Inferring Historical Introgression with Deep Learning

Journal Club | Mingyu Suo

2025-04-23

Outline

1. Background | What was the ERICA model designed to do?

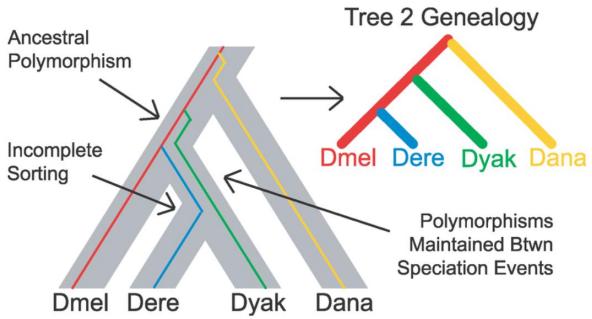
2. Method | How was ERICA trained, and how does it address the problem of introgression analysis?

3. Result | How well does ERICA perform in introgression analysis?

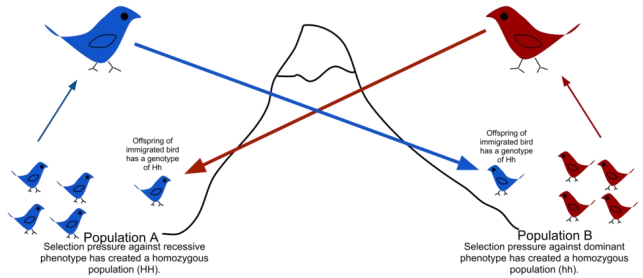
Resolving the relationships among taxa

Evolution isn't always a family tree with neat branches

- incomplete lineage sorting (ILS)

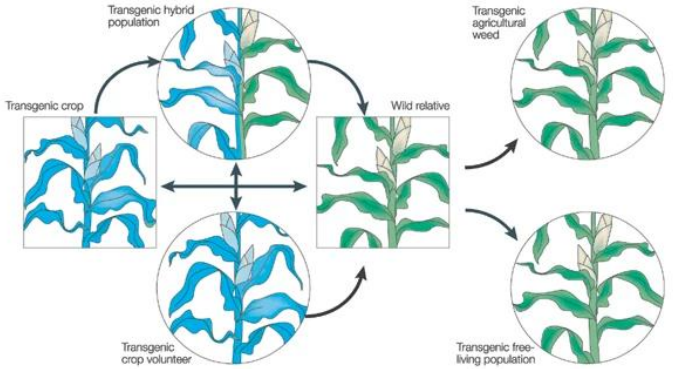


- gene flow after divergence



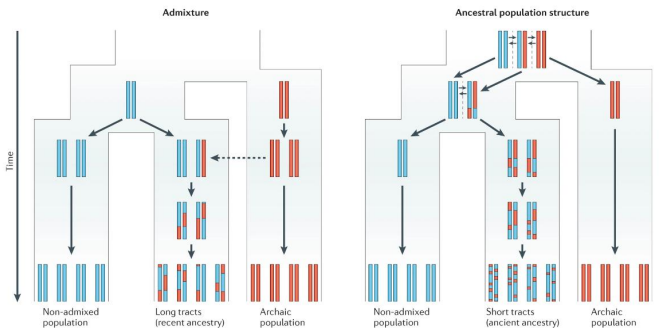
genome-wide patterns of admixture have been characterized in different organisms

- between crops and wild relatives



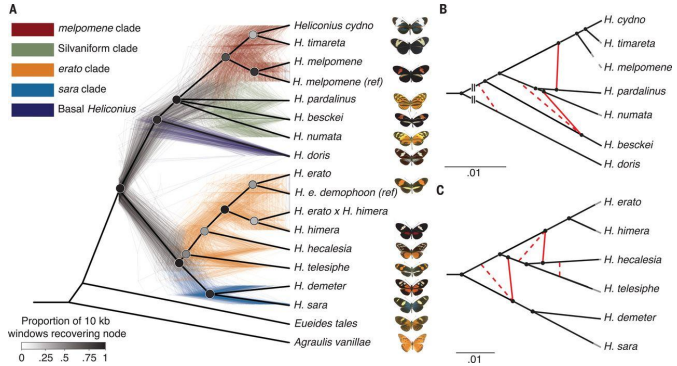
(Stewart et al., ***Nat Rev Genet***, 2003)

- between archaic[古人类] and modern humans



(Racimo et al.,***Nat Rev Genet***, 2015)

- *Heliconius* butterflies [袖蝶]



(Edelman et al., ***Science***, 2019)

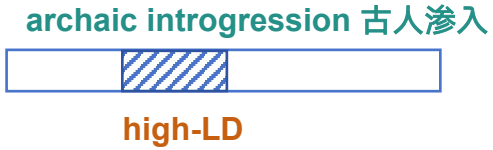
Algorithm for Admixutre | 4 categories

- **depict an overall demographic pattern of admixture**

  G-PhoCS, Treemix and PhyloNet

- **compare scales of linkage disequilibrium by detecting the structure of haplotypes from fine-scale and sufficient genomic data**

  HAPMIX, ELAI, S* and Sprime



- **performe window-based scans and describing relationships quantitatively among focal taxa according to the allele frequencies of given patterns**

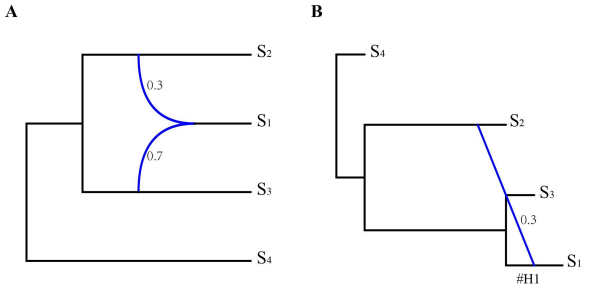
  Patterson's D-statistic, f_d statistic, D_{FOIL}

- **Combining multiple genomic features using machine learning models**

  conditional random fields (CRF)

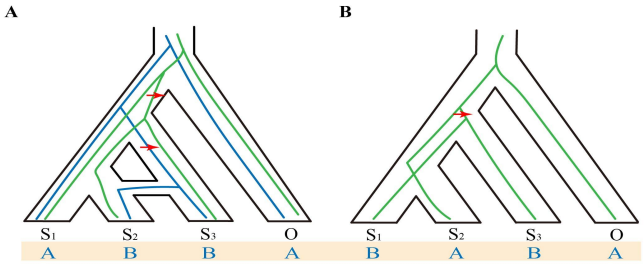
  hidden Markov models (HMMs)

PhyloNet

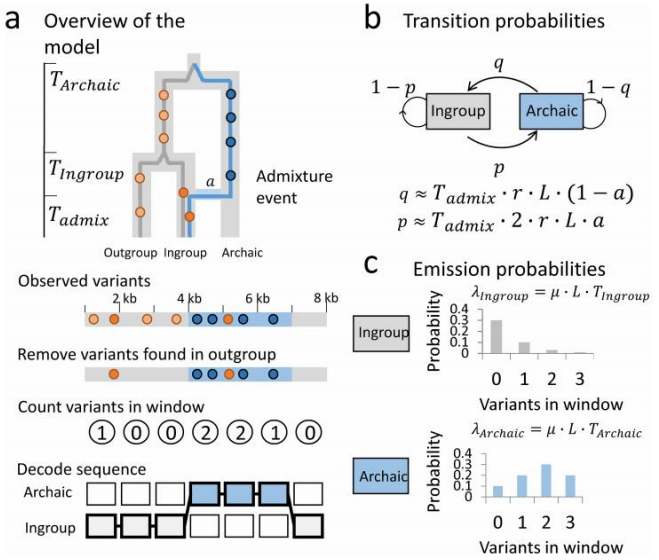


D-statistics (ABBA-BABA)

D = [sum(ABBA) – sum(BABA)] / [sum(ABBA) + sum(BABA)]



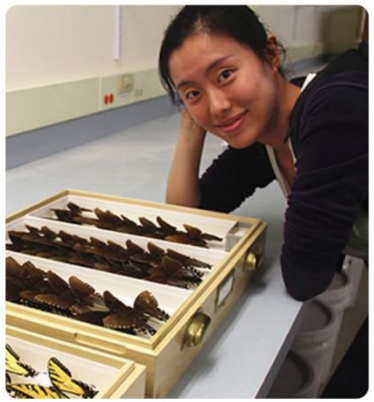
HMM (Skov,2018)



Research Aim

- allow **direct processing of sequence data** instead of requiring predefined population genetic statistics or inferred gene trees
- be capable of **resolving local introgression signals** in genomes with **heterogeneous gene flow**
- be applicable to model and non-model systems, and be **robust across taxa**
- **low computational complexity** and should be capable of handling genome-scale data in an acceptable amount of time

corresponding author



张蔚

Wei Zhang
Principal Investigator

Email:weizhangvv(AT)pku.edu.cn
Phone:010-62767697

Educational experience

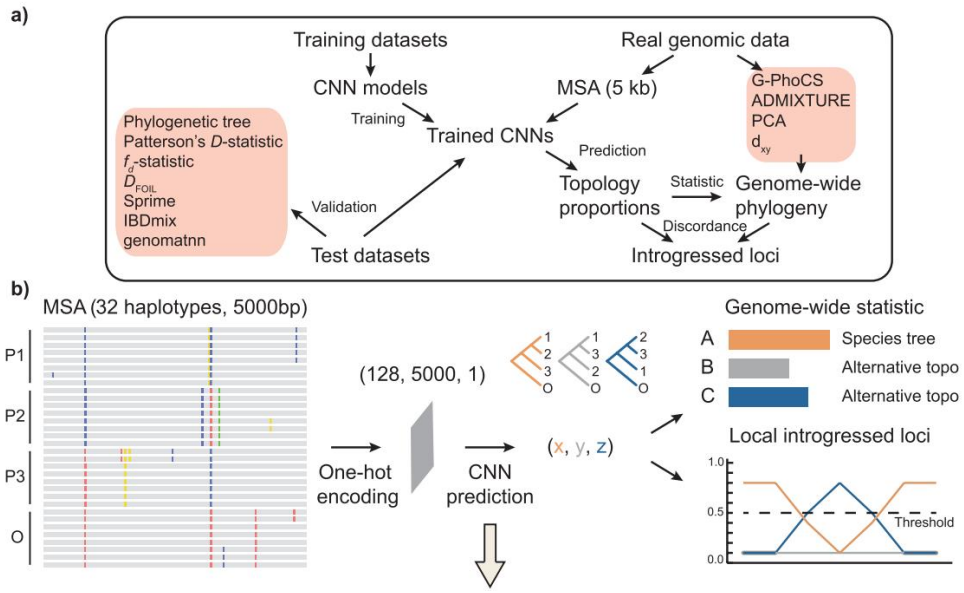
2005 BS in Biotechnology, Shandong University, China
2011 PhD in Botany, Peking University, China

Work experience

2024-Present Principal Investigator（with Tenure）, School of Life Sciences, Peking University
2024-Present Principal Investigator（with Tenure）, Peking-Tsinghua Center for Life Sciences, Peking University
2018-2024 Principal Investigator, School of Life Sciences, Peking University
2018-2024 Principal Investigator, Peking-Tsinghua Center for Life Sciences, Peking University
2012-2017 Postdoctoral Scholar, Department of Ecology and Evolution, University of Chicago
2017 Postdoctoral Scholar, Department of Pediatrics, University of Chicago

Introgression events allow the sharing of genetic information among different species, illustrated here as the *cross-linking of branches on the Tree of Life.* Zhang and collaborators propose a deep-learning approach to study these evolutionary events, and apply it to different organisms shown in the image

Method



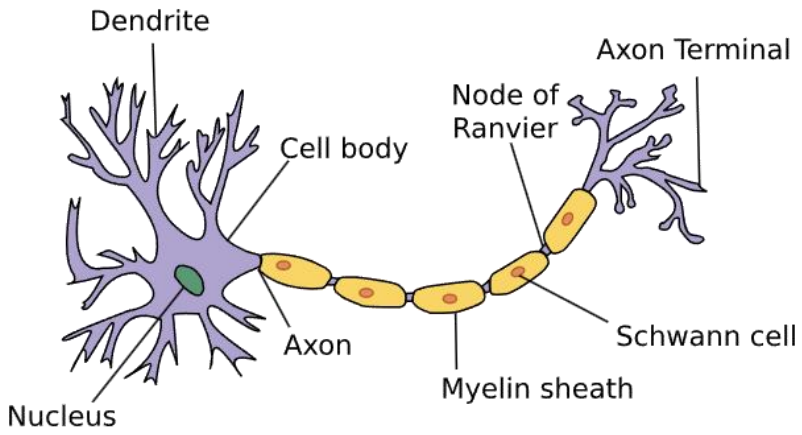
Outline

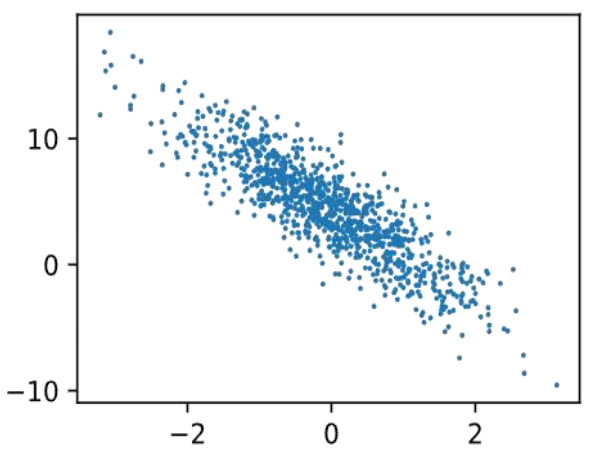
1. Background | What was the ERICA model designed to do?

2. Method | How was ERICA trained, and how does it address the problem of introgression analysis?

3. Result | How well does ERICA perform in introgression analysis?

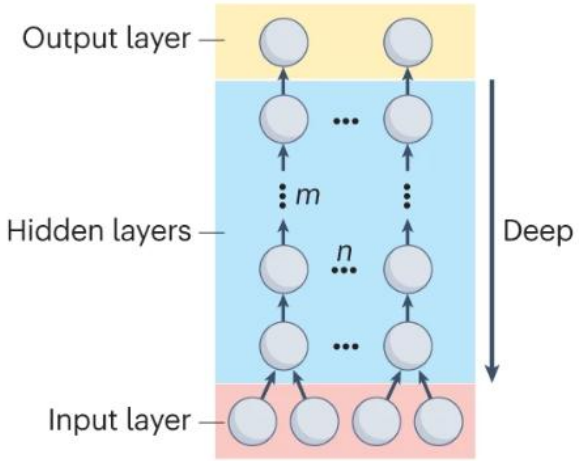
Deep Learning



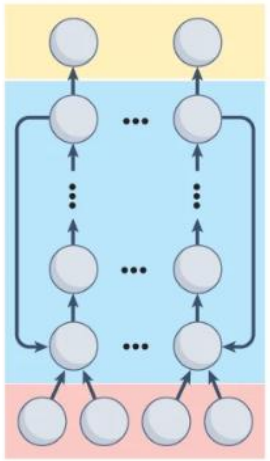


linear regression
y = ax + b
a=? b=?

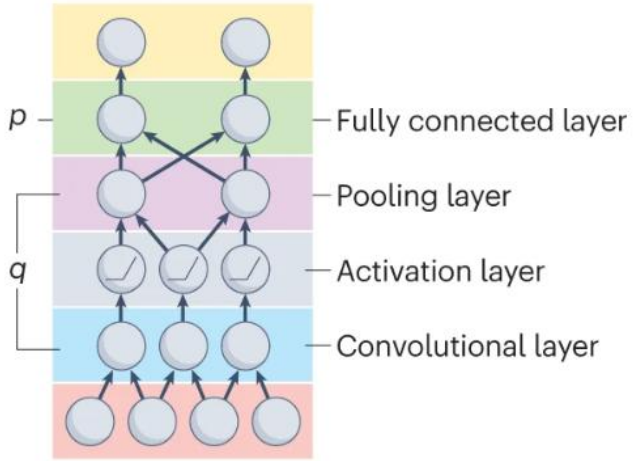
Feedforward neural network



Recurrent neural network

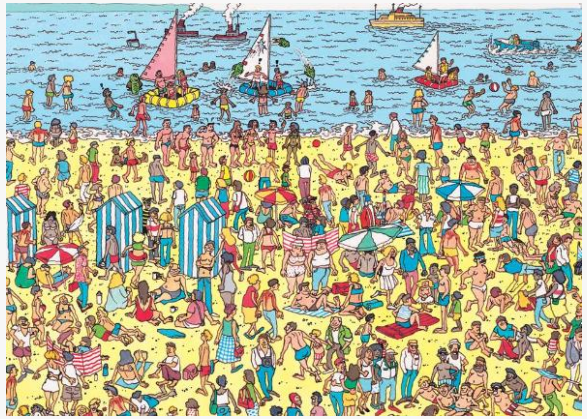


Convolutional neural network



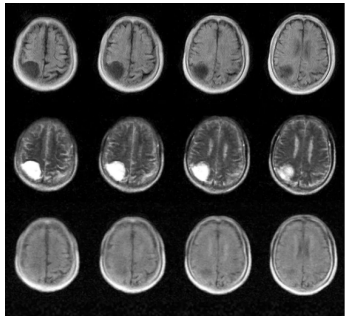
Deep Learning | CNN

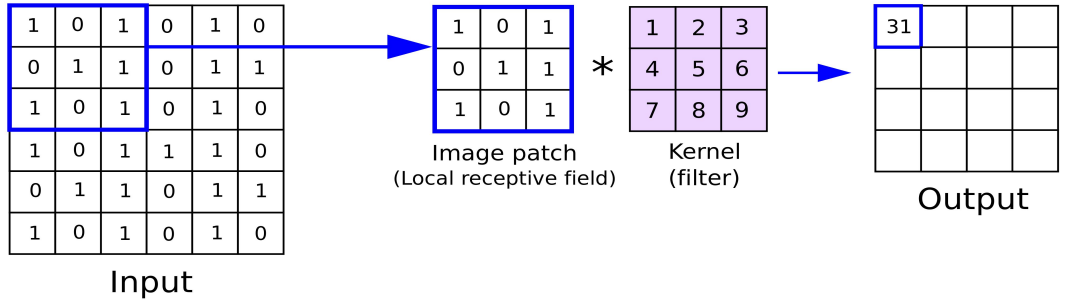
《Where's Wally?》Games

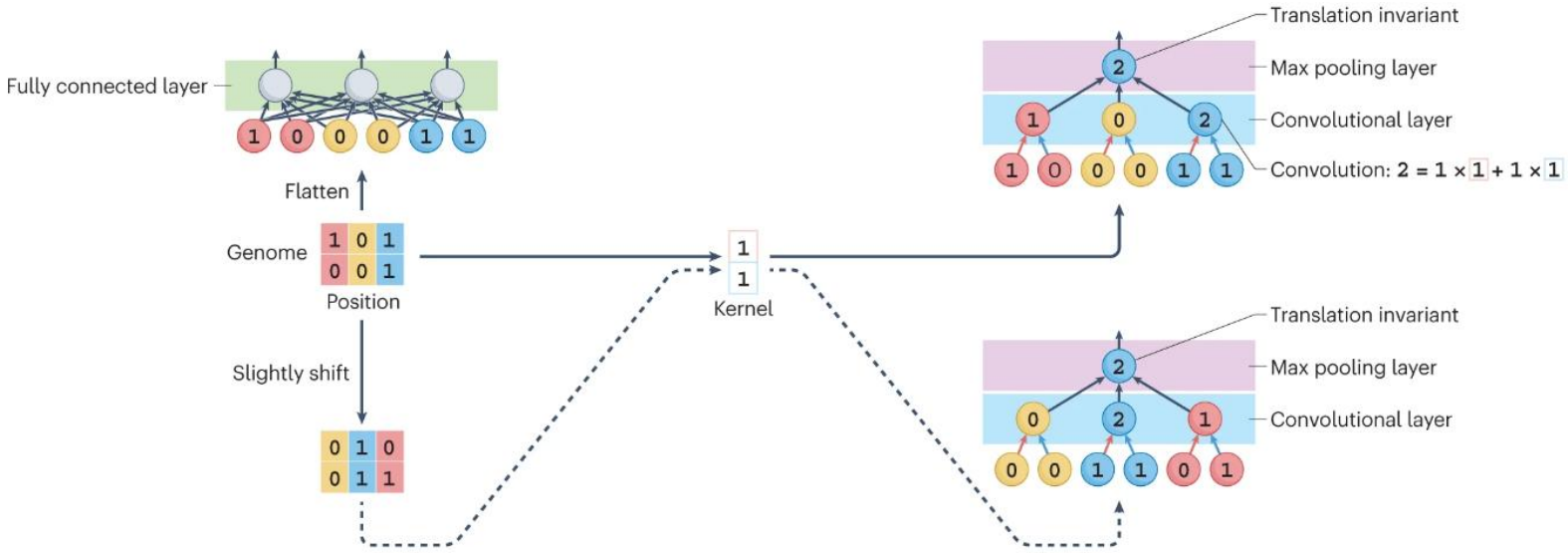


Features：

- translation invariance [平移不变性]
- locality [局部性]



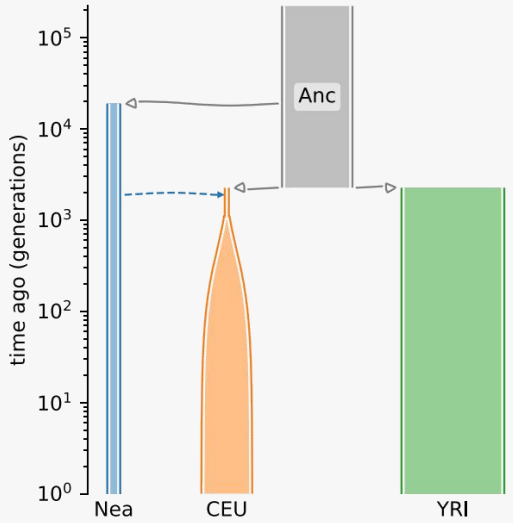




Huang et al., ***Nat Rev Genet***, 2024

Design Principles of ERICA

previous studies:
distinguished using classification tasks

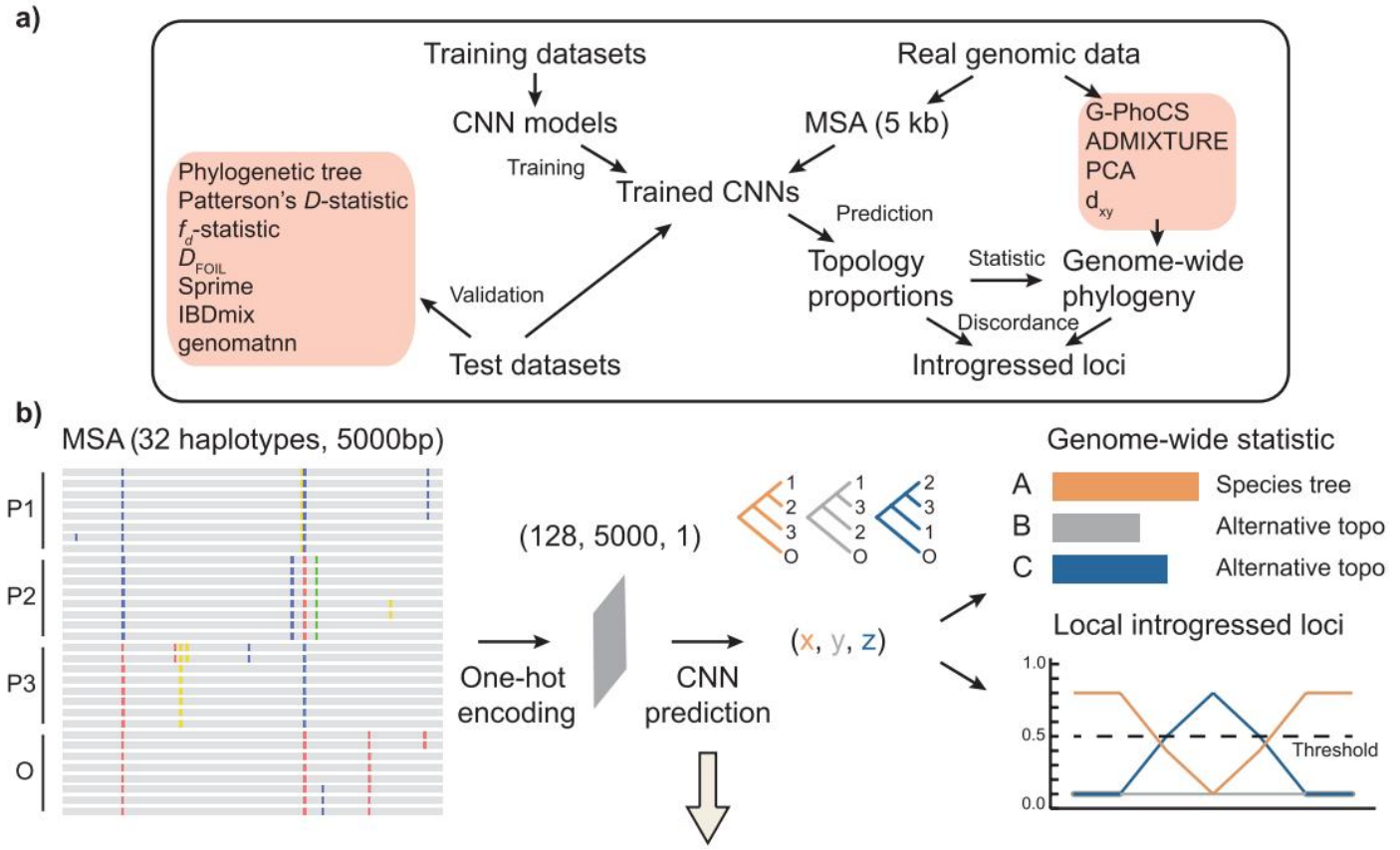


(Gower et al.,***eLife***, 2021)

problem occurs:
when more than two species were considered, there were multiple potential gene flow events

ERICA (**E**volutionary **R**elationship **I**nference using a **C**NN-based **A**pproach)

- the discordance between the gene trees and the species tree provides a way to identify introgressed regions



Design Principles of ERICA

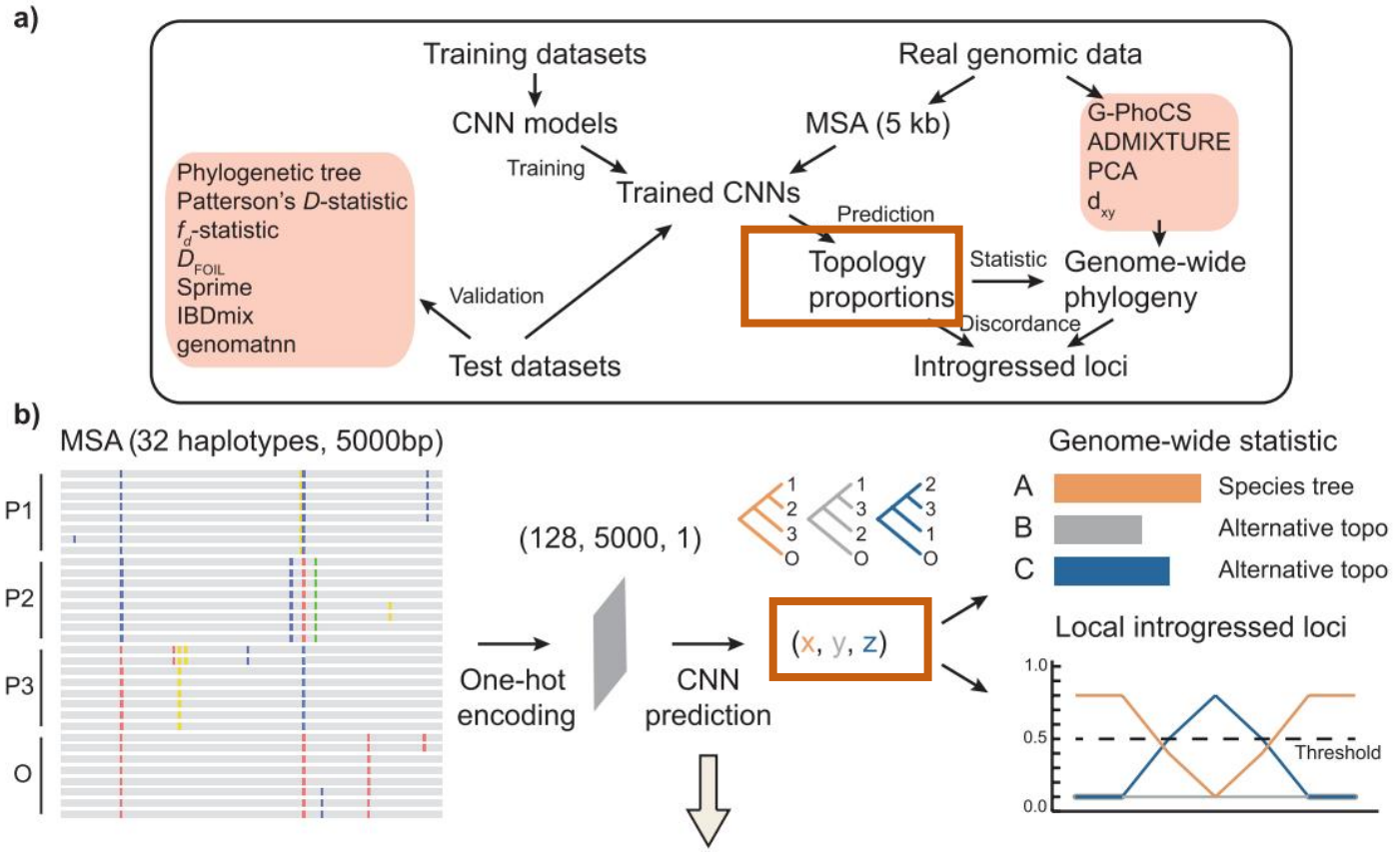
Phylogenetic relationship encoding for CNN

1. (((P1, P2), P3), O)
2. (((P1, P3), P2), O)
3. (((P2, P3), P1), O)

multi-dimensional vector:
derived from quartet sampling (Estabrook et al.1985) and topology weighting (Martin and Van Belleghem 2017)

ERICA (**E**volutionary **R**elationship **I**nference using a **C**NN-based **A**pproach)

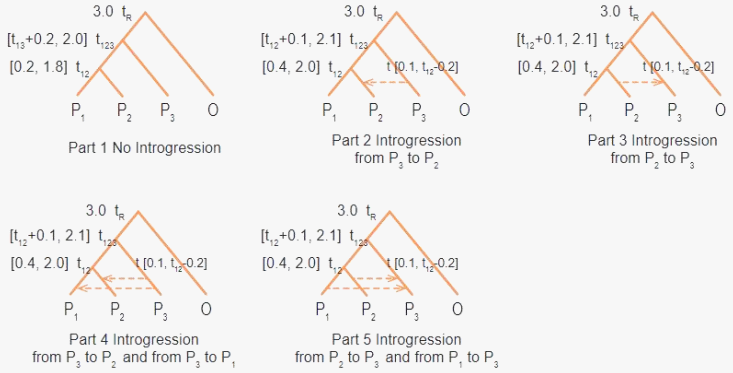
- the discordance between the gene trees and the species tree provides a way to identify introgressed regions



Design Principles of ERICA

Data simulation for CNN model training

generated a training dataset covering scenarios with varying degrees of ILS and gene flow

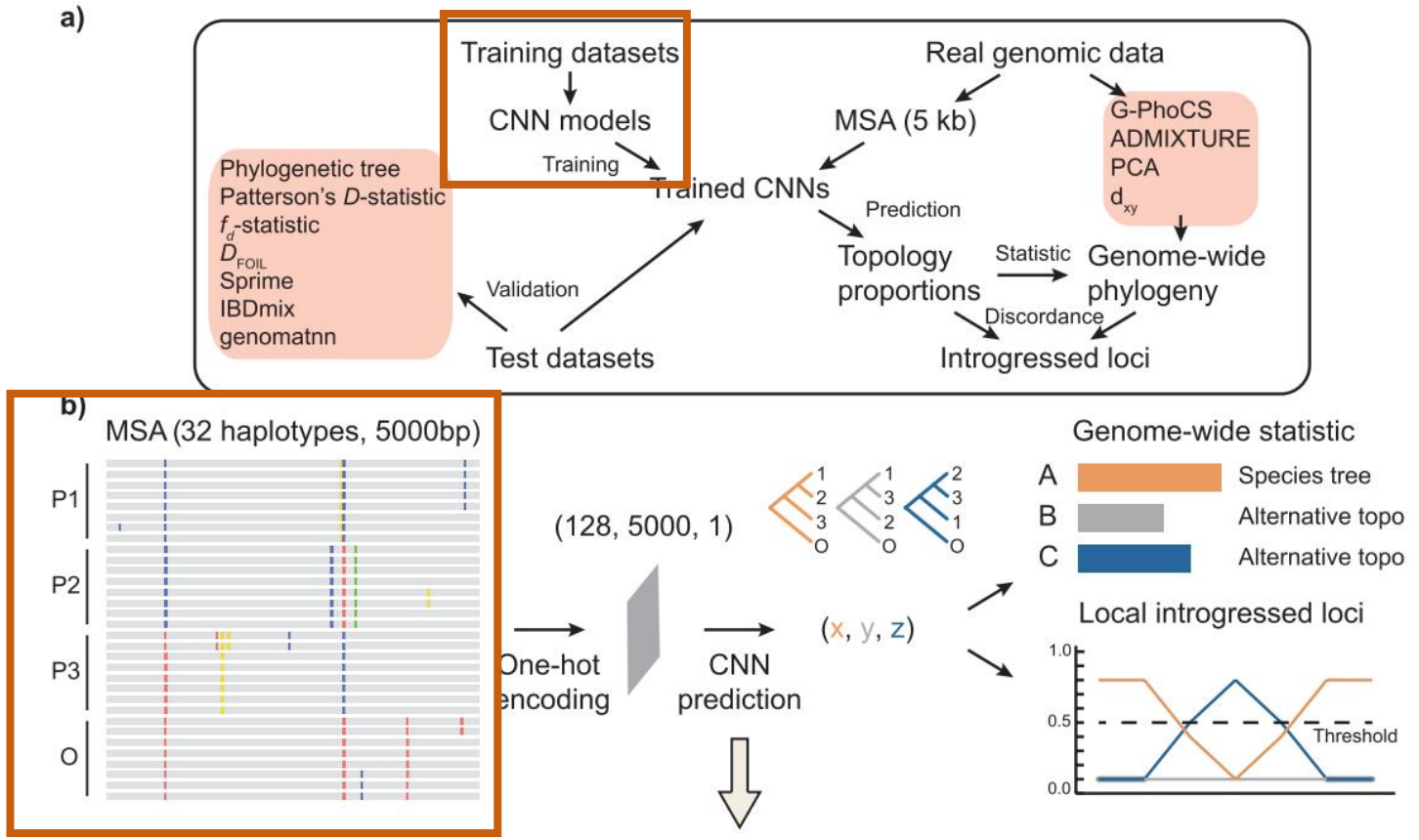


generated multiple sequence alignments (**MSAs**) for training and testing

- coalescent simulator *ms* & *Seq-Gen*
- 5000 bp in length
- eight haplotypes per taxon

ERICA (**E**volutionary **R**elationship **I**nference using a **C**NN-based **A**pproach)

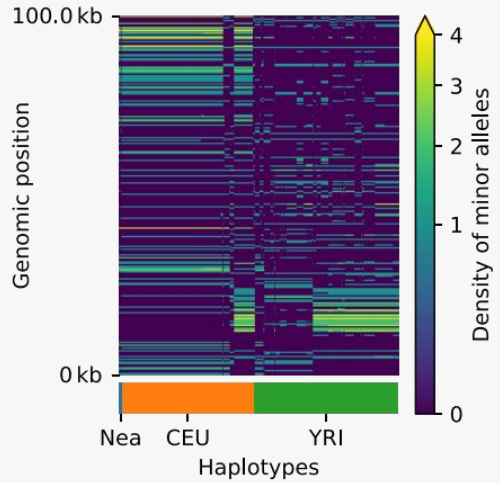
- the discordance between the gene trees and the species tree provides a way to identify introgressed regions



Design Principles of ERICA

Sequence Encoding

genomatnn: divided a sequence into a fixed number of bins and counted the number of minor alleles in each bin



(Gower et al.,***eLife***, 2021)

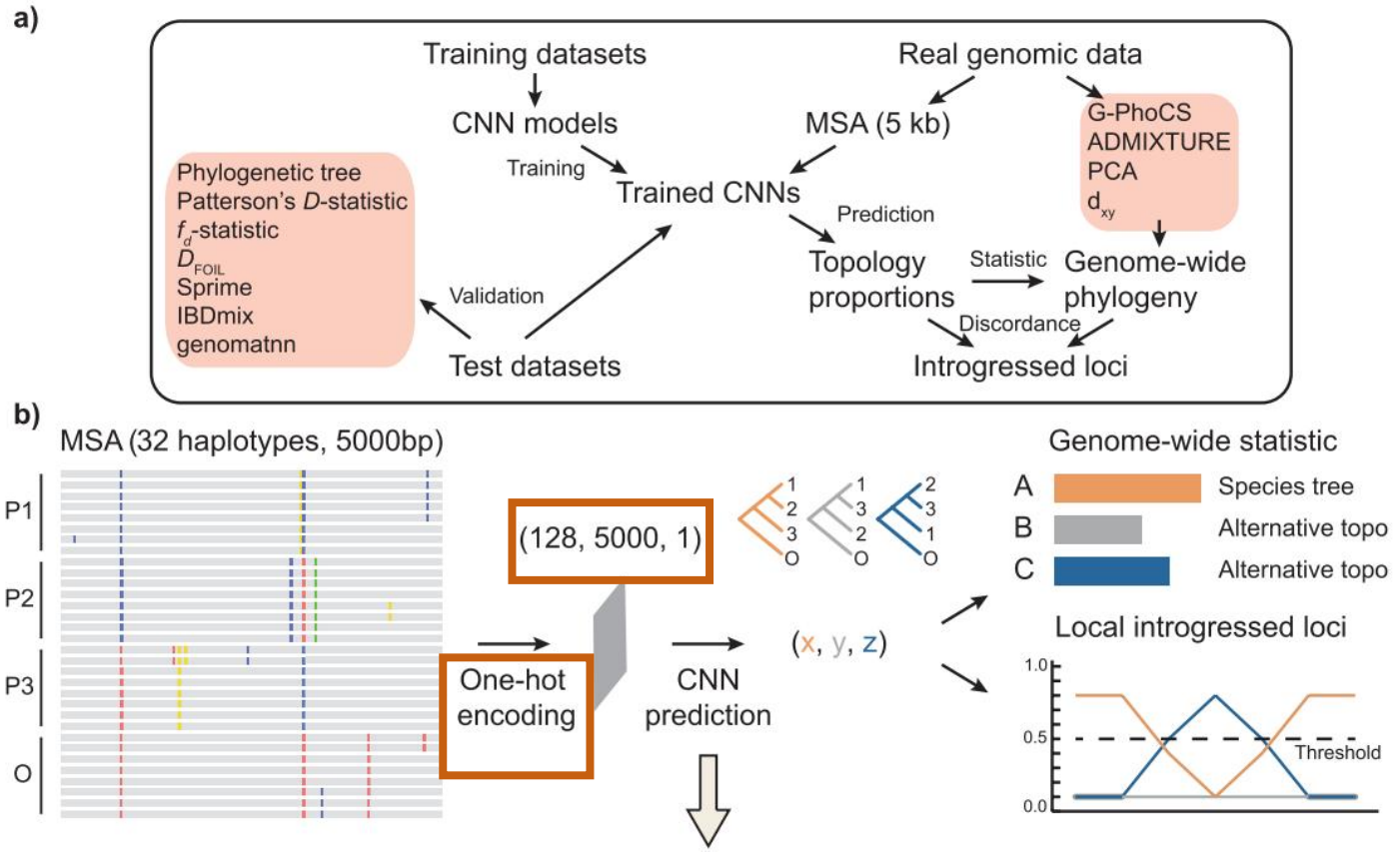
ERICA: one-hot encoding

- G:(1,0,0,0)
- T: (0,1,0,0)
- A: (0,0,1,0)
- C: (0,0,0,1)
- gap: (0,0,0,0)

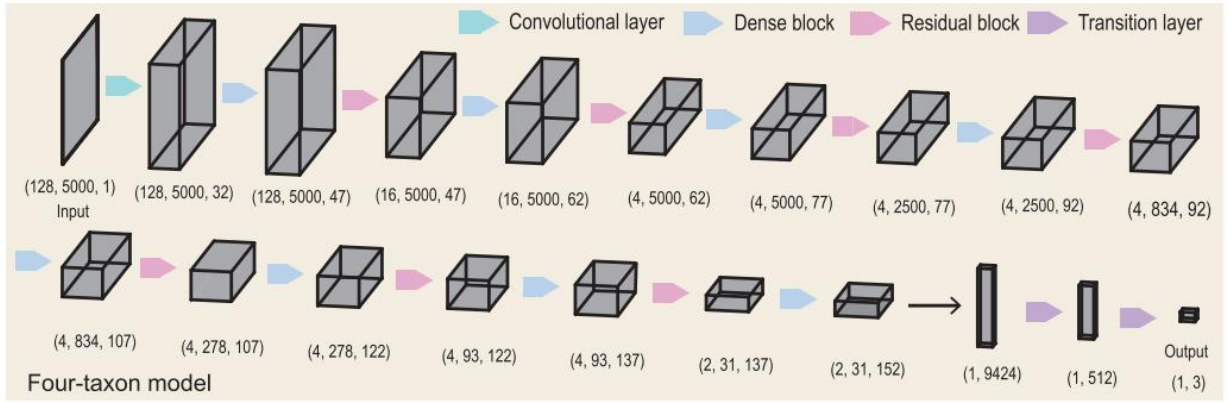
4x32 hap = 128 cols

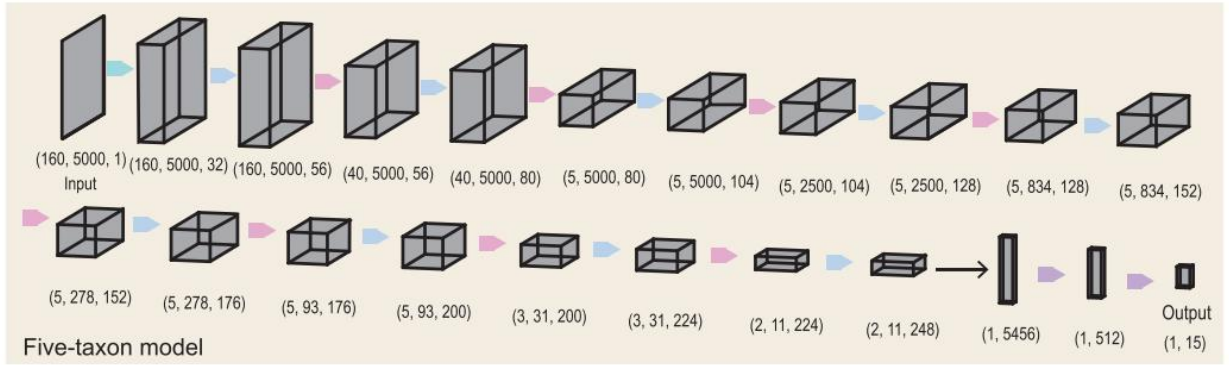
ERICA (**E**volutionary **R**elationship **I**nference using a **C**NN-based **A**pproach)

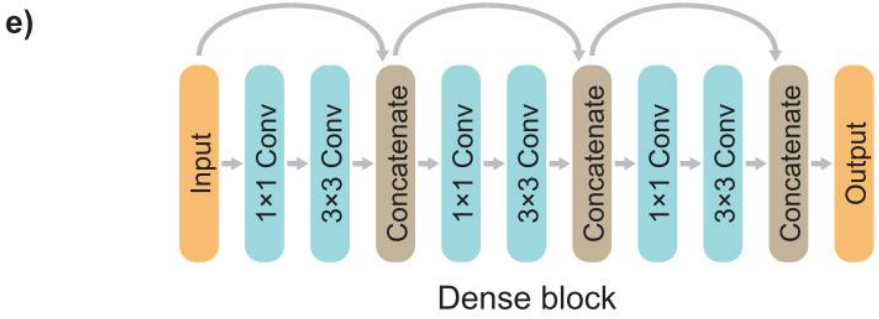
- the discordance between the gene trees and the species tree provides a way to identify introgressed regions

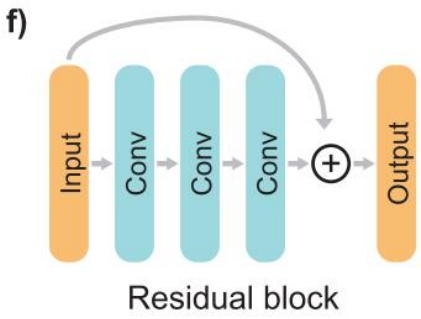


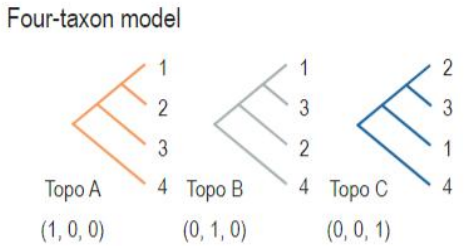
Design Principles of ERICA | CNN Architecture





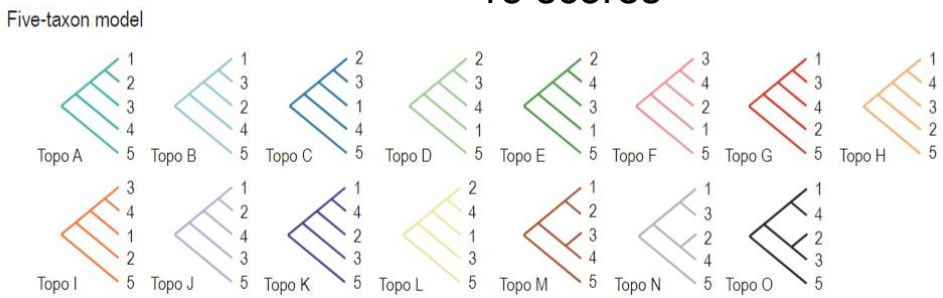






3 scores

15 scores



- Residual Networks [残差网络]:
  - skip connections to train deeper
- Dense Convolutional Networks [稠密网络]:
  - reduces the number of parameters by directly connecting all the layers in one dense block

Outline

1. Background | What was the ERICA model designed to do?

2. Method | How was ERICA trained, and how does it address the problem of introgression analysis?

3. Result | How well does ERICA perform in introgression analysis?

Performance evaluation of the four-taxon ERICA model | *Simulation*

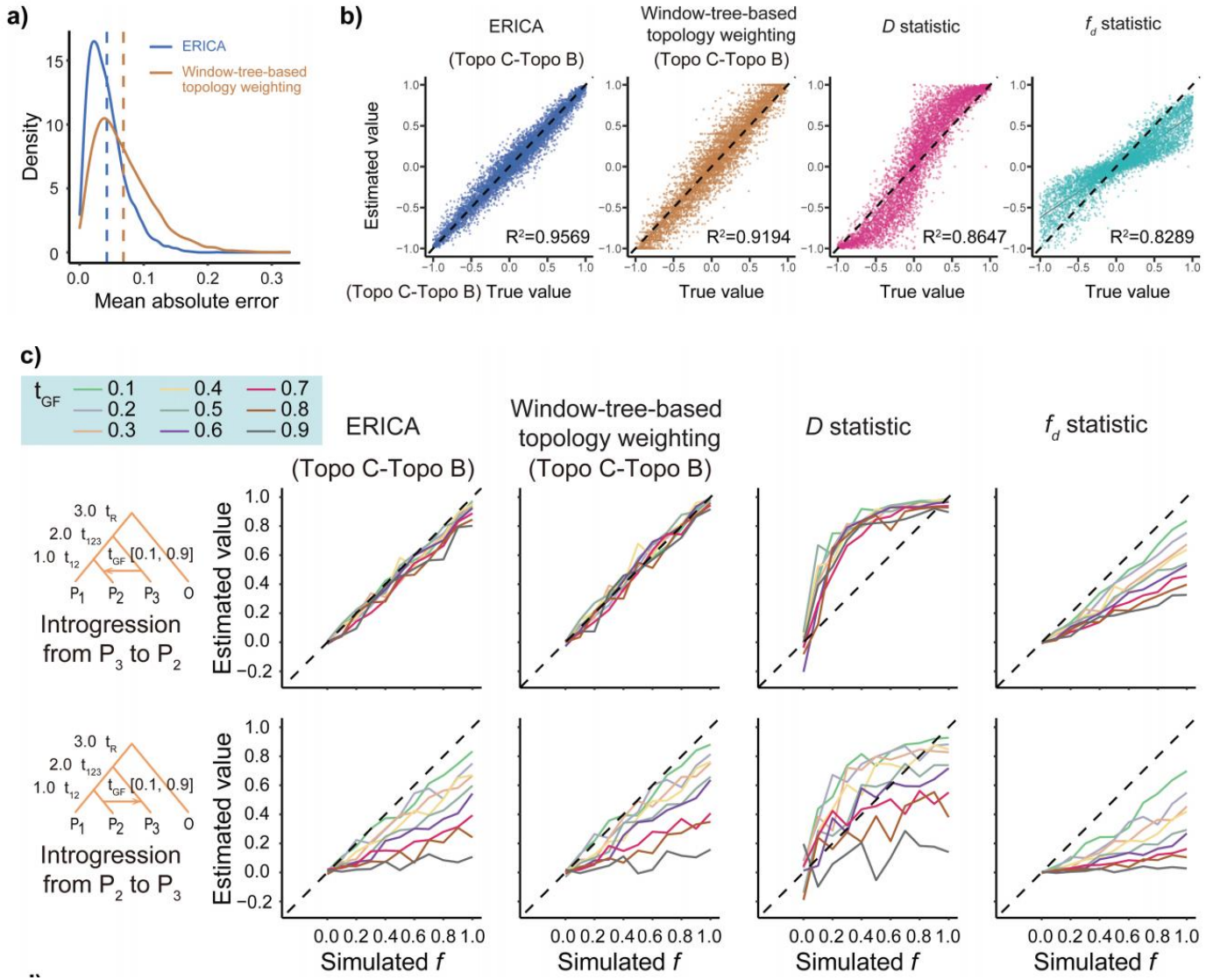
- **mean absolute error**

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

- **t_{GF}**
  the time of gene flow ranged from 10% to 90% of the split time (t12)

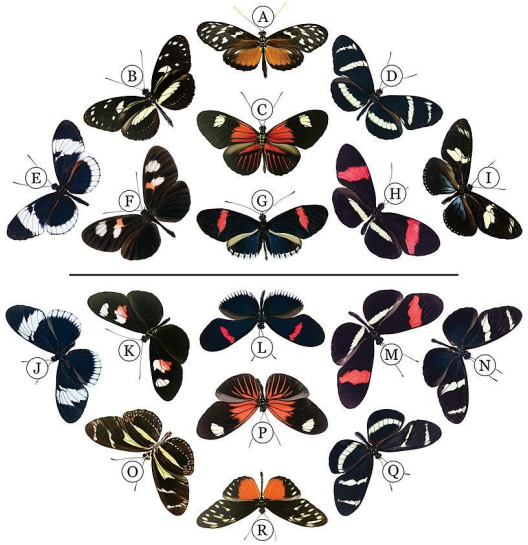
ERICA beat traditional tools:

- **Accuracy:** 95% vs. 70–80% for D-statistic in complex gene flow.
- **Speed:** Processes whole genomes in hours, not weeks.
- **Robustness:** Works with messy data (missing sequences, errors).

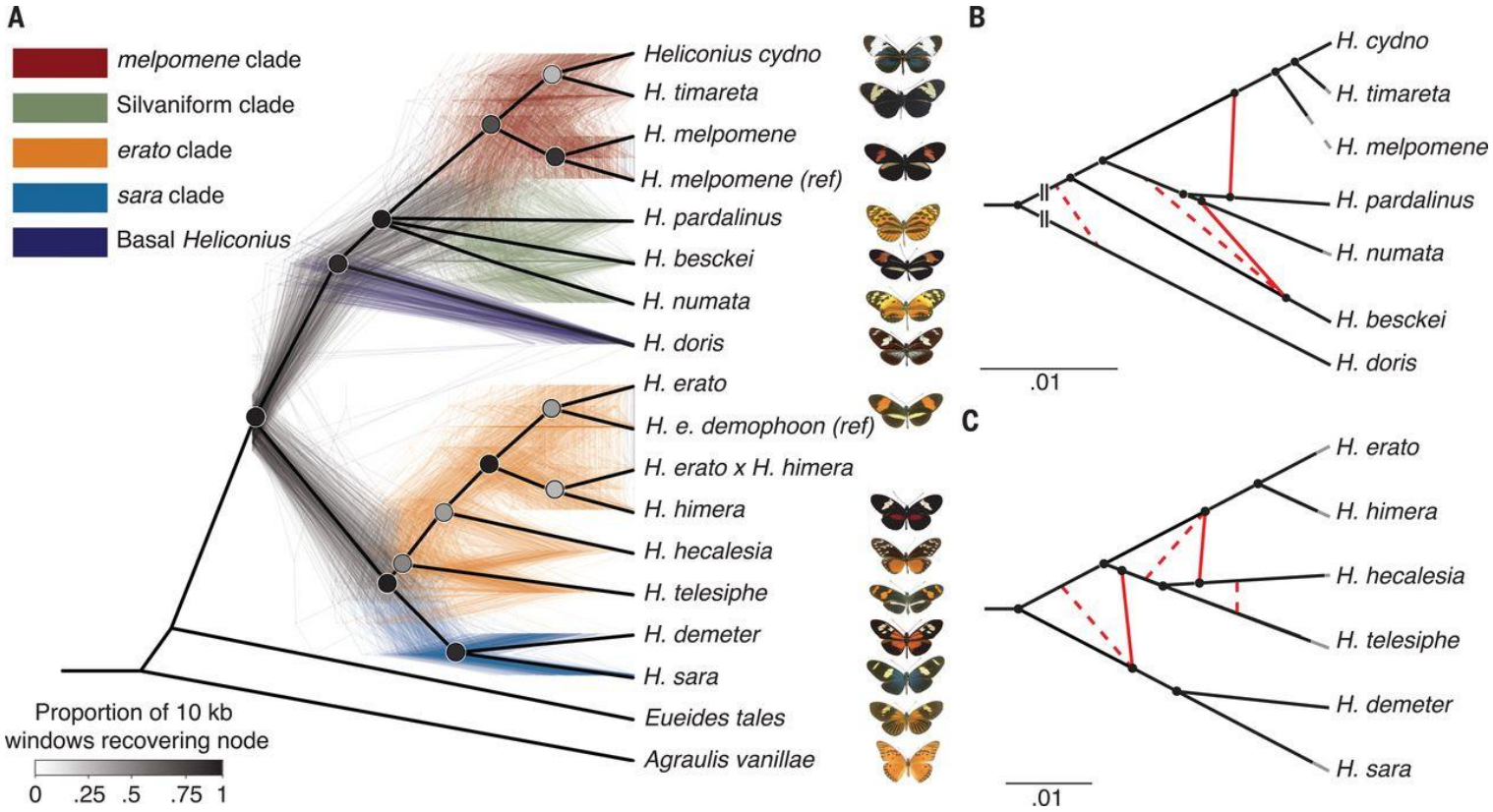


Inferring Gene Flow Based on Real Genomic Data | *Case 1*

Heliconius butterflies [袖蝶]



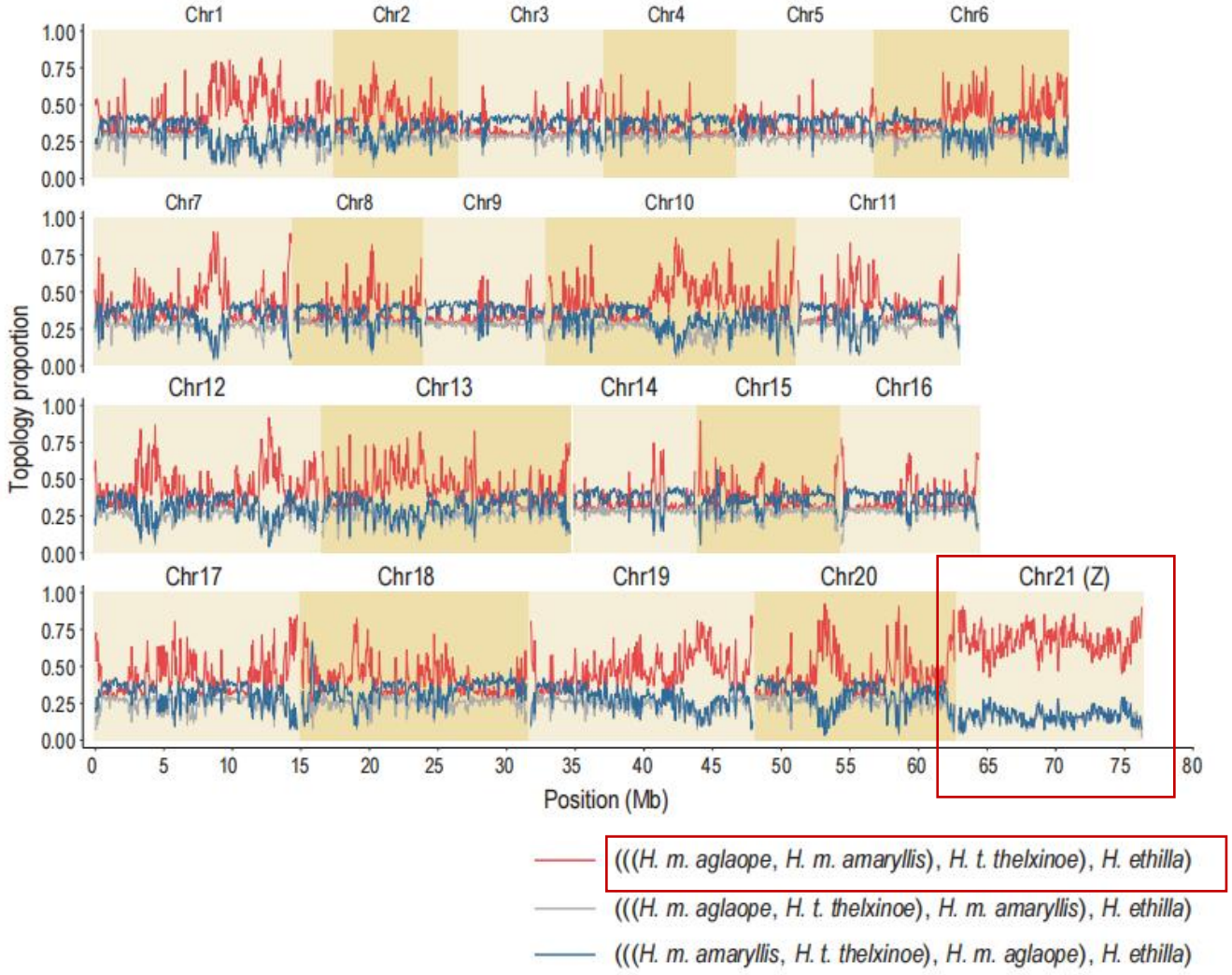
- distributed throughout the tropical and subtropical regions
- Adults exhibit bright wing color patterns which signal their distastefulness to potential predators
- Müllerian mimics [缪氏拟态]



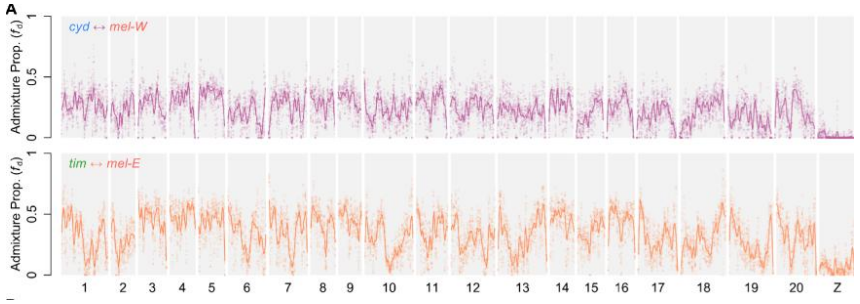
(Edelman et al., ***Science***, 2019)

- display complex relationships owing to intensive hybridization during adaptive radiation, even without an available bifurcating tree [分叉树]

Inferring Gene Flow Based on Real Genomic Data | *Case 1*



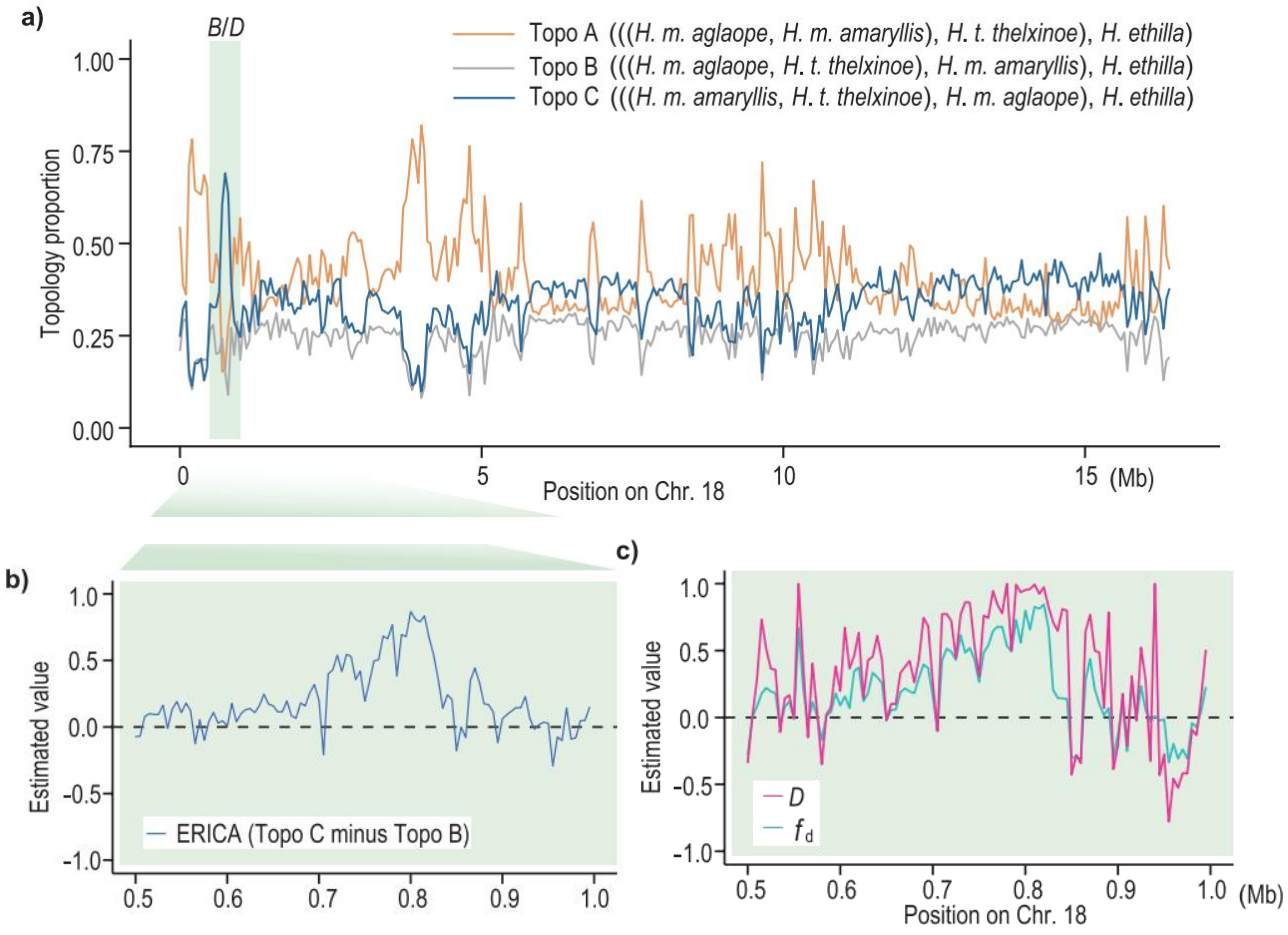
- the **highest proportion** was along the **Z chromosome (Chr21)**
- suggested that the **resistance** of the Z chromosome **to introgression** was greater than that of autosomes in *H. melpomene* and *H. timareta*

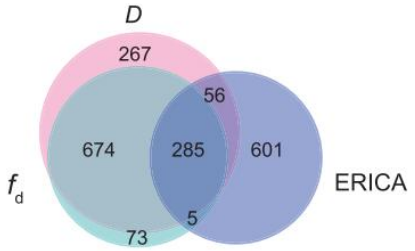


(Martin et al.,***PLoS Biol***,2019)

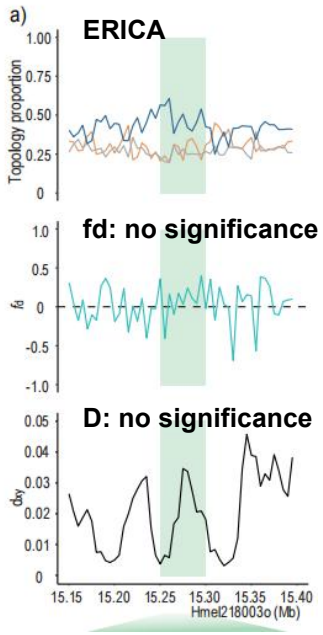
Consistent with the order of species differentiation

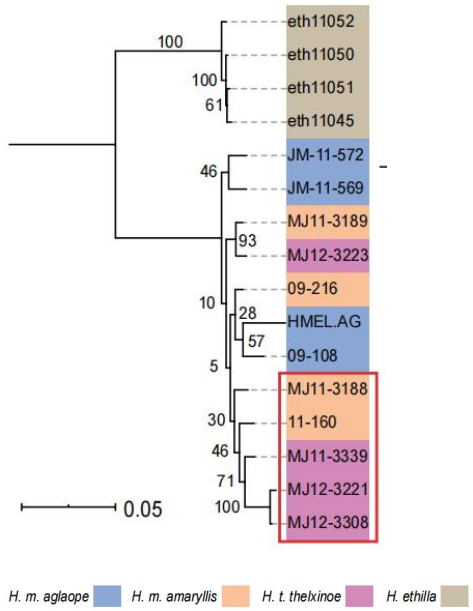
Inferring Gene Flow Based on Real Genomic Data | *Case 1*





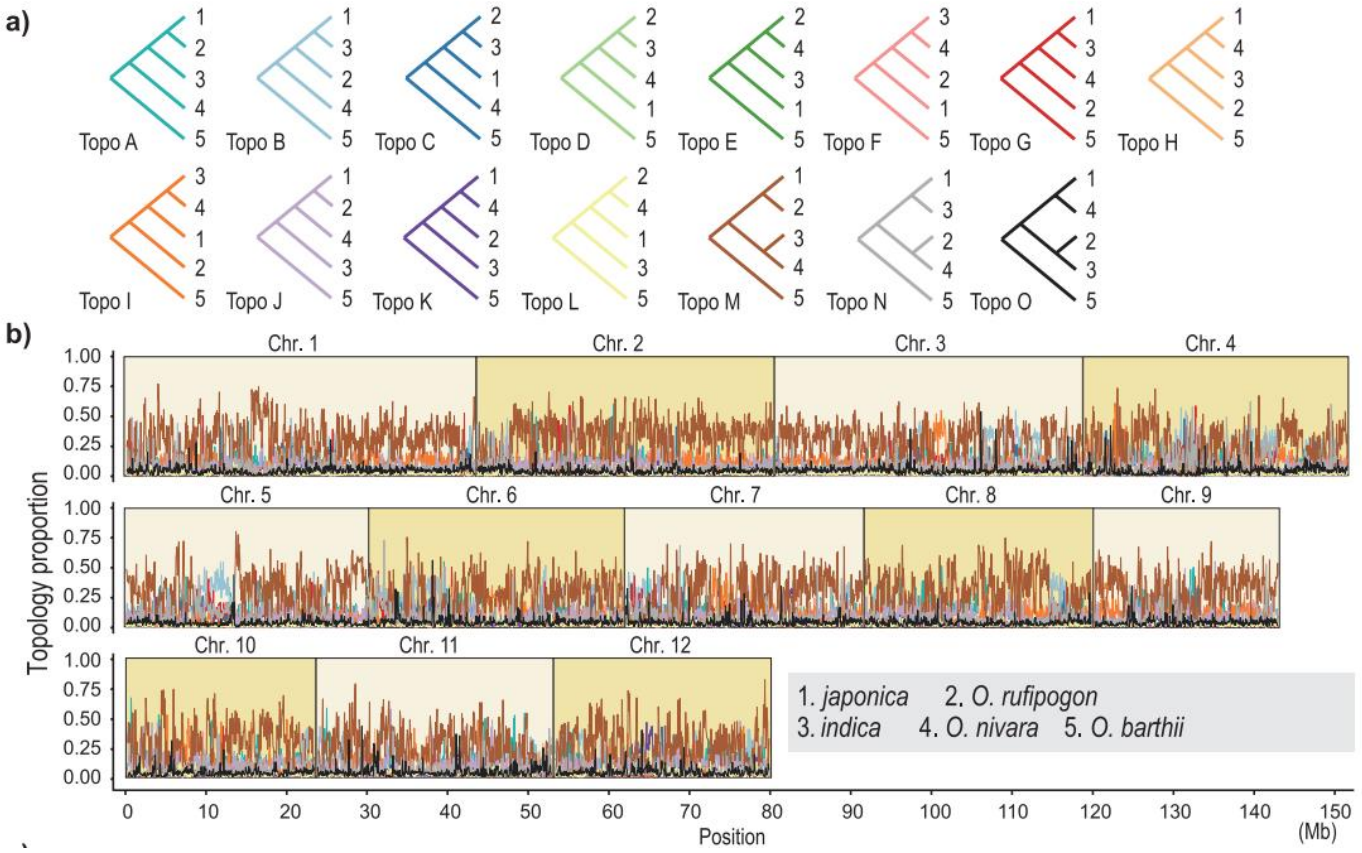
ERICA-specific results





- chromosome 18 (ranging from 700 kb to 850 kb)
- B/D locus: control wing color patterns

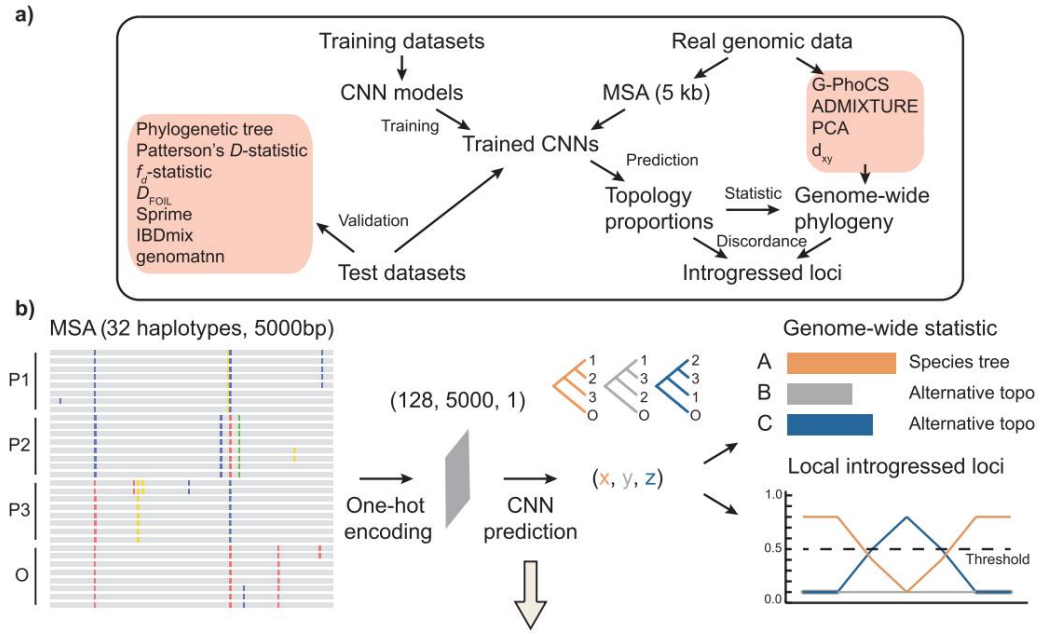
Inferring Gene Flow Based on Real Genomic Data | *Case 2*



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	

- Genome-wide patterns of admixture in rice domestication and adaptation

Summary



Why can use CNN?

- translation invariance [平移不变性]
- locality [局部性]

ERICA method

- validation: simulation & real genomic data
- limitation: fixed window size & no archaic introgression

Old Tool (D-statistic)	ERICA
Metal detector: beeps but can't locate the treasure.	MRI scanner: Maps exact locations of gene flow.
Fails with >2 species or small DNA chunks.	Handles 4–5 species and 5-kb fragments.
Needs species-specific tuning.	Train once, apply to any organism

Thanks for attention!

Journal Club | Mingyu Suo

2025-04-23