Syst. Biol. 72(5):1013–1038, 2023

https://doi.org/10.1093/sysbio/syad033
Advance Access Publication May 31, 2023

Inferring Historical Introgression with Deep Learning

Yubo Zhang[1], Qingjie Zhu[2], Yi Shao[2], Yanchen Jiang[1,3], Yidan Ouyang[4], Li Zhang[2,*], and Wei Zhang[1,3,*]

[1]*State Key Laboratory of Protein and Plant Gene Research, Peking-Tsinghua Center for Life Sciences, Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China*
[2]*Chinese Institute for Brain Research, Beijing 102206, China*
[3]*State Key Laboratory of Protein and Plant Gene Research, School of Life Sciences, Peking University, Beijing 100871, China*
[4]*National Key Laboratory of Crop Genetic Improvement and National Centre of Plant Gene Research (Wuhan), Hubei Hongshan Laboratory, Huazhong Agricultural University, Wuhan 430070, China*
**Correspondence to be sent to: Jin-guang Life Sciences Building, No. 5 Yiheyuan Road, Beijing 100871, China and Chinese Institute for Brain Research, Beijing 102206, China; Email: zhangli@cibr.ac.cn, weizhangvv@pku.edu.cn*
Yubo Zhang and Qingjie Zhu contributed equally to this article.
Li Zhang and Wei Zhang jointly supervised this project and contributed equally.

Received 1 November 2022; reviews returned 28 May 2023; accepted 30 May 2023
Associate editor: Manolo Perez

Abstract.—Resolving phylogenetic relationships among taxa remains a challenge in the era of big data due to the presence of genetic admixture in a wide range of organisms. Rapidly developing sequencing technologies and statistical tests enable evolutionary relationships to be disentangled at a genome-wide level, yet many of these tests are computationally intensive and rely on phased genotypes, large sample sizes, restricted phylogenetic topologies, or hypothesis testing. To overcome these difficulties, we developed a deep learning-based approach, named ERICA, for inferring genome-wide evolutionary relationships and local introgressed regions from sequence data. ERICA accepts sequence alignments of both population genomic data and multiple genome assemblies, and efficiently identifies discordant genealogy patterns and exchanged regions across genomes when compared with other methods. We further tested ERICA using real population genomic data from *Heliconius* butterflies that have undergone adaptive radiation and frequent hybridization. Finally, we applied ERICA to characterize hybridization and introgression in wild and cultivated rice, revealing the important role of introgression in rice domestication and adaptation. Taken together, our findings demonstrate that ERICA provides an effective method for teasing apart evolutionary relationships using whole genome data, which can ultimately facilitate evolutionary studies on hybridization and introgression. [Convolutional neural network; deep learning; evolutionary relationship; hybridization; introgression.]

Resolving the relationships among taxa is one of the fundamental tasks in evolutionary biology. The phylogenetic tree model with the assumption of strict bifurcation has been widely used for representing species' evolutionary history, allowing occasional discordance led by incomplete lineage sorting (ILS) or gene flow after divergence (Pamilo and Nei 1988; Currat et al. 2008). Accordingly, several computational algorithms using genetic distance, maximum likelihood, or Bayesian methods have been applied to phylogenetic reconstruction based on alignments of orthologous genes or regions (Yang and Rannala 2012). However, a growing body of literature suggests that genomic signatures of hybridization, the process of interbreeding, may be more common than expected and have significantly shaped the tree of life (Rieseberg 2019). The complex roles of hybridization, which may either accelerate or hinder speciation, or fuel adaptation by providing additional variability relative to new mutations and standing variations, have attracted considerable attention from evolutionary biologists (Arnold 2004; Mallet 2005; Hedrick 2013). Thus, assessing the complex relationships among species with widespread genetic admixture remains a challenging task, which leads to the assumption that strict divergence does not satisfactorily represent the full evolutionary history of an organism.

Accompanying rapid developments in sequencing technologies, a variety of algorithms have been developed and applied to whole genome sequencing data for demography inference, and genome-wide patterns of admixture have been characterized in different organisms, such as beneficial alleles shared between crops and wild relatives (Stewart et al. 2003), adaptive introgression between archaic and modern humans (Racimo et al. 2015) and pervasive hybridization during diversification of *Heliconius* butterflies (Edelman et al. 2019). To summarize, these algorithms fall into four categories: 1) depicting an overall demographic pattern of admixture instead of capturing local patterns by testing coalescent models using maximum likelihood or Bayesian methods such as G-PhoCS (Gronau et al. 2011), Treemix (Pickrell and Pritchard 2012) and PhyloNet (Than et al. 2008); 2) comparing scales of linkage disequilibrium by detecting the structure of haplotypes from fine-scale and sufficient genomic data such as HAPMIX (Price

et al. 2009), ELAI (Guan 2014), S* (Plagnol and Wall 2006), and Sprime (Browning et al. 2018); 3) performing window-based scans and describing relationships quantitatively among focal taxa according to the allele frequencies of given patterns, for example, Patterson's *D*-statistic (Durand et al. 2011), the f_d statistic (Martin et al. 2015), and the symmetry five-taxon analysis, that is, D_{FOIL} (Pease and Hahn 2015). In comparison with model testing and haplotype inference, these statistical tests are less computationally intensive, but their applications are limited, and they are only suitable for samples with specific phylogenetic topologies. In addition, usage of the allele frequencies of biallelic single nucleotide polymorphism (SNP) sites leads to omission of SNP positions and other types of variations in sequences, which are also meaningful for phylogenetic estimation. 4) Combining multiple genomic features using different machine learning models, including conditional random fields (CRF) (Sankararaman et al. 2014), hidden Markov models (HMMs) (Skov et al. 2018), and Extra-Trees classifiers (Schrider et al. 2018). Model training with these methods relies on appropriate and species-specific demographic models, which generally limits their applications to those involving less well-studied taxa with unknown population histories. Therefore, we aimed to develop a method of inferring evolutionary history and gene flow strength with greater accuracy and less computational complexity in comparison with existing methods, while also requiring less prior knowledge. Specifically, the ideal method we expect should yield performance improvements in the following aspects. First, it should allow direct processing of sequence data instead of requiring predefined population genetic statistics or inferred gene trees to reduce information loss in data pre-processing. Second, it should be capable of resolving local introgression signals in genomes with heterogeneous gene flow, which cannot be identified with demography modeling approaches. Third, it should be applicable to model and non-model systems, and it should be robust across taxa. Finally, it should have low computational complexity and should be capable of handling genome-scale data in an acceptable amount of time.

Machine learning, especially deep learning, has provided a powerful framework for feature extraction and classification. With rapid increases in computing power, deep learning has been widely used in computer vision, speech recognition, natural language processing and bioinformatics applications, such as medical image diagnoses and sequence feature recognition (Eraslan et al. 2019). Recently, the remarkable potential of deep learning algorithms for solving population genetic problems with high accuracy has been demonstrated by studies combining deep learning with Approximate Bayesian Computation to infer human history (Mondal et al. 2019), detecting selective sweep and estimating recombination rates (Flagel et al. 2019), and inferring four-taxon phylogenetic topologies (Suvorov et al. 2020; Zou et al. 2020). Unlike conventional population genetic statistics methods, deep learning-based approaches can directly extract features from high-dimensional data, which is an advantageous characteristic for methods used for analyses of genome-wide sequencing data. Some recent works have applied convolutional neural networks to detect local introgression segments in sister *Drosophila* species (Flagel et al. 2019), as well as between ancient and modern humans (Gower et al. 2021). These studies show that deep learning-based algorithms can effectively detect genomic admixture and introgression signals, but practical applications still have some limitations. First, the need to obtain detailed demographic history and gene flow information for the target species remains a challenge for species that are not well characterized. Also, the effects of differences between the demographic scenarios and true models on algorithm performance have not been comprehensively tested. Second, each specific demographic model requires sequence simulation and model training, which both consume computational resources. Third, while classification tasks are used for determining introgression regions, this method may limit the ability of deep learning-based algorithms to resolve more complex evolutionary histories, such as those involving multiple introgressions between different donor-recipient lineages.

Considering these issues, we report our development of a new pipeline for the inference of complex evolutionary history using convolutional neural networks (CNNs), named ERICA (Evolutionary Relationship Inference using a CNN-based Approach). ERICA accepts sequence alignments of both population genomic data and multiple genome assemblies, and can evaluate genome-wide evolutionary relationships, as well as local signatures of introgression. Unlike previous deep learning methods, species-specific demographic scenarios are not used in model training; therefore ERICA is applicable to different taxa, including species lacking detailed population histories. We report the performance of ERICA using both simulated and real genomic data in *Heliconius* butterflies, a classic model system with adaptive radiation and frequent interspecific hybridization. Notably, ERICA showed better performance than other methods in analyses of both simulated and real genomic data. Therefore, ERICA was employed to explore adaptive introgression between wild and domesticated rice, another classic model system with a history of intensive artificial selection and frequent hybridization. We used ERICA to characterize putative signatures of introgression and identify candidate loci of adaptive introgression between Asian cultivated rice *O. sativa* ssp. *japonica* and other tropical accessions. Our results suggest that introgression played a significant role in rice domestication and provide a list of genetic loci related to rice radiation and adaptation to fuel future agricultural research. In summary, ERICA provides an effective method for teasing apart evolutionary relationships using whole genome data, which has the potential to facilitate evolutionary studies involving hybridization and introgression.

Materials And Methods

Design Principles of ERICA

Detecting introgression based on topological discordance.—We aimed to establish a deep learning framework for inferring evolutionary relationships and identifying local introgression signals directly from sequence data. In previous studies, the windows with and without gene flow were distinguished using classification tasks (Flagel et al. 2019; Gower et al. 2021). However, when more than two species were considered, there were multiple potential gene flow events and it was difficult to set a category for each. In such cases, the discordance between the gene trees and the species tree provides a way to identify introgressed regions. Thus, we first focused on evaluating evolutionary relationships among taxa and then scanned the genome for regions supporting alternative topologies. Since the number of possible topologies increases double-factorially with taxa number (Felsenstein 1978), and is, therefore, directly related to computational complexity, we included cases with four and five taxa in the analysis. Compared with the four-taxon case, the five-taxon model produced more possible donor–receptor gene flow pairs, and the direction of some of these pairs was determined. We built two networks to quantify the relationships of four and five taxa from sequence alignments (Fig. 1). Since the relationships inferred by any phylogenetic analysis are hypothesized and approximated, they may differ from the real relationships; therefore, we trained the two models using simulated datasets generated according to predefined evolutionary scenarios. Data generation and encoding, network structure, and model training are described in the following sections.

Phylogenetic relationship encoding for CNN.—Given a four-taxon dataset including three ingroup taxa P1, P2, and P3, and one outgroup taxon O, there are three possible rooted topologies: (((P1, P2), P3), O), (((P1, P3), P2), O), and (((P2, P3), P1), O). In previous studies, the topological structures of taxon were classified into three categories (Suvorov et al. 2020; Zou et al. 2020). However, due to the effects of recombination and population structure, this classification method may not fully represent the evolutionary relationships among taxa. Thus, we adopted a multi-dimensional vector to represent the relative abundance of each possible rooted tree topology (Supplementary Fig. S1a,b), which was derived from quartet sampling (Estabrook et al. 1985) and topology weighting (Martin and Van Belleghem 2017) methods. For example, the vector (0.5, 0, 0.5) was used to label a window containing two segments, with one supporting (((P1, P2), P3), O) and the other supporting (((P2, P3), P1), O), which may represent a recombination breakpoint of two different demographic histories (Supplementary Fig. S1c). When the focal taxon was not monophyletic, the complicated topological structure of each sample was considered. For instance, in the case in which two samples of taxon P2 clustered with P1 and one individual clustered with P3, the vector (0.67, 0, 0.33) was used to represent the relationships of the three focal taxa (Supplementary Fig. S1d). Likewise, a fifteen-dimensional vector satisfying the sum-to-one constraint and corresponding to the 15 possible rooted topological structures was used to label the datasets of the five-taxon model (Supplementary Fig. S1a). This labeling strategy efficiently records information regarding real relationships, especially for genomic regions with different evolutionary histories or non-monophyletic sample groups, allowing the CNN models to evaluate both reference genome assemblies and population-level genetic datasets. The actual phylogenies generated by the coalescent simulator represented more complex scenarios and had topological structures more complex than those of the above examples. We, therefore, quantified the proportion of each topology for each segment lacking recombination with a topology weighting method (Martin and Van Belleghem 2017), which provided a quantitative measurement of the fractions of unique subtrees matching given topologies, and the mean values for all segments were used as the data labels.

Data simulation for CNN model training.—In previous studies, specific evolutionary scenarios of the focal species (e.g., *Drosophila* sister species (Schrider et al. 2018) and ancient and modern humans (Gower et al. 2021)) have been used in data simulation and model training. Thus, the CNN models needed to be retrained before being applied to new taxa, and the application to taxa without detailed population histories was limited. To address complex evolutionary histories and enhance the generalization ability of the CNN models, instead of using species-specific data for model training, we generated a training dataset covering scenarios with varying degrees of ILS and gene flow, including different genealogies, various divergence times, and introgressions between non-sister species (Supplementary Fig. S2).

We generated multiple sequence alignments (MSAs) for training and testing the CNN models using the coalescent simulator ms (Hudson 2002) and Seq-Gen (Rambaut and Grassly 1997). For the training and test datasets, the MSA data were 5000 bp in length and contained eight haplotypes per taxon. According to the simulation studies reported in other introgression detection methods (Supplementary Table S1), we used a population size (N) of 1 M and a recombination rate ($4Nr$) of 0.01, given the per site per generation rate (r) of 2.5×10^{-9} (0.25 cM/Mb).

Demographic scenarios representing all possible topologies were simulated, with species divergence times ranging from 0.2 to 2.7, in units of $4N$ generations. Multiple scenarios of possible introgressions between non-sister species were also included to generate complex population structures and enhance the generalizability of the model (Supplementary Fig. S2a).

Example commands for ms were: ms 32 1 -I 4 8 8 8 8 -ej t_{12} 2 1 -ej t_{123} 3 1 -ej 3 4 1 -r 50 5000 -T (for data

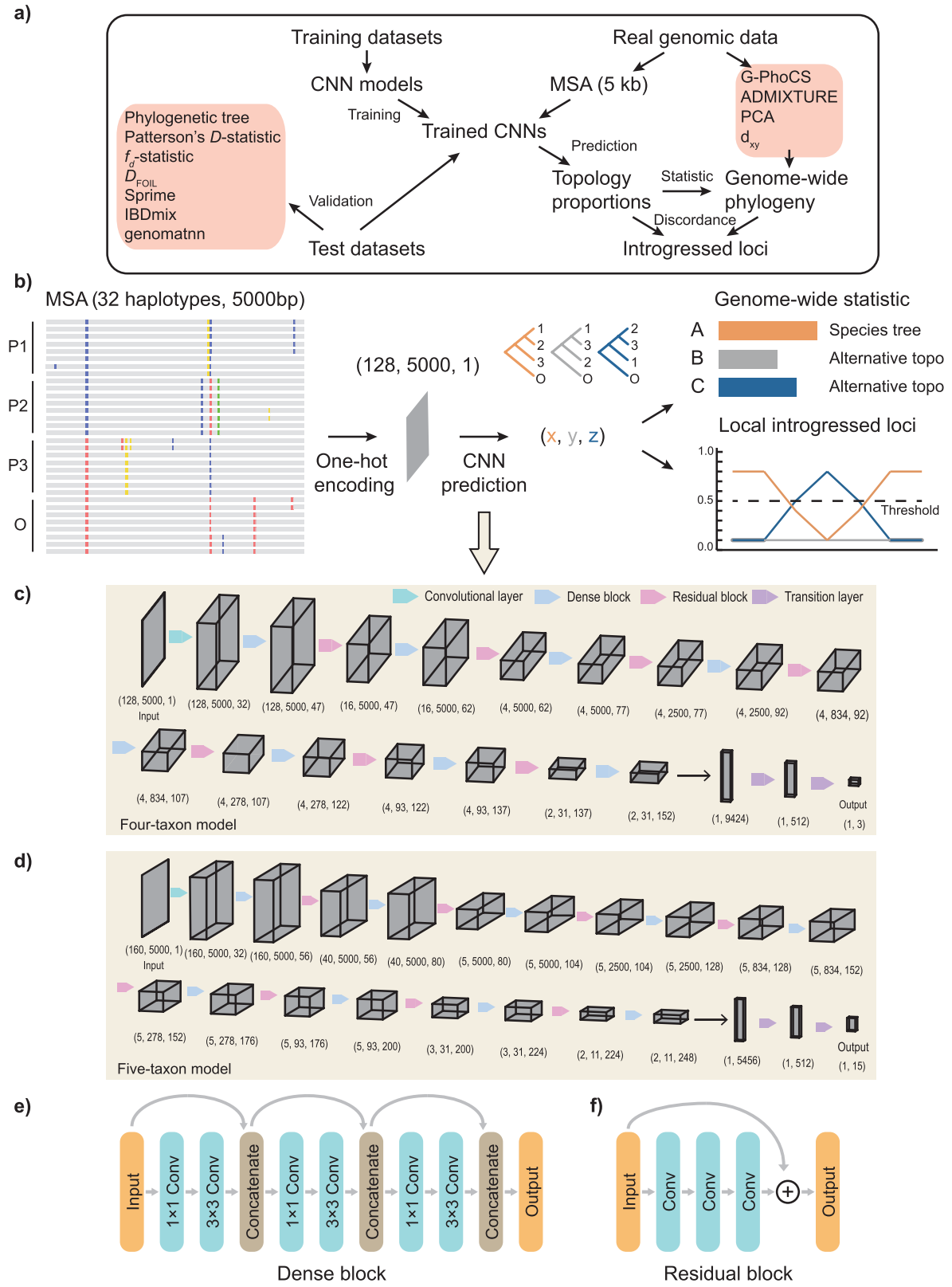


FIGURE 1. ERICA design principles, model architecture, and training. a) The flow diagram of ERICA model training, testing, and applying on simulated and real data. In brief, different simulated datasets were used to train the CNN, and the accuracy of topology inference and introgression detection was evaluated. The trained models were used in the analyses of real genomic data. b) Illustration of the main part of the ERICA pipeline. The multiple sequence alignment (MSA) data were first split into 5-kb windows and then encoded in a one-hot form. A

without admixture); and ms 32 1 -I 4 8 8 8 8 -ej t_{12} 2 1 -ej t_{123} 3 1 -ej 3 4 1 -es t_{GF} 2 0.5 -ej t_{GF} 5 3 -r 50 5000 -T (for scenarios with instantaneous gene flow from P3 to P2). t_{12}, t_{123}, and t_{GF} indicated the split time of (P1, P2), the split time of (P1, P2, P3), and the time of gene flow, respectively. Full commands and parameters are shown in Supplementary Table S2.

The gene trees of the sampled haplotypes were recorded with the "-T" option, and sequence data were simulated using Seq-Gen based on the genealogies, with the Hasegawa–Kishino–Yano (HKY) nucleotide substitution model. The branch length scaling factor was set to 0.01 ($4N\mu$), with a per site per generation substitution rate (μ) of 2.5×10^{-9}. In summary, a total of 120,600 MSAs (19.3 Gb) and 6030 MSAs (0.97 Gb) consisting of data representing 58,125 evolutionary scenarios (Supplementary Table S2) were generated for the four-taxon model as the training and test datasets (D1), respectively, whereas a total of 74,100 MSAs (14.8 Gb) and 7410 MSAs (1.48 Gb) representing 57,783 scenarios were generated as the training and test datasets (D1), respectively, for the five-taxon model (Supplementary Fig. S2b).

Sequence encoding, network architectures, and model training.—Different deep learning frameworks use different encoding models to process sequence data. For example, for the coding model used in genomatnn, Gower et al. (2021) divided a sequence into a fixed number of bins and counted the number of minor alleles in each bin, which solved the problem of genomic windows with different numbers of segregating sites. However, after compressing the information, this coding model cannot distinguish variations at different positions within a bin. For the coding model used in ERICA, we chose to retain the genotype information for every position in a sequence alignment, including sites with or without variations, which also maintained a fixed dimension for the input data. When referring to specific encoding methods, they also included binary encoding (i.e., "0" was assigned to the ancestral allele and "1" was assigned to the derived allele) (Flagel et al. 2019), label encoding (with a different integer value assigned to each nucleotide) (Suvorov et al. 2020), and one-hot encoding (Zou et al. 2020). Considering that binary encoding cannot handle data with multi-allelic and missing sites, and the distances between the four bases of A, T, C, and G were not equal in the label encoding, we used a one-hot format to encode the nucleotides of an input MSA, in which G, T, A, and C were encoded as (1, 0, 0, 0), (0, 1, 0, 0), (0, 0, 1, 0), and (0, 0, 0, 1), respectively, whereas the gap ("–") or missing data ("N") were encoded as (0, 0, 0, 0). For the input MSA used in ERICA, sequences of four or five populations were aligned from top to bottom, with eight haplotypes per population of length 5000 bp. Thus, the input MSA was converted into a 2-dimensional numerical matrix with dimensions of 128 × 5000 (four-taxon model) or 160 × 5000 (five-taxon model). Each column corresponded to a nucleotide position in the alignment, and every four rows corresponded to one haplotype.

We implemented our CNN models in Python using TensorFlow. Multiple dense blocks and residual blocks were combined to build deep neural networks and extract features. Residual Networks (ResNet) were designed for image recognition and have been widely used in deep neural networks. The depth is crucially important for model accuracy, and problems such as vanishing/exploding gradients and model degradation limit the depth increase. The ResNet uses residual functions for parameter learning by adding identity shortcut connections, which are easy to optimize and show greater accuracy than that achieved by simply stacking convolutional layers (He et al. 2016). Another architecture, Dense Convolutional Networks (DenseNet), alleviates the vanishing gradient problem and reduces the number of parameters by directly connecting all the layers in one dense block, such that each layer uses the feature maps of all previous layers as input (Huang et al. 2017). We employed both residual blocks and dense blocks to train deeper, more accurate, and more efficient networks. Eight dense blocks and seven residual blocks were connected one after the other in the four-taxon model (Fig. 1c and e–f, Supplementary Fig. S3a). Similarly, the five-taxon model used an alternately stacked formation of nine dense blocks and eight residual blocks (Fig. 1d and Supplementary Fig. S3b). The high-dimensional features were finally flattened and transformed into one dimension. After the generation of two fully connected layers using SoftMax activation, the output of the models included either 3 or 15 scores that added up to one, corresponding to the proportion of each topology.

The total datasets were first randomly split into the training (90%) and validation sets (10%), the latter of which was not used for parameter training. The training set was randomly shuffled for each epoch and split into batches of sizes 8 and 10 for the four-taxon and five-taxon models, respectively. The batches were used in turn as input for the networks, and the loss was calculated as the mean absolute error between network outputs and labels. The parameters were optimized and updated through

numeric vector with the shape of (128, 5000, 1) or (160, 5000, 1) was used as the input of the CNN models, which possessed both genotype and positional information. The output was a three-dimensional or fifteen-dimensional vector, which represented the proportion of each possible topology and added up to one. Data post-processing included calculating the genome-wide mean value, and the major topology provided a likely species tree. The local introgressed regions and directions of gene flow were identified based on the high support of alternative topologies, which was greater than the score caused by ILS. The architectures of used CNNs are shown for the four-taxon model c) and five-taxon model d). The architectures of the included Dense block e) and Residual block f) are also illustrated. In the schematic of the MSA, minor alleles are highlighted in different colors, with pink for "A," blue for "C," yellow for "G," and green for "T."

gradient descent and backpropagation with the Adam optimizer to minimize loss. The learning rate was set to 0.0001. Model training was stopped after 30,000 iterations (approximately two epochs and five epochs for the four-taxon and five-taxon models, respectively). Models were trained on two NVIDIA Tesla V100 SXM2 32GB GPUs, which took approximately 3.7 h and 6 h for the two models, respectively. The loss over time was visualized using TensorBoard, and it was synchronously reduced in both the training and validation datasets (Supplementary Fig. S3c,d), indicating that the CNN models successfully extracted features of different topologies. The trained models were used in the following analyses of simulated and real genomic data to predict the proportion of each possible topology from sequence alignments.

Comparing the Performance of ERICA with Other Approaches

Model evaluation for topology inference using simulated datasets.—We compared the performance of ERICA with that of the phylogenetic method using the test dataset (D1). We further evaluated its generalizability using several test datasets with different recombination rates (D2), substitution rates (D3), population sizes (D4), and sample sizes (D5), which were generated under the same demographic scenarios as D1 (Supplementary Table S3). For dataset D5, 1–8 haplotypes were sampled for each taxon during data simulation. The sequences were then randomly resampled to eight haplotypes per taxon to fit the limitation of input dimensions in ERICA prediction. In addition, due to the presence of sequencing errors or missing sites in the real genomic data, we also introduced those into the simulations (test datasets D6 and D7). For each sampled haplotype, we randomly selected a set of positions whose total number was the product of the sequence length and the preset error rate. The genotype at each position was randomly changed to A, T, C, or G to simulate a sequencing error (with the exception of the original base) or to N for a missing site. The proportion of each topology was predicted for the test datasets and compared with the data labels. The difference was calculated in two forms: the mean absolute error (MAE) and the double-scaled Euclidean distance.

The MAE was calculated as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\hat{v}_i - v_i|,$$

where n is the number of topologies (i.e., 3 for the four-taxon model and 15 for the five-taxon model), v_i is the true value for topology i, and $\hat{v}_i$ is the estimated value of ERICA for topology i.

The double-scaled Euclidean distance, which was used by Martin and Van Belleghem (2017), was calculated as follows:

$$\mathrm{Distance} = \frac{\sqrt{\sum_{i=1}^{n} \frac{(\hat{v}_i - v_i)^2}{max(1-v_i, v_i)^2}}}{\sqrt{(n-1)}}.$$

The difference between the estimated and true values was scaled using the maximum squared discrepancy for each variable. Since the variables had the sum-to-one restriction and thus were non-independent, (n–1) was used as the degree of freedom.

For the window-tree-based topology weighting method, maximum-likelihood trees were constructed using RAxML (Stamatakis 2014). As the accuracy of tree inference was affected by the numbers of SNPs and the recombination rates (Martin and Van Belleghem 2017), we used different window sizes to minimize the error. Topology weights were computed from these trees based on the Twisst pipeline (Martin and Van Belleghem 2017). MAEs and Euclidean distances were calculated between inferred weights and the data labels according to the equation described above. The window size of 500 bp (which had 75 SNPs on average) had the lowest error rate for the test dataset D1 (Supplementary Fig. S4); thus this parameter was used in subsequent analyses.

Model evaluation for introgression detection using simulated datasets.—For the comparison of ERICA with allele-frequency-based approaches, Patterson's D-statistic and f_d were employed to examine the signatures of introgression for four-taxon cases. For cases with one sequence per taxon, the D-statistic compares the phylogenetic distribution by weighting the binary counts of derived alleles supporting either an ABBA or BABA topology (Green et al. 2010). For cases with more than one sequence per taxon, allele frequencies are used instead of binary counts (Durand et al. 2011). Similarly, f_d is a modified version of the f-statistic that was originally developed to estimate the admixture fraction and has been shown to be less affected by low effective population size in comparison with the D-statistic (Martin et al. 2015). Both D-statistic and f_d calculations were conducted with test dataset D1 using ABBABABAwindows.py (Martin et al. 2015), with the assumption that P1 and P2 are sister species.

Considering that the weight of topology A in the phylogenetic method and the proportion of topology A are consistent with the null hypothesis of the ABBA–BABA test, while the weights/ERICA probabilities of alternative topologies C and B are related to the expected frequencies of ABBA and BABA sites, we calculated the relative difference between the probabilities of C and B using the following formula for the true values:

$$\frac{v_3 - v_2}{v_3 + v_2},$$

and the following formula for the estimated values:

$$\frac{\hat{v}_3 - \hat{v}_2}{\hat{v}_3 + \hat{v}_2}.$$

The calculated values were used for comparisons with D and f_d statistics. Considering that the proportions have been normalized on a scale of 0 to 1, we also tested the absolute difference between two topologies, that is, the numerator of the above equations.

Since test dataset D1 contained complex demographic models, we also simulated sequences under scenarios of an instant admixture event with different species split times, gene flow times, and proportions of introgression.

The command line inputs for the ms simulations were:

For the scenario with gene flow from P3 to P2: ms 32 1 -I 4 8 8 8 8 -ej t_{12} 2 1 -ej t_{123} 3 1 -ej t_R 4 1 -es t_{GF} 2 1-f -ej t_{GF} 5 3 -r 50 5000 -T.

For the scenario with gene flow from P2 to P3: ms 32 1 -I 4 8 8 8 8 -ej t_{12} 2 1 -ej t_{123} 3 1 -ej t_R 4 1 -es t_{GF} 3 1-f -ej t_{GF} 5 2 -r 50 5000 -T.

In these input strings, t_{12}, t_{123}, and t_R indicated the split times between species and were set to (1, 2, 3), (0.5, 1, 1.5), and (0.1, 0.2, 0.3) for test datasets D8, D9, and D10, respectively. t_{GF} indicated the time of gene flow, ranging from 10% to 90% of t_{12}. f indicated the admixture fraction, ranging from 0 to 1 with a step size of 0.1. Other parameters were the same as those used for test dataset D1.

Moreover, we also evaluated the performance of ERICA for adaptive introgression. As the ms simulator did not support dataset modeling with selection, datasets with scenarios of selection were simulated using msms (Ewing and Hermisson 2010). Four evolutionary scenarios were generated. The "Null model" scenario had the topology structure of (((P1, P2), P3), O) for the four-taxon case without migration or selection. The "Selective sweep" scenario included population-specific selection. The "Neutral introgression" scenario had continuous gene flow with a variety of migration rates, and the "Adaptive introgression" scenario included a positive selection of introgressed segments.

Some example command lines were:

"Null model": msms -ms 32 1 -I 4 8 8 8 8 -ej t_{12} 2 1 -ej t_{123} 3 1 -ej t_R 4 1 -r $4Nr \times L$ L -T (t_{12}, t_{123}, and t_R indicated the split time of P1 and P2, the split time of (P1, P2) and P3, and the time of the root, respectively. L indicated the window size, and r was set to 2.5×10^{-9}).

"Selective sweep" (in P2): msms -ms 32 1 -I 4 8 8 8 8 -ej t_{12} 2 1 -ej t_{123} 3 1 -ej t_R 4 1 -r $4Nr \times L$ L -Sp 0.5 -SI 0.5 $\times$ t_{12} 4 0 0 0 0 -Smu 0.01 -Sc 0 2 2 $\times$ 0.001 $\times$ N 2 $\times$ 0.001 $\times$ N 0 -N N -T. The time when selection started was set to 0.5 $\times$ t_{12}, and the selection coefficient (s) was set to 0.001, resulting in a selection strength of 2 $\times$ 0.001 $\times$ N for homozygote and heterozygote genotypes in P2.

"Neutral introgression" (from P3 to P2): msms -ms 32 1 -I 4 8 8 8 8 -dej t_{12} 2 1 -ej t_{123} 3 1 -ej t_R 4 1 -m 2 3 (migration rates) -r $4Nr \times L$ L -T. Migration rates were set to 0.05, 0.1, 0.5, 1, and 5, in units of $4Nm$, where m was the fraction of migrants per generation.

"Adaptive introgression" (from P3 to P2): msms -ms 32 1 -I 4 8 8 8 8 -ej t_{12} 2 1 -ej t_{123} 3 1 -ej t_R 4 1 -m 2 3 (migration rates) -r $4Nr \times L$ L -SAA 2 $\times$ 0.001 $\times$ N -SAa 2 $\times$ 0.001 $\times$ N -Sp 0.5 -SI 0.5 $\times$ t_{12} 4 0 0 1 0 -N N -T.

We also simulated scenarios with gene flow in the opposite direction. For test dataset D11, t_{12}, t_{123}, and t_R were set to 1, 2, and 3, respectively, N was set to 1 M, and L was set to 5000. We also used different species split times (0.5, 1, and 1.5 for D12; 0.1, 0.2, and 0.3 for D13), population sizes (0.5 M for D14, 0.1 M for D15), and window sizes (50-kb for both D16 and D17) to obtain more comprehensive results. In addition, a sequencing error rate of 2% (D18) and a missing site rate of 10% (D19) were incorporated, with other parameters the same as those of D11.

For the ERICA models, a window with a topological proportion greater than a preset threshold indicates the presence of gene flow. The threshold is set to the topological proportion of a false positive rate (FPR) of less than 5% for a neutral, non-introgressed simulated dataset. Note that this rule was not applied to the topology of the species tree. As there were multiple possible directions of gene flow, only the topology corresponding to the given gene flow was identified as a true positive signal, for example, topo C for gene flow between P2 and P3. The ERICA pipeline can be found at http://erica.cibr.ac.cn/ and includes multiple steps, such as data pre-processing, evaluating evolutionary relationships, post-processing, and visualization of the results. For D and f_d statistics, Z-tests were performed to determine the potential introgression regions, and the standard deviations were calculated using the moving block bootstrap method. As with the ERICA models, all windows that significantly deviated from 0 (P value < 0.05) were identified as false positive signals, while only gene flow between P2 and P3 ($D/f_d > 0$) was recognized as a true positive signal.

Sprime (Browning et al. 2018) and IBDmix (Chen et al. 2020) are two other well-known introgression detection methods, which were initially designed to study admixture between modern and ancient humans. Sprime was developed from the S* algorithm (Plagnol and Wall 2006), which analyzes patterns of linkage disequilibrium (LD), while IBDmix is based on the probability of identity by descent (IBD). We also used Sprime and IBDmix to detect gene flow in our datasets. Specifically, as Sprime needs a closely related population without admixture, only the data with gene flow from P3 to P2 were used in the analysis. Therefore, sequence variations of P1 and P2 were used as the outgroup and the introgression recipient, respectively, with the recombination rate and substitution rate set according to the simulation parameters. Given that IBDmix was designed to detect gene flow between modern and archaic humans, we treated the recipient population as the modern samples, and since IBDmix only supports one archaic sample, we randomly selected one sample from the donor populations. The error rate of the data was set to 0, except for the test with a sequence error rate of 0.02. Sprime and IBDmix were applied for each window, and each window with at least one segment with a Sprime/LOD score greater than the preset threshold was recognized as a positive signal.

D_{FOIL} is an extension of the D-statistic that is designed for a symmetric five-taxon model (Pease and Hahn 2015). The D_{FOIL} test contains four statistics using the counts of biallelic sites supporting given patterns and infers introgression events from the significance of each

D_{FOIL} component. There are eight possible introgression patterns among current populations, and each of them can lead to a discordant and unique genealogy. For example, under an assumption of a null model with (((P1, P2), (P3, P4)), O), the probability of topology D ((((P2, P3), P4), P1), O) increases with introgression patterns from P3 to P2, and an excess of topology C ((((P2, P3), P1), P4), O) is observed when the directions are opposite. Therefore, the ERICA results can be used to determine the direction of gene flow and can be compared with the results of the D_{FOIL} test. Test datasets containing a pair of introgression events between P2 and P3 were used for model evaluation, and other introgression events should show the same tendency according to the topological symmetry. D_{FOIL} tests were conducted using fasta2dfoil.py and dfoil.py (Pease and Hahn 2015), with one individual sampled from each population, the true negative windows having an introgression type of "none," and the true positive windows having an introgression type of simulated gene flow ("32" or "23"). The other algorithms were used in the same way as the four-taxon case.

Model evaluation using human demographic scenarios.—We also compared ERICA with another deep learning-based approach, genomatnn, which was designed primarily to detect adaptive introgression in the human genome (Gower et al. 2021), and we followed its workflow for data simulation and model training. In brief, we used the same demographic scenarios described by Gower et al. (2021), with scenario A simulating the gene flow from Neanderthal to Europeans and scenario B simulating the gene flow from Denisovans to Melanesians. An African population was also sampled as a sister taxon of the recipient population. The genealogies and genotype information were generated using the SLiM simulator (Haller and Messer 2019) under the stdpopsim framework (Adrion et al. 2020). We first simulated datasets using the same parameters as the original study and evaluated them with the pre-trained CNNs from the genomatnn software ("Nea_to_CEU_af-0.25" and "Den_to_Melanesian_af-0.25" from https://github.com/grahamgower/genomatnn). Since more individuals were sampled in the pre-trained CNNs of genomatnn in comparison with ERICA, for a fair comparison, we resimulated the datasets but reduced the sample size to 8 for each current population, while keeping the default values for the other parameters. Datasets including scenarios under neutral evolution, selective sweep, and adaptive introgression with varying selection coefficients and times, as well as 10,000 simulations with a length of 100 kb, were generated for each scenario. After dividing the datasets into training datasets (90%) and test datasets (10%), the CNNs of genomatnn were trained with parameters num_rows = 256, epochs = 10, AF = 0, and phased = true. We also evaluated the robustness of genomatnn by validating the datasets using the networks trained based on the other demographic scenario.

As the original demographic scenarios only included three taxa, we added the chimpanzee as the outgroup for ERICA evaluation to meet its minimum taxa requirement, with a split time of 6.6 Ma (Besenbacher et al. 2019) and a population size of 20 k. The other simulation parameters were the same as those used for the genomatnn datasets. The sample size of the test dataset was also comparable to the size of the test dataset used to evaluate genomatnn, with 1000 100-kb simulations in each scenario. The reference and alternative genotypes of the raw simulations were converted to sequence alignments by assigning a randomly chosen base. The topological probabilities were predicted using the ERICA model with the African population as P1, the recipient population (European/Melanesian) as P2, the donor population as P3 (Neanderthal/Denisovan), and the chimpanzee as the outgroup. For each 100 kb simulation, the average value of twenty 5-kb windows was used. For consistency, we also applied IBDmix and Sprime methods to analyze these four-taxon datasets, although IBDmix and Sprime did not require outgroup information.

Using ERICA to Detect Introgression in Real Genomic Data

Data collection and SNP calling of Heliconius butterflies.—Genome-resequencing datasets of three populations of the *Heliconius melpomene-cydno* clade, *H. m. aglaope* (Peru), *H. m. amaryllis* (Peru), and *H. t. thelxinoe* (Peru), and one outgroup, *H. ethilla* (Brazil), were downloaded from NCBI PRJNA308754 (Zhang et al. 2016), PRJEB1749 (Martin et al. 2013), PRJNA73595 (*Heliconius* Genome Consortium 2012), and PRJEB11772 (Davey et al. 2016). Raw reads trimmed by Trimmomatic v0.38 (Bolger et al. 2014) were aligned to the *H. melpomene* v2.5 reference genome (Davey et al. 2016) using Bowtie2 v2.3.4 (Langmead and Salzberg 2012), and PCR duplicates were removed by Picardtools v1.96 function "MarkDuplicates" (http://broadinstitute.github.io/picard/). SNP genotypes were called using GATK v3.7 "UnifiedGenotyper" (DePristo et al. 2011), and genotypes with Qual <50 were removed (Supplementary Table S4).

Whole-genome reciprocal alignment of the genus Oryza.—We generated a whole-genome alignment (WGA) of *Oryza* species with chromosome-level assemblies of *japonica* (cv. Nipponbare), *indica* (cv. 93-11), *O. rufipogon*, *O. nivara*, *O. glaberrima*, *O. barthii*, *O. glumaepatula*, *O. meridionalis*, *O. brachyantha*, *O. punctata*, and *Leersia perrieri*, which were obtained from the OGE/IOMAP 13-genome package (Stein et al. 2018). We performed whole-genome reciprocal alignment according to a previously described pipeline (Zhang et al. 2019). In brief, pairwise alignments were generated using LASTZ (Harris 2007). Alignment blocks that were sufficiently close were joined into chains, and the longest chains were kept and grouped using the UCSC Kent

Utilities (Kent et al. 2003). The genome of *O. sativa* ssp. *japonica* was used as the reference, and the final multiple sequence alignment was generated using aligner Multiz/TBA (Blanchette et al. 2004). Genome annotations from OGE/IOMAP were also transformed into the new coordinate of WGA based on the alignment blocks using in-house scripts.

Demographic modeling of the genus Oryza using G-PhoCS.—We used G-PhoCS v1.3 (Gronau et al. 2011) to estimate the demographic histories of domesticated rice. Coding sequences with flanking 1-kb intervals and repeat regions were masked to satisfy the neutral assumption. A total of 2267 independent loci with a length of 1 kb were chosen for the Bayesian inference. The number of iterations was set to 200,000 for each MCMC run, and the initial 20,000 iterations were ignored in the post-processing using Tracer v1.6 (http://tree.bio.ed.ac.uk/software/tracer/). Three repeat runs for all possible migration bands between current species were tested independently, and only the significant migration bands were included in the full model test. The mutation rate was set to 6.5×10^{-9} per site per generation (Choi et al. 2017). The raw estimates and calibrated values of population size, divergence times, and migration rates are shown in Supplementary Table S5.

To distinguish introgression from ILS in rice domestication, we simulated and labeled 1000 loci in a region with a length of 50 kb using the demographic histories estimated using G-PhoCS, but without gene flow, with the following command: ms 5 1 -I 5 1 1 1 1 1 -n 2 1 -ej 1.25 2 1 -n 3 1 -ej 5.5 3 1 -n 4 1 -ej 5 4 3 -n 5 1 -ej 25 5 1 -en 0.025 5 17.25 -en 1.25 1 11 -en 5 3 1.5 -en 5.5 1 54 -en 25 1 17.25 -r 8 50000 -T, where populations 1, 2, 3, 4, and 5 are *japonica*, *O. rufipogon*, *indica*, *O. nivara*, and *O. barthii*, respectively. Multiple sequence alignments were generated using Seq-Gen (Rambaut and Grassly 1997), with a scale factor of 0.0001, and analyzed using ERICA.

Genotype calling of the rice pan-genome data.—Whole-genome assemblies of *indica* (cv. 93-11), *O. rufipogon*, *O. nivara*, *O. barthii*, and 64 other accessions of domesticated and wild rice (Zhao et al. 2018) were aligned to the reference genome of *japonica* using minimap2 v2.17 (Li 2018) with the parameters -ax asm5 –secondary = no. Alignments with a map quality lower than 60 were removed using SAMtools v0.1.19 (Li et al. 2009), and read groups were added using the Picardtools v1.96 function "AddOrReplaceRead" (http://broadinstitute.github.io/picard/). A Variant Call Format (VCF) file was produced using GATK v3.7 "UnifiedGenotyper" with the parameter—defaultBaseQualities 60 (Supplementary Table S4).

Phylogenetic and population structure analyses of rice.—8.2 Mb SNPs were used to reconstruct a genome-wide maximum likelihood tree using RAxML (Stamatakis 2014) with the GTRGAMMA model and 20 bootstrap replicates. The tree was visualized using iTOL v5 (Letunic and Bork 2019).

Population structures were identified using ADMIXTURE (Alexander et al. 2009) and principal components analysis (Price et al. 2006). SNP sites were pruned according to linkage disequilibrium using PLINK v1.9 (Purcell et al. 2007) with the parameter -indep -pairwise 50 10 0.1. With the exception of the outgroup taxon *O. barthii*, 67 samples were included in the ADMIXTURE analysis, and *k*-values were set 2–6. GCTA v1.93 (Yang et al. 2011) was used to calculate the first two principal components using the genome-wide SNP data.

We applied ERICA to a subset of WGA containing the sequences of *japonica*, *indica*, *O. rufipogon*, *O. nivara*, and *O. barthii*, and we identified the introgressed 50-kb windows between domesticated rice with a cutoff of proportion > 0.4. The borders of these windows were converted to the genomic coordinates of *japonica* for downstream analyses. For the pan-genome dataset, up to eight samples were chosen to represent each focal taxon (Supplementary Table S4), and consensus sequences were generated using vcf2MSA.py following the ERICA pipeline.

The absolute divergence (d_{xy}) between populations was calculated as follows:

$$d_{xy} = \frac{1}{n}\sum_{i=1}^{n} p_{ix}\left(1 - p_{iy}\right) + p_{iy}\left(1 - p_{ix}\right),$$

where n is the window size and p_{ix} and p_{iy} are the reference allele frequencies for base i in populations x and y, respectively. Only candidate windows with d_{xy} less than the chromosome-wide mean values were retained and merged.

Functional Annotation of Candidate Introgressed Loci in Oryza

Gene function enrichment.—The introgressed genes were extracted based on annotations from the MSU Rice Genome Annotation Project (MSU-RAP) (Kawahara et al. 2013). GO enrichment analysis was performed using CARMO (Wang et al. 2015) with a cutoff of $P < 0.05$. Plant Experimental Conditions Ontology (PECO) and Plant Trait Ontology (TO) information was downloaded from Planteome (Cooper et al. 2018), and the hypergeometric test was used to calculate the P values for enrichment terms.

Gene correlation analysis.—We obtained a rice expression matrix from MSU-RAP (Kawahara et al. 2013). In brief, raw RNA sequencing reads of nine tissues from *japonica* (cv. Nipponbare), including leaves at 20 days after sowing, primordial inflorescences at 10 days before flower emergence, whole inflorescences at the time of flower emergence, anthers and pistils at the time of anthesis, whole seeds at 5 days after pollination (DAP), whole seeds at 10 DAP, embryos at 25 DAP, and endosperm at 25 DAP, were aligned to the reference genome using

Tophat (Trapnell et al. 2009). FPKM (fragments per kilobase of exon model per million fragments mapped) values were calculated using Cufflinks (Trapnell et al. 2010).

For co-expression clustering, low-expression genes with FPKM values <2 across all tissues were removed and 24,919 genes were retained. FPKM values were log2 transformed, and the *k*-means clustering algorithm in Multiple Experiment Viewer v4.9.0 (Howe et al. 2011) was performed with $k = 9$ and maximum iterations = 100. Genes assigned to different clusters in five independent runs were dropped according to Davidson et al. (2012).

Differential gene expression analysis.—RNA-seq data for 18 accessions of temperate *japonica*, 24 accessions of tropical *japonica*, and 25 accessions of *indica* with two biological replicates for each accession were downloaded from NCBI PRJNA385135 (Campbell et al. 2020). The shoot tissues from seedlings at 10 days after transplant were sampled and sequenced. Raw reads were aligned to the reference genome of *japonica* using Tophat2 v2.1.1 (Kim et al. 2013) with gene models from MSU-RAP (Kawahara et al. 2013) in union mode. Count normalization and differential expression analyses were carried out using DESeq2 (Love et al. 2014). Only genes expressed in at least one clade (samples with non-zero counts greater than 50%) were included in the downstream analyses. The cutoffs of differentially expressed genes (DEGs) were set to an adjusted *P* value less than 0.01 and an absolute value of log2 fold change greater than 1.

Results

Efficiency and Robustness of ERICA for Topological Inference Based on Simulated Data

Performance comparison of ERICA and the window-tree-based approach.—We first evaluated the performance of ERICA for topology inference using a smaller-scale test dataset (D1) that was simulated independently with the same evolutionary scenarios used for the training dataset (Supplementary Table S2). We analyzed test dataset D1 using ERICA and other approaches, including the window-tree-based topology weighting methods ("Methods"). The mean absolute errors (MAEs) and scaled Euclidean distances (EDs) of the four-taxon and five-taxon ERICA models were significantly lower than those of the window trees approach (Mann–Whitney *U* test $P < 0.001$, Fig. 2a and Supplementary Fig. S5). The results showed that ERICA efficiently extracted phylogenetic information with a degree of accuracy higher than that of the window trees approach. When applying the window-tree-based topology weighting approach, the genotypes of diploid genomes should first be phased, then the phylogenic trees are inferred for each window using the maximum likelihood, neighbor-joining, or Bayesian methods, and the topology weights are calculated from the trees (Martin and Van Belleghem 2017). Therefore, a smaller window size likely leads to low resolution owing to insufficient patterns, whereas a larger window size creates difficulty in resolving local heterogeneity, both of which increase the bias of phylogenetic inferences. However, ERICA avoided the multiple steps and directly estimated the topology proportions from sequence data without reconstructing a bifurcation tree, reducing the errors.

Considering that the order of taxa in the MSAs will not affect the relative topology proportions, we further examined whether the prediction results would be affected by the order of input data by exchanging the order of (P1, P2), (P1, P3), and (P2, P3). The results showed that there was no significant difference in the errors when swapping the order of (P1, P2) and (P2, P3) (Mann–Whitney *U* test $P > 0.05$), despite the average errors of combination (P1, P3) increased by 4%. Similarly, randomly changing the sequence order in each taxon did not affect the results (Mann–Whitney *U* test $P = 0.97$ for MAEs and $P = 0.94$ for EDs), indicating that the model was hardly affected by the input order.

Robustness tests using demographic parameters not included in the model training.—To further investigate the generalizability and application range of ERICA models, we calculated error rates using datasets with different simulation parameters, including recombination rates (dataset D2), substitution rates (dataset D3), effective population sizes (dataset D4), and sample numbers (dataset D5) (Supplementary Table S3). For the four-taxon model, there were significant differences among MAEs yielded from datasets with different recombination rates (dataset D2, Kruskal–Wallis test $P < 0.001$, Supplementary Fig. S6a,b). Datasets with higher recombination rates ($4Nr$) such as 0.01, 0.05, and 0.1 yielded the same MAEs (Kruskal–Wallis test $P = 0.60$), whereas datasets with lower recombination rates such as 0, 0.001, and 0.005 yielded MAEs significantly greater than those generated with the recombination rate of the training datasets (0.01) (Mann–Whitney *U* test $P < 0.001$, Bonferroni correction), with more extreme values (Supplementary Fig. S6a). In contrast, the error rates of the tree-based topology weighting method presented an opposite tendency, which was roughly positively correlated with the recombination rate (Supplementary Fig. S6b). The MAEs of the window trees approach were significantly greater than those of ERICA in most cases (Mann–Whitney *U* test $P < 0.001$), with the exceptions of the datasets for recombination rates of 0 and 0.001 (Supplementary Fig. S6a). These results indicated that the traditional window trees approach performed well for phylogeny reconstruction of sequence data with homogenous evolutionary history. In contrast, for ERICA models, the phylogeny inference did not rely on a long haplotype structure, and the presence of a sufficient number of independent variation sites improved the model performance. These results suggested that ERICA was particularly suitable for data with recombination events, which are more similar to real experimental data.

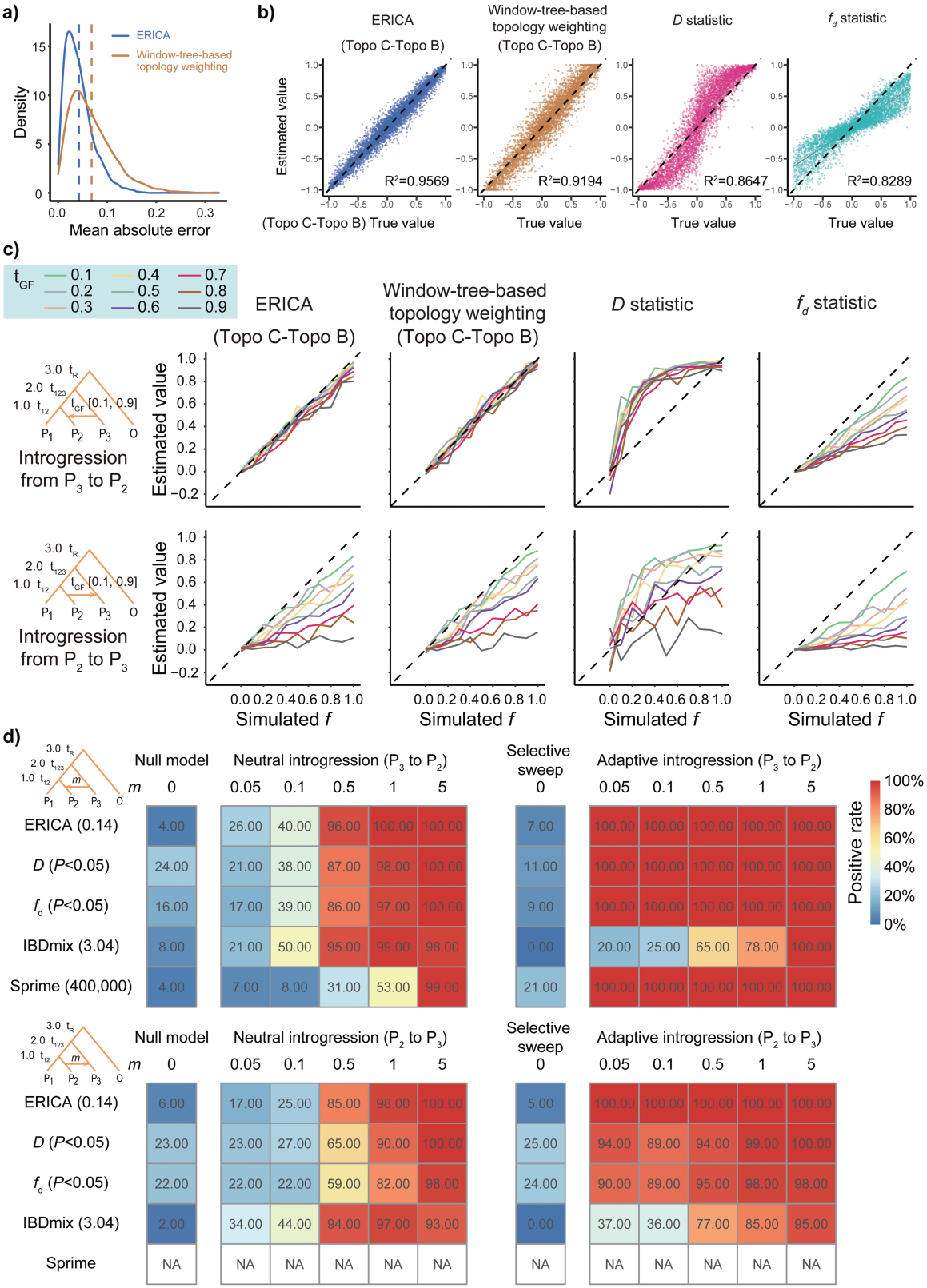


FIGURE 2. Performance evaluation of the four-taxon ERICA model. a) The mean absolute error distributions of ERICA and the window-tree-based topology weighting method. The dashed lines indicate average values for all data (evaluated using test dataset D1, n = 6030). b) The relationships between true values and estimated values of the absolute differences between two alternative topologies (evaluated using

With regard to tests with different substitution rates (dataset D3), we observed that larger and smaller substitution rates ($4N\mu$) led to errors greater than those achieved with the substitution rate of the training datasets (0.01) (Mann–Whitney *U* test $P < 0.001$, Bonferroni correction, Supplementary Fig. S6c,d), indicating that the sequence absolute divergence may affect the performance of the current model. Nevertheless, the errors of ERICA were significantly smaller than those of the window trees approach when substitution rates were less than 0.01 (Mann–Whitney *U* test $P < 0.001$), suggesting that ERICA favored samples with small and moderate differentiation, for example, ingroup taxa with absolute divergence ranging from 0.01 to 0.06, which matched that of the training datasets.

Similarly, changes in the effective population size (N) also led to increased error rates (dataset D4, Supplementary Fig. S6e–f), because the scaled recombination and substitution rates varied. For example, for smaller population sizes, both the number of variants and the independence between loci were reduced, suggesting that a larger window may yield better performance for phylogeny inference. We calculated the mean values of multiple 5-kb windows and compared them with the true values. The results showed that the error decreased as the window size increased for all population sizes and stabilized at a window size of about 100 kb (Supplementary Fig. S7). Thus, when the models were applied to real data that had demography extremely different from that of the training data, large windows could be used to obtain higher accuracy, at the cost of reduced resolution. For instance, MAEs consistent with or less than those obtained with the training data (5-kb window, $N = 1$ M) were observed when a 200-kb window was used for population sizes ranging from 0.1 M and 10 M.

When test datasets containing different sample sizes per taxon (dataset D5) were analyzed, ERICA displayed outstanding robustness, which consistently yielded significantly lower MAEs than those yielded by the window trees approach, even for the case with one individual per taxon (Mann–Whitney *U* test $P < 0.001$, Supplementary Fig. S6g,h). For ERICA, only a slight decrease in accuracy was observed when the sample size was decreased to one individual per taxon (the mean MAE increased from 0.043 to 0.056), whereas the mean MAE of the window trees approach increased from 0.070 to 0.083. We, therefore, speculated that either a chromosome-level reference genome or aligned population genomic data would be particularly suitable for ERICA. The performance of the five-taxon model was influenced to a degree similar to that of the four-taxon model by changes in the recombination rate, substitution rate, and sample size (Kruskal–Wallis test $P < 0.001$, Supplementary Fig. S8).

Another difference between real sequence data and simulations was the presence of sequence errors, missing sites, and alignment gaps. To evaluate the effect of incorrect genotype information, we randomly introduced sequence errors (ranging from 0% to 2%) and missing data (ranging from 0% to 20%) into the simulated dataset (datasets D6–D7, Supplementary Table S3, "Methods" section). The differences between the ERICA results and the true values were positively correlated with the rates of sequence errors and missing data, and the adverse impacts of sequence errors were greater than those of equal proportions of missing sites, while the window trees approach was not affected by either change (Supplementary Fig. S9). Nevertheless, the data showed that the MAEs of ERICA were still significantly smaller than or equal to those of the tree-based method for an error rate of 2% (Mann–Whitney *U* test $P < 0.001$ for the four-taxon model and $P = 0.25$ for the five-taxon model), and ERICA had relatively better performance when the missing data rate was less than about 10%.

Performance of ERICA and Other Algorithms for Introgression Detection Based on Simulated Data

Using asymmetric alternative topologies to detect introgression.—To infer local introgression fragments, we focused on solving the relationship between topological structure and historical gene flow. For the four-taxon model with three genealogies, the two alternative topologies that were different from the species tree had equal frequencies under evolutionary scenarios without gene flow (Durand et al. 2011), and thus the existence of gene flow was indicated when the difference between the proportions of alternative topologies deviated from 0. The *D*-statistic (also called the ABBA–BABA test) used the relative counts of derived alleles supporting specific patterns to represent the probabilities of alternative topologies (Durand et al. 2011). We followed the same assumption, but we replaced the site counts with the true values of topology weights and estimated values predicted by deep learning models, and the corresponding statistic was calculated as the difference between the proportions of two alternative topologies divided by their sum. In test dataset D1, the results of the ERICA model were well correlated with

test dataset D1, "Methods" section, $n = 6030$). c) The estimated values of introgression fractions (f) (evaluated using test dataset D8). Eleven different values were simulated. The split times t_{12}, t_{123}, and t_R were set to 1, 2, and 3 (in units of $4N$ generations), respectively, and the time of gene flow ranged from 10% to 90% of the split time (t_{12}). Gene flows from P3 to P2 (the top panel) and in the opposite direction (the bottom panel) were tested. Twenty replicates were used for each scenario. d) The performance of different methods for detecting introgression signals with or without selection (evaluated using test dataset D11). The heatmaps show false positive rates for scenarios in the absence of gene flow ("null model" and "selective sweep") and TPR for neutral and adaptive introgression at different migration rates (0.05–5 migrants per generation). The number in brackets is the threshold used for each method. Sprime was not available for detecting gene flow from P2 to P3. Hundred replicates were tested for each case.

the true values, with a regression coefficient comparable to the *D*-statistic (0.9253 and 0.9031, respectively, Supplementary Fig. S10a). Since dataset D1 contained multiple demographic scenarios with and without gene flow, to further explore the relationship between ERICA results and the introgression intensity under an explicit population history, we simulated scenarios of an instant admixture event, with different intensities, times, and directions of gene flow (datasets D8–D10, "Methods" section). The estimated value of ERICA had a trend consistent with the *D*-statistic, which was greater than 0 when gene flow existed (evaluated using dataset D8, Supplementary Fig. S10b), suggesting that ERICA was also sensitive for detecting introgression. In addition, the estimated value of ERICA was positively correlated with the proportion of introgression (*f*, which represents the fraction of shared haplotypes), but it tended to overestimate the true value, which was also a characteristic of the *D*-statistic (Martin et al. 2015). To solve this problem, we evaluated the performance of ERICA directly using the difference between two alternative topologies, considering that the topology proportions had been normalized to fall between 0 and 1. Under these conditions, the correlation with the true value increased (0.9569, evaluated using dataset D1, Fig. 2b), and more importantly, it was a good estimator of the introgression fraction in general, although it was still affected by the time and direction of introgression (evaluated using datasets D8–D10, Fig. 2c and Supplementary Fig. S11a,b). When the gene flow was from P3 to P2, the estimated values of ERICA were almost identical to the simulated *f* value (Fig. 2c); however, for gene flow from P2 to P3, ERICA tended to underestimate the true value. Several theoretical and simulation studies have shown that the direction of gene flow affects the power of *D* and other statistics (Martin et al. 2015; Hibbins and Hahn 2019, 2022), since the internal branch lengths of the introgressed fragments can be affected by the direction of gene flow. For example, when the direction of gene flow was from P2 to P3, the expected coalescent time of P1 and P3 was t_{12}, whereas it was t_{123} when the direction of gene flow was opposite. Thus, the differences between the time of population split and gene flow from P2 to P3 (t_{12}–t_{GF}) were relatively smaller than those in the opposite direction (t_{123}–t_{GF}) (Supplementary Fig. S11c), which led to an increased intensity of ILS and a decreased proportion of the introgression topology. For the same reason, when the population divergence time was decreased overall, both ERICA and the f_d statistic underestimated the *f* value, with the ERICA model showing less variance (Supplementary Fig. S11).

Using proportions of discordant topologies to detect introgression.—Calculating the differences between alternative topologies provided a way to detect introgression and estimate the fraction in the four-taxon case. However, extending the method to more populations is a difficult task. Because the gene flow between non-sister species changed the relative relationships among taxa, the existence and direction of gene flow could be identified according to the discordant patterns between local and genome-wide topologies (Supplementary Fig. S12). Thus, we wanted to further confirm whether the proportion of a given topology can be used to directly detect introgression. Validation was performed using a series of test datasets with continuous gene flow (datasets D11–D19) and a variety of migration rates, species split times, effective population sizes, window sizes, and sequencing error rates (Supplementary Table S6, "Methods" section). The effect of selection was also considered in the analyses, and each dataset contained four scenarios, with the "Null model" and "Selective sweep" scenarios comprising the negative category, and the "Neutral introgression" and "Adaptive introgression" scenarios comprising the positive category. We also compared the performance of different approaches, including allele frequency-based statistics and other widely used methods, like Sprime (Browning et al. 2018) and IBDmix (Chen et al. 2020). Since accuracy and precision scores can be affected by the ratio of both categories, we mainly focused on the true positive rate (TPR, also known as sensitivity/recall) and the FPR (equal to one minus specificity) for each method. For ERICA, Sprime, and IBDmix, the sensitivity and specificity varied with the threshold. To facilitate comparison, we used thresholds that resulted in "Null model" FPRs that were less than or equal to 5% and obtained the TPRs for each dataset. For the *D* and f_d statistics, the windows with values significantly deviated from 0 at a significance level of 0.05 represented the false positive introgression signals in the absence of gene flow. Similarly, the D/f_d values that were statistically significant and consistent with the introgression direction were the true positive signals.

We first evaluated the four-taxon data. Evaluated using dataset D11, IBDmix had the best performance for detecting neutral introgression (Fig. 2d and Supplementary Fig. S13a and Table S6). Specifically, the TPRs of ERICA and IBDmix were both approximately 100% for data with higher migration rates ($4Nm$, where m is the fraction of migrants in the recipient population), for example, 1 and 5, but, for lower migration rates, the TPRs of ERICA were smaller than those of IBDmix, especially for data with gene flow from P2 to P3. Nevertheless, the sensitivity of ERICA was greater than that of other methods; the average TPR of ERICA for all migration rates was 68.7%, compared to 64.9% for the *D*-statistic, 62.2% for the f_d-statistic, and 39.6% for Sprime. For adaptive introgression, ERICA and Sprime showed greatly improved classification performance in comparison with their performance for neutral cases, but the performance of IBDmix did not change markedly (Supplementary Fig. S13a and Table S6). The TPR of ERICA was 100%, with an FPR of about 6%, while IBDmix had a relatively small TPR (61.8%). Other methods, including the *D*-statistic, f_d-statistic, and Sprime, were also highly sensitive for introgressed regions, but their FPRs ranged from 9% to 25% and were thus

higher than that of ERICA. In addition, the sensitivities of the *D*-statistic and f_d-statistic were slightly reduced when the gene flow was from P2 to P3 (Fig. 2d and Supplementary Table S6).

We also used more demographic parameters to evaluate the performance tendencies of different methods under varied demographic histories. For neutral introgression, the sensitivities of all methods decreased with smaller divergence times (evaluated using datasets D12 and D13, Supplementary Figs. S13b,c and S14 and Table S6) and population sizes (evaluated using datasets D14–D15, Supplementary Fig. S15 and Table S6), and detecting the introgression signals in datasets with smaller divergence times were more difficult than identifying those in datasets with smaller population sizes. The performance of the ERICA model was comparable to that of allele frequency-based statistics and better than that of the Sprime method, except for the dataset with the smallest divergence time (D13). However, for all adaptive introgressions, ERICA consistently had the highest TPR, which was greater than 95% with the exception of dataset D13, while other methods had greater decreases in their TPRs as the divergence time and population size were reduced, and allele frequency-based statistics were strongly affected by the direction of introgressions. In addition, the performance of all methods increased when a large window (50-kb) was used (evaluated using datasets D16–D17), and the resulting sensitivity was close to or greater than that achieved with the dataset with a large population size and a small window size, suggesting that there may be a trade-off between resolution and accuracy. However, the relative performance of the different methods did not change when a large window (50-kb) was used (Supplementary Fig. S16 and Table S6).

Developed from *D*, D_{FOIL} was designed for detecting introgression for a five-taxon dataset (Pease and Hahn 2015). Since D_{FOIL} is not suitable for asymmetric topologies, we simulated datasets to represent possible interspecific gene flow in a symmetric topology (((P1, P2), (P3, P4)), O) and with different divergence times, introgression times and directions. We summarized the performance of different methods and replaced the *D*- and f_d-statistics with D_{FOIL} (Supplementary Figs. S17 and S18 and Table S6). Consistent with the four-taxon dataset, the sensitivities of the five-taxon cases were influenced by demographic histories and population sizes (evaluated using datasets D11–D15). More specifically, the performance of IBDmix changed only slightly in comparison with its performance for the corresponding datasets with four taxa, probably because this analysis only involved recipient and donor populations without outgroups. As more alternative topologies increased the complexity of the model, the sensitivities of ERICA and the D_{FOIL} method decreased in the five-taxon case. Again, increasing the window size improved the sensitivities of ERICA and other methods (evaluated using datasets D16–D17, Supplementary Fig. S19 and Table S6). Nevertheless, ERICA showed the best in performance for detecting adaptive introgressions in all datasets with the exception of D13.

Considering that sequencing errors can lead to an increased rate of errors in phylogeny inference, we also tested their impact on introgression detection (evaluated using datasets D18–D19). The performance of ERICA, the *D*-statistic, and the f_d-statistic did not change greatly with the addition of sequencing errors and missing data, but IBDmix, Sprime, and D_{FOIL} were sensitive to both sequencing errors and missing data (Supplementary Figs. S20 and S21 and Table S6). In particular, IBDmix showed poor performance for datasets with 10% missing sites (D19).

The threshold of candidate introgressed segments is a key parameter in the models, because an inappropriate threshold may reduce the sensitivity of introgression detection. For example, when the default threshold scores suggested by the software were used, the FPRs of IBDmix and Sprime were strongly increased in some cases, which may have resulted in the identification of all windows without gene flow as positive loci (evaluated on datasets D11–D13, Supplementary Fig. S22). Using the simulated data, we evaluated the distributions under the null hypothesis and set the threshold according to the preset FPR, and the thresholds were influenced by the given population histories. However, applying this method to real data was difficult, because the true introgressed regions were not known. Therefore, we determined a threshold for the ERICA model that was less affected by the population history and applicable to real genomic data by teasing apart the signatures of introgression and ILS. Basically, the windows with strong support for alternative topologies, which had significantly higher proportions than those caused by ILS, were considered as introgressed loci. Although the intensity of ILS is influenced by demographic history, it is always smaller than one-third theoretically (**Supplementary Results**). Given the existence of stochastic errors, we evaluated the distribution of topology proportions under the strongest ILS using simulated datasets and used the 95% quantile (0.5 for 5-kb windows and 0.4 for 50-kb windows) as the cutoff in further analyses of real data, unless otherwise noted (**Supplementary Results**, "Discussion" section). This relatively strict threshold ensured a low FPR under different population histories, which was 0 for all tested data sets and was highly sensitive to adaptive introgression, despite the undetectable weak neutral gene flow (evaluated using datasets D11–D13, Supplementary Fig. S22).

Performance based on specific, real demographic models with human data as an example.—To further evaluate the generalization ability of ERICA, we compared the performance of ERICA with that of genomatnn (Gower et al. 2021), a deep-learning-based approach that includes CNNs that can be trained using prior demographic information and pre-trained CNNs based on human demographic history. Therefore, we generated

two additional datasets according to human demographic history with adaptive introgressions and also fitted two of genomatnn's pre-trained CNNs: human demographic scenario A (including gene flow from Neanderthal to Europeans) and scenario B (including a more complex history and focusing on introgressions from Denisovans to Melanesians) (Supplementary Fig. S23a, "Methods" section). Since genomatnn's CNNs required a specified number of samples for the input, and its pre-trained CNNs include more samples than the maximum number that ERICA can input, we also trained new CNNs for genomatnn with a comparable number of samples and compared its performance to that of the pre-trained CNNs. Our results showed that genomatnn was robust for inferences with small sample sizes (Supplementary Fig. S23b), and we thus compared it with ERICA and other methods (Supplementary Fig. S23b). For human demographic scenario A, genomatnn performed best, with a TPR of 79.0% (with neutral datasets as the negative category and FPR ≤ 5%), followed by ERICA (65.1%), Sprime (39.2%), and IBDmix (15.1%) (Supplementary Fig. S23b). In the case of scenario B, with a more complex history, all methods showed decreased performance, although the TPR of ERICA (48.4%) was slightly higher than that of genomatnn (45.1%) (Supplementary Fig. S23d). Furthermore, since the demographic information used in genomatnn's CNN training may differ from the real condition, that is, model misspecification, we also evaluated the impact of this difference on the genomatnn method by exchanging the test datasets of the two demographic scenarios. Both results showed performance declines; the TPR of scenario A (68.5%) was similar to that of ERICA, while the TPR of scenario B declined even more sharply and was much lower than that of ERICA. We further explored the relationship between accuracy and the selection coefficient, as well as the time at which the selection started (Supplementary Fig. S23c and e). The results showed that the TPRs of both methods were highly affected by the selection intensity and time. The TPRs of ERICA were greater than 95% for some of the subsets under ancient and strong selection, and the TPR for scenario B was especially high (12.5% and 31.25% for scenarios A and B, respectively), suggesting that ERICA can effectively detect strong signatures of adaptive introgression. In conclusion, our results revealed the different performance and characteristics of two deep learning-based approaches. On the one hand, genomatnn had better performance in some cases, such as demographic scenario A, but was also affected by the demographic histories of the focal taxa and by the errors between the pre-set and real demographic scenarios. On the other hand, ERICA showed TPRs that were comparable or higher than those of genomatnn with model misspecification. When ERICA was applied to pre-trained CNNs, it showed good performance for scenarios A and B, which both differed greatly from its training dataset, including the inclusion of different simulators, as well as differences in parameters such as population sizes, mutation rates, and selection pressures, suggesting greater generalization ability in comparison with that of genomatnn.

Time and memory costs of ERICA.—We also compared the run times of ERICA models with allele-frequency-based methods (Supplementary Table S7). ERICA models were faster than the traditional methods in handling large datasets, but their speed advantages were not seen with smaller datasets, given that loading the parameters of neural networks comprised the majority of the execution time and became a rate-limiting step. The minimum memory required for running ERICA was 3.3 Gb, and the memory consumption increased about 9-fold with the size of the MSA file. Since the MSA data were first divided into non-overlapping 5-kb windows and then the following prediction processes of different windows were independent, the genome-wide or chromosome-wide sequences could be split into subsets with appropriate data size according to the hardware resource limit.

Inferring Gene Flow Based on Real Genomic Data

Disentangling local and partial introgression in Heliconius butterflies.—Next, ERICA was evaluated using a representative genome-resequencing dataset of *Heliconius* butterflies, which are known as Müllerian mimics and display complex relationships owing to intensive hybridization during adaptive radiation, even without an available bifurcating tree (Edelman et al. 2019). We analyzed the dataset at the whole genome level using ERICA. Consistent with the order of species differentiation, our results showed that the two *H. melpomene* races, *H. m. aglaope* and *H. m. amaryllis*, were grouped together overall as sister taxa relative to *H. timareta thelxinoe* (Topo A in **Fig. 3**), and the highest proportion was along the Z chromosome (Chr21) (Supplementary Fig. S24), which was consistent with previous results, and suggested that the resistance of the Z chromosome to introgression was greater than that of autosomes in *H. melpomene* and *H. timareta* (Martin et al. 2013, 2019). In addition, the Z chromosome may also be less affected by ILS owing to its smaller effective population size in comparison with those of autosomes. The highest proportion of grouping for the sympatric co-mimics, *H. m. amaryllis* and *H. t. thelxinoe* (Topo C in Fig. 3), was observed on chromosome 18 (ranging from 700 kb to 850 kb), where it is located at a known locus controlling wing color patterns, named *B/D* (*Heliconius* Genome Consortium 2012) (Fig. 3a,b and Supplementary Fig. S24 and Table S8), suggesting that ERICA efficiently captured the known signature of introgression from the real data. Consistent with previous studies, the results of f_d were smoother than those of the *D*-statistic along the *B/D* locus (Fig. 3c), suggesting that the *D*-statistic may not be suitable for detecting local introgressed signals (Martin et al. 2015), whereas the results of ERICA were similar and comparable to those of f_d. To more accurately estimate the random errors of different methods for real genomic data, we compared the results of their analyses of adjacent 5-kb windows by assuming

a similar evolutionary history within a linkage disequilibrium block in *Heliconius* butterflies (*Heliconius* Genome Consortium 2012). The differences had a mean value of zero, and the variance of ERICA was smaller than that of f_d (Fig. 3d). We subsequently detected loci of putative introgression between *H. m. amaryllis* and *H. t. thelxinoe* using ERICA, *D* and f_d. ERICA characterized 947 50-kb loci suggesting introgression between *H. m. amaryllis* and *H. t. thelxinoe*, with 37% of these loci also supported by *D* or f_d (Fig. 3e). We also evaluated some ERICA-specific results by focusing on the three outlier regions with the highest proportions of topology C that did not have significant *D* and f_d statistics. Among these regions, two loci showed reduced absolute divergence between *H. m. amaryllis* and *H. t. thelxinoe*, and the phylogenetic patterns of discordance were limited to a subset of samples (Supplementary material Fig. S25a–b and Table S8). The *D*, f_d, and d_{xy} methods suggested that the third locus was a typical, but not significant, signature of introgression due to heterogeneity (Supplementary Fig. S25c and Table S8). Taken together, these findings demonstrate that ERICA is a remarkably accurate and efficient method for detecting signatures of introgression in species with complex evolutionary history.

Detecting genome-wide hybridization in rice using multiple reference genomes.—Given that ERICA showed excellent performance with both simulated and real datasets, we used it to investigate potential introgression between domesticated and wild rice in the genus *Oryza*, including the Asian cultivated rice *Oryza sativa* and its wild relatives *O. rufipogon* and *O. nivara*, in addition to the African cultivated rice *O. glaberrima* and its wild progenitor *O. barthii*. As two major subspecies of Asian cultivated rice, *O. sativa* ssp. *japonica* (with tropical and temperate subgroups) and *O. sativa* ssp. *indica*, can be distinguished by morphological and genetic differences, but their origins and relationship are controversial. Two other populations, aromatic rice and *aus* rice, are considered to be subgroups of the *japonica* and *indica* subspecies, respectively (Garris et al. 2005). Genome-wide phylogeny studies suggest that *japonica* and *indica* have different wild progenitors along with several shared crucial domestication genes, indicating the existence of gene flow (Huang et al. 2012; Civáň et al. 2015; Huang and Han 2015; Choi et al. 2017).

We first applied both ERICA and a Bayesian inference approach, G-PhoCS, to detect global signals of gene flow using *Oryza* reference assemblies. Three independent runs of G-PhoCS suggested a consistent demographic model with slightly later divergence times for cultivated and wild rice (about 4.2 ky for Asian rice (95% HPD interval: 1.1–7.7 ky) and 0.4 ky for African rice (95% HPD interval: 0–1.3 ky)) in comparison with the ages of the oldest archaeological remains (about 9 ky in Asia (Zheng et al. 2016) and 3 ky in Africa (Wang et al. 2014)) with strong migration from *japonica* to *indica* (Supplementary Fig. S26 and Table S5). As G-PhoCS considers only independent neutral loci, it detected global gene flow between *O. sativa* populations, but it did not resolve local and adaptive introgressions that were likely present near coding regions. Based on the results described above, we inferred both signatures of global and local introgression in reference assemblies of Asian cultivated rice using ERICA, and ERICA yielded the highest proportion of topology M, which was consistent with the species tree (Fig. 4a–b and Supplementary Fig. S26). Given that both ILS and gene flow might lead to alternative topologies and that the influences of both of these processes can be determined based on divergence time and population size (Pamilo and Nei 1988; Rosenberg 2002), we determined the baseline intensity of ILS by modeling the rice demography without gene flow using simulated sequences and analyzing the data using ERICA (Fig. 5a–c, "Methods" section). According to the null modeling results, ILS caused four different levels of probabilities for alternative topologies, and the ERICA results for simulated data showed a similar, but not completely identical, pattern of distribution due to random errors. In comparison with the simulated null model, three categories, B, G and C, showed a significantly higher proportion and yielded a list of introgressed loci between *japonica* and *indica*, as well as from *O. rufipogon* to *indica* (Fig. 5d), supporting the idea that introgression played a significant role in the domestication of *indica* subspecies (Choi et al. 2017). To further dissect the introgression regions involved in rice domestication, we identified loci that supported the cluster of *japonica* and *indica* with high confidence (proportion of topology B, G, or N > 0.4) and extracted putative introgression regions with absolute genetic divergence lower than the chromosomal mean divergence, because such introgression regions have a younger divergence time than other genomic regions, which also distinguishes introgression from ancestral variation (Smith and Kronforst 2013). We obtained a list of 71 candidate introgressed loci scattered around the 12 chromosomes, including 1174 genes, 50 of which were well annotated in the Rice Annotation Project Database (RAP-DB) (Sakai et al. 2013) (Supplementary Table S9). Among these 71 candidate loci, 40 loci overlapped with 19 previously reported selective sweep regions related to domestication traits such as panicle length, germination rate, hull color, stigma exsertion, stigma color, tiller angle, and awn length (Huang et al. 2012). A few known domestication genes, including *OsSh1* (Lin et al 2012), *PROG1* (Jin et al 2008; Tan et al 2008), *OsC1* (Saitoh et al 2004), and *Rc* (Sweeney et al 2006), were also located within four introgressed loci or within the range of LD (approximately 100–200 kb in cultivated rice (Huang et al. 2010)). A Gene Ontology (GO) enrichment analysis for all 1174 genes in the candidate introgressed loci in comparison with all rice genes suggested that these genes are involved in the response to stimulus and bacterium, biosynthetic process and photorespiration (Supplementary Fig. S27a and Table S9). For example, one remarkable locus on chromosome five containing a

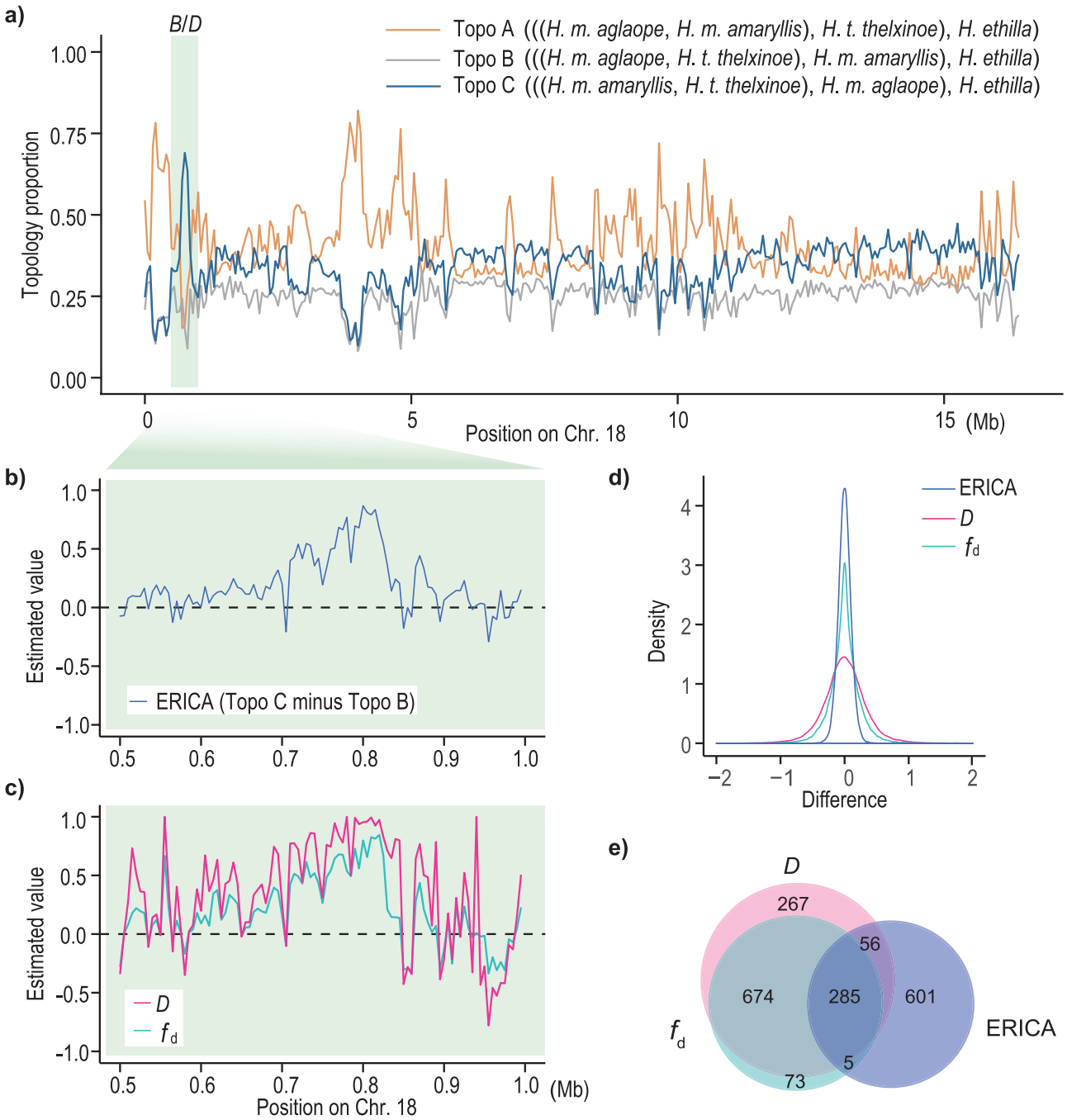


FIGURE 3. Signatures of introgression in *Heliconius* butterflies. The proportions of three topologies for each 50-kb adjacent window were inferred across chromosome 18 using ERICA a). Color pattern locus *B/D*, which is known to be introgressed between *H. m. amaryllis* and *H. t. thelxinoe*, is highlighted in green a–b). Zooming in to the *B/D* locus, the signature of introgression was evaluated for each 5-kb window using ERICA b), the *D*-statistic and f_d c). The ERICA results are shown in the form of proportion of topology ((((*H. m. amaryllis*, *H. t. thelxinoe*), *H. m. aglaope*), *H. ethilla*)) minus proportion of topology ((((*H. m. aglaope*, *H. t. thelxinoe*), *H. m. amaryllis*), *H. ethilla*)) to make them comparable with the *D*-statistic and f_d results. For the three approaches, the differences between the values of two adjacent 5-kb windows were calculated, and their distributions were plotted to indicate the intensity of the random error d). A Venn diagram was plotted to show the overlapping 50-kb windows detected by ERICA, the *D*-statistic and f_d e).

protein-coding gene controlling plant innate immunity, *EBR1* (You et al 2016), showed both the strongest and longest signature of introgression between *japonica* and *indica* (up to 0.45 Mb in length), demonstrating the high functionality of this introgression locus (Supplementary Table S9). These results suggest that ERICA is a useful tool for identifying adaptive introgression among complex patterns of introgression.

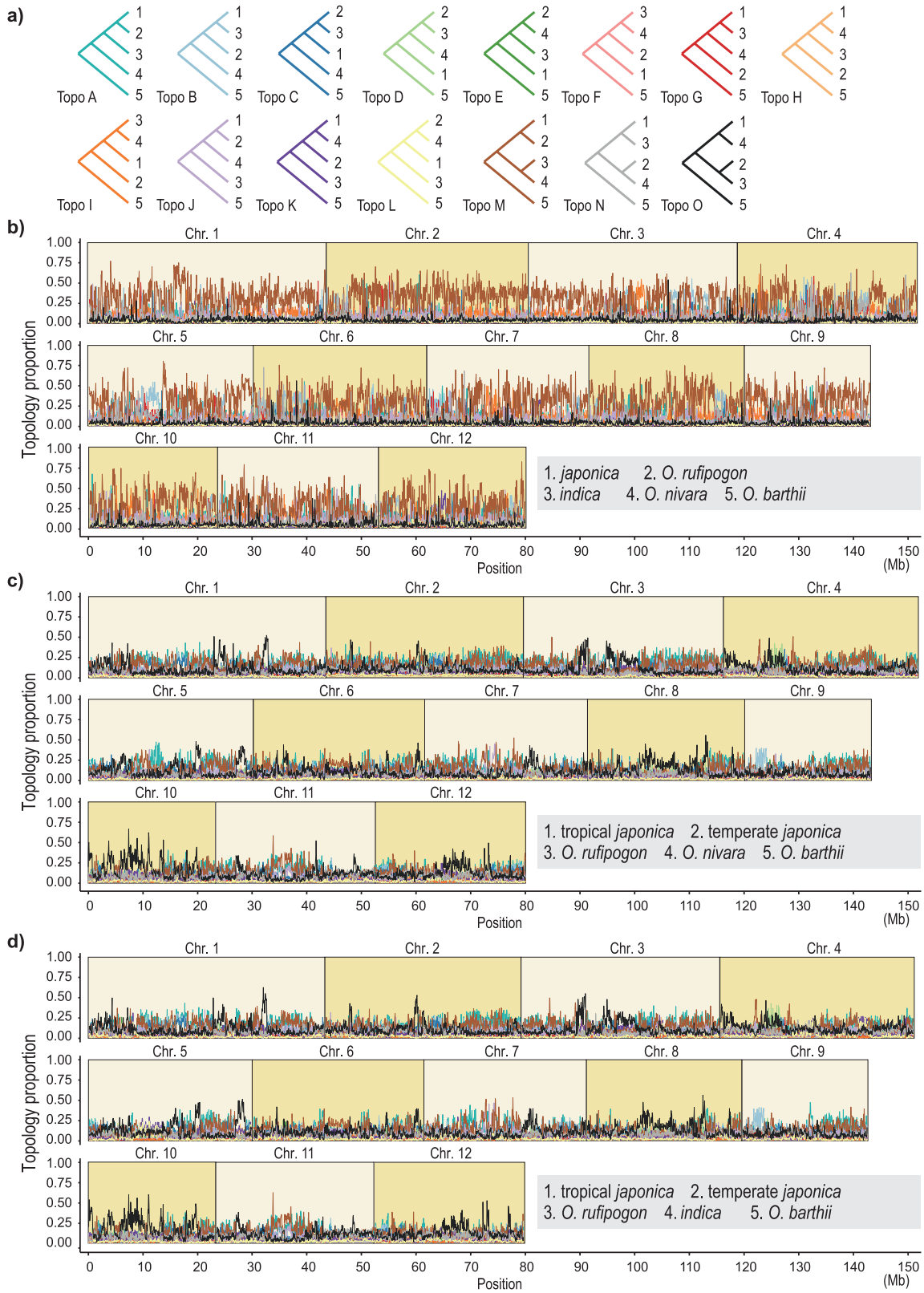
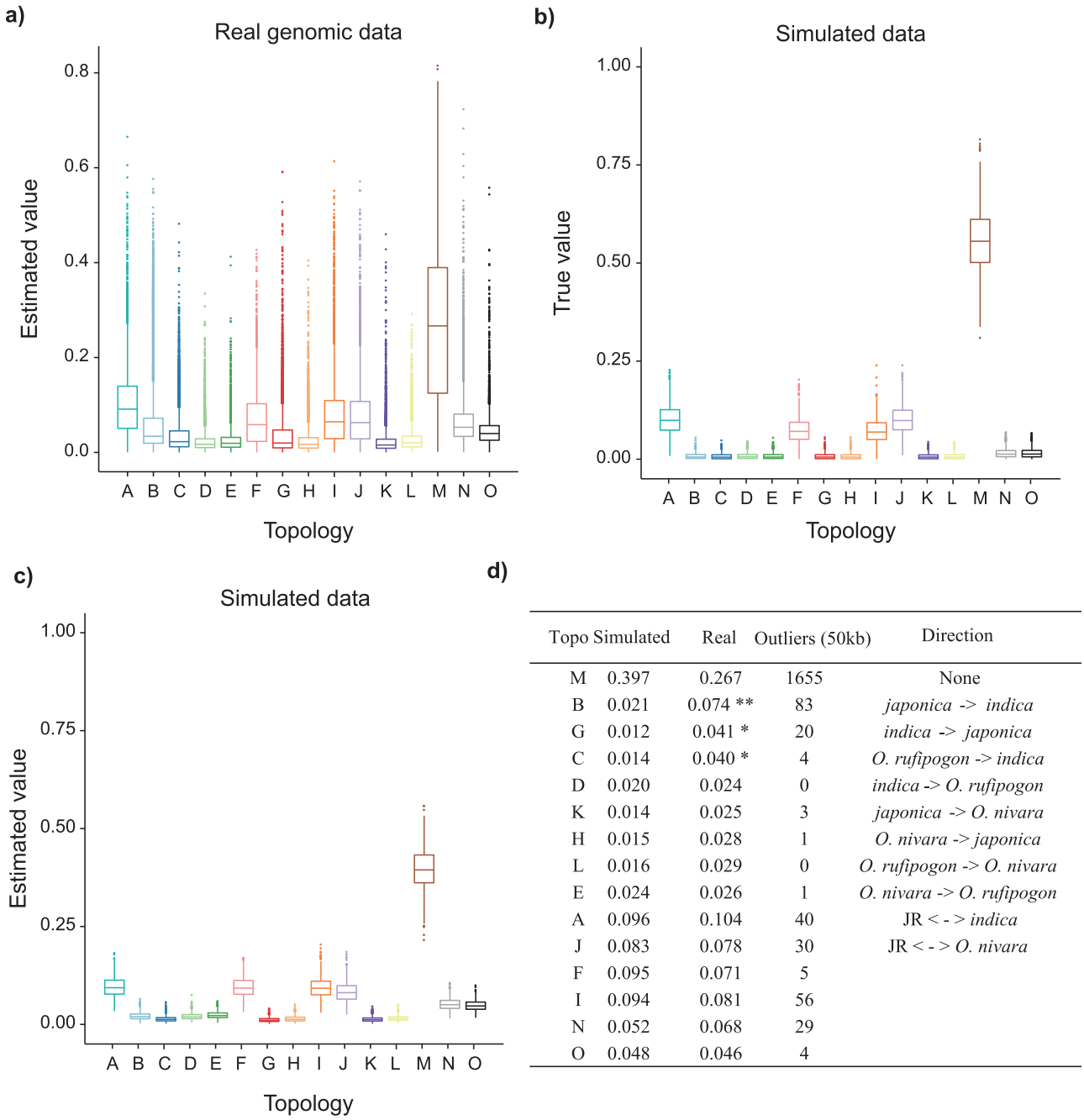


FIGURE 4. Genome-wide patterns of admixture in rice domestication and adaptation. There are 15 representative topologies suggesting possible relationships for four ingroup rice accessions, given *O. barthii* as an outgroup taxon a). Genome-wide evolutionary relationships were predicted using ERICA for Asian cultivated and wild rice (*japonica*, *O. rufipogon*, *indica*, *O. nivara*) b) and for tropical and temperate rice accessions (tropical *japonica*, temperate *japonica*, *O. rufipogon*, *O. nivara* c) and tropical *japonica*, temperate *japonica*, *O. rufipogon*, *indica* d)). The color codes correspond to the topologies shown in a).



d)

Topo	Simulated	Real	Outliers (50kb)	Direction
M	0.397	0.267	1655	None
B	0.021	0.074 **	83	*japonica* -> *indica*
G	0.012	0.041 *	20	*indica* -> *japonica*
C	0.014	0.040 *	4	*O. rufipogon* -> *indica*
D	0.020	0.024	0	*indica* -> *O. rufipogon*
K	0.014	0.025	3	*japonica* -> *O. nivara*
H	0.015	0.028	1	*O. nivara* -> *japonica*
L	0.016	0.029	0	*O. rufipogon* -> *O. nivara*
E	0.024	0.026	1	*O. nivara* -> *O. rufipogon*
A	0.096	0.104	40	JR < - > *indica*
J	0.083	0.078	30	JR < - > *O. nivara*
F	0.095	0.071	5	
I	0.094	0.081	56	
N	0.052	0.068	29	
O	0.048	0.046	4	

FIGURE 5. Disentangling introgression and ILS in rice domestication. The genome-wide proportion of each topology was inferred using ERICA for every 50-kb window across the genomes of Asian cultivated and wild rice (*japonica*, *O. rufipogon*, *indica*, *O. nivara*) given *O. barthii* as an outgroup taxon. The results are summarized as a boxplot. The 15 topologies correspond to topological structures shown in Fig. 4a a). To estimate the strength of ILS, a dataset including 1000 50-kb windows was simulated according to rice demography without gene flow, which showed four different levels among possible topologies in topology weighting b). The simulated dataset was analyzed using ERICA and showed four levels of proportion c). The putative introgressed loci were summarized by comparing the ERICA results of the real and simulated data. The *P* values were calculated using Chebyshev's Theorem, with a sample mean of 0.017 and a standard deviation of 0.004 for eight topologies that had the same distribution. ** indicates *P* value < 0.01 and * indicates *P* value < 0.05. d).

Dissecting local adaptive introgression in rice based on pan-genomic data.—We further dissected the signatures of introgression at a population level by applying ERICA to a rice pan-genome dataset including 66 divergent rice accessions of both wild and cultivated rice (Zhao et al. 2018). We observed that the 66 accessions formed distinct clades, but with patterns of admixture in *O. rufipogon* and among tropical accessions, for example, *O. nivara*, tropical *japonica*, and *aus* rice, and most accessions of *aus* rice were assigned to the *O. nivara* clade (Supplementary Fig. S28). We, therefore, focused on detecting introgression between tropical *japonica*

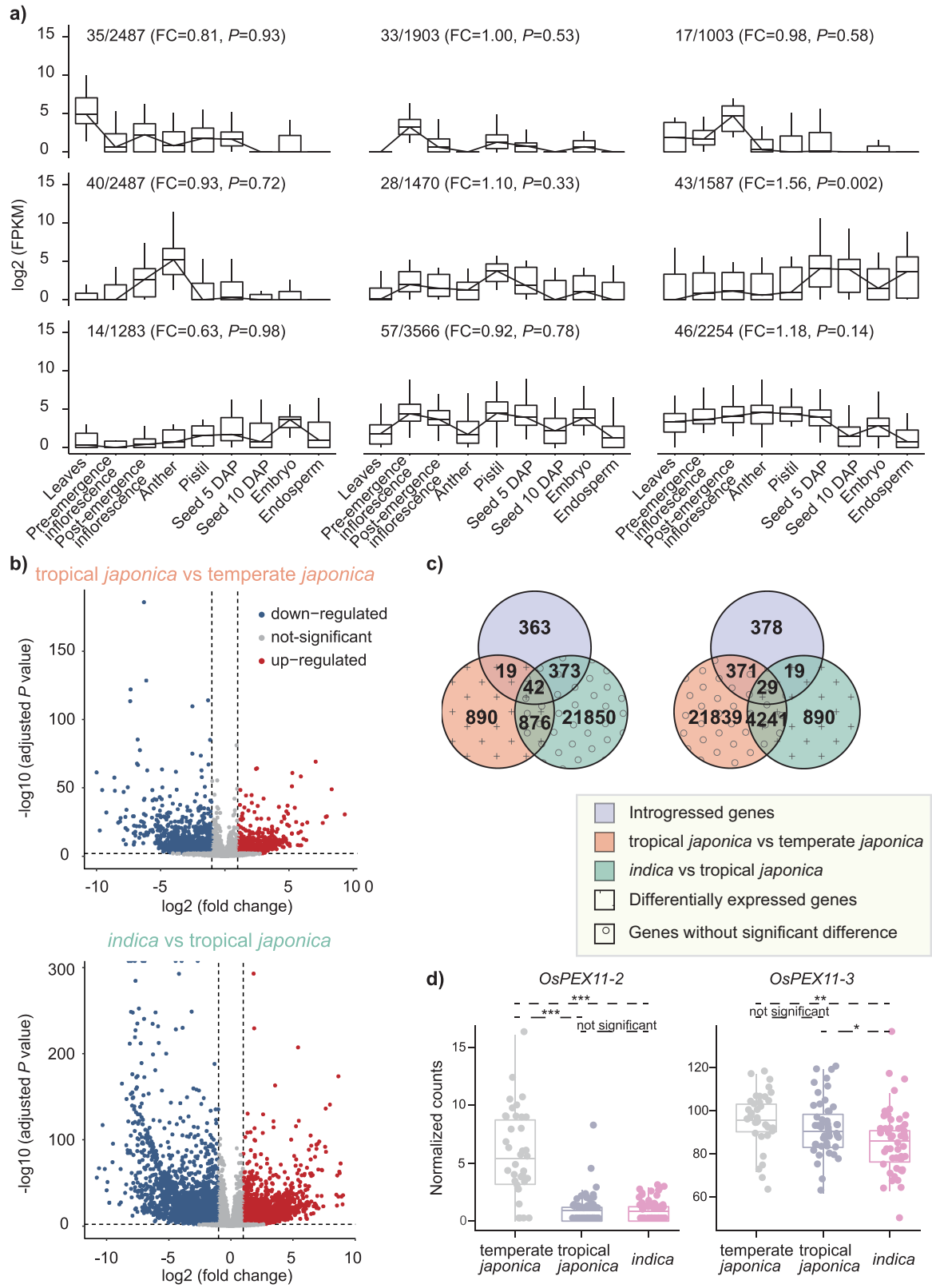


FIGURE 6. Expression patterns of introgressed genes in different rice tissues and clades. a) The expression levels of introgressed genes were assigned to nine co-expression clusters. Boxplots show the ranges of log2 FPKM. Line plots show the medians across nine tissues, including leaves at 20 days after sowing, primordial inflorescences at 10 days before flower emergence, whole inflorescences at the time of flower emergence, anthers and pistils at the time of anthesis, whole seeds at 5 days after pollination (DAP), whole seeds at 10 DAP, embryos at 25 DAP, and endosperm at 25 DAP. The numbers of introgressed genes/all expressed genes in each cluster are annotated above the relevant cluster,

and other tropical accessions using ERICA. Our results suggest a genome-wide pattern of introgression from *indica* and *O. nivara* to tropical *japonica* (Fig. 4c and Supplementary Tables S9–S10). We identified 55 putative introgressed loci with high confidence (proportion of topology O > 0.4) and absolute genetic divergence lower than the chromosomal mean divergence, ranging from 50 kb to 300 kb in length and including 797 genes (Supplementary Table S9). Notably, the results of GO and Plant Experimental Conditions Ontology (PECO) analyses suggested that these introgression loci were likely involved in tropical adaptation and stress resistance. The top terms yielded by the GO analysis included response to light stimulus, salicylic acid mediated signaling pathway, cellular response to phosphate starvation and homoiothermy, whereas the significant terms yielded by the PECO analysis included continuous dark exposure, *Magnaporthe grisea* exposure, sodium chloride exposure and cold or sub-optimal temperature exposure (Supplementary Fig. S27b and Table S9). In addition, some known introgressed genes were also located within the range of LD to the introgression hotspots (Supplementary Table S9), such as the grain size locus *OsSPL13* and the thermotolerance gene *OsTT1*, which were introgressed from *indica* to tropical *japonica* and were selected for large grain size (Si et al. 2016) and local adaptation to tropical temperatures (Li et al. 2015), respectively, indicating that both artificial selection and natural selection were involved in the fixation of introgressed loci. In addition, we found that tropical *japonica* received another thermotolerance gene, *OsCaM1-1*, which plays a crucial role in the Ca^{2+} signal-mediated heat shock response in rice (Wu et al. 2012), suggesting a potential role in the heat adaptation process of tropical accessions.

Some of the introgressed genes show similar expression patterns.—To explore the expression patterns of the introgressed genes, we also investigated the co-expression clusters of all rice genes. After the removal of low-expression genes and *k*-means clustering, 18,040 genes were divided into nine clusters, seven of which showed tissue-specific up-regulation patterns. Among the 797 introgressed genes, 313 genes with a detection rate higher than the genome average were assigned to the nine clusters described above and found to be significantly enriched in a set of genes expressed in early and late seed development (hypergeometric test *P* value = 0.002) (Fig. 6a). We also analyzed differentially expressed genes (DEGs) among the three clades of temperate *japonica*, tropical *japonica*, and *indica* using transcriptome data of 67 rice accessions. We found that the similarity of the expression patterns of temperate *japonica* and tropical *japonica* was greater than the similarity of either of these patterns to that of *indica* (**Fig. 6b**), suggesting a closer evolutionary relationship between temperate *japonica* and tropical *japonica*. We further focused on genes with the same expression level in tropical *japonica* and *indica*, but not in temperate *japonica*, which were likely involved in the local adaptation of tropical *japonica*, yielding 918 genes, among which 42 genes (4.6%) overlapped with the introgressed loci (**Fig. 6c**). In contrast, for the 4270 genes expressed similarly in tropical *japonica* and temperate *japonica*, but differentially in *indica*, only 29 genes (0.7%) overlapped with the introgressed loci, indicating that gene flow may have played an important role in both the genomic and transcriptomic differentiation of Asian cultivated rice. Note that only seedlings were sampled in our analyses, and the DEGs may not represent all transcriptomic differences. One of the remarkably introgressed DEGs was *OsPEX11-2*, belonging to the rice *PEX11* gene family, which is implicated in peroxisome biogenesis and maintenance. *OsPEX11-2* was down-regulated in both tropical *japonica* and *indica*, whereas its neighboring paralogue, the salt stress tolerance gene *OsPEX11-3* (Cui et al. 2016), showed no significant expression difference (**Fig. 6d**). Both *OsPEX11-2* and *OsPEX11-3* were included in a putative introgressed locus from *indica*/*O. nivara* to tropical *japonica* inferred by ERICA. However, the two genes had differential expression patterns under normal and stress situations (Nayidu et al. 2008), indicating that they may have divergent functions to suit different types of peroxisomes. Although *OsPEX11-2* did not respond to common stresses and its role in peroxisome organization remains unclear, our results show that it was selected in tropical *japonica* during adaptation to the local environment, probably by influencing photorespiration.

Discussion

Data Labeling and Model Architectures of CNNs

In this study, we present an efficient and robust CNN-based pipeline to infer complex evolutionary history, demonstrating the remarkable potential of deep learning for generating sequence-based evolutionary inferences. Owing to the intrinsic property of universal

with the fold changes (FC) of the introgressed sets and the *P* values of hypergeometric tests enclosed in parentheses. b) Volcano plots show differentially expressed genes (DEGs) between tropical *japonica*/temperate *japonica* and between *indica*/tropical *japonica*. Log2 transformed FC were plotted against log10 transformed statistical significances. Red and blue points represent up-regulated and down-regulated genes, respectively. c) Venn diagrams show the overlaps between sets of introgressed genes and DEGs for three different rice accession clades: temperate *japonica*, tropical *japonica*, and *indica*. d) Examples of introgressed genes with and without expression differences. *OsPEX11-2* has a lower expression level in tropical *japonica* and *indica* in comparison with that of temperate *japonica*, whereas *OsPEX11-3* shows no significant difference in expression among these accession clades. The box plots summarize the normalized counts (DESeq2's median of ratios) for each clade. *** indicates adjusted *P* value < 0.001, ** indicates adjusted *P* value < 0.01 and * indicates adjusted *P* value < 0.05.

function approximators, CNNs are suitable for feature extraction in a growing number of classification tasks. However, the data labeling strategy can limit the model performance and data capacity. For example, the maximum data capacities of two previously reported CNN models were limited to four sequence alignments, and the task is more likely to be classification instead of quantification (Suvorov et al. 2020; Zou et al. 2020). Instead, we generated a vector recording the proportions of all possible topologies of genomic windows, which were related to both spatial heterogeneity and sample heterogeneity, as a data label. This labeling strategy recorded more information than simple classification for downstream analyses, and its relatively lower dimensionality decreased the amount of computation and the risk of overfitting, consequently increasing the generalization capability and robustness of the model. In short, this labeling strategy allowed ERICA to accommodate both genome assemblies and population genomic data in multiple taxa, while enabling ERICA to quantify the evolutionary history across the genome, demonstrating the potential of deep learning in handling population genomic data and quantifying complex history.

For the network architecture, we referred to previous studies (Suvorov et al. 2020; Zou et al. 2020) and tried different models, including simple convolution layers and Residual Networks. We finally selected the current networks with multiple residual and dense blocks and the best performance in the training process. Although the increase in model parameters may introduce the risk of over-fitting, we found that the losses were decreased in all of the training, validation and independent test datasets, indicating that the networks were not over-fitted. In addition, the losses approached the minimum rapidly, which were convergent within one epoch in both the four-taxon and five-taxon models, suggesting that the size of training datasets was sufficient to train our networks. Other hyper-parameters, including the number of convolution layers, kernel size, activation function, learning rate, and batch size may also be related to the accuracy, training time, and resource consumption of the model. Optimizing these parameters will aid further improvement of the performance of ERICA models.

Model Generalization and Specialization

As the effectiveness of population genetic inference is related to demographic histories and there is a trade-off between the model generalization ability and accuracy under a specific evolutionary scenario, previous machine learning and deep learning approaches, such as CRF (Sankararaman et al. 2014), HMM (Skov et al. 2018), FILET (Schrider et al. 2018), and genomatnn (Gower et al. 2021), require a well-studied and species-specific demographic model for parameter training, and thus are difficult to apply to non-model organisms and increase computational complexity. Instead of using simulations from a specific demographic scenario for model training (example.g., the demographic histories of *Drosophila* sister species used in FILET (Schrider et al. 2018) and the human demographic models used in genomatnn (Gower et al. 2021)), our training datasets covered a large number of evolutionary scenarios with variable divergence times and gene flow intensities, and we evaluated the performance on test datasets with large variations in simulation parameters. Our results suggested that the performance of ERICA may be influenced by demographic history, which may be the result of a combination of two factors. The first factor influencing demographic history is the differences between the simulation parameters of the test and training datasets, just as the error rates of ERICA's topology inference varied with population sizes; the second factor is the demographic history of the focal taxa, as the performance of other methods for detecting introgression showed the same trend as ERICA. Nevertheless, we found that ERICA models had accuracy comparable to that of other established methods, even with large changes in population sizes and divergence times.

In particular, for the detection of adaptive introgression, ERICA was found to have the highest sensitivity in most cases (Supplementary Table S6). We also found that the error decreased with increasing window size, suggesting that a larger window size can be used when the real demographic scenario differs greatly from that of the training dataset. Evaluations using datasets under human demographic scenarios also showed that ERICA can efficiently detect introgressions under ancient and strong selection. Thus, when the real population history is broadly consistent with the training dataset, ERICA's trained models can be used directly without retraining and can be applied to different taxa, including animals and plants. In future research, we aim to optimize ERICA through model training in two directions. On the one hand, the generalization ability of ERICA can be further improved by increasing the range of simulation parameters and the number of demographic models for training datasets, such as by introducing a wider range of population sizes, population dynamics over time and species, different selection pressures, and more sequence errors. To efficiently generate more realistic scenarios, flexible and programmable simulators, such as the coalescent simulator msprime (Kelleher et al. 2016) and the forward simulator SLiM (Haller and Messer 2019) can be used as alternatives to ms, which can facilitate the simulation of scenarios with larger sample sizes, recombination hotspots, and selection. However, it is important to note that more attempts are needed due to the higher heterogeneity of the training dataset, which may increase the difficulty of model training. On the other hand, for species with a well-characterized evolutionary history, species-specific data can be used as training datasets to obtain higher accuracy. Considering the generalization ability of current models, in both cases, fine-tuning the network parameters starting from the pre-trained models may help to reduce training time and training data size in comparison with training from scratch.

In addition, for the approaches based on sequence features without model training, the strength of the

background noise and introgression signal can also be affected by the population history; for example, the IBD score of the null hypothesis without gene flow increases with decreasing divergence time, whereas it decreases with decreasing effective population size. Therefore, it is difficult to assess the specificity and sensitivity of preset thresholds under unknown population histories, and inappropriate thresholds can strongly reduce the accuracy of the analyses (Supplementary Fig. S22). These findings also indicate that distinguishing local signatures of introgression from the genomic background under complex demographic scenarios is an inevitable challenge for general approaches. We have determined a threshold by evaluating the impact of ILS under extreme conditions to simplify the question, which could help identify introgressed loci that are highly credible. For most scenarios under rapid speciation or with closely related taxa, the intensity of ILS is strong and comparable with our hypothesis. Nevertheless, we still cannot rule out possible false-negative errors, because the proportions of discordant topologies caused by ILS are likely less than this threshold in other real genomic data. Therefore, some introgressed loci with weaker signals may be filtered out by the stringent threshold, owing to either having undergone recombination or becoming unfixed in the recipient population, which indicates that these loci may have relatively small effects on adaptation. For these scenarios, we suggest that the filtering threshold should be flexible and customizable. For example, with more prior knowledge such as the demographic history of the focal taxa, the features of ILS can be modeled and evaluated specifically, which is helpful in determining a more appropriate threshold for a given system. In conclusion, the ERICA pipeline is a general and easy-to-use utility, which is ready to use and can have applications in a wide range of scenarios. As a new attempt to make such deep learning-based applications available for broad usage, both an online submission portal and a local version are offered to meet different needs.

More Features and More Taxa

The reduction of absolute divergence is considered another signature of introgression; and a recent method, QuIBL (Edelman et al. 2019), uses internal branch lengths instead of gene tree discordance to infer introgression. Therefore, we employed DNA sequence divergence as an additional filtering step, and we considered learning features of the distance between sequences using CNN as a potential way to further improve the accuracy of ERICA models. In addition, current ERICA models can deal with up to five taxa, which we consider an appropriate and applicable number for many tasks. When studying more than five taxa, possible combinations of four or five taxa can be used for the analyses. To reduce the consumption of computing resources, taxon combinations can be selected according to biological problems, prior knowledge, or results of other population genetic analyses.

Conclusions

We present a feasible scheme to detect introgression signals using deep learning algorithms with DNA sequence data as input, which includes data labeling, sequence encoding, neural network structure, data simulation, and model training. We designed two CNNs with multiple dense and residual blocks and trained the models with simulated data under various demographic scenarios of phylogenies, divergence times, and gene flow events. The pre-trained models can be used to predict relationships of four and five focal taxa from MSAs, and the local signals of introgression between non-sister species can be identified via discordant topologies. We evaluated the accuracy and robustness using several test datasets and show that the ERICA approach performs well in inferring topology proportions and gene flow in most cases. The adaptive introgression regions detected in *Heliconius* and *Oryza* suggest that ERICA is applicable to multiple taxa, which may contribute to studies of introgression on a variety of organisms. We provide the source code, trained models and a web server to aid the use of ERICA. We also suggest that the models can be further fine-tuned by covering more demographic models in the training dataset to improve generalization or by providing a species-specific training dataset to improve accuracy.

Supplementary Information

Data available from the Dryad Digital Repository: (https://doi.org/10.5061/dryad.m905qfv6d).

Data Availability

We downloaded Illumina paired-end raw reads from the NCBI Sequence Archive (SRA) with accession numbers NCBI PRJNA308754, PRJEB1749, PRJNA73595, and PRJEB11772. Illumina single-end RNA-seq reads were downloaded from the NCBI SRA with the accession number PRJNA385135. Genome assemblies were downloaded from NCBI Genbank with the following accession numbers: GCA_001433935.1, GCA_000004655.2, GCA_000817225.1, GCA_000576065.1, GCA_000147395.2, GCA_000182155.3, GCA_000576495.1, GCA_000338895.2, GCA_000231095.2, GCA_000573905.1, and GCA_000325765.3. The pan-genome dataset was downloaded from the RicePanGenome database (http://db.ncgr.ac.cn/RicePanGenome/). The simulated datasets used for model training and testing are available at http://erica.cibr.ac.cn. The trained models and web server are available at http://erica.cibr.ac.cn. The source code and scripts are hosted on GitHub (https://github.com/YuboZhangPKU/ERICA). The supplementary information, scripts and datasets for model training, and dataset used for G-PhoCS analyses are available on Dryad (https://doi.org/10.5061/dryad.m905qfv6d).

Acknowledgments

We would like to thank the Computing Platforms of the Center for Life Sciences and the School of Life Sciences at Peking University for their assistance with computation.

Funding

This project was supported by grants from the National Natural Science Foundation of China (32170420 to WZ, 32170642 to LZ, 32170622 to YO, 31871271 to WZ), the Beijing Natural Science Foundation (JQ19021 to WZ), the Peking-Tsinghua Center for Life Science, the State Key Laboratory of Protein and Plant Gene Research, Qidong-SLS Innovation Fund, and Benyuan Charity Young Investigator Exploration Fellowship in Life Science to WZ.

Conflict of interests

The authors declare that they have no competing interests.

Author contributions

WZ and LZ conceived and designed the study; YZ, YJ, WZ, and YO generated simulated datasets and performed population genomic analyses; QZ, YS, and LZ trained the CNN-based models and constructed the online tools. YZ, WZ, QZ, and LZ wrote the manuscript with input from YJ, YS, and YO; all authors proofread and approved the manuscript.

ETHICAL APPROVAL

No ethical approval was required.

REFERENCES

Adrion J.R., Cole C.B., Dukler N., Galloway J.G., Gladstein A.L., Gower G., Kyriazis C.C., Ragsdale A.P., Tsambos G., Baumdicker F., Carlson J., Cartwright R.A., Durvasula A., Gronau I., Kim B.Y., McKenzie P., Messer P.W., Noskova E., Ortega-Del Vecchyo D., Racimo F., Struck T.J., Gravel S., Gutenkunst R.N., Lohmueller K.E., Ralph P.L., Schrider D.R., Siepel A., Kelleher J., Kern A.D. 2020. A community-maintained standard library of population genetic models. Elife 9:e54967.

Alexander D.H., Novembre J., Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 19:1655–1664.

Arnold M.L. 2004. Transfer and origin of adaptations through natural hybridization: were Anderson and Stebbins right? Plant Cell 16:562–570.

Besenbacher S., Hvilsom C., Marques-Bonet T., Mailund T., Schierup M.H. 2019. Direct estimation of mutations in great apes reconciles phylogenetic dating. Nat. Ecol. Evol. 3:286–292.

Blanchette M., Kent W.J., Riemer C., Elnitski L., Smit A.F., Roskin K.M., Baertsch R., Rosenbloom K., Clawson H., Green E.D., Haussler D., Miller W. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res. 14:708–715.

Bolger A.M., Lohse M., Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120.

Browning S.R., Browning B.L., Zhou Y., Tucci S., Akey J.M. 2018. Analysis of human sequence data reveals two pulses of archaic Denisovan admixture. Cell 173:53–61.e9.

Cooper L., Meier A., Laporte M.A., Elser J.L., Mungall C., Sinn B.T., Cavaliere D., Carbon S., Dunn N.A., Smith B., Qu B., Preece J., Zhang E., Todorovic S., Gkoutos G., Doonan J.H., Stevenson D.W., Arnaud E., Jaiswal P. 2018. The Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics. Nucleic Acids Res. 46:D1168–D1180.

Campbell M.T., Du Q., Liu K., Sharma S., Zhang C., Walia H. 2020. Characterization of the transcriptional divergence between the subspecies of cultivated rice (*Oryza sativa*). BMC Genom. 21:394.

Chen L., Wolf A.B., Fu W., Li L., Akey J.M. 2020. Identifying and interpreting apparent Neanderthal ancestry in African individuals. Cell 180:677–687.e16.

Choi J.Y., Platts A.E., Fuller D.Q., Hsing Y.I., Wing R.A., Purugganan M.D. 2017. The rice paradox: multiple origins but single domestication in Asian rice. Mol. Biol. Evol. 34:969–979.

Civáň P., Craig H., Cox C.J., Brown T.A. 2015. Three geographically separate domestications of Asian rice. Nat. Plants 1:15164.

Cui P., Liu H., Islam F., Li L., Farooq M.A., Ruan S., Zhou W. 2016. OsPEX11, a peroxisomal biogenesis factor 11, contributes to salt stress tolerance in *Oryza sativa*. Front. Plant Sci. 7:1357.

Currat M., Ruedi M., Petit R.J., Excoffier L. 2008. The hidden side of invasions: massive introgression by local genes. Evolution 62:1908–1920.

Davey J.W., Chouteau M., Barker S.L., Maroja L., Baxter S.W., Simpson F., Merrill R.M., Joron M., Mallet J., Dasmahapatra K.K., Jiggins C.D. 2016. Major improvements to the *Heliconius melpomene* genome assembly used to confirm 10 chromosome fusion events in 6 million years of butterfly evolution. G3 (Bethesda) 6:695–708.

Davidson R.M., Gowda M., Moghe G., Lin H., Vaillancourt B., Shiu S.H., Jiang N., Robin Buell C. 2012. Comparative transcriptomics of three Poaceae species reveals patterns of gene expression evolution. Plant J. 71:492–502.

DePristo M.A., Banks E., Poplin R., Garimella K.V., Maguire J.R., Hartl C., Philippakis A.A., del Angel G., Rivas M.A., Hanna M., McKenna A., Fennell T.J., Kernytsky A.M., Sivachenko A.Y., Cibulskis K., Gabriel S.B., Altshuler D., Daly M.J. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. 43:491–498.

Durand E.Y., Patterson N., Reich D., Slatkin M. 2011. Testing for ancient admixture between closely related populations. Mol. Biol. Evol. 28:2239–2252.

Edelman N.B., Frandsen P.B., Miyagi M., Clavijo B., Davey J., Dikow R.B., García-Accinelli G., Van Belleghem S.M., Patterson N., Neafsey D.E., Challis R., Kumar S., Moreira G.R.P., Salazar C., Chouteau M., Counterman B.A., Papa R., Blaxter M., Reed R.D., Dasmahapatra K.K., Kronforst M., Joron M., Jiggins C.D., McMillan W.O., Di Palma F., Blumberg A.J., Wakeley J., Jaffe D., Mallet J. 2019. Genomic architecture and introgression shape a butterfly radiation. Science 366:594–599.

Eraslan G., Avsec Z., Gagneur J., Theis F.J. 2019. Deep learning: new computational modelling techniques for genomics. Nat. Rev. Genet. 20:389–403.

Estabrook G.F., McMorris F.R., Meacham C.A. 1985. Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. Syst. Biol. 34:193–200.

Ewing G., Hermisson J. 2010. MSMS: a coalescent simulation program including recombination, demographic structure and selection at a single locus. Bioinformatics 26:2064–2065.

Felsenstein J. 1978. The number of evolutionary trees. Syst. Biol. 27:27–33.

Flagel L., Brandvain Y., Schrider D.R. 2019. The unreasonable effectiveness of convolutional neural networks in population genetic inference. Mol. Biol. Evol. 36:220–238.

Garris A.J., Tai T.H., Coburn J., Kresovich S., McCouch S. 2005. Genetic structure and diversity in *Oryza sativa* L. Genetics 169:1631–1638.

Gower G., Picazo P.I., Fumagalli M., Racimo F. 2021. Detecting adaptive introgression in human evolution using convolutional neural networks. Elife 10:e64669.

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

Gronau I., Hubisz M.J., Gulko B., Danko C.G., Siepel A. 2011. Bayesian inference of ancient human demography from individual genome sequences. Nat. Genet. 43:1031–1034.

Guan Y. 2014. Detecting structure of haplotypes and local ancestry. Genetics 196:625–642.

Haller B.C., Messer P.W. 2019. SLiM 3: forward genetic simulations beyond the Wright-Fisher model. Mol. Biol. Evol. 36:632–637.

Harris R.S. 2007. Improved pairwise alignment of genomic DNA [Ph.D. thesis]. The Pennsylvania State University.

He K., Zhang X., Ren S., Sun J. 2016. Deep residual learning for image recognition. Proc. IEEE Conf. Computer Vis. Pattern Recogn. 770–778.

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.

Heliconius Genome Consortium. 2012. Butterfly genome reveals promiscuous exchange of mimicry adaptations among species. Nature 487:94–98.

Hibbins M.S., Hahn M.W. 2019. The timing and direction of introgression under the multispecies network coalescent. Genetics 211:1059–1073.

Hibbins M.S., Hahn M.W. 2022. Phylogenomic approaches to detecting and characterizing introgression. Genetics 220:iyab173.

Howe E.A., Sinha R., Schlauch D., Quackenbush J. 2011. RNA-Seq analysis in MeV. Bioinformatics 27:3209–3210.

Huang G., Liu Z., Van Der Maaten L., Weinberger K.Q. 2017. Densely connected convolutional networks. Proc. IEEE Conf. Computer Vis. Pattern Recogn. 4700–4708.

Huang X., Wei X., Sang T., Zhao Q., Feng Q., Zhao Y., Li C., Zhu C., Lu T., Zhang Z., Li M., Fan D., Guo Y., Wang A., Wang L., Deng L., Li W., Lu Y., Weng Q., Liu K., Huang T., Zhou T., Jing Y., Li W., Lin Z., Buckler E.S., Qian Q., Zhang Q.-F., Li J., Han B. 2010. Genome-wide association studies of 14 agronomic traits in rice landraces. Nat. Genet. 42:961–967.

Huang X., Kurata N., Wei X., Wang Z.X., Wang A., Zhao Q., Zhao Y., Liu K., Lu H., Li W., Guo Y., Lu Y., Zhou C., Fan D., Weng Q., Zhu C., Huang T., Zhang L., Wang Y., Feng L., Furuumi H., Kubo T., Miyabayashi T., Yuan X., Xu Q., Dong G., Zhan Q., Li C., Fujiyama A., Toyoda A., Lu T., Feng Q., Qian Q., Li J., Han B. 2012. A map of rice genome variation reveals the origin of cultivated rice. Nature 490:497–501.

Huang X., Han B. 2015. Rice domestication occurred through single origin and multiple introgressions. Nat. Plants 2:15207.

Hudson R.R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics 18:337–338.

Jin J., Huang W., Gao J.P., Yang J., Shi M., Zhu M.Z., Luo D., Lin H.X. 2008. Genetic control of rice plant architecture under domestication. Nat. Genet. 40:1365–1369.

Kawahara Y., de la Bastide M., Hamilton J.P., Kanamori H., McCombie W.R., Ouyang S., Schwartz D.C., Tanaka T., Wu J., Zhou S., et al. 2013. Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. Rice (NY) 6:4.

Kelleher J., Etheridge A.M., McVean G. 2016. Efficient coalescent simulation and genealogical analysis for large sample sizes. PLoS Comput. Biol. 12:e1004842.

Kent W.J., Baertsch R., Hinrichs A., Miller W., Haussler D. 2003. Evolution's cauldron: duplication, deletion, and rearrangement in the mouse and human genomes. Proc. Natl. Acad. Sci. U.S.A. 100:11484–11489.

Kim D., Pertea G., Trapnell C., Pimentel H., Kelley R., Salzberg S.L. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14:R36.

Langmead B., Salzberg S.L. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9:357–359.

Letunic I., Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 47:W256–W259.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R., 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics 34:3094–3100.

Li X.M., Chao D.Y., Wu Y., Huang X., Chen K., Cui L.G., Su L., Ye W.W., Chen H., Chen H.C., Dong N.-Q., Guo T., Shi M., Feng Q., Zhang P., Han B., Shan J.-X., Gao J.-P., Lin H.-X. 2015. Natural alleles of a proteasome α2 subunit gene contribute to thermotolerance and adaptation of African rice. Nat. Genet. 47:827–833.

Lin Z., Li.X., Shannon L.M., Yeh C.T., Wang M.L., Bai G., Peng Z., Li J., Trick H.N., Clemente T.E., et al. 2012. Parallel domestication of the *Shattering1* genes in cereals. Nat. Genet. 44:720–724.

Love M.I., Huber W., Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 15:550.

Mallet J. 2005. Hybridization as an invasion of the genome. Trends Ecol. Evol. 20:229–237.

Martin S.H., Dasmahapatra K.K., Nadeau N.J., Salazar C., Walters J.R., Simpson F., Blaxter M., Manica A., Mallet J., Jiggins C.D. 2013. Genome-wide evidence for speciation with gene flow in *Heliconius* butterflies. Genome Res. 23:1817–1828.

Martin S.H., Davey J.W., Jiggins C.D. 2015. Evaluating the use of ABBA-BABA statistics to locate introgressed loci. Mol. Biol. Evol. 32:244–257.

Martin S.H., Van Belleghem S.M. 2017. Exploring evolutionary relationships across the genome using topology weighting. Genetics 206:429–438.

Martin S.H., Davey J.W., Salazar C., Jiggins C.D. 2019. Recombination rate variation shapes barriers to introgression across butterfly genomes. PLoS Biol. 17:e2006288.

Mondal M., Bertranpetit J., Lao O. 2019. Approximate Bayesian computation with deep learning supports a third archaic introgression in Asia and Oceania. Nat. Commun. 10:246.

Nayidu N.K., Wang L., Xie W., Zhang C., Fan C., Lian X., Zhang Q., Xiong L. 2008. Comprehensive sequence and expression profile analysis of *PEX11* gene family in rice. Gene 412:59–70.

Pamilo P., Nei M. 1988. Relationships between gene trees and species trees. Mol. Biol. Evol. 5:568–583.

Pease J.B., Hahn M.W. 2015. Detection and polarization of introgression in a five-taxon phylogeny. Syst. Biol. 64:651–662.

Pickrell J.K., Pritchard J.K. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. 8:e1002967.

Plagnol V., Wall J.D. 2006. Possible ancestral structure in human populations. PLoS Genet. 2:e105.

Price A.L., Patterson N.J., Plenge R.M., Weinblatt M.E., Shadick N.A., Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. Nat. Genet. 38:904–909.

Price A.L., Tandon A., Patterson N., Barnes K.C., Rafaels N., Ruczinski I., Beaty T.H., Mathias R., Reich D., Myers S. 2009. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. PLoS Genet. 5:e1000519.

Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M.A., Bender D., Maller J., Sklar P., de Bakker P.I., Daly M.J., Sham P.C. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81:559–575.

Racimo F., Sankararaman S., Nielsen R., Huerta-Sánchez E. 2015. Evidence for archaic adaptive introgression in humans. Nat. Rev. Genet. 16:359–371.

Rambaut A., Grassly N.C. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput. Appl. Biosci. 13:235–238.
Rieseberg L.H. 2019. Mapping footprints of past genetic exchange. Science 366:570–571.
Rosenberg N.A. 2002. The probability of topological concordance of gene trees and species trees. Theor. Popul. Biol. 61:225–247.
Saitoh K., Onishi K., Mikami I., Thidar K., Sano Y. 2004. Allelic diversification at the *C* (*OsC1*) locus of wild and cultivated rice: nucleotide changes associated with phenotypes. Genetics 168:997–1007.
Sakai H., Lee S.S., Tanaka T., Numa H., Kim J., Kawahara Y., Wakimoto H., Yang C.C., Iwamoto M., Abe T., Yamada Y., Muto A., Inokuchi H., Ikemura T., Matsumoto T., Sasaki T., Itoh T. 2013. Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics. Plant Cell Physiol. 54:e6.
Sankararaman S., Mallick S., Dannemann M., Prüfer K., Kelso J., Pääbo S., Patterson N., Reich D. 2014. The genomic landscape of Neanderthal ancestry in present-day humans. Nature 507:354–357.
Schrider D.R., Ayroles J., Matute D.R., Kern A.D. 2018. Supervised machine learning reveals introgressed loci in the genomes of *Drosophila simulans* and *D. sechellia*. PLoS Genet. 14:e1007341.
Si L., Chen J., Huang X., Gong H., Luo J., Hou Q., Zhou T., Lu T., Zhu J., Shangguan Y., Chen E., Gong C., Zhao Q., Jing Y., Zhao Y., Li Y., Cui L., Fan D., Lu Y., Weng Q., Wang Y., Zhan Q., Liu K., Wei X., An K., An G., Han B. 2016. *OsSPL13* controls grain size in cultivated rice. Nat. Genet. 48:447–456.
Skov L., Hui R., Shchur V., Hobolth A., Scally A., Schierup M.H., Durbin R. 2018. Detecting archaic introgression using an unadmixed outgroup. PLoS Genet. 14:e1007641.
Smith J., Kronforst M.R. 2013. Do *Heliconius* butterfly species exchange mimicry alleles? Biol. Lett. 9:20130503.
Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.
Stein J.C., Yu Y., Copetti D., Zwickl D.J., Zhang L., Zhang C., Chougule K., Gao D., Iwata A., Goicoechea J.L., Wei S., Wang J., Liao Y., Wang M., Jacquemin J., Becker C., Kudrna D., Zhang J., Londono C.E.M., Song X., Lee S., Sanchez P., Zuccolo A., Ammiraju J.S.S., Talag J., Danowitz A., Rivera L.F., Gschwend A.R., Noutsos C., Wu C.-C., Kao S.-M., Zeng J.-W., Wei F.-J., Zhao Q., Feng Q., El Baidouri M., Carpentier M.-C., Lasserre E., Cooke R., Rosa Farias D., da Maia L.C., Dos Santos R.S., Nyberg K.G., McNally K.L., Mauleon R., Alexandrov N., Schmutz J., Flowers D., Fan C., Weigel D., Jena K.K., Wicker T., Chen M., Han B., Henry R., Hsing Y.-I. C., Kurata N., de Oliveira A.C., Panaud O., Jackson S.A., Machado C.A., Sanderson M.J., Long M., Ware D., Wing R.A. 2018. Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus *Oryza*. Nat. Genet. 50:285–296.
Stewart C.N., Halfhill M.D., Warwick S.I. 2003. Transgene introgression from genetically modified crops to their wild relatives. Nat. Rev. Genet. 4:806–817.
Suvorov A., Hochuli J., Schrider D.R. 2020. Accurate inference of tree topologies from multiple sequence alignments using deep learning. Syst. Biol. 69:221–233.
Sweeney M.T., Thomson M.J., Pfeil B.E., McCouch S. 2006. Caught red-handed: *Rc* encodes a basic helix-loop-helix protein conditioning red pericarp in rice. Plant Cell 18:283–294.
Tan L., Li X., Liu F., Sun X., Li C., Zhu Z., Fu Y., Cai H., Wang X., Xie D., Sun C. 2008. Control of a key transition from prostrate to erect growth in rice domestication. Nat. Genet. 40:1360–1364.
Than C., Ruths D., Nakhleh L. 2008. PhyloNet: a software package for analyzing and reconstructing reticulate evolutionary relationships. BMC Bioinf. 9:322.
Trapnell C., Pachter L., Salzberg S.L. 2009. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25:1105–1111.
Trapnell C., Williams B.A., Pertea G., Mortazavi A., Kwan G., van Baren M.J., Salzberg S.L., Wold B.J., Pachter L. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat. Biotechnol. 28:511–515.
Wang J., Qi M., Liu J., Zhang Y. 2015. CARMO: a comprehensive annotation platform for functional exploration of rice multi-omics data. Plant J. 83:359–374.
Wang M., Yu Y., Haberer G., Marri P.R., Fan C., Goicoechea J.L., Zuccolo A., Song X., Kudrna D., Ammiraju J.S., Cossu R.M., Maldonado C., Chen J., Lee S., Sisneros N., de Baynast K., Golser W., Wissotski M., Kim W., Sanchez P., Ndjiondjop M.-N., Sanni K., Long M., Carney J., Panaud O., Wicker T., Machado C.A., Chen M., Mayer K.F.X., Rounsley S., Wing R.A. 2014. The genome sequence of African rice (*Oryza glaberrima*) and evidence for independent domestication. Nat. Genet. 46:982–988.
Wu H.C., Luo D.L., Vignols F., Jinn T.L. 2012. Heat shock-induced biphasic $Ca^{(2+)}$ signature and OsCaM1-1 nuclear localization mediate downstream signalling in acquisition of thermotolerance in rice (*Oryza sativa* L.). Plant Cell Environ. 35:1543–1557.
Yang J., Lee S.H., Goddard M.E., Visscher P.M. 2011. GCTA: a tool for genome-wide complex trait analysis. Am. J. Hum. Genet. 88:76–82.
Yang Z., Rannala B. 2012. Molecular phylogenetics: principles and practice. Nat. Rev. Genet. 13:303–314.
You Q., Zhai K., Yang D., Yang W., Wu J., Liu J., Pan W., Wang J., Zhu X., Jian Y., Liu J., Zhang Y., Deng Y., Li Q., Lou Y., Xie Q., He Z. 2016. An E3 ubiquitin ligase-BAG protein module controls plant innate immunity and broad-spectrum disease resistance. Cell Host Microbe 20:758–769.
Zhang L., Ren Y., Yang T., Li G., Chen J., Gschwend A.R., Yu Y., Hou G., Zi J., Zhou R., Wen B., Zhang J., Chougule K., Wang M., Copetti D., Peng Z., Zhang C., Zhang Y., Ouyang Y., Wing R.A., Liu S., Long M. 2019. Rapid evolution of protein diversity by de novo origination in *Oryza*. Nat. Ecol. Evol. 3:679–690.
Zhang W., Dasmahapatra K.K., Mallet J., Moreira G.R., Kronforst M.R. 2016. Genome-wide introgression among distantly related *Heliconius* butterfly species. Genome Biol. 17:25.
Zhao Q., Feng Q., Lu H., Li Y., Wang A., Tian Q., Zhan Q., Lu Y., Zhang L., Huang T., Wang Y., Fan D., Zhao Y., Wang Z., Zhou C., Chen J., Zhu C., Li W., Weng Q., Xu Q., Wang Z.-X., Wei X., Han B., Huang X. 2018. Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. Nat. Genet. 50:278–284.
Zheng Y., Crawford G.W., Jiang L., Chen X. 2016. Rice domestication revealed by reduced shattering of archaeological rice from the lower Yangtze valley. Sci. Rep. 6:28136.
Zou Z., Zhang H., Guan Y., Zhang J. 2020. Deep residual neural networks resolve quartet molecular phylogenies. Mol. Biol. Evol. 37:1495–1507.