approach, *S′*, modifies the calculation of *S** by incorporating local mutation and recombination rates [15].

Instead of frequency-based or linkage-based information, Twisst and its upgraded version Twisst2 summarize subtree topologies, each formed by selecting one sample per species, to assess how often gene trees support alternative species tree relationships [16,17]. This topology weighting approach indirectly captures discrepancies between gene trees and the species tree, which are explicitly tested by *D* [9]. While *D* is limited to four taxa, Twisst can in principle handle any number, but becomes increasingly impractical with more than six due to the large number of possible species tree topologies [16].

Probabilistic Model-based Methods

Probabilistic modeling enables model-based inference of introgressed loci by defining the relationship between observed genetic variation and underlying evolutionary processes through probability distributions and performing inference under likelihood-based or Bayesian frameworks (Figure 2B). For example, IBDmix estimates the probability of **identity-by-descent** between the target and source populations at each locus, thereby eliminating the need for a reference population [19]. It surpasses *S** on simulated data, especially in scenarios where introgressed fragments are also present in the reference population [19].

Probabilistic methods can also address compound scenarios such as adaptive introgression from a ghost population. VolcanoFinder is a likelihood-based method designed for this setting [20]. It identifies genomic regions exhibiting a characteristic “volcano” pattern of genetic diversity, which is marked by reduced diversity at the selected site flanked by elevated diversity, reflecting a selective sweep on an introgressed, highly divergent haplotype. However, it often fails to distinguish adaptive introgression from classic selective sweeps in both real and simulated data [21,22]. This limitation may stem from its reliance on a specific demographic model that assumes a beneficial allele first undergoes fixation via a selective sweep in the source population before introgressing into the target population, where it sweeps again. A recent study suggests that VolcanoFinder performs well only under conditions of strong selection and high divergence between source and target populations [22].

Probabilistic models can further capture gene tree discordance with fine resolution. For example, **ancestral recombination graphs** (ARGs) provide a more detailed representation of ancestry and

can be interpreted as a sequential model of local genealogies along the genome, which is also known as a tree sequence [23,24]. ARGweaver-D is a recent method that applies ARGs to detect introgressed loci. In addition to identifying Neanderthal and Denisovan introgressed loci in modern human genomes, it also detects putative introgressed loci from a super-archaic ghost population into Denisovans [25]. However, its Bayesian framework results in high computational cost, limiting its scalability to larger sample sizes and more complex demographic models. Although recent methods have improved the scalability of ARG inference [23], most are designed to reconstruct genealogies rather than directly detecting introgressed loci, which requires additional implementation to extract such signals.

While ARGs explicitly represent the full genealogical history, hidden Markov models (HMMs)—a classical probabilistic model for sequential data in **machine learning** [26]—treat ancestry as a latent state to be inferred from observed sequences. HMMs have long been used for **local ancestry inference** (LAI) [27]; early applications for detecting introgressed loci adapted these methods, which rely on a reference panel (i.e., the source population in this review) to label ancestry [28]. However, they are unsuitable for detecting ghost introgression where the source population is unavailable and their performance may be also questionable when the source population is represented by only a few samples. More recent work has designed HMMs specifically for introgression detection, including applications without source populations, in multi-source scenarios, and using low-coverage, contaminated sequencing data, as well as for jointly inferring adaptive introgression and the strength of natural selection acting on the introgressed loci [29–33].

Supervised Learning-based Methods

The emergence of machine learning, particularly **deep learning**, reflects a broader trend across disciplines, including genetics [34]. As large-scale genomic data become available for an expanding range of populations and taxa, traditional model-based methods increasingly struggle with both processing feasibility and the challenge of constructing detailed models for each group. Moreover, as sample sizes and variant densities grow, the curse of dimensionality may further limit the effectiveness of traditional methods [26]. As these approaches do not scale well, data-driven alternatives that forgo explicit mechanistic models of the evolutionary process are gaining importance for their scalability and flexibility in analyzing high-dimensional genomic data.

Supervised learning, a machine learning paradigm that maps inputs to labeled outputs, has attracted growing interest in population genetics method development [35]. For detecting introgressed loci, current approaches comprise two main groups: those that use predefined summary statistics, such as the ones discussed above, as input features, and those that automatically extract features from raw data. These methods frame the detection of introgressed loci as a classification problem, aiming to determine whether a given genomic region, variant, or allele is introgressed. Methods such as ArchIE, FILET, and MaLAdapt belong to the first group, employing **logistic regression** or **Extra-Trees classifiers** [36–38]. In contrast, genomatnn, ERICA, and IntroUNET represent the second group, using **convolutional neural networks** to directly learn from genotype matrices [39–41].

A major challenge for applying supervised learning in evolutionary biology is the lack of labeled data, that is, ground truth indicating whether a locus is introgressed. To overcome this limitation, simulated datasets are used to train machine learning models, which are then applied to empirical data for prediction (Figure 2C). Although simulated data may not perfectly reflect reality, several supervised methods have demonstrated promising results. For example, ArchIE has been shown to surpass S^* and S' by incorporating genetic distance between genomes from the reference and target populations [36]. Similarly, MaLAdapt outperforms the U statistic, the Q statistic, and VolcanoFinder under a Neanderthal introgression model, and have revealed novel candidates for adaptive introgression in modern human populations [38].

By intersecting introgressed regions predicted by summary statistic-based outlier detection, MaLAdapt, and genomatnn, circadian loci have been implicated as adaptively introgressed from archaic hominins [42]. Moreover, genomatnn, MaLAdapt, and the U and Q statistics all support *BNC2*, a gene associated with human pigmentation and previously identified as a target of positive selection in modern Europeans [43,44], whereas VolcanoFinder does not detect such a signal [20]. While ERICA employs deep learning, it shares a core principle with Twisst by predicting the proportions of gene tree topologies within a genomic region to identify introgressed loci through gene tree discordance [40].

Among these approaches, IntroUNET is particularly interesting, as it frames the identification of introgressed alleles from the ghost population as a **semantic segmentation** task, which is a fundamental problem in modern compute vision [45], and thus, in principle, enables high-

resolution predictions at the allele level (Figure 3A), which may be difficult to achieve using other approaches. This capability may be valuable for precisely delineating the boundaries of introgressed segments.

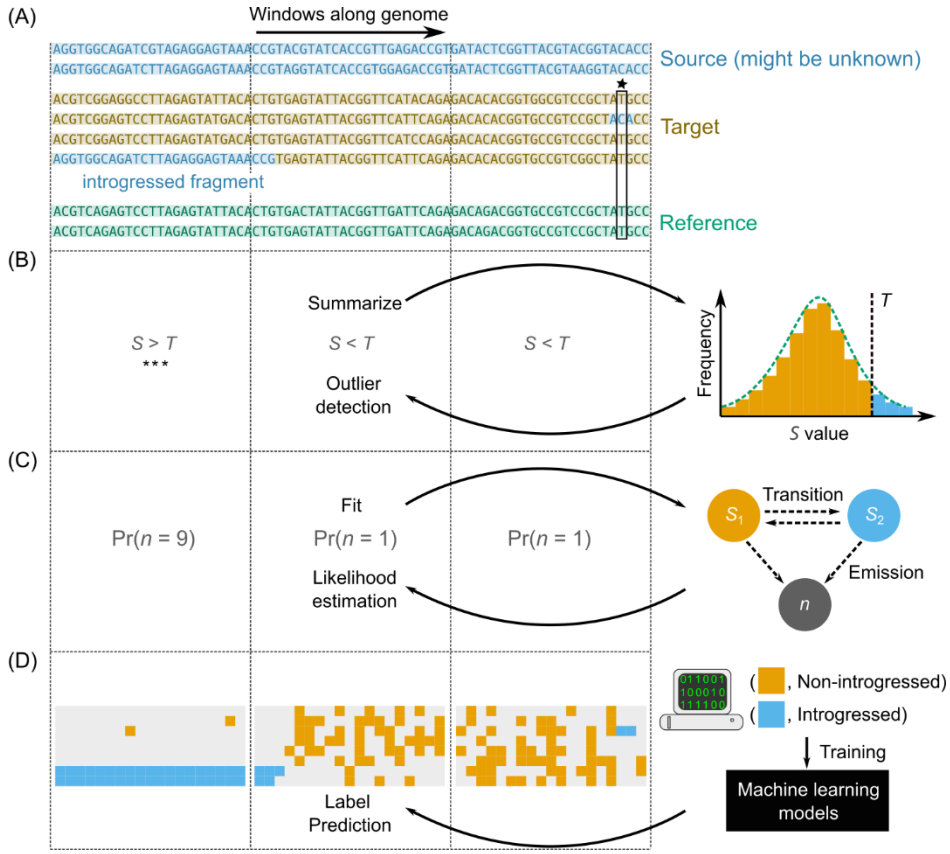

Figure 2. Conceptual overview of computational approaches for detecting introgressed loci. (A) Underlying genomic data from different individuals representing different populations, where variants can be observed within sliding windows along the genome. The star denotes a private variant that is observed in genomes of the target population and absent in the reference population. (B) Summary statistics-based methods summarize genomic information into statistic values (*S*) from the reference and target genomes, and optionally from a source genome. They typically apply outlier detection to identify putative introgressed loci based on a threshold (*T*). An outlier is highlighted with three asterisks. (C) Probabilistic model-based methods describe how the data are generated under a probabilistic framework, based on various strategies to determine different patterns. For example, an HMM represents transitions between hidden states, where S_1 denotes the non-introgressed state and S_2 denotes the introgressed state. The model defines how these states emit observations, such as number of private variants, enabling likelihood estimation and model fitting to

observed data. (D) Supervised learning-based methods rely on labeled training data to learn models that predict the introgression status of regions or variants in target genomes. For example, in genotype matrices of ancestral (grey) and derived (colored) alleles, an artificial neural network predicts whether alleles are introgressed or non-introgressed. Labels are typically generated from computer simulations.

Recent Applications beyond Humans

Recent methodological innovations have provided a variety of tools for decoding the genomic landscapes of introgression beyond humans (Figure 3B). A wide range of taxa have been investigated with summary statistics-based methods [46–55]. These approaches are now widely used thanks to recent implementations like Dsuite [56], and a comprehensive survey is beyond scope. Still, recent studies offer notable examples of adaptive introgression: f_{dM} plus selection scans identified flowering-time genes in *Brassica napus*; and the *U* and *Q* statistics detected sperm function genes in sticklebacks and high-altitude candidates in Tibetan cattle [46,49,54]. These findings highlight recurrent targets among genes involved in key biological functions.

Probabilistic model-based approaches are also extensively utilized in non-human species, demonstrating their versatility across diverse introgression scenarios. For example, IBDmix has inferred introgressed fragments in baboons and bears [58–60]. hmmix has identified loci from ghost introgression in orcas, canids, and the extinct Columbian mammoth, and in great apes when combined with the *S** statistic, underscoring the benefits of integrative analyses [3,50,53,61–63]. In Tibetan canids, the high-altitude adaptation gene *EPAS1*, previously linked to Denisovan introgression in humans, may also derive from a ghost lineage [62,64]. admixfrog, leveraging both ancient and modern genomes, has estimated ancestry proportions in ancient bears, detected immunity-related introgression in Alpine ibex, and resolved fine-scale introgression patterns in chimpanzees using non-invasive fecal samples resembling degraded ancient DNA [65–67]. For adaptive introgression, VolcanoFinder has identified candidate loci from ghost lineages in gorillas, hickories, and pigs, associated with bitter taste perception, defense response, and commercial traits, respectively [3,53,68]. AHMM-S has detected insecticide-resistance loci in fruit flies, though its multi-locus extension, AHMM-MLS, suggests AHMM-S may overestimate selection coefficients [32,33].

To date, supervised learning-based methods have been applied to a limited number of non-human taxa, such as fruit flies, butterflies, and rice [37,40,41]. In these applications, the primary goal

has been to assess whether model predictions align with results from previous approaches, rather than to investigate new biological questions. Nonetheless, ERICA has identified multiple domestication-related loci in rice as candidates for adaptive introgression. This limited scope is likely due to the technical complexity of machine learning, as most implementations are tailored to specific datasets and are not easily applied beyond their original training context, reflecting broader challenges in software development within evolutionary biology [69,70]. As population-scale genomic datasets and machine learning algorithms continue to develop, broader application across diverse lineages is expected. Such efforts will help refine our understanding of introgression landscapes and population interactions throughout evolutionary history.

Beyond analyzing empirical datasets, some studies have explored how different approaches perform in non-human scenarios using simulated data. For example, both S^* and S' perform well under a Neanderthal introgression model, but only S^* remains effective in a bonobo ghost introgression scenario [18]. This difference may be due to its recent implementation, sstar, which provides a flexible computational framework applicable to diverse demographic scenarios, whereas SPrime, the implementation of S', is hard-coded with parameters specific to an out-of-Africa Neanderthal-admixture model [18,71]. Furthermore, a recent study suggests that the Q statistic performs comparably or better than genomatnn, MaLAdapt, and VolcanoFinder under non-human demographic models, indicating that summary statistics continue to be valuable even when more advanced methods are available [22,39].

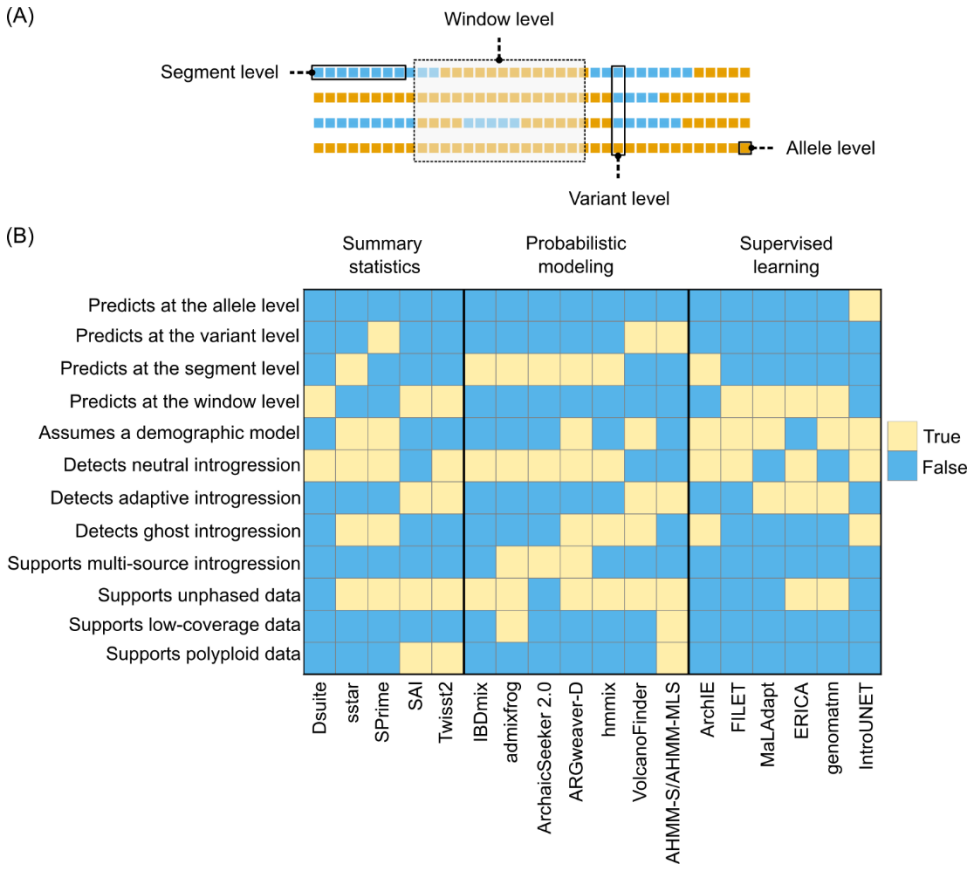

Figure 3. Methodological features of different open-source implementations. (A) Different prediction levels across genomes refer to the resolution at which an introgressed label is assigned. Window level: fixed-length genomic windows, defined by base pairs or by number of variants, are typically summarized across all variants and samples without individual-level resolution. Segment level: continuous introgressed haplotypes of varying lengths are inferred for each individual genome. Variant level: individual genomic variants, such as single nucleotide polymorphisms, are classified as introgressed or not. Allele level: the status of each allele at a segregating site is individually determined. (B) Feature assessment of implementations representing different methodological approaches. Implementations with summary statistics-based methods include Dsuite (https://github.com/millanek/Dsuite) for the d_f, f_d, and f_{dM} statistics; sai (https://github.com/xin-huang/sai) for the U and Q statistics; sstar (https://github.com/xin-huang/sstar) for the S^* statistic; SPrime (https://github.com/browning-lab/sprime) for the S' statistic; and Twisst2 (https://github.com/simonhmartin/twisst2) for topology weighting. Implementations with probabilistic model-based methods comprise IBDmix (https://github.com/PrincetonUniversity/IBDmix), VolcanoFinder

(https://degiorgiogroup.fau.edu/vf.html), ARGweaver-D (http://compgen.cshl.edu/ARGweaver/doc/argweaver-d-manual.html), admixfrog (https://github.com/BenjaminPeter/admixfrog), ArchaicSeeker 2.0 (https://github.com/Shuhua-Group/ArchaicSeeker2.0), AHMM-S (https://github.com/jesvedberg/Ancestry_HMM-S), AHMM-MLS (https://github.com/genicos/ahmm_mls), hmmix (https://github.com/LauritsSkov/Introgression-detection). Implementations with supervised learning-based methods encompass ArchIE (https://github.com/sriramlab/ArchIE), FILET (https://github.com/kr-colab/FILET), MaLAdapt (https://github.com/xzhang-popgen/maladapt), ERICA (https://github.com/YuboZhangPKU/ERICA), genomatnn (https://github.com/grahamgower/genomatnn), IntroUNET (https://github.com/SchriderLab/introNets). For methods with multiple implementations such as the d_f, f_d, and S^* statistics, whether by the same or different authors, only the most recent version is assessed. "Assumes a demographic model" refers to using a specific model to tune parameters (SPrime), condition inference (ARGweaver-D and VolcanoFinder), or simulate training data (sstar and supervised learning-based methods).

Challenges

Despite substantial methodological progress, decoding genomic landscapes of introgression is still fraught with challenges due to confounding factors, model misspecification, and analysis opacity that can bias or obscure inference. One major source of confounding arises from evolutionary processes that mimic the genomic signatures of introgression. For example, **incomplete lineage sorting** (ILS) can produce gene tree discordance similar to that expected under introgression. Although the *D* statistic is expected to distinguish ILS and introgression at the whole-genome level, this is not the case at the locus level [9]. Also, population structure in unsampled or ancestral lineages can generate spurious signals resembling ghost introgression, even in the absence of gene flow [72]. A long-standing debate in human evolution concerns whether ghost introgression occurred in African populations, with different conclusions from different demographic inference approaches, although one study nonetheless applied ArchIE to examined putatively introgressed fragments [73–75]. Furthermore, **long-term balancing selection** may give rise to patterns that resemble adaptive introgression [76,77].

Another challenge is model misspecification, which can arise in several forms. First, some methods embed rigid assumptions in their model design. For instance, hmmix assumes the

presence of both introgressed and non-introgressed hidden states in the data [29]. If introgression is absent, the model may nonetheless infer false signals simply by fitting its predefined state structure. Second, methods like SPrime or VolcanoFinder often assume simplified or idealized demographic models. Such assumptions may not hold in more complex or less ideal evolutionary scenarios, potentially limiting applicability. Third, in supervised learning, performance may degrade if the demographic model used to simulate training data differs substantially from the test scenario [78]. Finally, some deep learning architectures, such as convolutional neural networks, have architectural constraints such as requiring fixed input shapes and being sensitive to the order of input samples, which may limit their flexibility across various datasets [34].

In practice, a third challenge is analysis opacity. While current critiques of machine learning often focus on their interpretability [79], a lack of transparency can also arise from the analysis procedure, including undocumented preprocessing steps, hard-coded parameters, or discrepancies between published methods and their actual implementations [57,70,77]. These issues, not only relevant for machine learning [57], frequently force researchers to inspect source code directly to verify correctness, thereby impeding reproducibility and slowing scientific progress. For example, the performance of IntroUNET may be affected by training data that inadvertently retained information of polymorphic sites from an unintended fourth population, as it reused the demographic model from ArchIE for training, which was described as a three-population model but in fact included a fourth, and by repeated training datasets caused by unexpected behavior in its modified simulator [70,77]. This challenge can be addressed through transparent reporting, reproducible workflows, and community standards. Furthermore, the lack of accessible and robust implementations has hindered consistent benchmarking, which is essential for method evaluation. For example, the performance of the Q statistic differs between two studies [22,38]. The lack of standardized Q statistic implementations and transparent documentation makes it difficult to determine whether discrepancies result from implementation, data processing, or demographic models. Moreover, studies have used different approaches to detect adaptive introgression, either by combining introgression signals with selection scans or by applying dedicated methods [21,38]. However, it is still uncertain which approach performs best, or under what circumstances each should be applied. The recent emergence of machine learning benchmarks provides a valuable reference and highlights the importance of structured

comparison and shared standards when evaluating the performance of different methods (https://iclr.cc/virtual/2024/invited-talk/21799).

Outlook

As introgression detection continues to mature as a methodological field, future progress will rely on extending current approaches—including statistical, probabilistic, and supervised learning-based methods—to accommodate increasingly complex evolutionary scenarios and data types. Owing to their simplicity, interpretability, and computational efficiency, summary statistics-based methods remain attractive. In particular, those incorporating **linkage disequilibrium** (LD) information may prove useful in more complex cases. However, no existing summary statistic can currently distinguish loci resulting from multi-source introgression (Figure 3B). Extending such approaches toward locus-level detection, especially under selection or multiple pulses of gene flow, remains an open and important challenge.

By modeling introgression under explicitly defined evolutionary scenarios, probabilistic methods allow key processes such as mutation, recombination, and natural selection to be incorporated into a unified framework. This capacity enables not just detection of introgressed loci but also quantitative characterization of their properties, such as estimating the age of introgressed variants, the length distribution of introgressed fragments, or the strength of selection acting on them. Extending these probabilistic models to decode genomic landscapes of introgression shaped by multiple evolutionary forces continues to be a key area of development [58,80,81]. Further efforts may focus on improving scalability to large datasets and enhancing robustness to model misspecification.

Supervised learning, and machine learning more broadly, holds great potential. In principle, these approaches can integrate diverse data types and achieve high-resolution predictions. However, current applications require high-quality input data and do not support polyploid datasets (Figure 3B). While other approaches are also limited in this regard, machine learning approaches, being data-driven, may offer greater flexibility for accommodating such complexities in the future. Extending these implementations to support low-coverage data and to unify the analysis of neutral, adaptive, ghost, and multi-source introgression would be highly desirable. Most importantly, such implementations should be accessible to the community and not tailored to a specific species. Beyond model performance, software engineering is critical for ensuring that

machine learning methods are reproducible, maintainable, and broadly usable across datasets and research groups.

Another open question is how well machine learning models can generalize across demographic conditions. Although supervised learning-based methods typically employ simulations to generate labeled training data but indeed make no demographic assumptions during inference. Performance degradation under mismatched test scenarios is likely caused by current implementations relying on training data simulated under a specific demographic model, which may lead to overfitting. To mitigate this, training on a diverse set of demographic models may improve generalizability, allowing models to integrate signals from multiple evolutionary processes. For example, ERICA was trained on data with a range of ILS and gene flow settings, which makes it adaptable to diverse scenarios, although it may underperform compared to models trained under the exact test scenario [40].

Alongside supervised learning, **unsupervised learning** and **self-supervised learning** also show promise, as these paradigms do not use labeled data and therefore avoid the requirement to generate simulated training data. For instance, outlier detection, which is central to summary statistics-based methods, can be naturally extended using **deep generative models** such as **variational autoencoders** [82]. Similarly, recent extensions of LAI have incorporated deep learning architectures [27,34], raising the possibility that these approaches could be repurposed for detecting introgressed loci. Additionally, recent trends in genetics involve developing **genomic language models** using self-supervised learning [83,84]. This paradigm has demonstrated strong generalization ability across diverse scenarios and could help improve the robustness and transferability of machine learning models for detecting introgressed loci [85].

The integration of different methodological paradigms also presents an important opportunity. For example, summary statistics-based and probabilistic model-based approaches, which are typically interpretable and grounded in explicit evolutionary assumptions, can serve as valuable baselines for assessing the performance and reliability of emerging machine learning-based methods, while also helping to improve interpretability and robustness. Furthermore, developing standardized benchmark datasets that span a range of demographic scenarios and evolutionary processes will be crucial for systematic comparisons across methods. Such integrative efforts will not only support methodological advancement but also accelerate biological discovery.

Concluding Remarks

The detection of introgressed loci has evolved into a diverse methodological field encompassing summary statistics, probabilistic modeling, and supervised learning. Each class of methods offers distinct advantages: summary statistics remain efficient and interpretable for initial scans, probabilistic modeling enables principled inference under explicit evolutionary assumptions, and machine learning offer scalability and potential for discovering patterns beyond predefined models. Rather than converging on a single best method, future progress will likely depend on leveraging the complementarity between approaches, improving transparency and benchmarking, and developing tools that are robust to real-world complexity (see Outstanding questions). As evolutionary datasets continue to expand in scale and scope, refining the approaches for decoding genomic landscapes of introgression is essential for understanding how gene flow has shaped genomes across the tree of life.

Outstanding Questions

- Are there shared patterns of introgression landscapes across species that could inform general evolutionary principles?
- How can the biological interpretability and analytical transparency of introgressed loci inferred by complex models, especially those using machine learning, be improved?
- How can the potential of machine learning, including genomic language models, be fully leveraged to decode introgression landscapes, and under what conditions do these approaches outperform traditional methods?
- How can computational tools be developed to be accessible, generalizable across species, and robust under variation in data quality, confounding factors, and model misspecification?
- Can different methods be systematically evaluated under diverse demographic models and confounding factors to clarify performance discrepancies across studies, establish consistent benchmarks, and identify which methods are best suited to specific scenarios for accurate and reliable inference?

Glossary

Adaptive introgression: the transfer of genetic variants from one distant lineage into another via gene flow, followed by natural selection favoring those variants in the recipient population.

Ancestral recombination graph: a graph-based data structure that exhaustively represents the evolutionary relationships between a set of genomes, accounting for both coalescent and recombination events.

Convolutional neural network: a type of artificial neural network that is particularly effective for grid-like data.

Deep generative model: generative models that use deep neural networks to learn and sample from complex data distributions.

Deep learning: a machine learning approach that utilizes deep neural networks to learn hierarchical representation from data.

Extra-Trees classifier: an ensemble machine learning algorithm that builds multiple decision trees using random splits and the entire training dataset to reduce variance and enhance generalization.

Generalized additive model: a statistical model that captures non-linear relationships between inputs and outputs by combining smooth functions additively, while preserving interpretability.

Genomic language model: a machine learning model adapted from natural language processing to investigate genomic problems.

Ghost introgression: gene flow from a population which is not directly represented in genomic data, as either unsampled or extinct but inferred from recipient populations.

Identity-by-descent: identical genomic segments shared among individuals that are inherited from a common ancestor without being broken by recombination.

Incomplete lineage sorting (ILS): the phenomenon where gene trees fail to match the species tree because ancestral polymorphisms are retained and randomly sorted through rapid speciation events.

Introgression: the phenomenon of transferring genetic material across genetically divergent populations.

Linkage disequilibrium (LD): the discrepancy between the probability to observe two alleles at two loci together and the probability to observe them independently.

Local ancestry inference (LAI): the process that determines the ancestral origin of genomic segments along the genome in admixed individuals.

Logistic regression: a statistical model that estimates the probability that a given input belongs to one of two categories using a logistic function.

Long-term balancing selection: a form of natural selection that maintains ancestral genetic variants over long evolutionary timescales, even across speciation events, resulting in trans-species polymorphisms.

Machine learning: an algorithmic approach that automatically learns patterns or structures from data to make predictions or decisions.

Positive selection: a form of natural selection that increases the frequency of beneficial mutations.

Self-supervised learning: a machine learning paradigm that learns patterns or structures from data through automatically generated labels derived from the data itself.

Semantic segmentation: a machine learning task that determines the category of each individual pixel in an image.

Unsupervised learning: a machine learning paradigm that learns patterns or structures from data without using labeled examples.

Variational autoencoder: a deep generative model that learns a probabilistic latent representation of data by combining artificial neural networks with variational inference.

Acknowledgements

X.H. thanks Jie Wang for helpful discussions during the preparation of this manuscript. This project has been funded by the Vienna Science and Technology Fund (WWTF) [10.47379/VRG20001] to M.K.

Resources List

[i] https://github.com/millanek/Dsuite

[ii] https://github.com/xin-huang/sai

[iii] https://github.com/xin-huang/sstar

[iv] https://github.com/browning-lab/sprime

[v] https://github.com/simonhmartin/twisst2

[vi] https://github.com/PrincetonUniversity/IBDmix

[vii] https://degiorgiogroup.fau.edu/vf.html

[viii] http://compgen.cshl.edu/ARGweaver/doc/argweaver-d-manual.html

[ix] https://github.com/BenjaminPeter/admixfrog

[x] https://github.com/Shuhua-Group/ArchaicSeeker2.0

[xi] https://github.com/jesvedberg/Ancestry_HMM-S

[xii] https://github.com/genicos/ahmm_mls

[xiii] https://github.com/LauritsSkov/Introgression-detection

[xiv] https://github.com/sriramlab/ArchIE

[xv] https://github.com/kr-colab/FILET

[xvi] https://github.com/xzhang-popgen/maladapt

[xvii] https://github.com/YuboZhangPKU/ERICA

[xviii] https://github.com/grahamgower/genomatnn

[xix] https://github.com/SchriderLab/introNets

[xx] https://iclr.cc/virtual/2024/invited-talk/21799

Declaration of interests

The authors declare no conflict of interests.

Declaration of generative AI in scientific writing

During the preparation of this work the authors used ChatGPT in order to edit the language. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

1. Hibbins, M.S. and Hahn, M.W. (2022) Phylogenomic approaches to detecting and characterizing introgression. *Genetics* 220, iyab173.
2. Ongaro, L. and Huerta-Sanchez, E. (2024) A history of multiple Denisovan introgression events in modern humans. *Nat. Genet.* 56, 2612–2622.
3. Zhang, W. *et al.* (2024) Uncovering ghost introgression through genomic analysis of a distinct eastern Asian hickory species. *Plant J.* 119, 1386–1399.
4. Wang, G. *et al.* (2025) Genomic evidence for hybridization and introgression between blue peafowl and endangered green peafowl and molecular foundation of leucistic plumage of blue peafowl. *GigaScience* 14, giae124.
5. North, H.L. *et al.* (2024) Rapid adaptation and interspecific introgression in the North American crop pest *Helicoverpa zea*. *Mol. Biol. Evol.* 41, msae129.
6. Patterson, N. *et al.* (2012) Ancient admixture in human history. *Genetics* 192, 1065–1093.
7. Martin, S.H. *et al.* (2015) Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Mol. Biol. Evol.* 32, 244–257.
8. Pfeifer, B. and Kapan, D.D. (2019) Estimates of introgression as a function of pairwise distances. *BMC Bioinform.* 20, 207.
9. Fang, L.L. *et al.* (2024) Leveraging shared ancestral variation to detect local introgression. *PLoS Genet.* 20, e1010155.
10. Malinsky, M. *et al.* (2015) Genomic islands of speciation separate cichlid ectomorphs in an East African crater lake. *Science* 350, 1493–1498.
11. Racimo, F. *et al.* (2017) Signatures of archaic adaptive introgression in present-day human populations. *Mol. Biol. Evol.* 34, 296–317.
12. Plagnol, V. and Wall, J.D. (2006) Possible ancestral structure in human populations. *PLoS Genet.* 2, e105.

13. Vernot, B. and Akey, J.M. (2014) Resurrecting surviving Neanderthal lineages from modern human genomes. *Science* 343, 1017–1021.
14. Vernot, B. *et al.* (2016) Excavating Neanderthal and Denisovan DNA from the genomes of Melanesian individuals. *Science* 352, 235–239.
15. Browning, S.R. *et al.* (2018) Analysis of human sequence data reveals two pulses of archaic Denisovan admixture. *Cell* 173, 53–61.
16. Martin, S.H. and van Belleghem, S.M. Exploring evolutionary relationships across the genome using topology weighting. *Genetics* 206, 429–438.
17. De-Kayne, R. *et al.* (2024) Incomplete recombination suppression fuels extensive haplotype diversity in a butterfly color pattern supergene. *bioRixv*, Published online July 26, 2024. https://doi.org/10.1101/2024.07.26.605145
18. Huang, X. *et al.* (2022) sstar: a Python package for detecting archaic introgression from population genetic data with *S**. *Mol. Biol. Evol.* 39, msac212.
19. Chen, L. *et al.* (2020) Identifying and interpreting apparent Neanderthal ancestry in African individuals. *Cell* 180, 677–687.e16.
20. Setter, D. *et al.* (2020) VolcanoFinder: genomic scans for adaptive introgression. *PLoS Genet.* 16, e1008867.
21. Moest, M. *et al*. (2020) Selective sweeps on novel and introgressed variation shape mimicry loci in a butterfly adaptive radiation. *PLoS Biol.* 18, e3000597.
22. Romieu, J. *et al.* (2024) Performance evaluation of adaptive introgression classification methods. *bioRixv*, Published online June 14, 2024. https://doi.org/10.1101/2024.06.12.598278
23. Nielsen, R. *et al.* (2025) Inference and applications of ancestral recombination graphs. *Nat. Rev. Genet.* 26, 47–58.
24. Wong, Y. *et al.* (2024) A general and efficient representation of ancestral recombination graphs. *Genetics* 228, iyae100.
25. Hubisz, M.J. *et al.* (2020) Mapping gene flow between ancient hominis through demography-aware inference of the ancestral recombination graph. *PLoS Genet.* 16, e1008895.
26. Bishop, C.M. (2006) *Pattern Recognition and Machine Learning*. Springer.
27. Sun, Q. *et al*. (2025) Opportunities and challenges of local ancestry in genetic association analyses. *Am. J. Hum. Genet.* 112, 727–740.

28. Sankararaman, S. (2020) Methods for detecting introgressed archaic sequences. *Curr. Opin. Genet. Dev.* 62, 85–90.
29. Skov, L. *et al.* (2018) Detecting archaic introgression using an unadmixed outgroup. *PLoS Genet.* 14, e1007641.
30. Peter, B.M. (2020) 100,000 years of gene flow between Neandertals and Denisovans in the Altai mountains. *bioRixv*, Published online, March 15, 2020. https://doi.org/10.1101/2020.03.13.990523
31. Yuan, K. *et al.* (2021) Refining models of archaic admixture in Eurasia with ArchaicSeeker 2.0. *Nat. Commun.* 12, 6232.
32. Svedberg, J. *et al.* (2021) Inferring adaptive introgression using hidden Markov model. *Mol. Biol. Evol.* 38, 2152–2165.
33. Ayala, N.M. *et al.* (2023) Inferring multi-locus selection in admixed populations. *PLoS Genet.* 19, e1011062.
34. Huang, X. *et al.* (2024) Harnessing deep learning for population genetic inference. *Nat. Rev. Genet.* 25, 61–78.
35. Schrider, D.R. and Kern, A.D. (2018) Supervised machine learning for population genetics: a new paradigm. *Trends. Genet.* 34, 301–312.
36. Durvasula, A. and Sankararaman, S. (2019) A statistical model for reference-free inference of archaic local ancestry. *PLoS Genet.* 15, e1008175.
37. Schrider, D.R. *et al.* (2018) Supervised machine learning reveals introgressed loci in the genomes of *Drosophila simulans* and *D. sechellia*. *PLoS Genet.* 14, e1007341.
38. Zhang, X. *et al.* (2023) *MaLAdapt* reveals novel targets of adaptive introgression from Neanderthals and Denisovans in worldwide human populations. *Mol. Biol. Evol.* 40, msad001.
39. Gower, G. *et al.* (2021) Detecting adaptive introgression in human evolution using convolutional neural networks. *eLife* 10, e64669.
40. Zhang, Y. *et al.* (2023) Inferring historical introgression with deep learning. *Syst. Biol.* 72, 1013–1038.
41. Ray, D.D. *et al.* (2024) IntroUNET: identifying introgressed alleles via semantic segmentation. *PLoS Genet.* 20, e1010657.

42. Velazquez-Archelay, K. *et al.* (2023) Archaic introgression shaped human circadian traits. *Genome Biol. Evol.* 15, evad203.
43. Cheng, J.Y. *et al.* (2021) Detecting selection in multiple populations by modeling ancestral admixture components. *Mol. Biol. Evol.* 39, msab294.
44. Huang, X. *et al.* (2021) Dissecting dynamics and differences of selective pressures in the evolution of human pigmentation. *Biol. Open* 10, bio056523.
45. Hao, S. *et al.* (2020) A brief survey on semantic segmentation with deep learning. *Neurocomputing* 406, 302–321.
46. Wang, T. *et al.* (2023) Interploidy introgression shaped adaptation during the origin and domestication history of *Brassica napus*. *Mol. Biol. Evol.* 40, msad199.
47. Stone, B.W. and Wessinger, C.A. (2024) Ecological diversification in an adaptive radiation of plants: the role of *de novo* mutation and introgression. *Mol. Biol. Evol.* 41, msae007.
48. Coughlan, J.M. *et al.* (2022) Patterns of population structure and introgression among recently differentiated *Drosophila melanogaster* populations. *Mol. Biol. Evol.* 39, msac223.
49. Feng, X. *et al.* (2024) Secondary contact, introgressive hybridization, and genome stabilization in sticklebacks. *Mol. Biol. Evol.* 41, msae031.
50. Kuhlwilm, M. *et al.* (2019) Ancient admixture from an extinct ape lineage into bonobos. *Nat. Ecol. Evol.* 3, 957–965.
51. Evans, B.J. *et al.* (2021) Mitonuclear interactions and introgression genomics of macaque monkeys (*Macaca*) highlight the influence of behaviour on genome evolution. *Proc. R. Soc. B Biol. Sci.* 288, 20211756.
52. Jensen, A. *et al.* (2023) Complex evolutionary history with extensive ancestral gene flow in an African primate radiation. *Mol. Biol. Evol.* 40, msad247.
53. Pawar, H. *et al.* (2023) Ghost admixture in eastern gorillas. *Nat. Ecol. Evol.* 7, 1503–1514.
54. Lyu, Y. *et al.* (2024) Recent selection and introgression facilitated high-altitude adaptation in cattle. *Sci. Bull.* 69, 3415–3424.
55. van der Valk, T. *et al.* (2024) Comparative genomic analyses provide new insights into evolutionary history and conservation genomics of gorillas. *BMC Ecol. Evol.* 24, 14.
56. Malinsky, M. *et al.* (2021) Dsuite – Fast D-statistics and related admixture evidence from VCF files. *Mol. Ecol. Resour.* 21, 584–595.

57. Huang, X. *et al.* (2025) SAI: a Python package for statistics for adaptive introgression. *bioRixv*, Published online April 22, 2025. https://doi.org/10.1101/2025.04.19.649497
58. Vilgalys, T.P. *et al.* (2022) Selection against admixture and gene regulatory divergence in a long-term primate field study. *Science* 377, 635–641.
59. Wang, M.S. *et al.* (2022) A polar bear paleogenome reveals extensive ancient gene flow from polar bears into brown bears. *Nat. Ecol. Evol.* 6, 936–944.
60. Puckett, E.E. (2025) Phylogeography of introgression: spatial and temporal analyses identify two introgression events between brown and American black bears. *Heredity*, Published online April 19, 2025. https://doi.org/10.1038/s41437-025-00762-0
61. Foote, A.D. *et al.* (2019) Killer whale genomes reveal a complex history of recurrent admixture and vicariance. *Mol. Ecol.* 28, 3427–3444.
62. Wang, M.S. *et al.* (2020) Ancient hybridization with an unknown population facilitated high-altitude adaptation of canids. *Mol. Biol. Evol.* 37, 2616–2629.
63. van der Valk, T. *et al.* (2021) Million-year-old DNA sheds light on the genomic history of mammoths. *Nature* 591, 265–269.
64. Huerta-Sánchez, E. *et al.* (2014) Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. *Nature* 512, 194–197.
65. Segawa, T. *et al.* (2024) The origins and diversification of Holarctic brown bear populations inferred from genomes of past and present populations. *Proc. R. Soc. B* 291, 20232411.
66. Münger, X. *et al.* (2024) Facilitated introgression from domestic goat into Alpine ibex at immune loci. *Mol. Ecol.* 33, e17429.
67. Fontsere, C. *et al.* (2022) Population dynamics and genetic connectivity in recent chimpanzee history. *Cell Genom.* 2, 100133.
68. Peng, Y. *et al.* (2023) Distinct traces of mixed ancestry in western commercial pig genomes following gene flow from Chinese indigenous breeds. *Front. Genet.* 13, 1070783.
69. Nekrutenko, A. *et al.* (2018) Biology needs evolutionary software tools: let's build them right. *Mol. Biol. Evol.* 35, 1372–1375.
70. Huang, X. (2024) Developing machine learning applications for population genetic inference: Ensuring precise terminology and robust implementation. *EcoEvoRixv*, Pulished online September 20, 2024. https://doi.org/10.32942/X2N90M

71. Zhou, Y. and Browning, S.R. (2021) Protocol for detecting introgressed archaic variants with SPrime. *STAR Protoc.* 2, 100550.
72. Tournebize, R. and Chikhi, L. (2025) Ignoring population structure in hominin evolutionary models can lead to the inference of spurious admixture events. *Nat. Ecol. Evol.* 9, 225–236.
73. Durvasula, A. and Sankararaman, S. (2020) Recovering signals of ghost archaic introgression in African populations. *Sci. Adv.* 6, eaax5097.
74. Fan, S. *et al.* (2023) Whole-genome sequencing reveals a complex African population demographic history and signatures of local adaptation. *Cell* 186, 923–939.e14.
75. Ragsdale, A.P. *et al*. (2023) A weakly structured stem for human origins in Africa. *Nature* 617, 755–763.
76. Hedrick, P.W. (2013) Adaptive introgression in animals: examples and comparison to new mutation and standing variation as sources of adaptive variation. *Mol. Ecol.* 22, 4606–4618.
77. Hackl, J. and Huang, X. (2025) Revisiting adaptive introgression at the HLA genes in Lithuanian genomes with machine learning. *Infect. Genet. Evol.* 127, 105708.
78. Mo, Z. and Siepel, A. (2023) Domain-adaptive neural networks improve supervised machine learning based on simulated population genetic data. *PLoS Genet.* 19, e1011032.
79. Azodi, C.B. *et al.* (2020) Opening the black box: interpretable machine learning for geneticists. *Trends Genet.* 36, 442–455.
80. Duranton, M. and Pool, J.E. (2022) Interactions between natural selection and recombination shape the genomic landscape of introgression. *Mol. Biol. Evol.* 39, msac122.
81. Glasenapp, M.R. and Pogson, G.H. (2024) Selection shapes the genomic landscape of introgressed ancestry in a pair of sympatric sea urchin species. *Genome Biol. Evol.* 16, evae124.
82. Neloy, A.A. and Turgeon, M. (2024) A comprehensive study of auto-encoders for anomaly detection: efficiency and trade-offs. *Mach. Learn. Appl.* 17, 100572.
83. Benegas, G. *et al.* (2025) Genomic language models: opportunities and challenges. *Trends Genet.* 41, 286–302.
84. Consens, M.E. *et al.* (2025) Transformers and genome language models. *Nat. Mach. Intell.* 7, 346–362.

85. Zhou, C. *et al.* (2024) A comprehensive survey on pretrained foundation models: a history from BERT to ChatGPT. *Int. J. Mach. Learn. & Cyber.*, Published online November 24, 2024. https://doi.org/10.1007/s13042-024-02443-6