Putting the HAL in Haldane: leveraging deep learning for population genetics

Andrew Kern
Institute of Ecology and Evolution
University of Oregon

Leo Breiman's Two Cultures

the logic of data analysis

L. Breiman, Statistical Science (2001)

Leo Breiman's Two Cultures

Data Modeling Culture

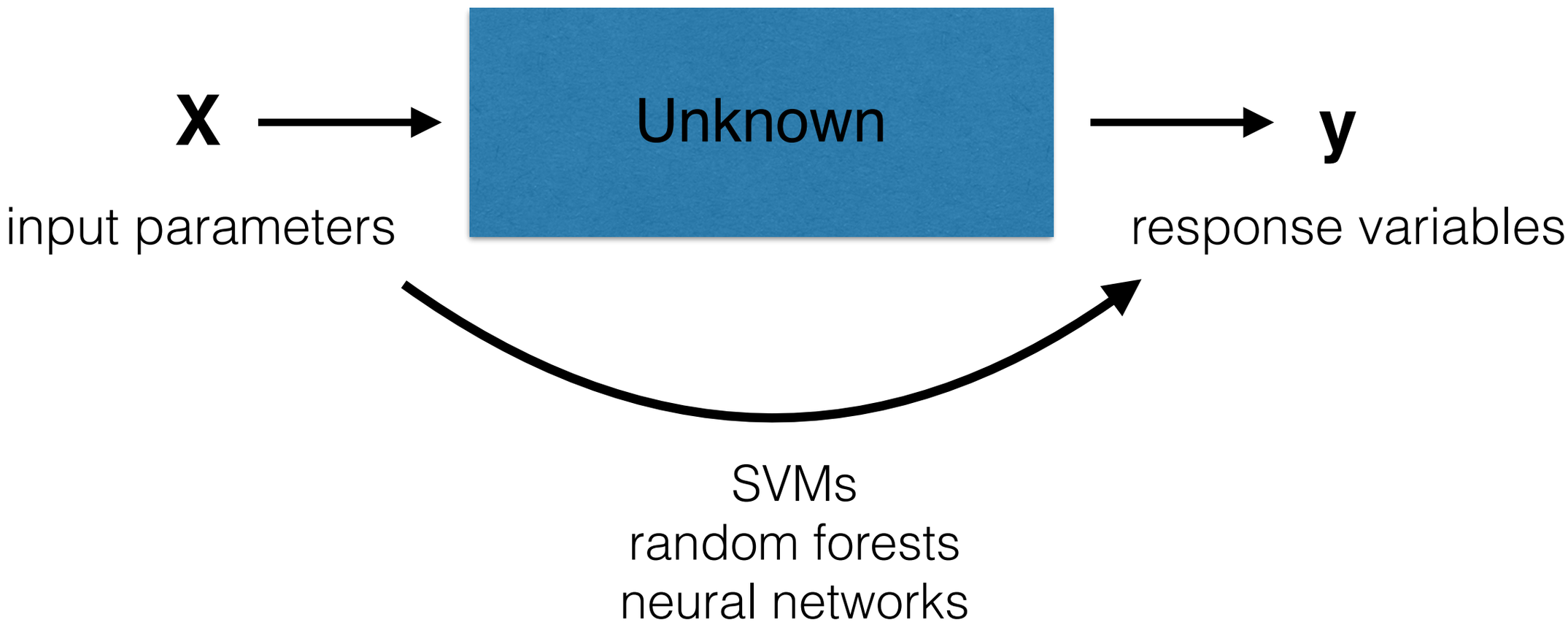
e.g. linear regression

Focus on stochastic model to explain how f(x)-> y

98% of Statistics

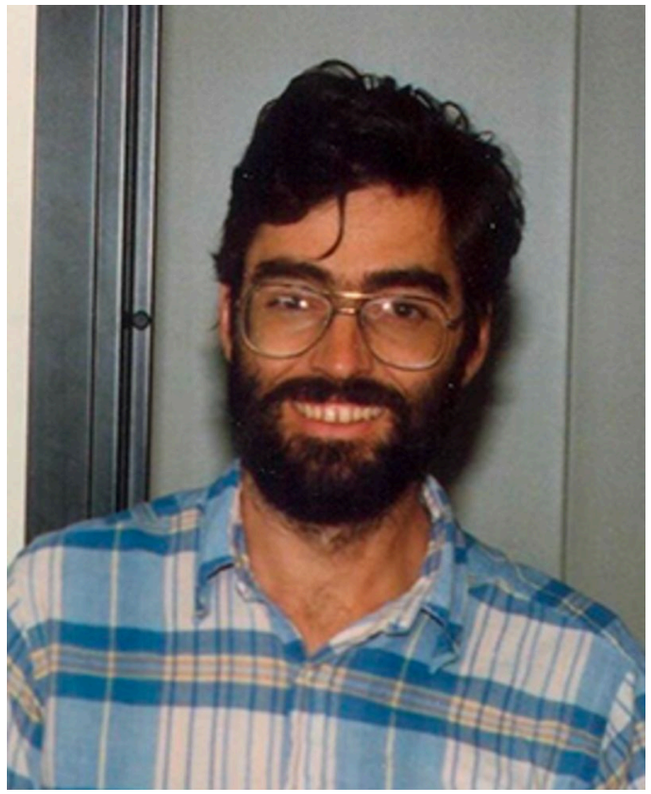
Leo Breiman's Two Cultures

Algorithmic Modeling Culture
(machine learning)



Ignore probabilistic generative model f(x)-> y

Analysis in Population Genetics? One culture!



Major focus on stochastic models, param estimation from those models

Analysis in Population Genetics? One culture!

A Genomic Map of the Effects of Linked Selection in Drosophila

Eyal Elyashiv ✉, Shmuel Sattath, Tina T. Hu, Alon Strutsovsky, Graham McVicker, Peter Andolfatto, Graham Coop, Guy Sella ✉

$$\pi(x) = \frac{\pi_0 \cdot (u(x)/\bar{u})}{\pi_0 \cdot (u(x)/\bar{u}) + 1/B(x) + S(x; \bar{N}_e, T)},$$

Model		Background selection and classic sweeps with added "missing substitutions"	Background selection and classic sweeps without "missing substitutions"
ΔCL		3.9×10^{-4}	3.8×10^{-4}
R^2	1 Mb	0.71	0.70
	100 kb	0.44	0.44
	10 kb	0.26	0.25
	1 kb	0.20	0.20

Deep learning has a lot to offer to popgen

- variable level of model dependance

- Powerful simulation-based inference

- automated feature extraction

Outline

- RNNs for estimating recombination — model-based
- DNNs for estimating geographic origin — model-free
- CNNs for estimating spatial params — model-based

Jeff Adrion

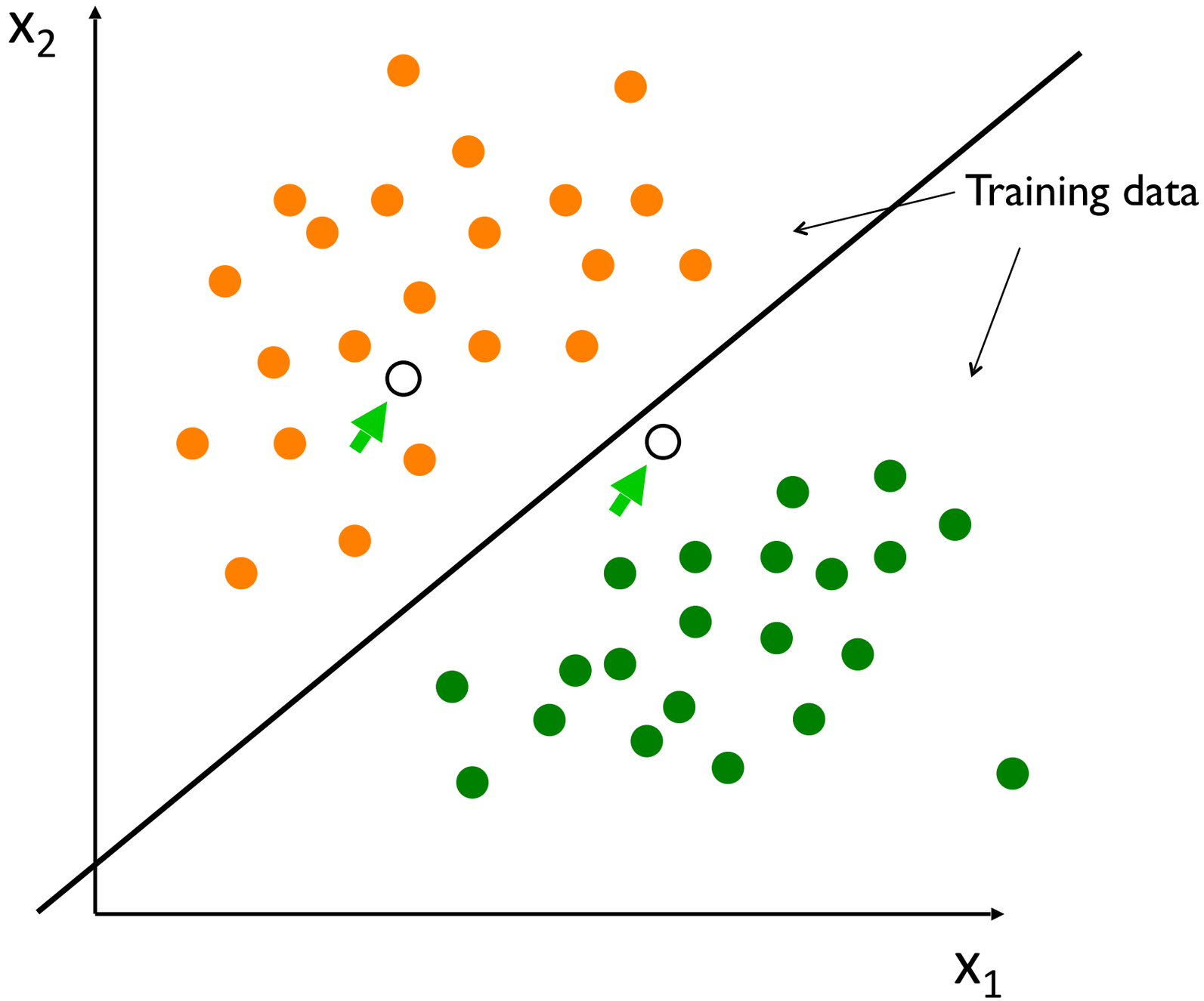
CJ Battey

Clara Rehmann

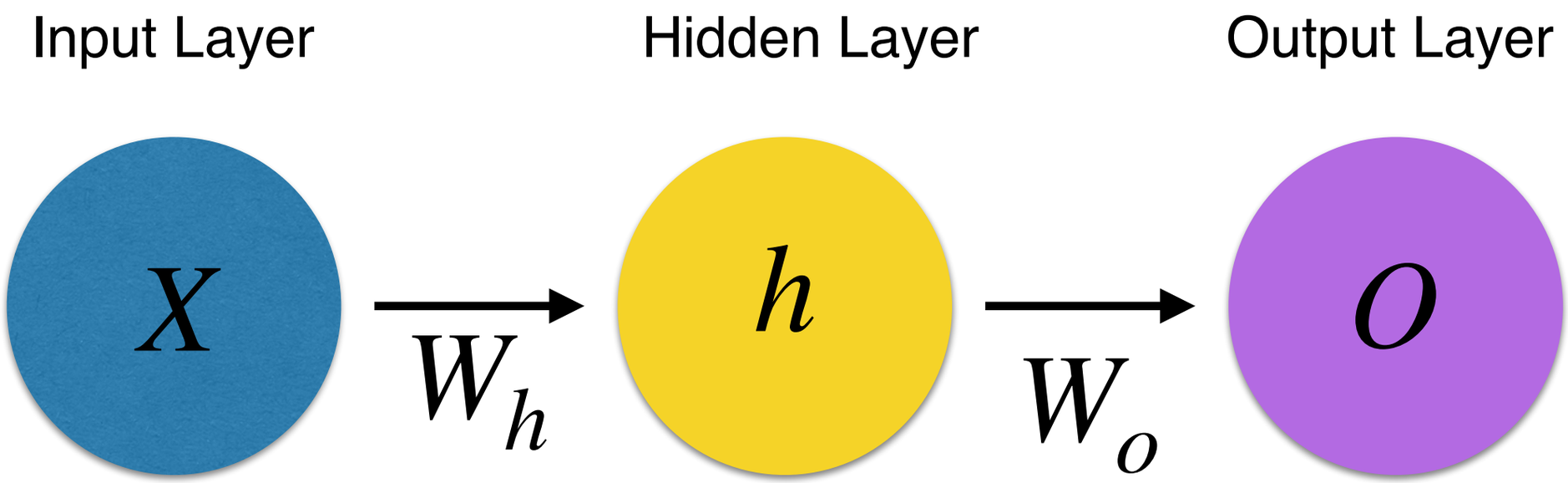
Jordan Rodriguez

Chris Smith

Supervised Machine Learning



toy neural network

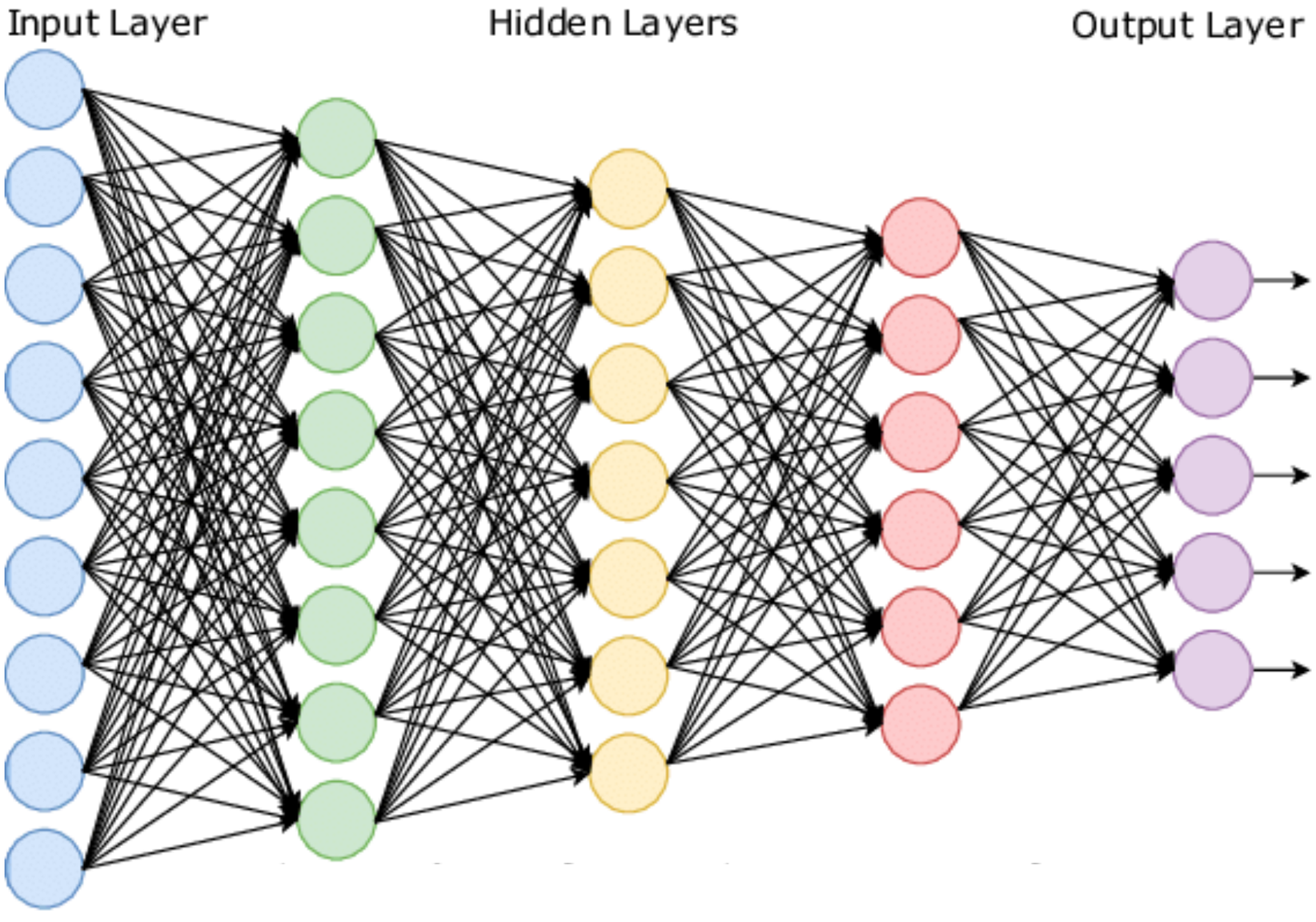


Feed-forward

$$O = f(f(X \cdot W_h + b_h) \cdot W_o + b_o)$$

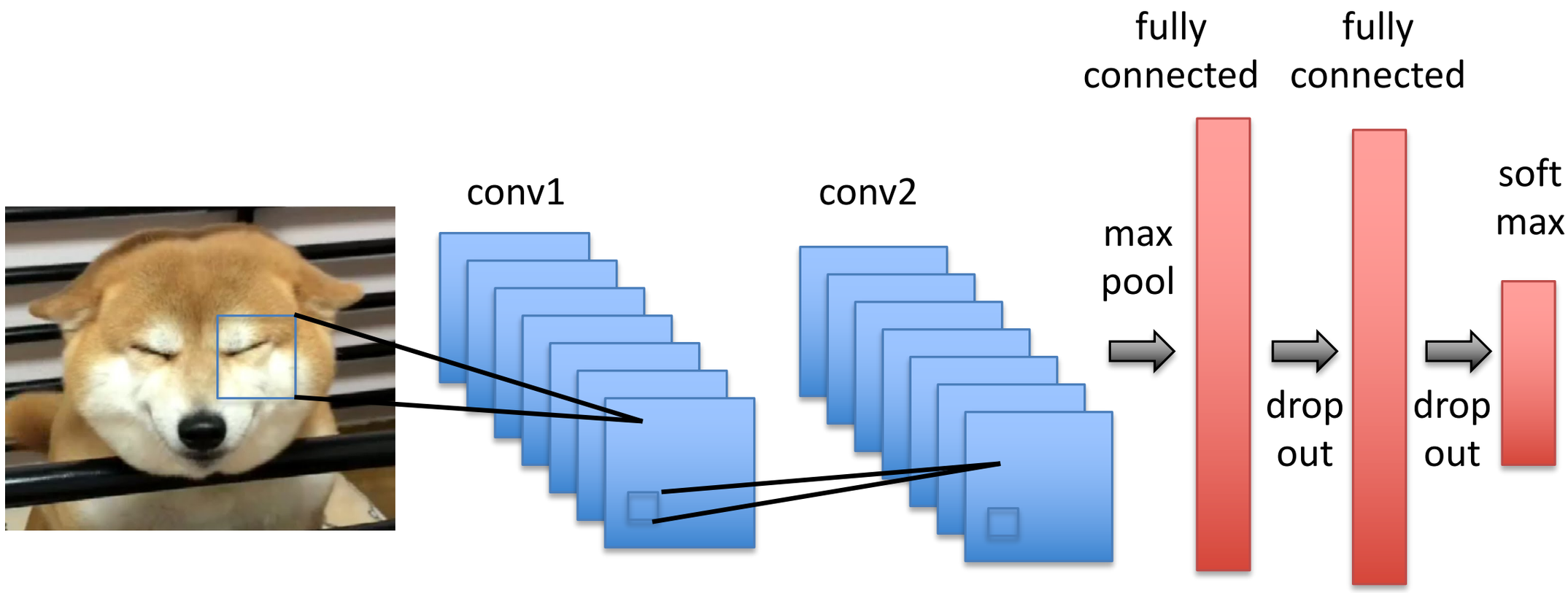
Linear algebra, no magic

Deep Neural Networks



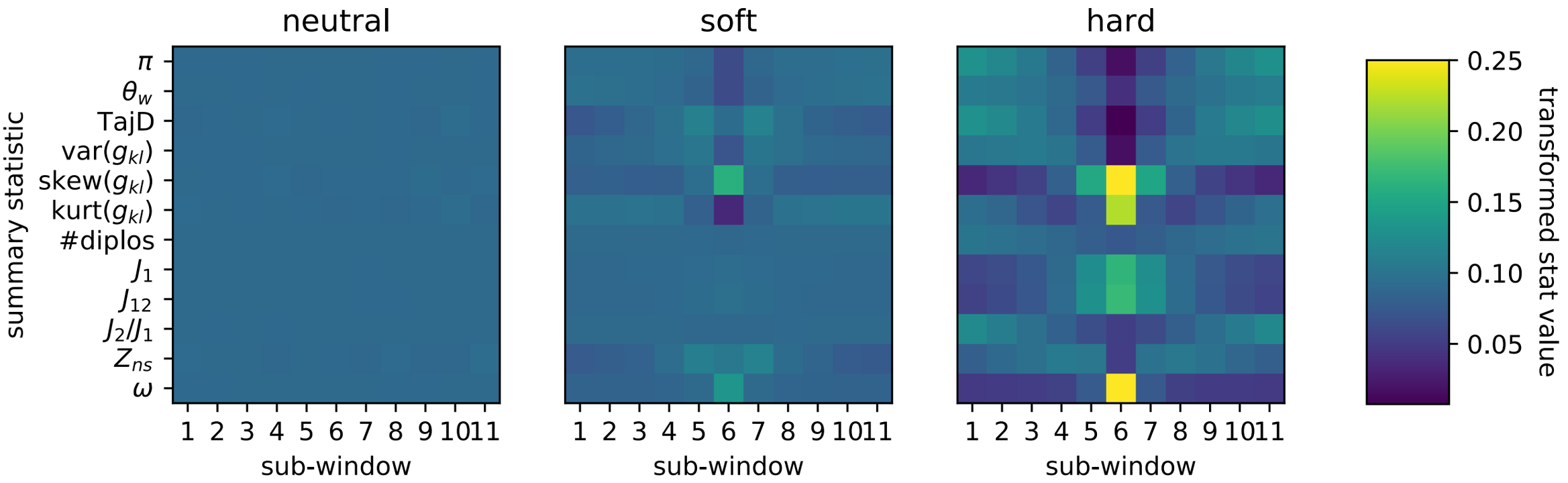
“multilayer perceptron”

Convolutional Neural Network



Slide “receptive filter” along image, take dot product, repeat

popgen as image recognition

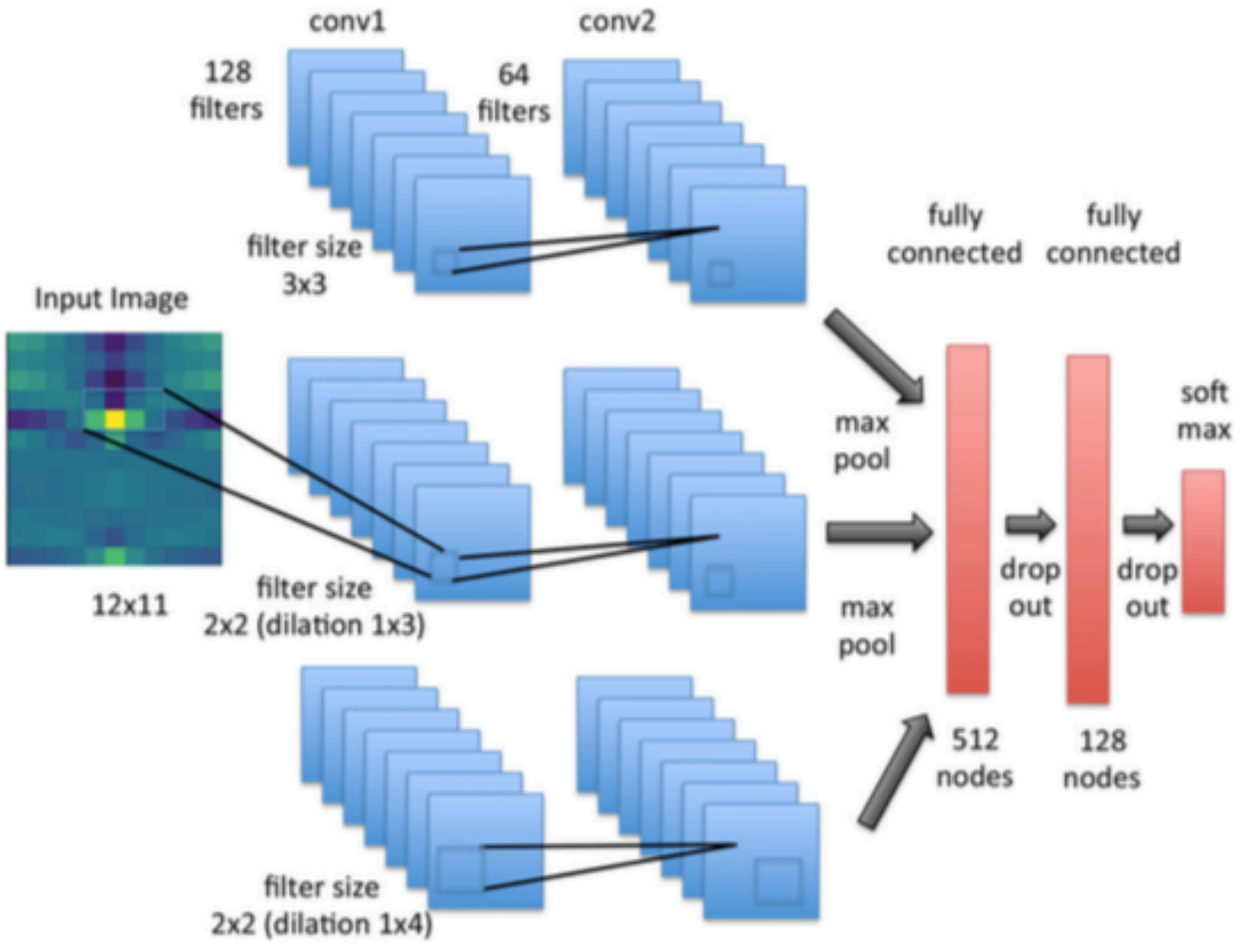


G3 Genes | Genomes | Genetics

diploS/HIC: An Updated Approach to Classifying Selective Sweeps

Andrew D. Kern[1] and Daniel R. Schrider
Department of Genetics, Rutgers University, Piscataway, NJ 08854
ORCID IDs: 0000-0003-4381-4680 (A.D.K.); 0000-0001-5249-4151 (D.R.S.)

diploS/HIC



G3
Genes | Genomes | Genetics

diploS/HIC: An Updated Approach to Classifying Selective Sweeps

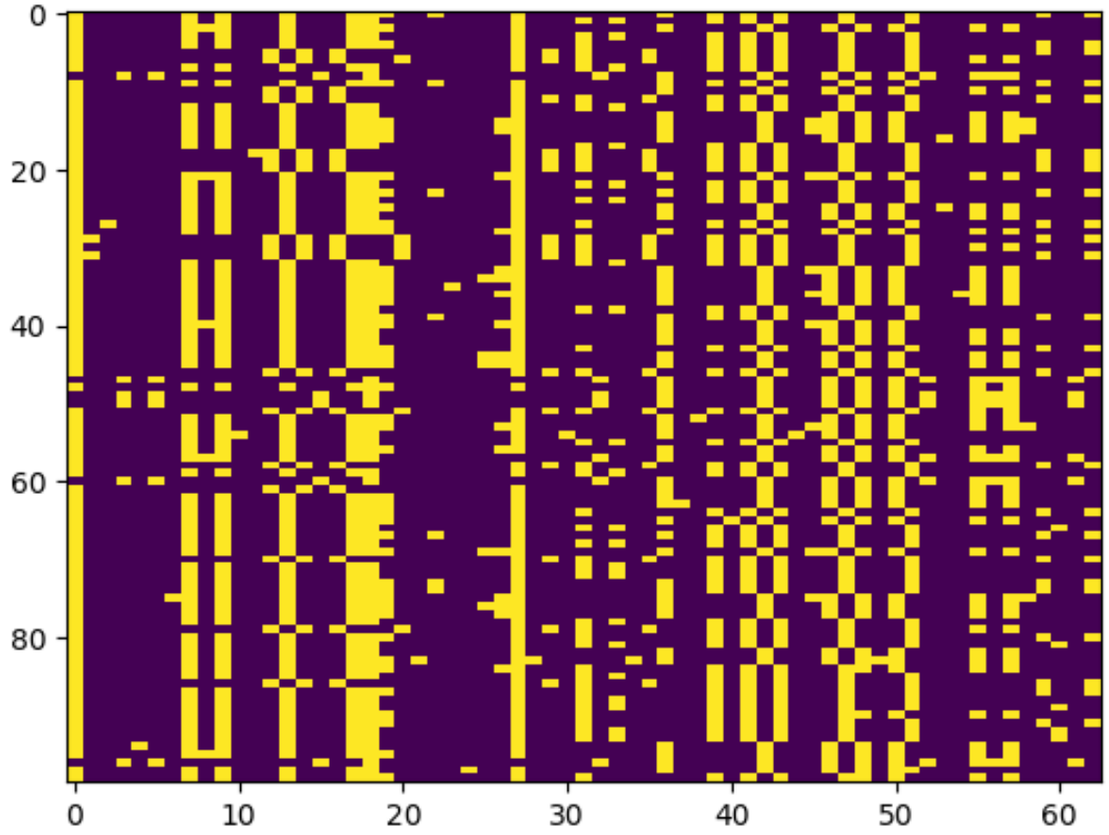
Andrew D. Kern[1] and Daniel R. Schrider
Department of Genetics, Rutgers University, Piscataway, NJ 08854
ORCID IDs: 0000-0003-4381-4680 (A.D.K.); 0000-0001-5249-4151 (D.R.S.)

using CNNs provides added accuracy & robustness

Alignment “images” for Deep Learning

- Automated feature extraction

- Using convolutional neural nets (deep learning) quickly have gained new traction on many problems



Dan Schrider (now UNC)

A Likelihood-Free Inference Framework for Population Genetic Data using Exchangeable Neural Networks

Jeffrey Chan, Valerio Perrone, Jeffrey P. Spence, Paul A. Jenkins, Sara Mathieson, Yun S. Song

(Submitted on 16 Feb 2018)

The Unreasonable Effectiveness of Convolutional Neural Networks in Population Genetic Inference

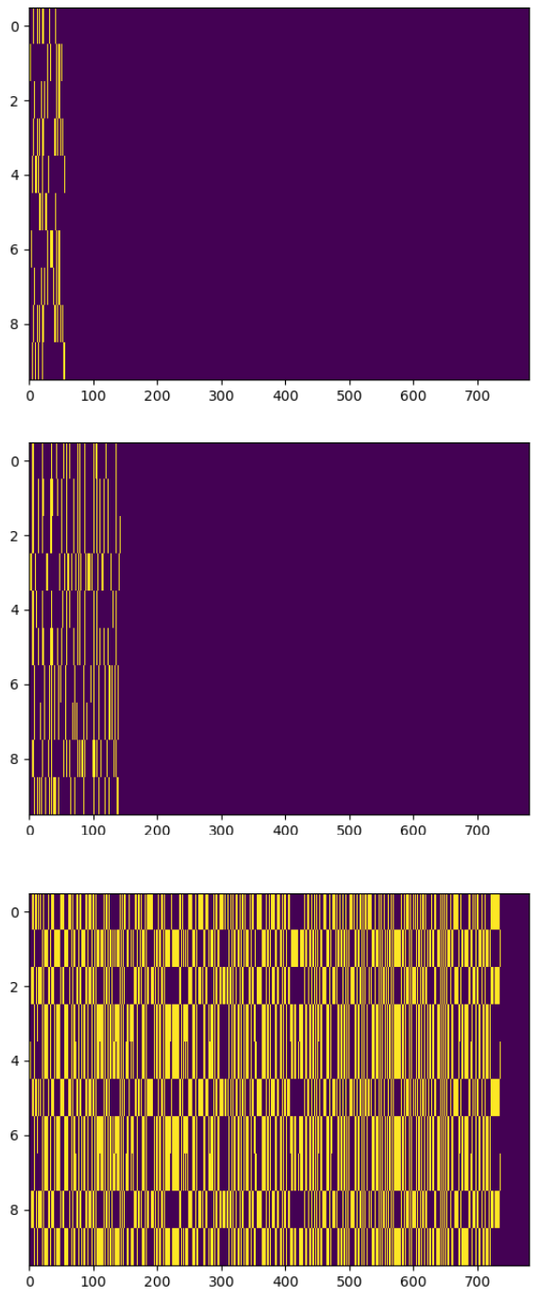
Lex Flagel, Yaniv J Brandvain, Daniel R Schrider

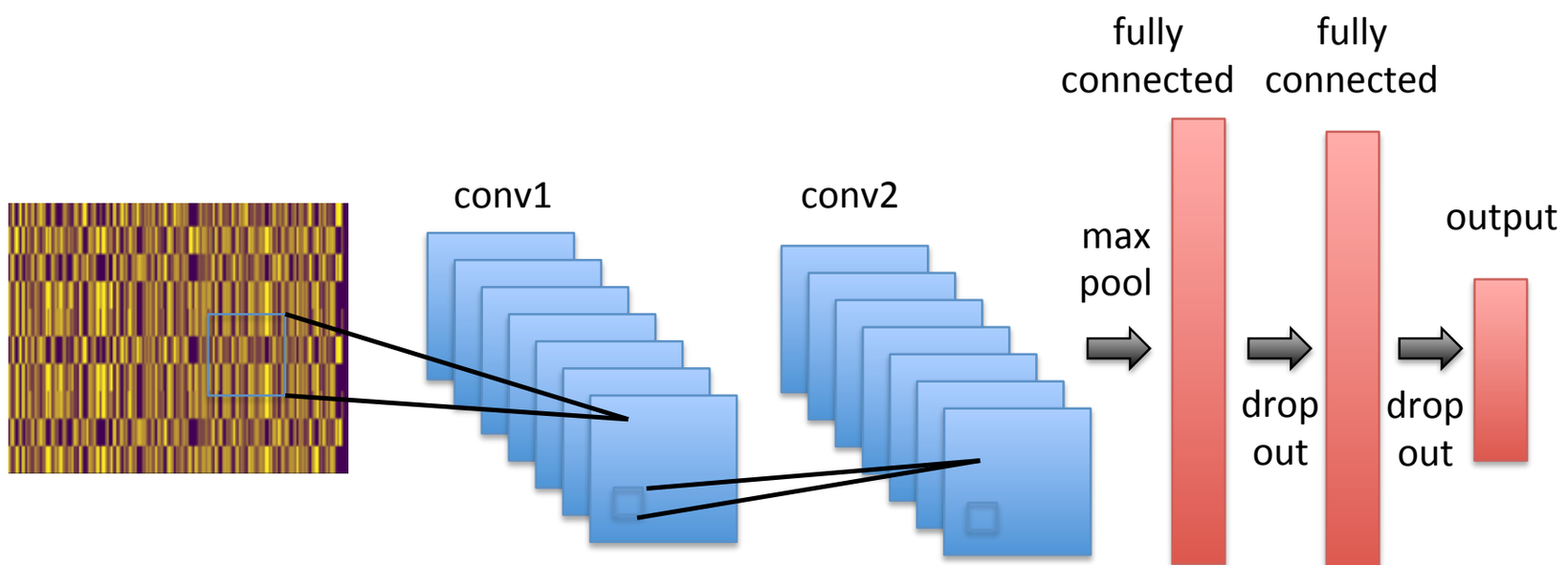
doi: https://doi.org/10.1101/336073

Convolutional Neural Nets

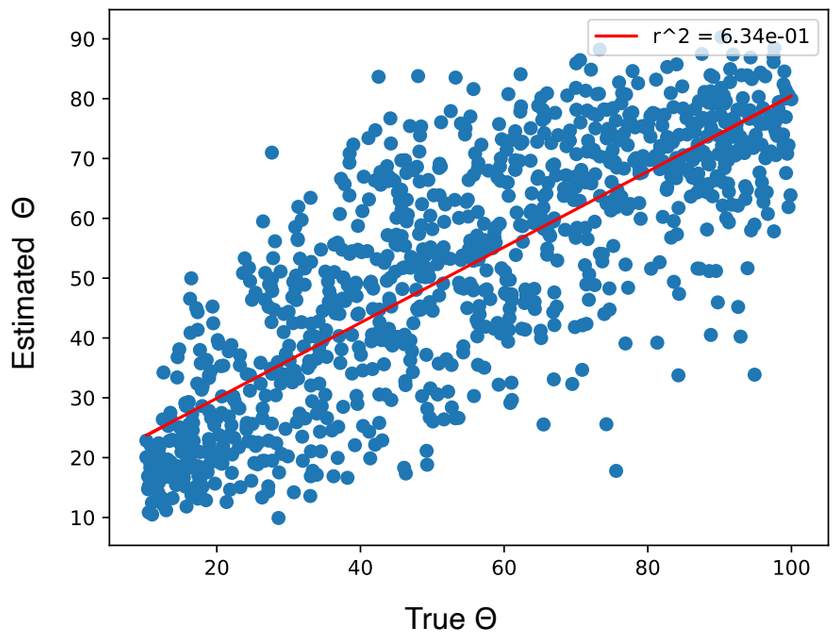
Population genetic parameter estimation: Θ

Alignment images

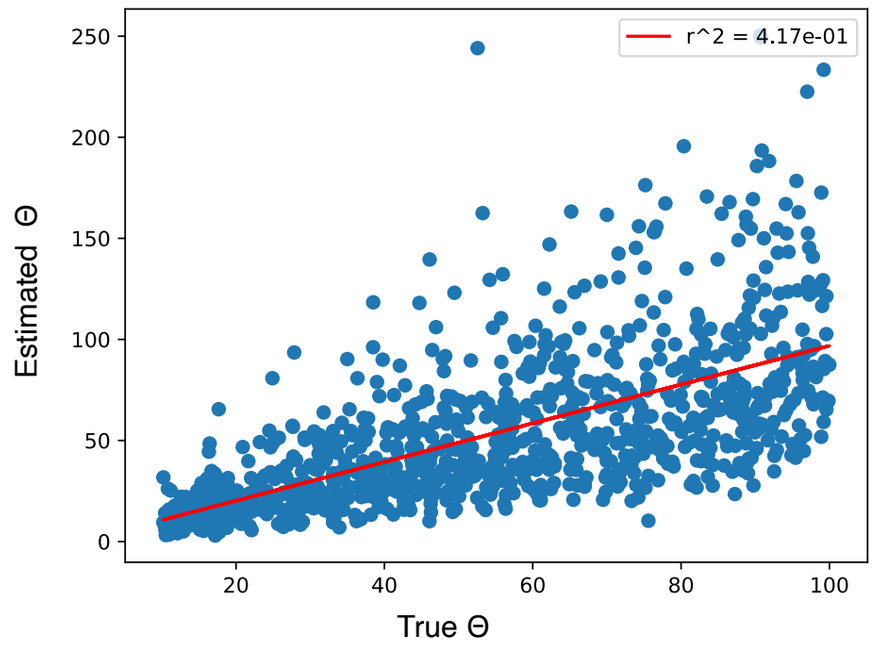




CNN estimator



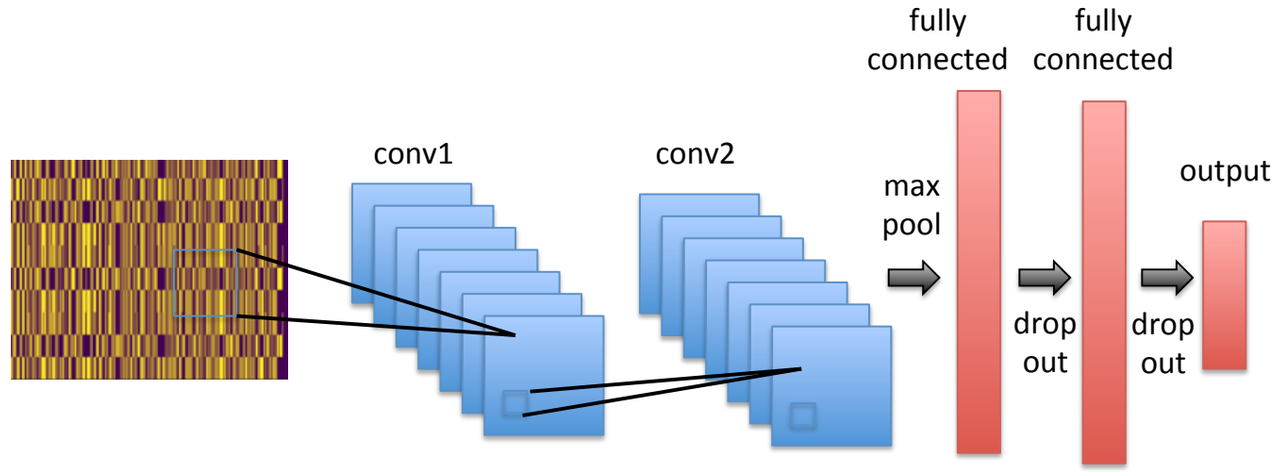
Tajima's π estimator



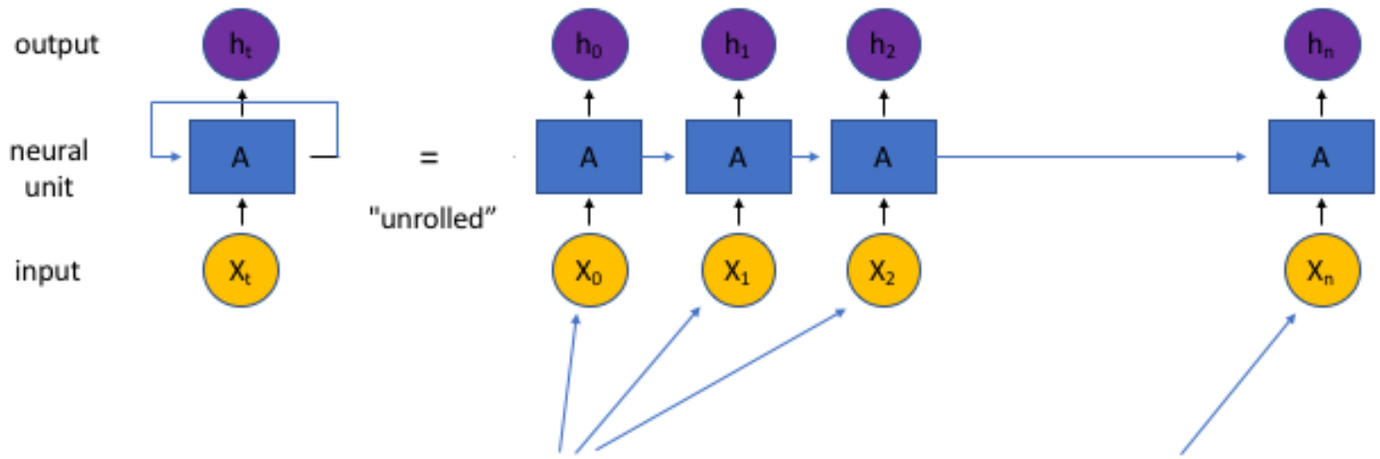
Deep Neural Nets

What is the best architecture to use?

CNNs?

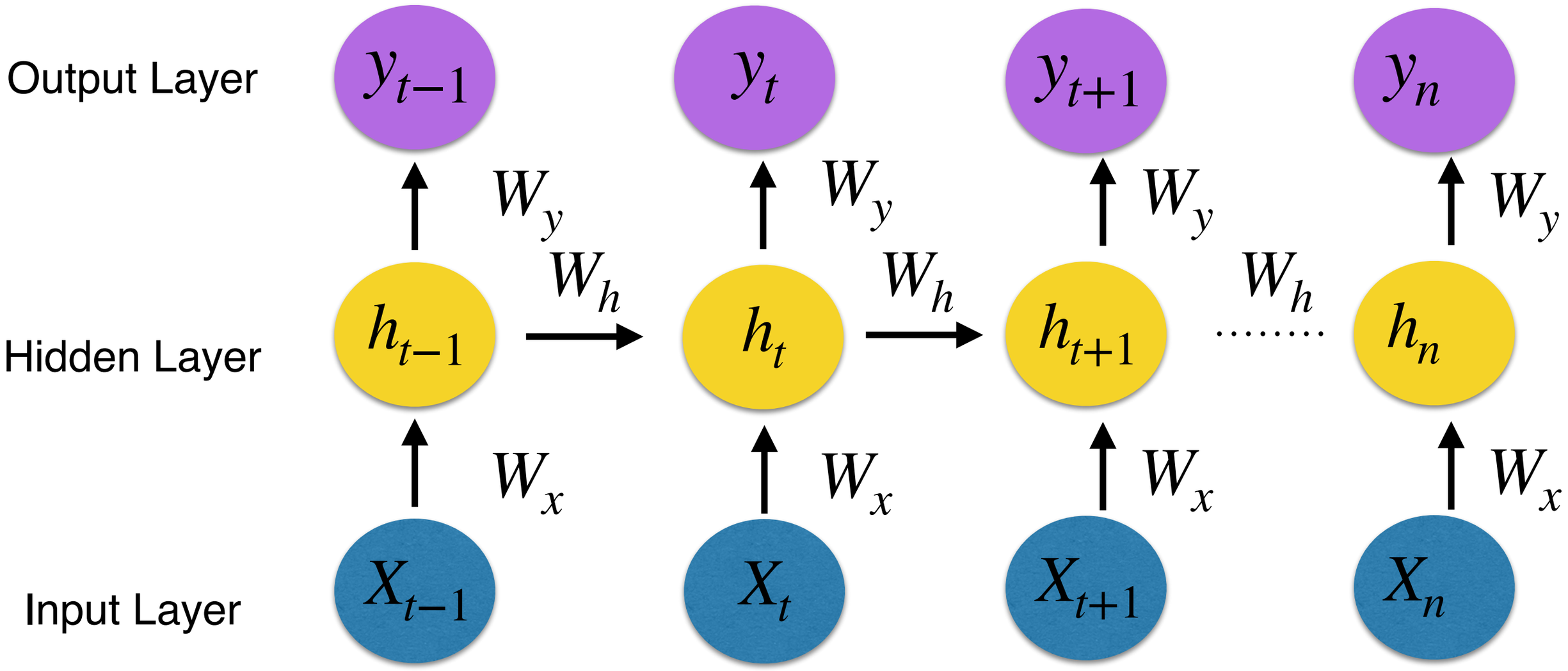


RNNs?



Jeff Adrion

recurrent neural network

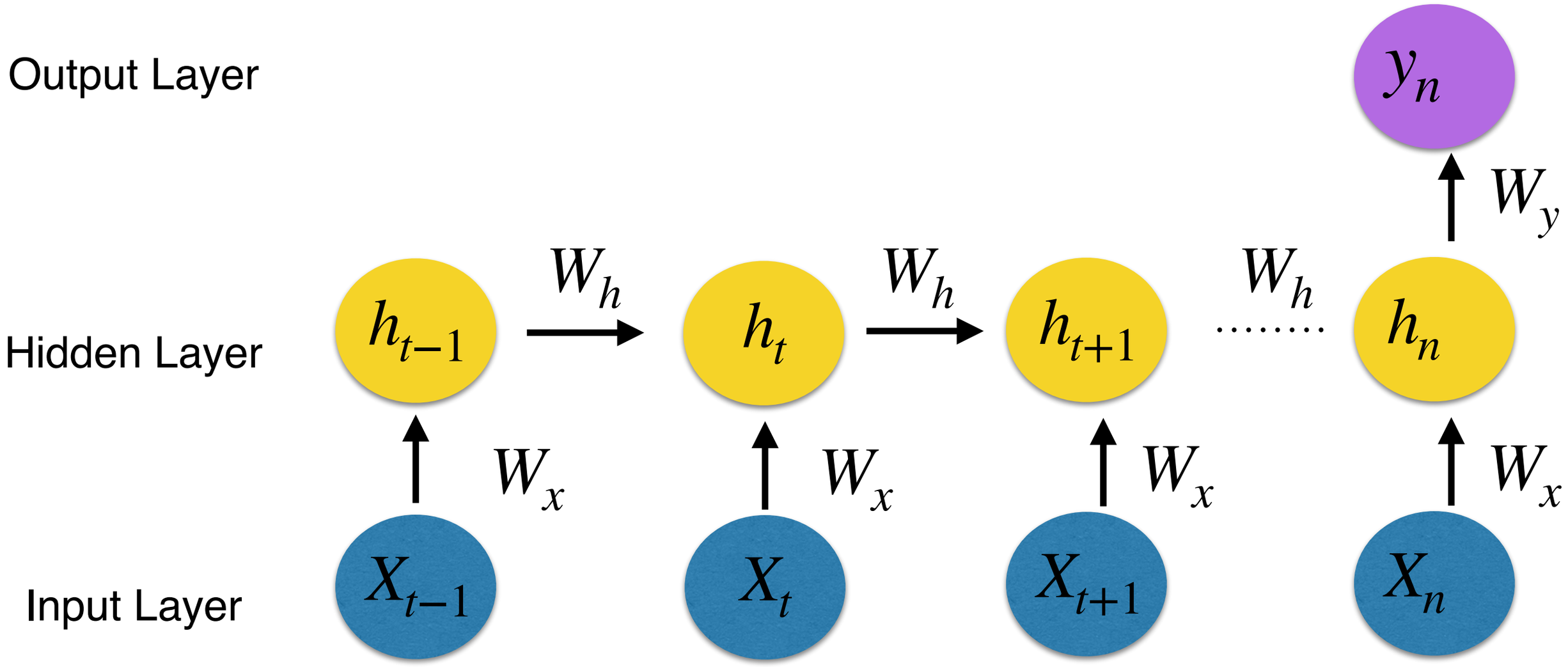


Feed-forward

$$h_t = f(W_h \cdot h_{t-1} + W_x \cdot x_t)$$

$$\hat{y}_t = softmax(W_y \cdot h_t)$$

recurrent neural network



Feed-forward

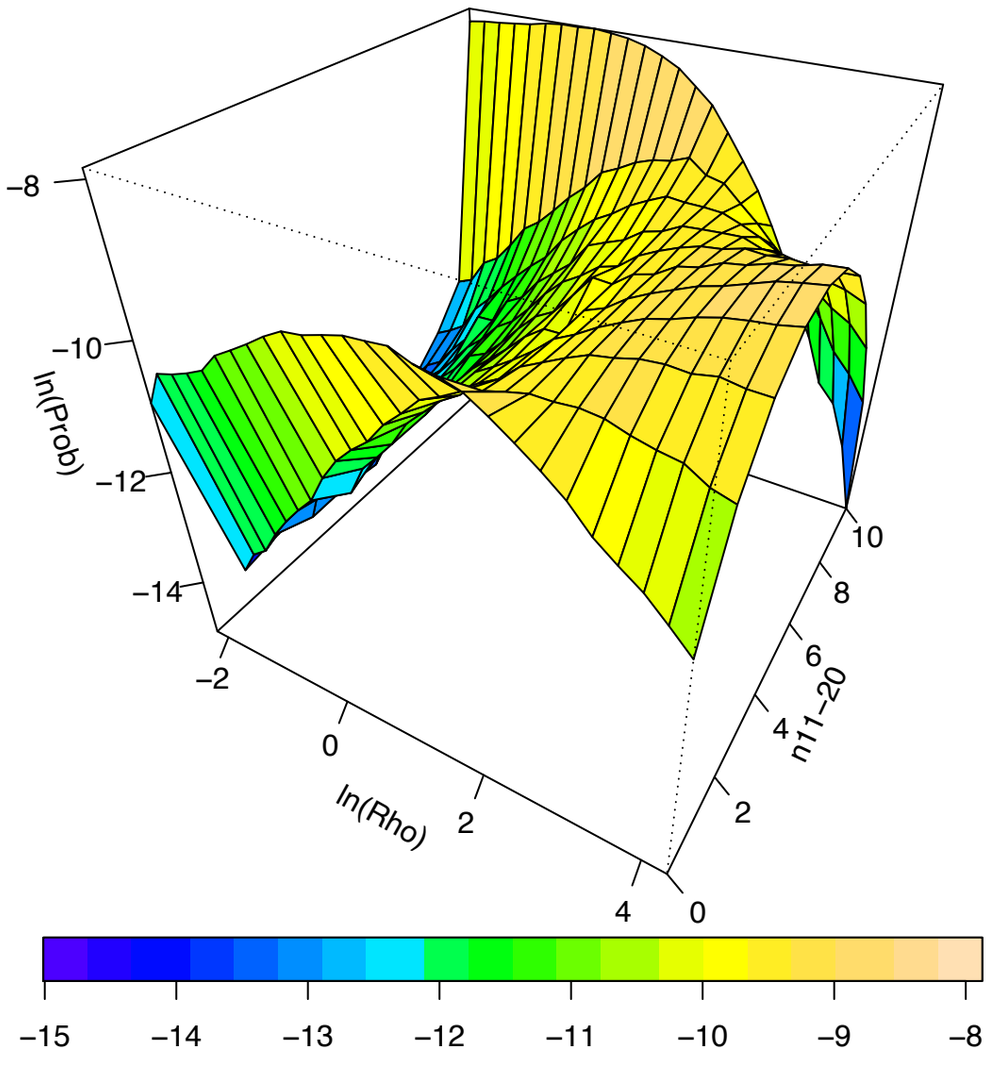
$$h_t = f(W_h \cdot h_{t-1} + W_x \cdot x_t)$$

$$\hat{y}_t = softmax(W_y \cdot h_t)$$

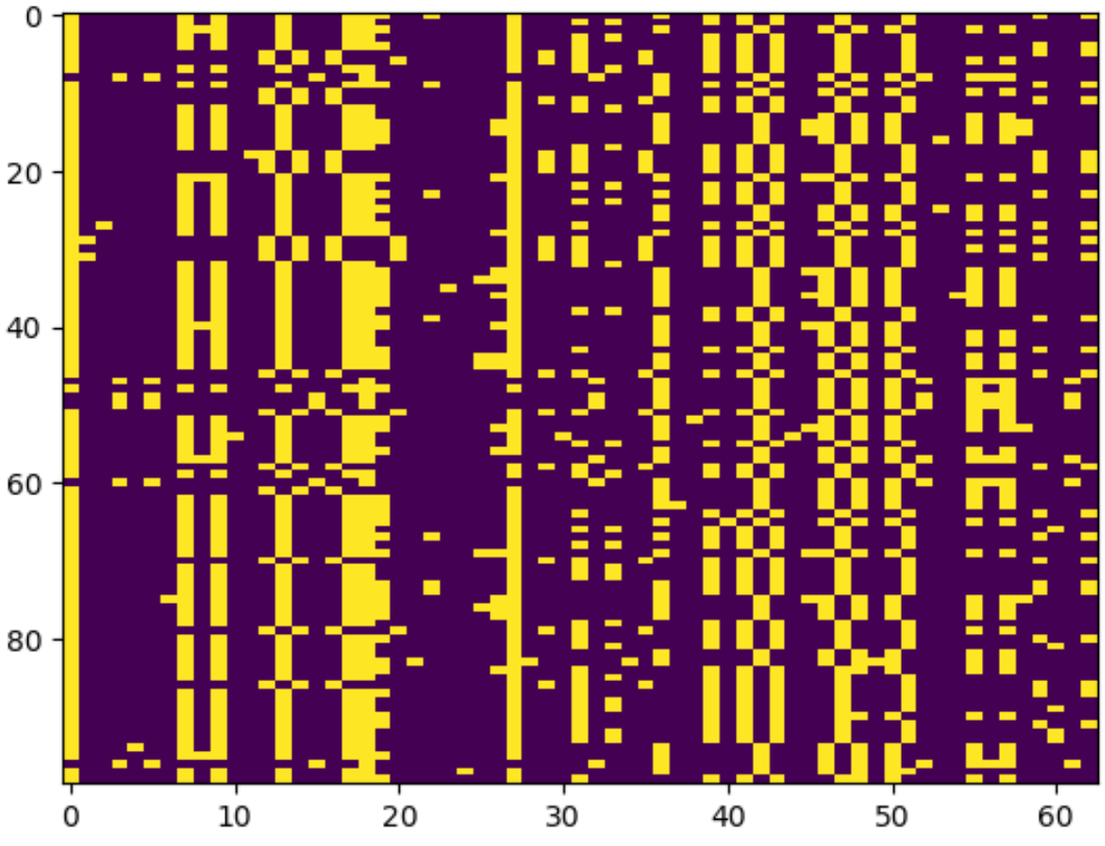
Estimating Recombination Rate with Deep Learning

Population genetic parameter estimation: 4Nr

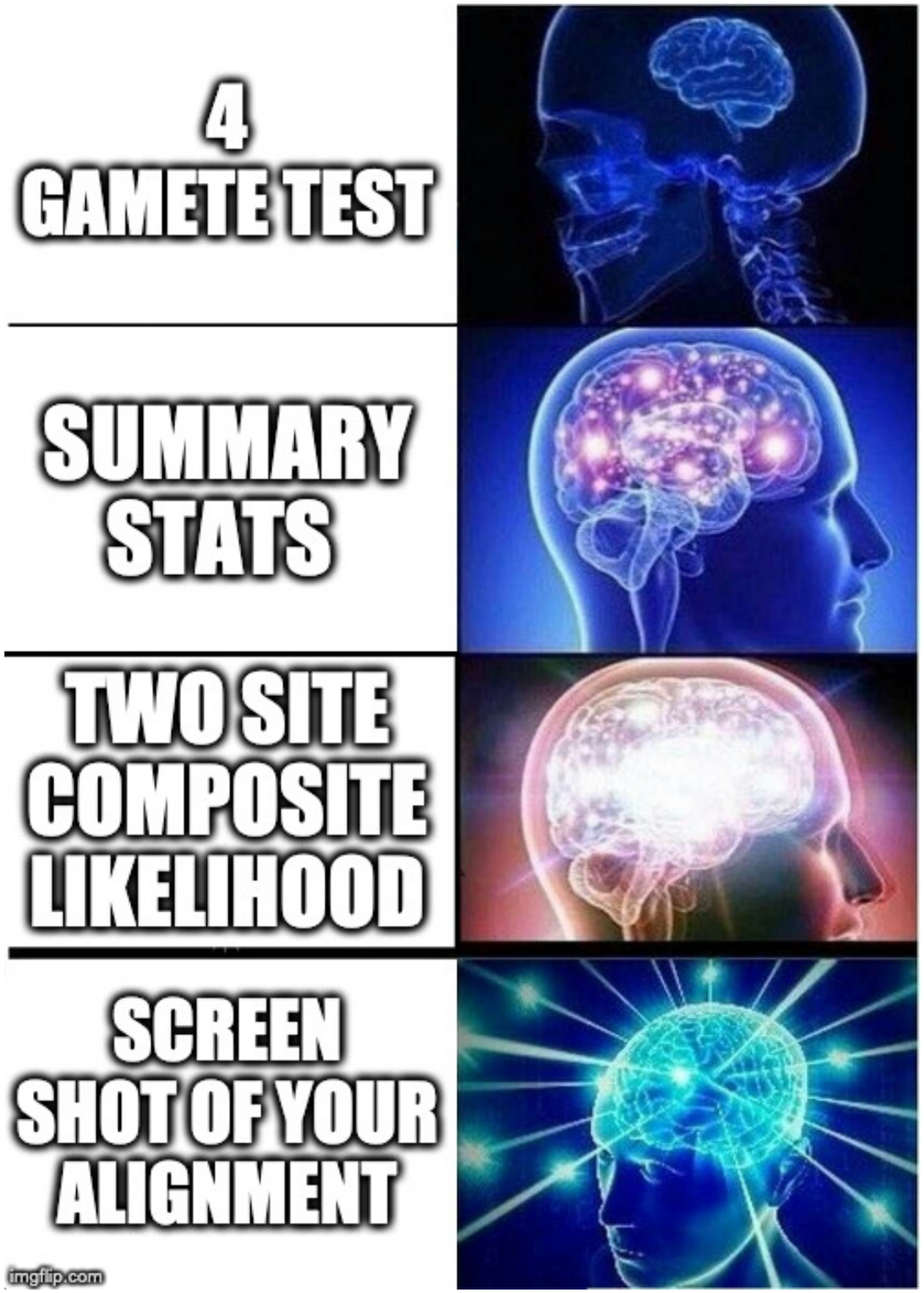
Composite Likelihood— LDhat



Deep Neural Nets

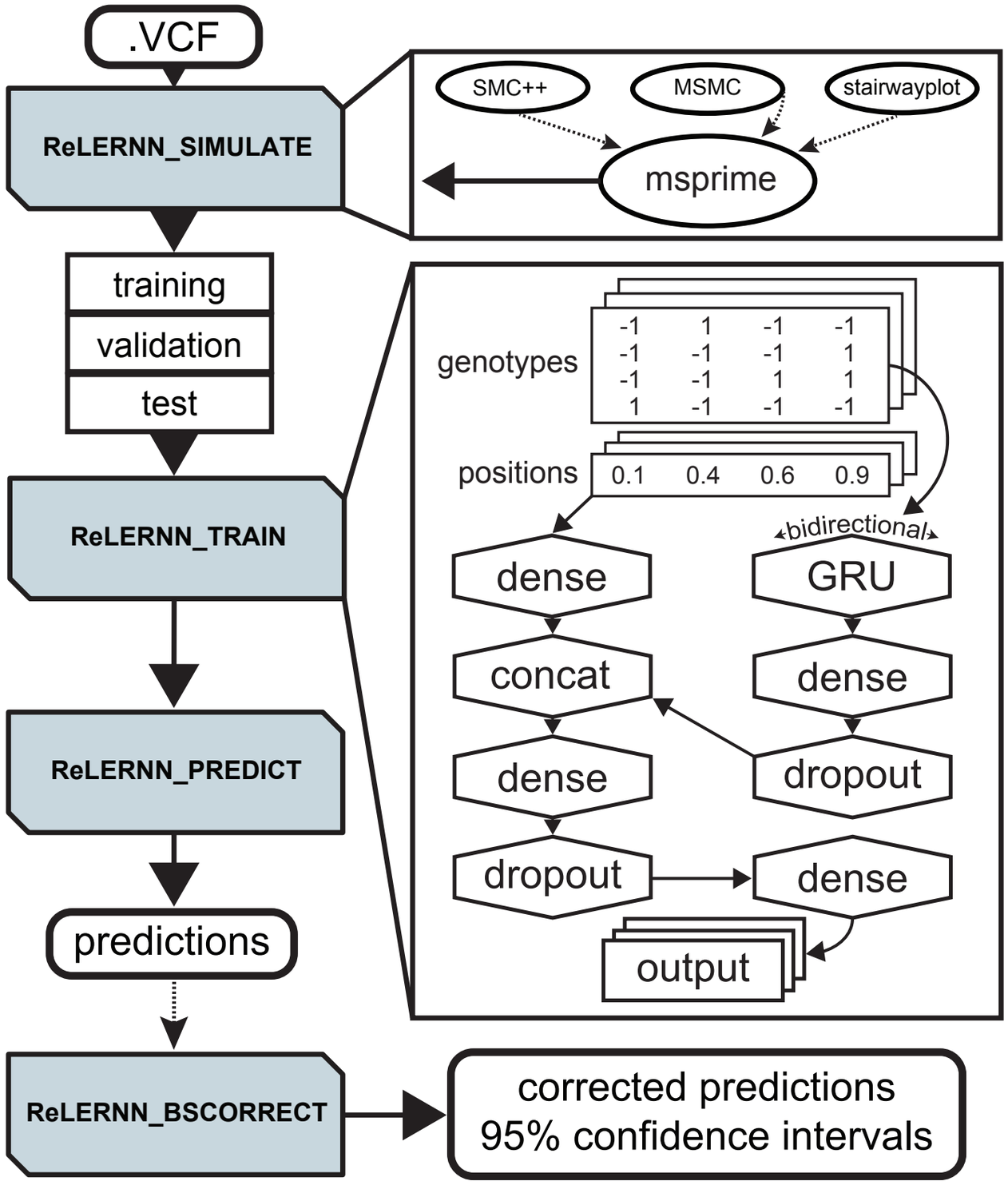


Meme history of estimating recombination rate



Where we are now

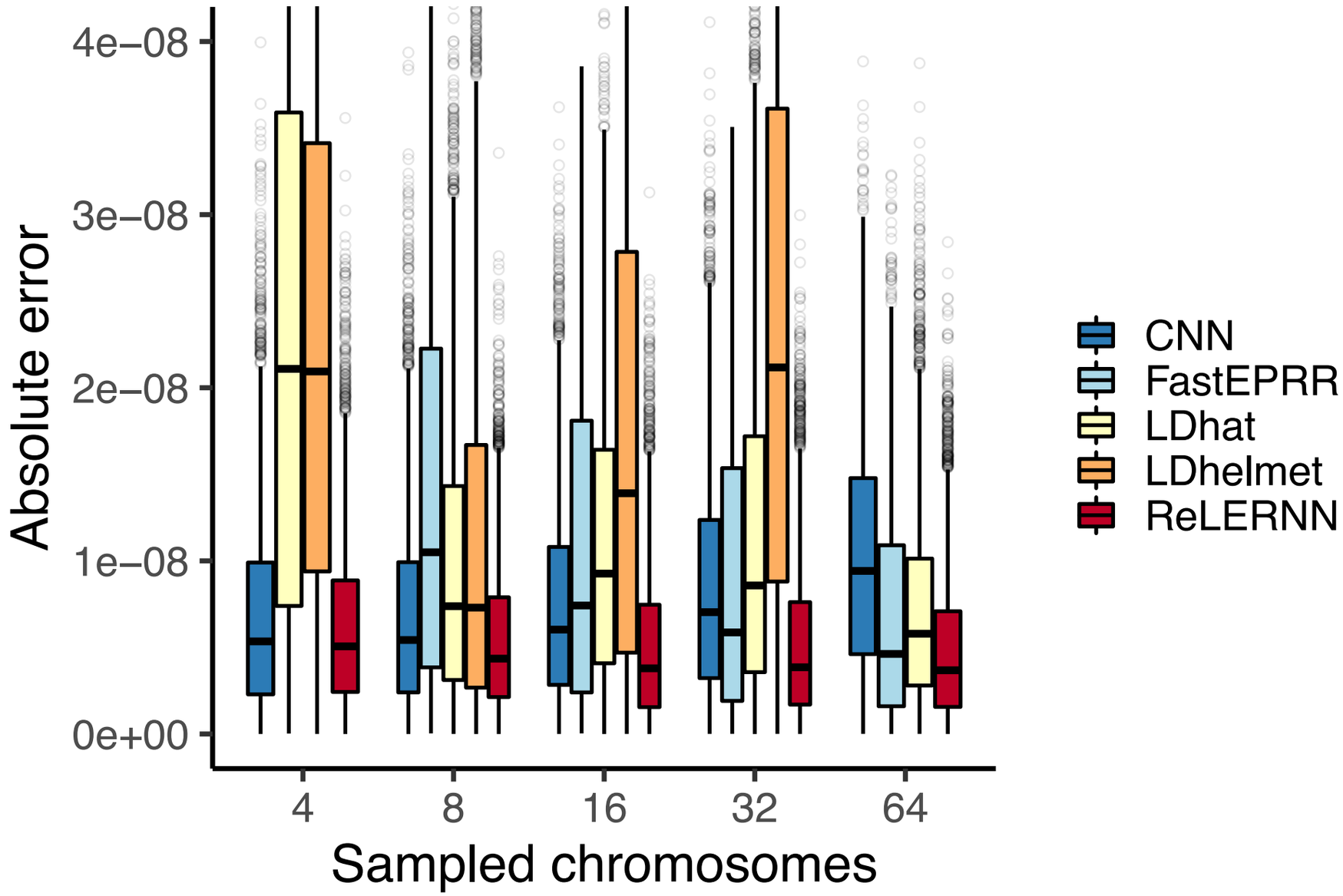
ReLERNN



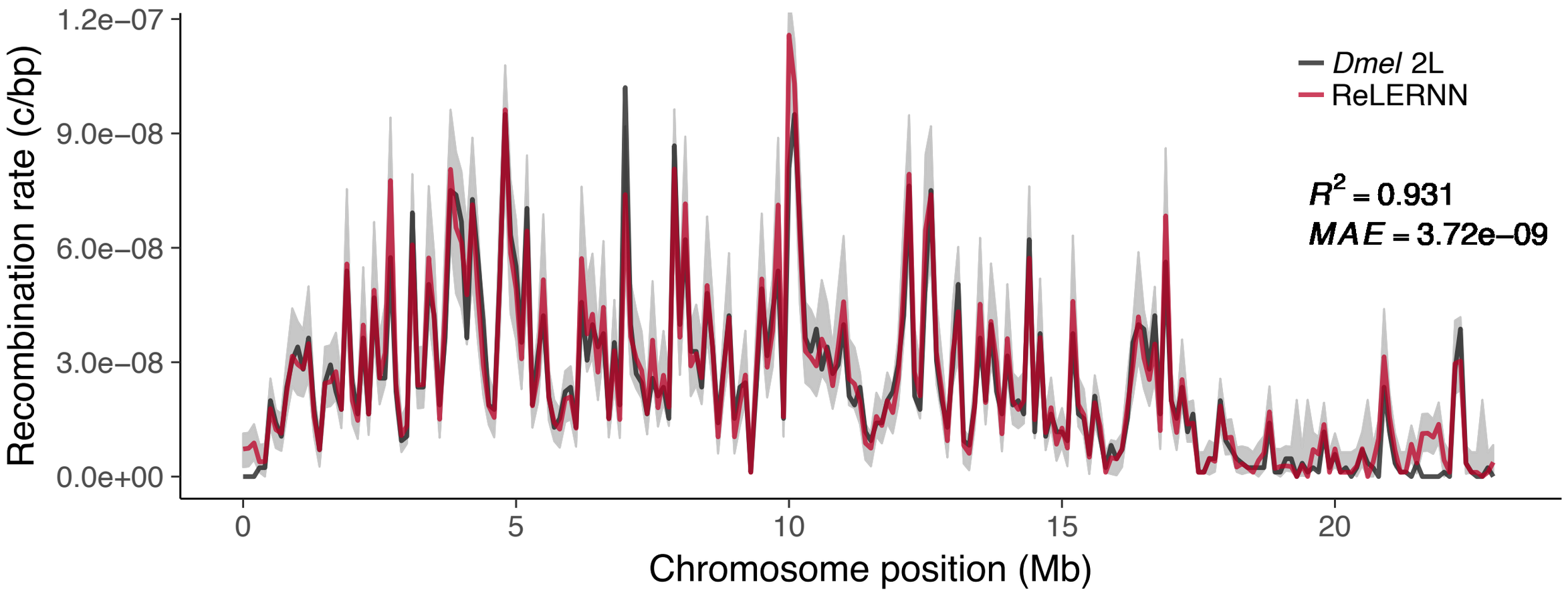
Jeff Adrion

Estimating Recombination Rate with Deep Learning

Population genetic parameter estimation: 4Nr

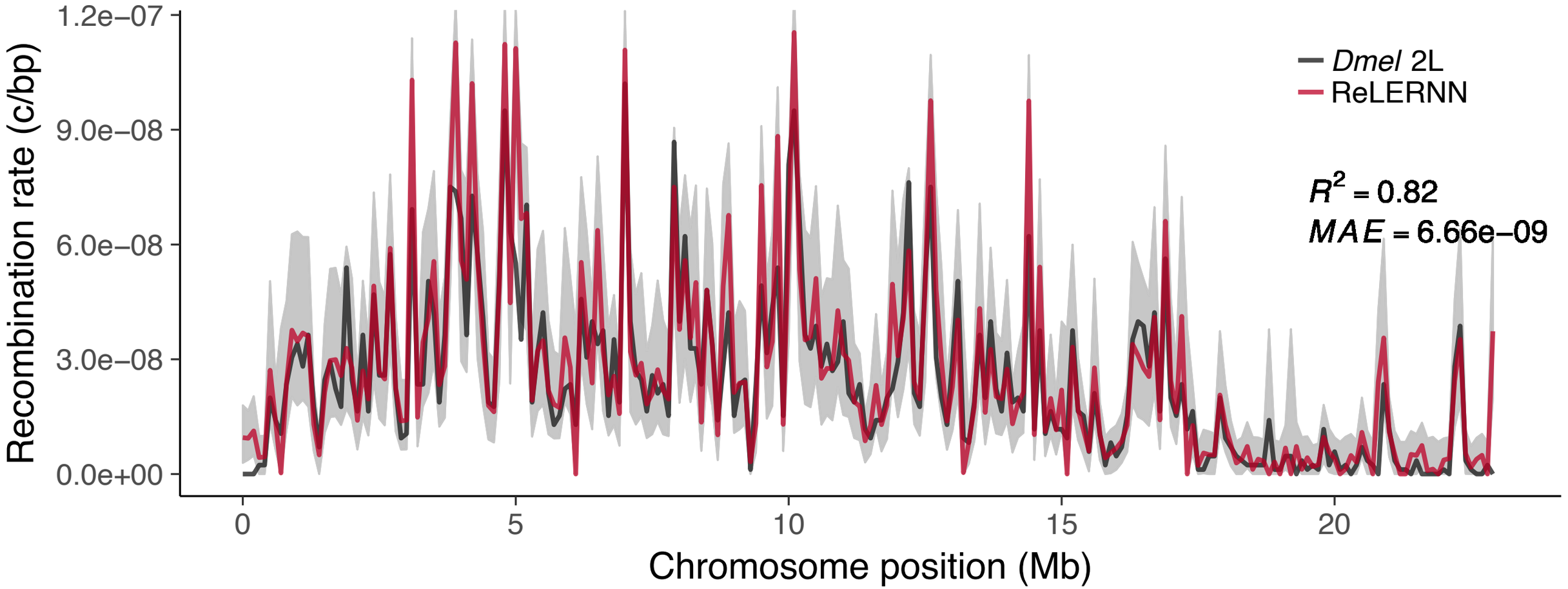


Estimating Recombination Rate with Deep Learning



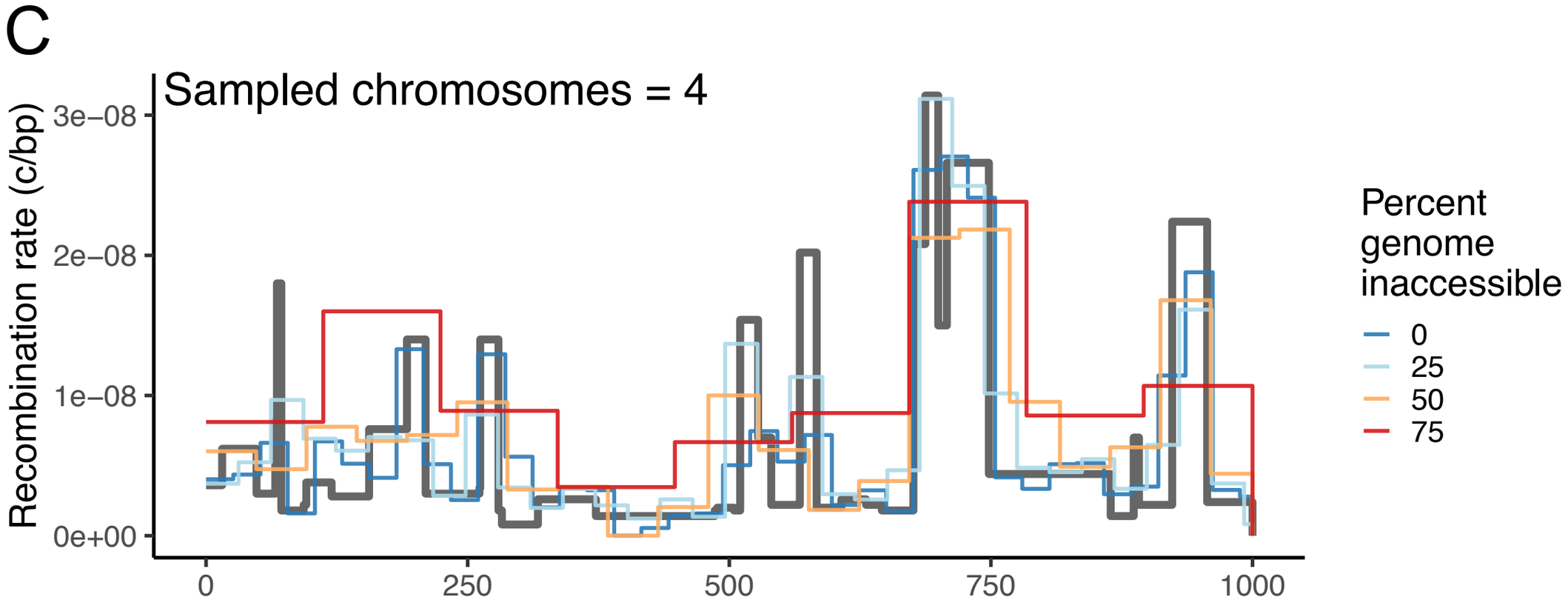
Comeron et al. (2012) genetic map from Drosophila (n=20)

Estimating Recombination Rate with Deep Learning



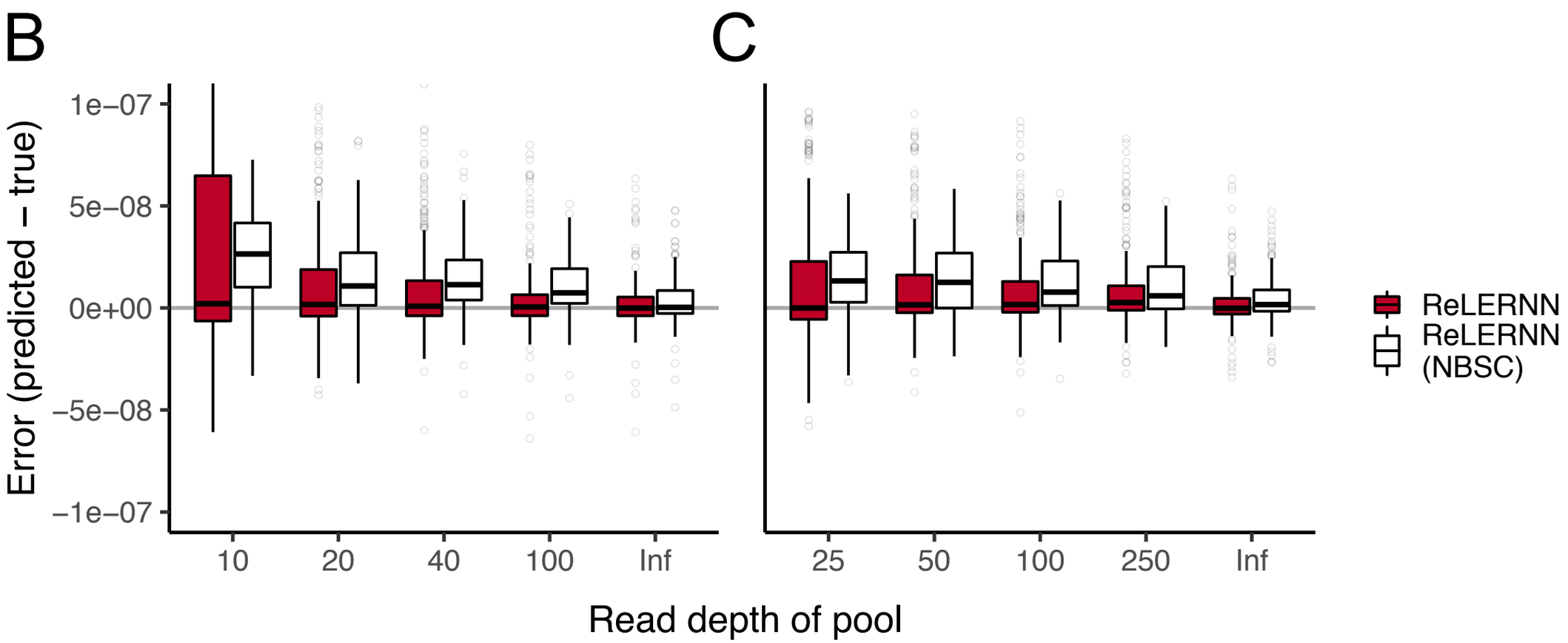
Estimating Recombination Rate with Deep Learning

ReLERNN does well with ‘bad data’



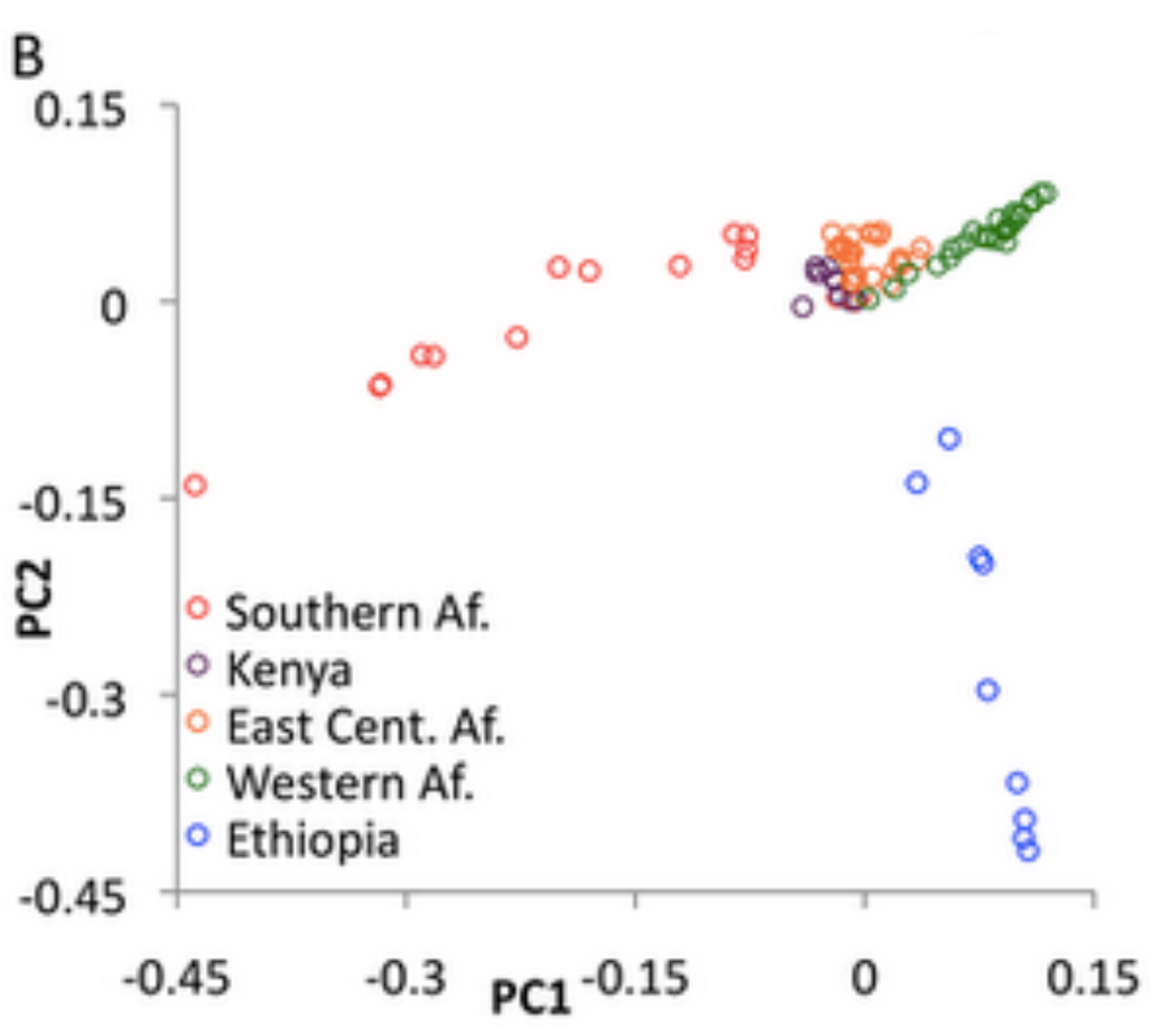
Estimating Recombination Rate with Deep Learning

ReLERNN does well with ‘bad data’



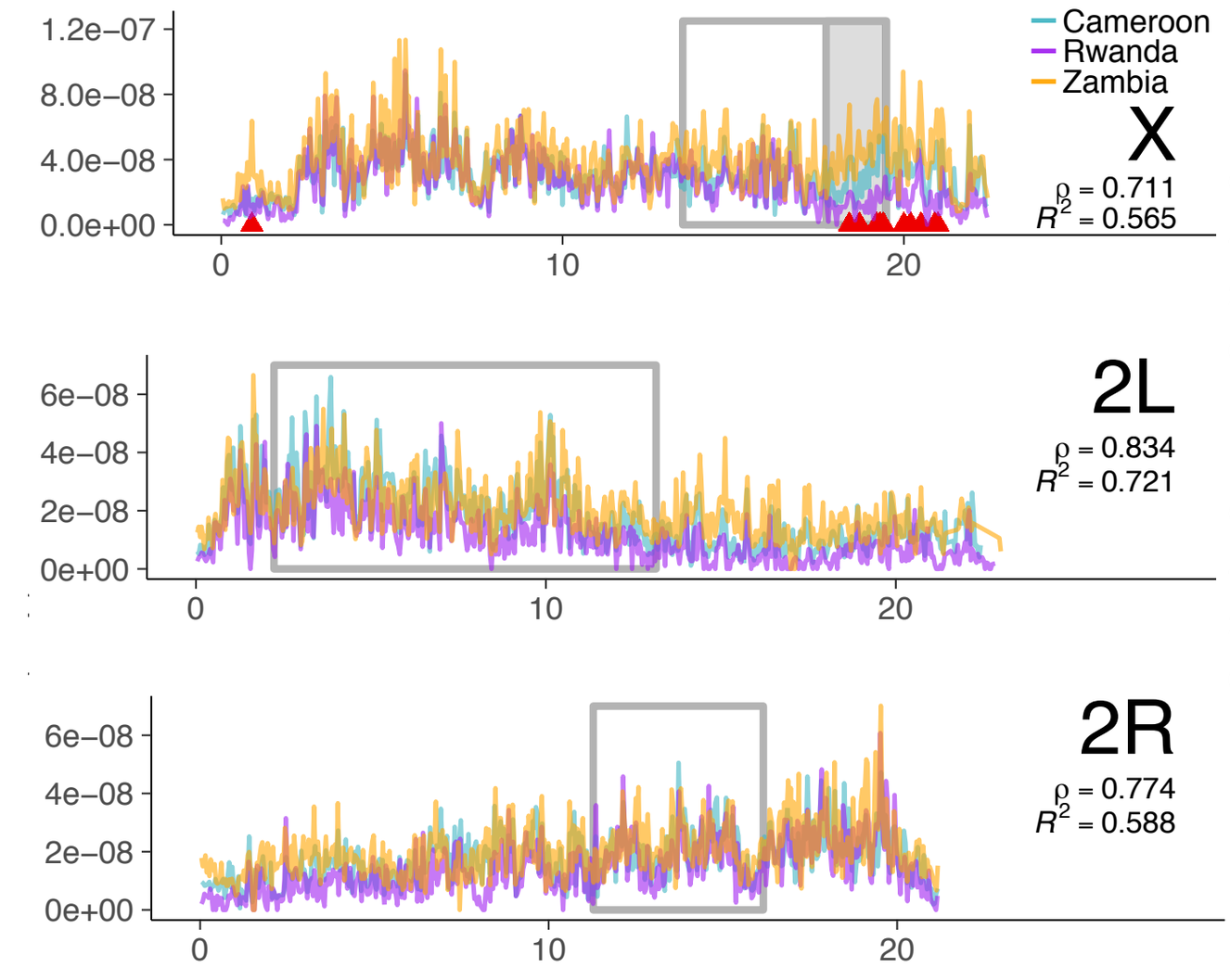
Can even use ReLERNN for pool-seq data!

Application to African Drosophila Samples



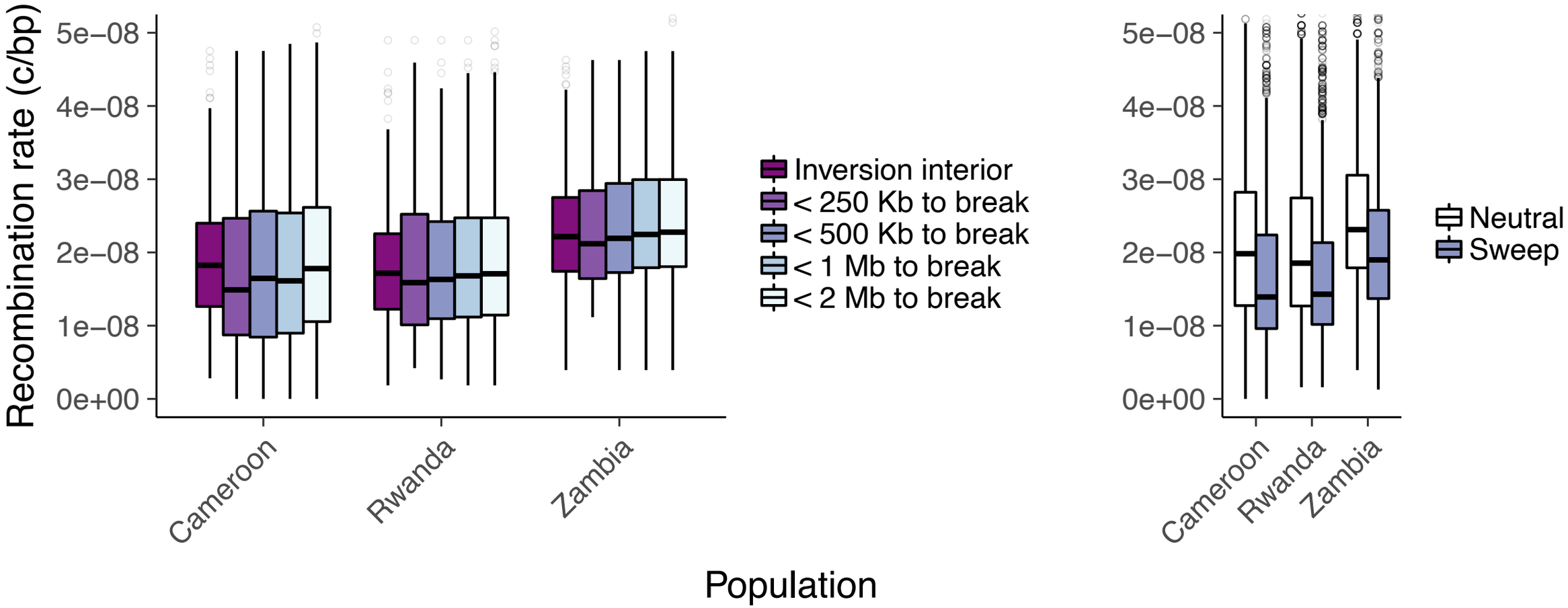
Data from 3 populations sampled in Pool et al (2012)

Application to African Drosophila Samples



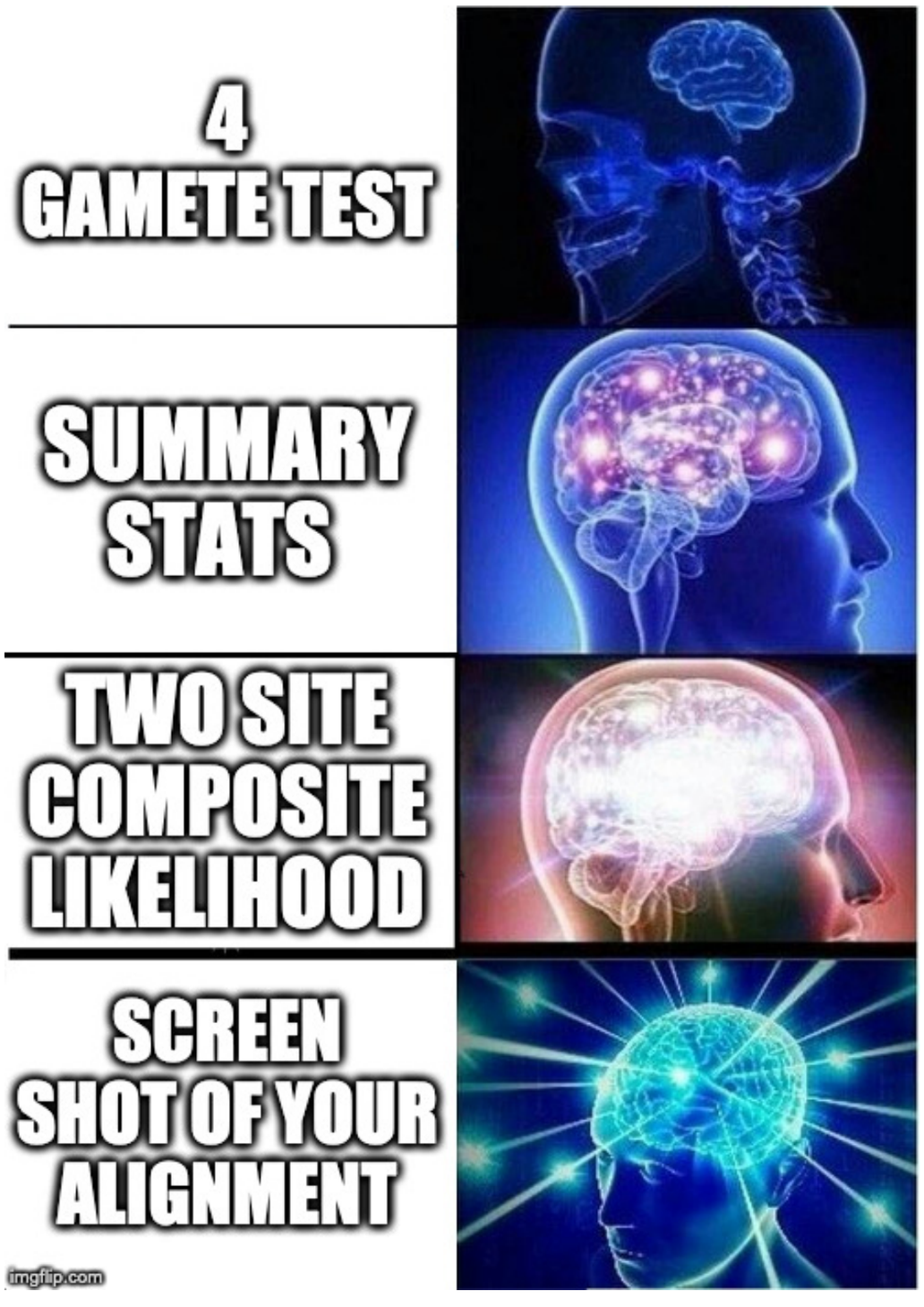
Major Outliers Associated with Segregating Inversions

Both inversions and sweeps affect inferred recombination



Effect of cosmopolitan inversion *In(2L)t*

ReLERNN Summary



- Early days for automated feature extraction for popgen

- To do: which architectures? which tasks? uncertainty?

Predicting the landscape of recombination using deep learning

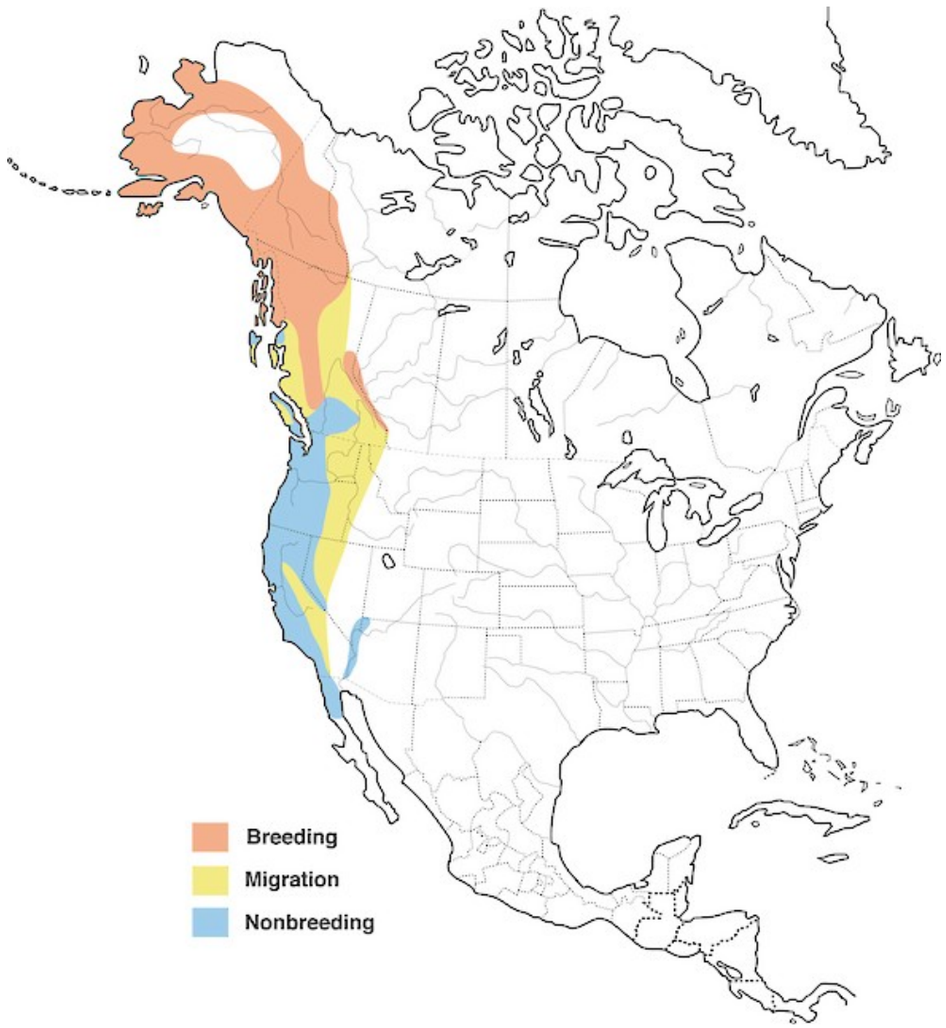
Jeffrey R Adrion ✉, Jared G Galloway, Andrew D Kern Author Notes

Molecular Biology and Evolution, msaa038, https://doi.org/10.1093/molbev/msaa038

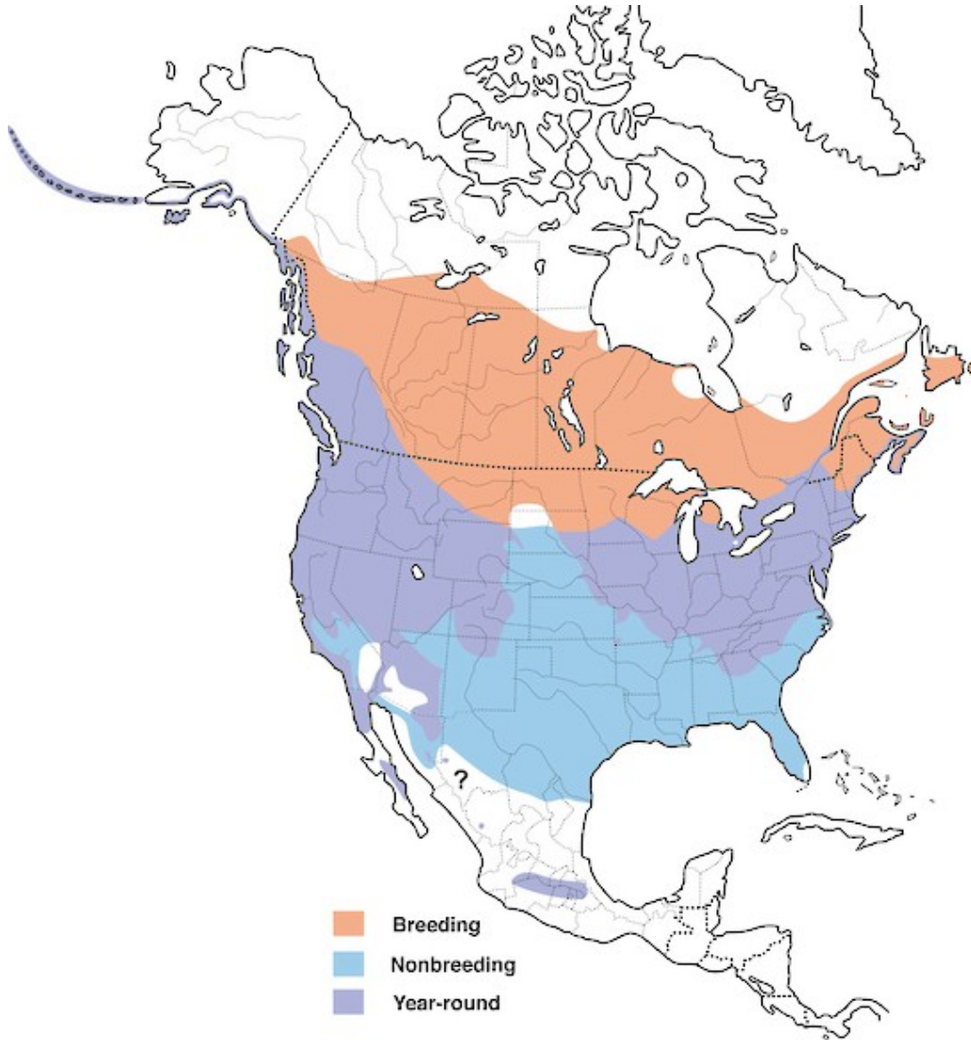
Published: 20 February 2020 **Article history** ▾

CJ Battey

Organisms live in space

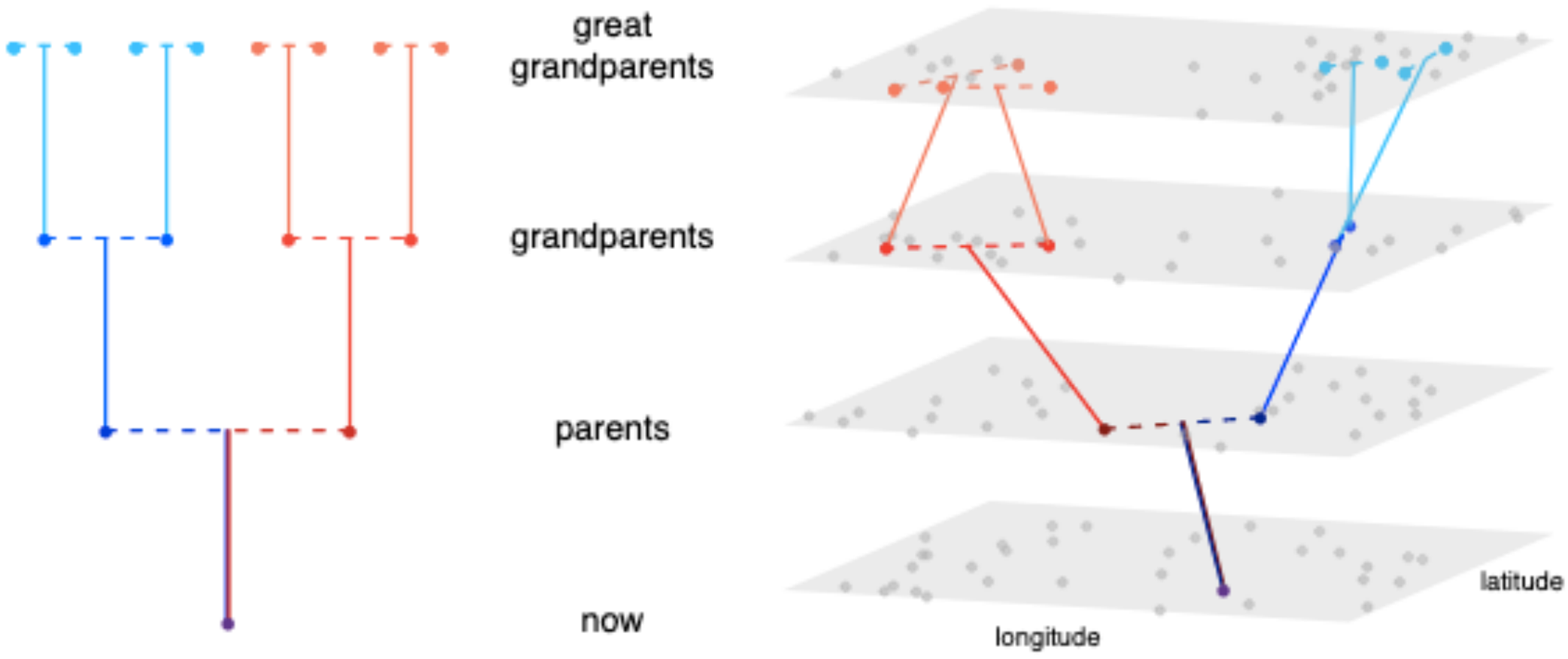


Golden crowned sparrow



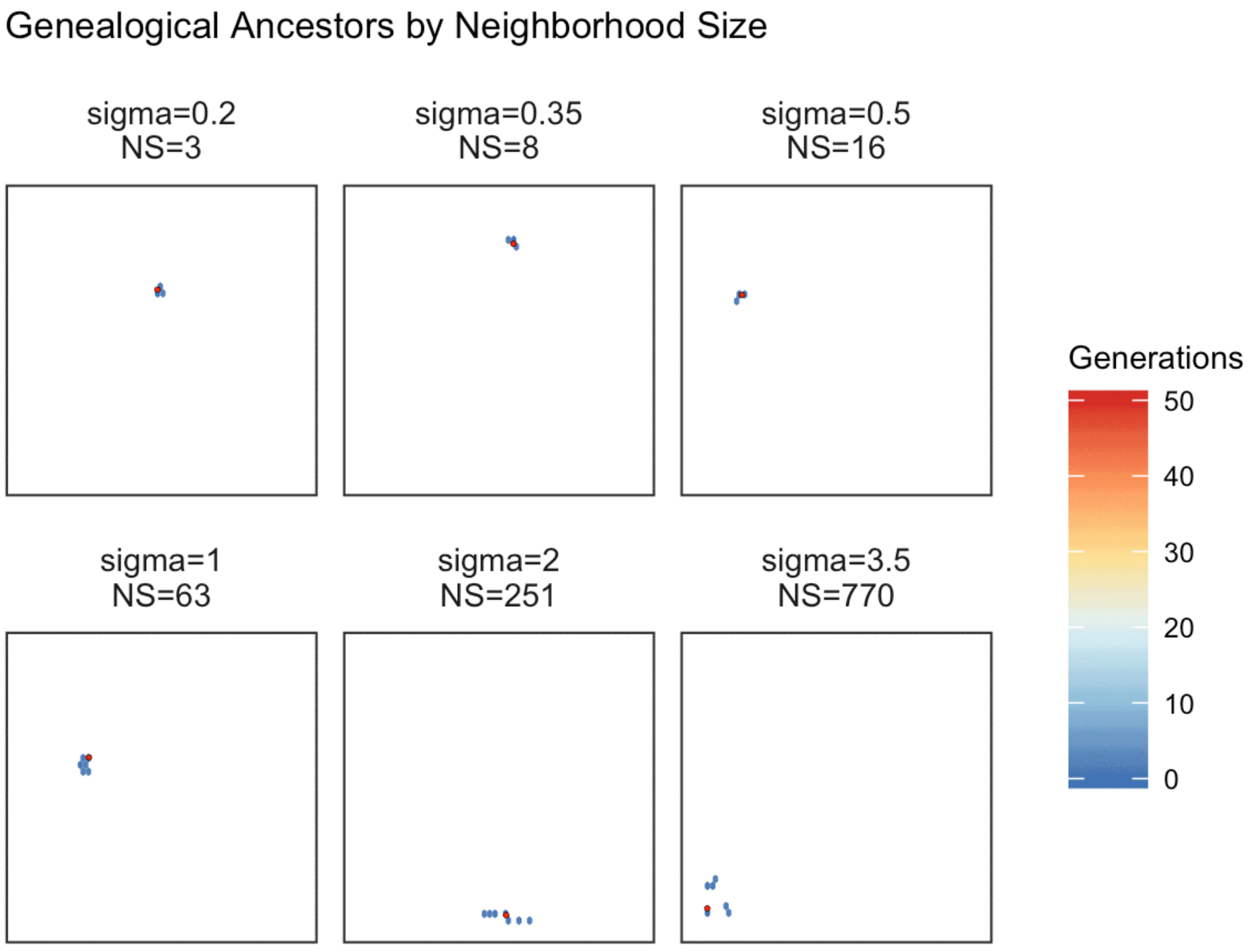
Song sparrow

Space is the Place

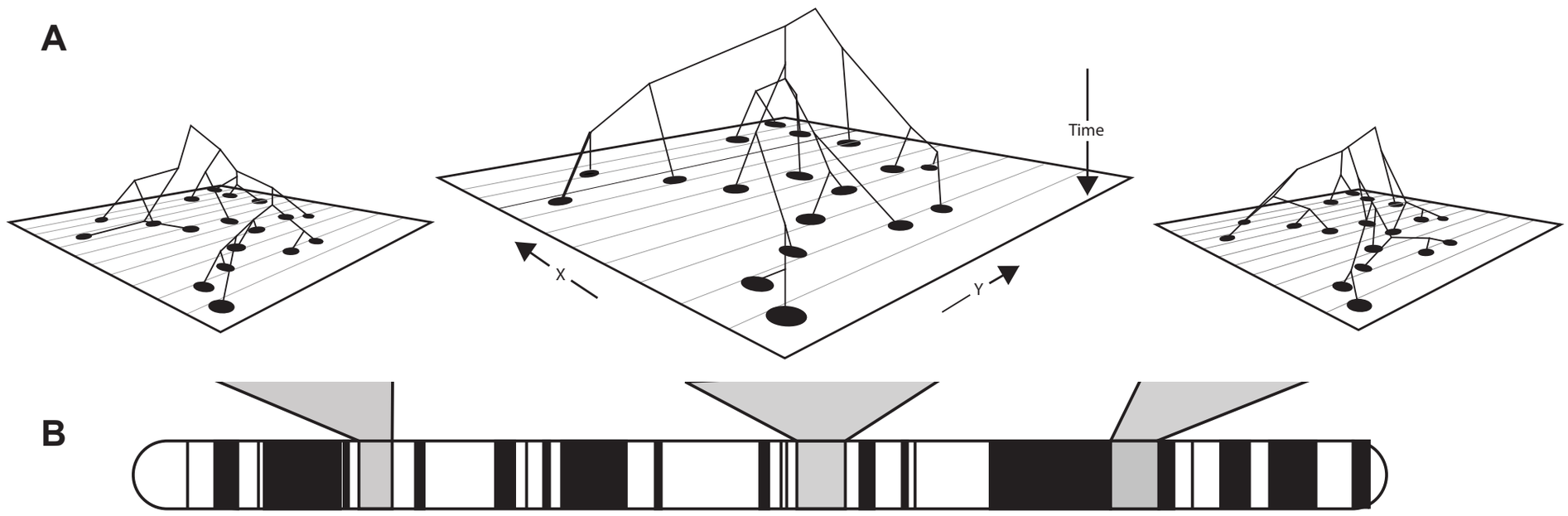


From Bradburd and Ralph (2019)

Space is the Place

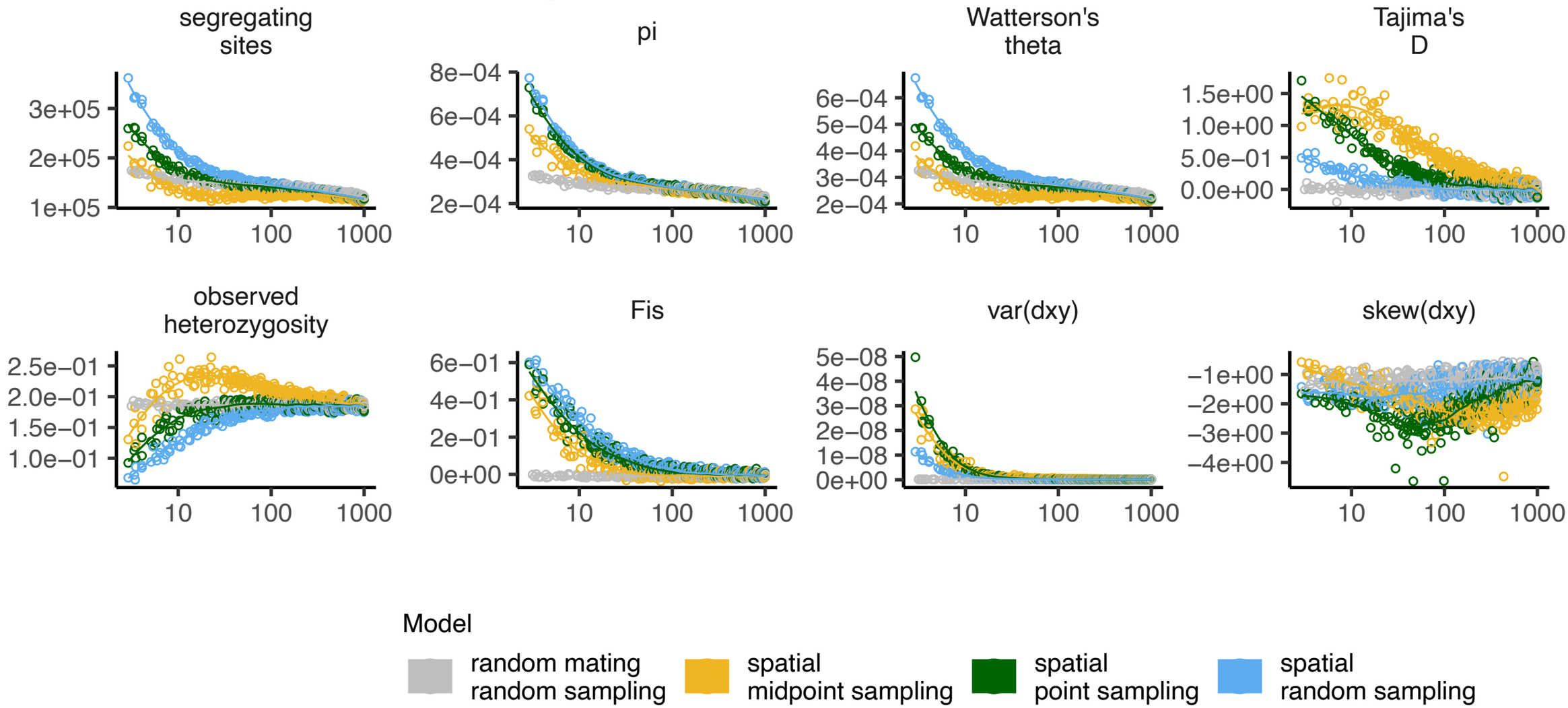


Space is the Place



Recombination means different spatial ancestry
at different points along chromosomes

Space is the Place



Space matters for genetic variation

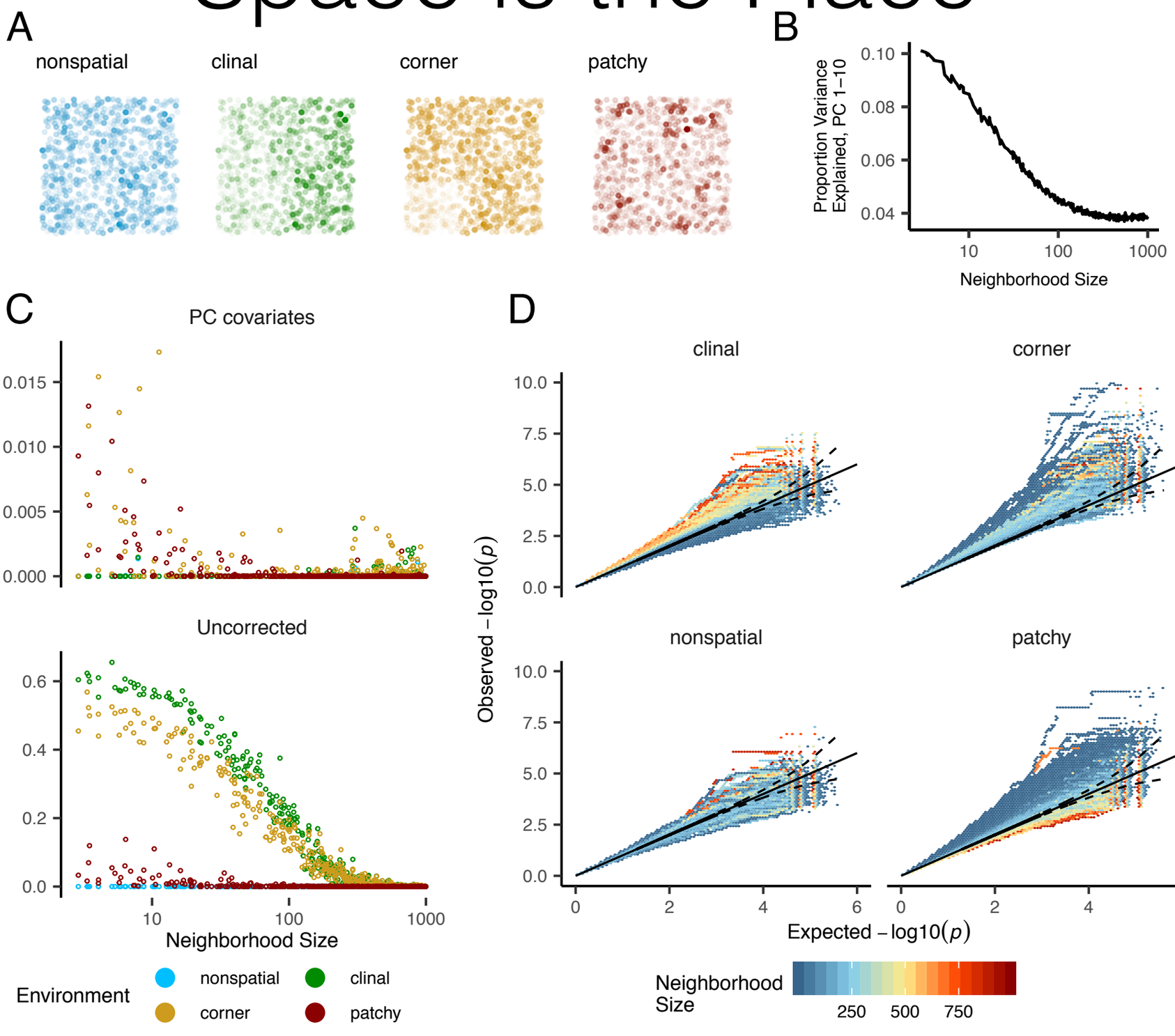
JOURNAL ARTICLE

Space is the Place: Effects of Continuous Spatial Structure on Analysis of Population Genetic Data FREE

C J Battey ✉, Peter L Ralph, Andrew D Kern

Genetics, Volume 215, Issue 1, 1 May 2020, Pages 193–214,

Space is the Place



from Battey et al 2020

Geography in our genes

POLITICS AUGUST 7, 2012

Gibson Guitars and Feds Settle in Illegal Wood Case

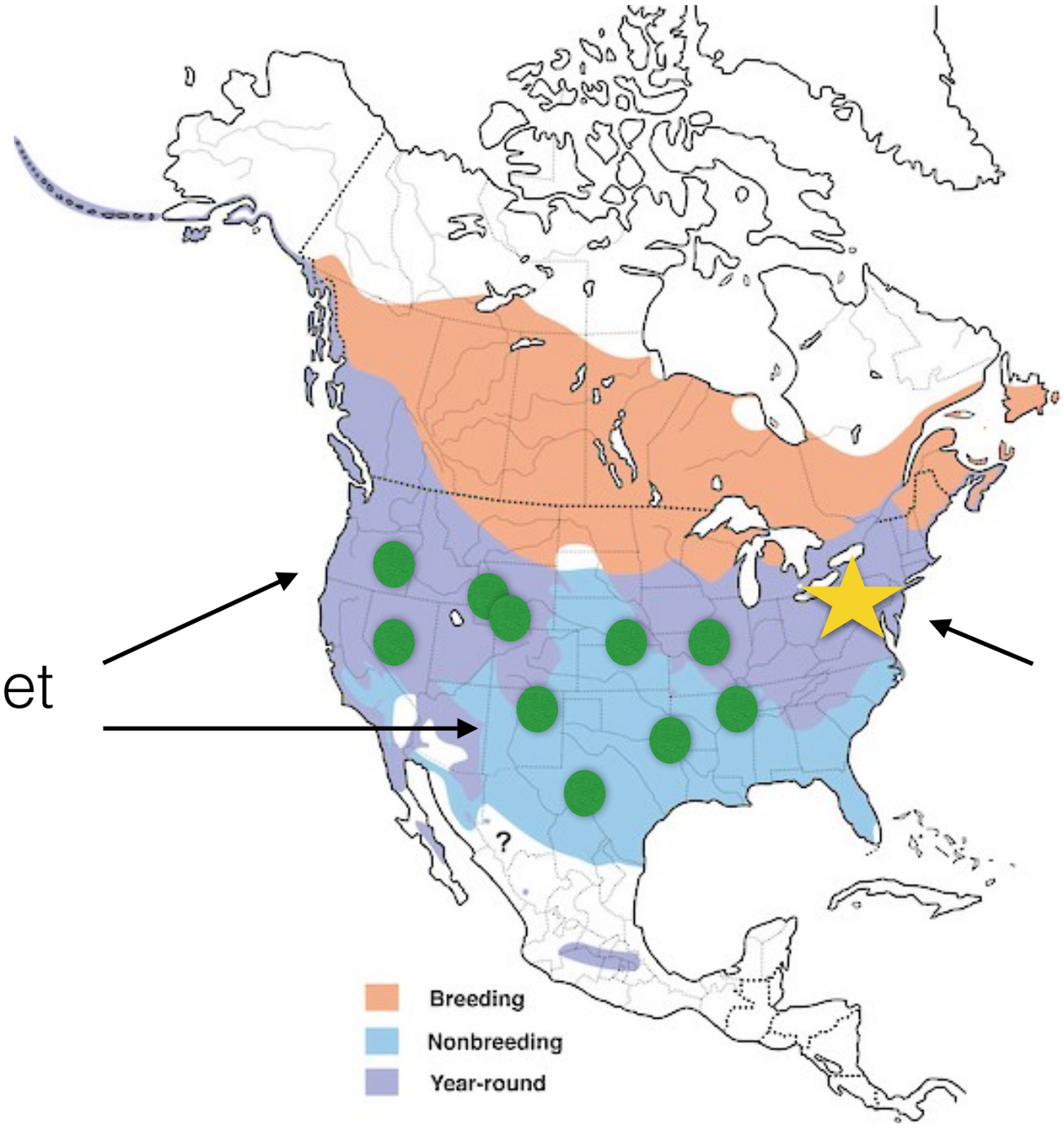
KATE SHEPPARD
Bio | Follow

Can we exploit this?

Predicting geography

Song sparrow

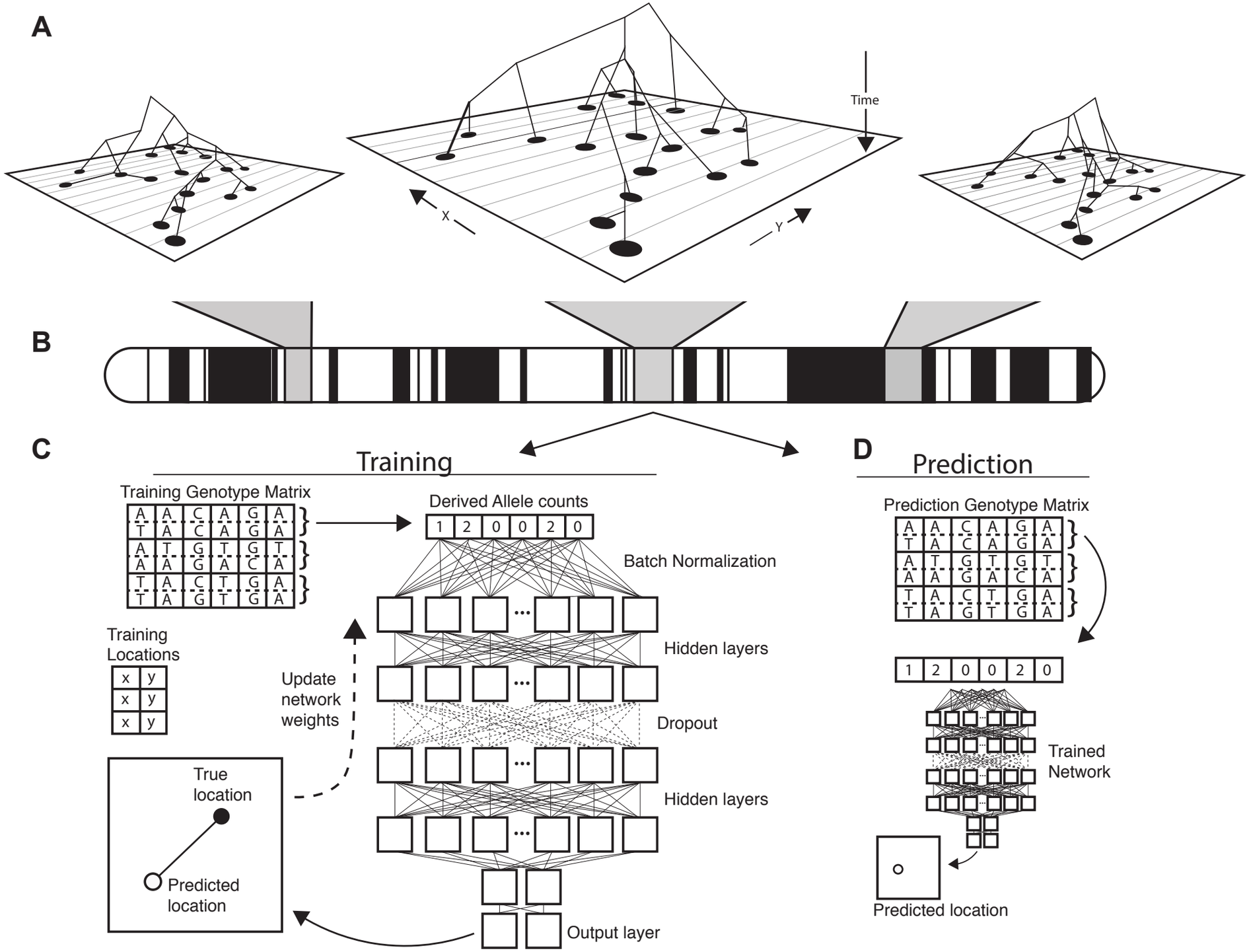
Training set



Predict locations of new genomes?

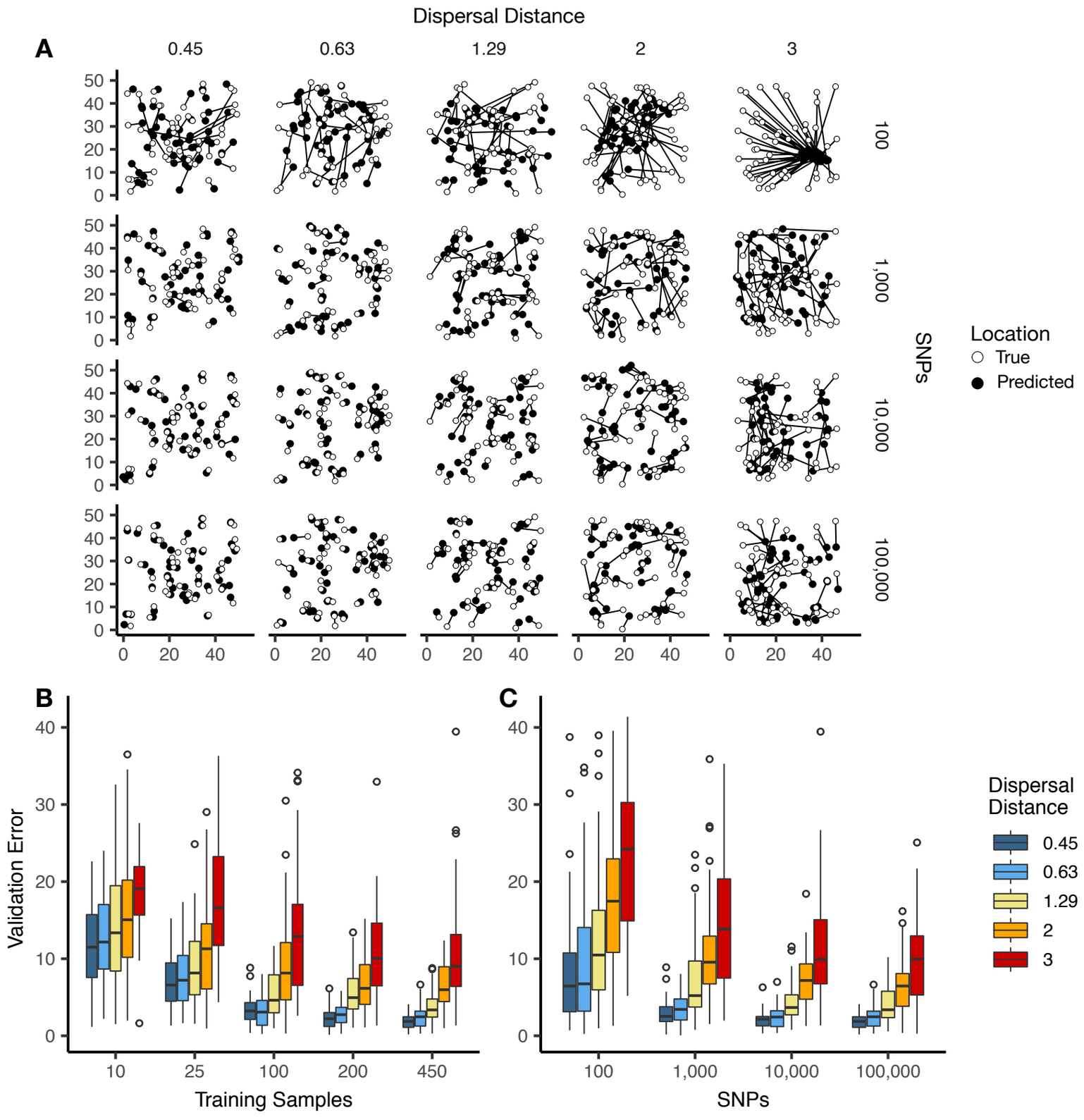
MODEL FREE

Locator— (deep) learning space

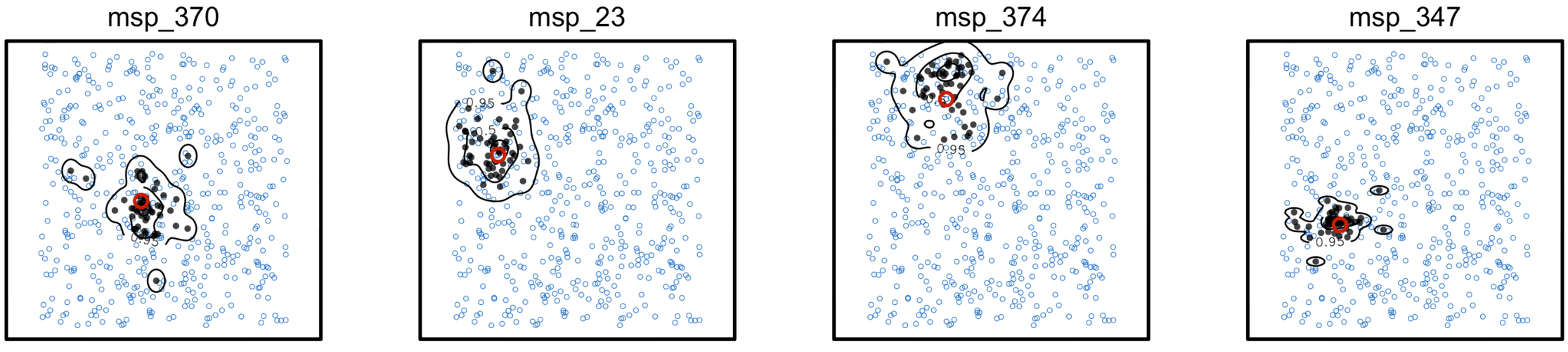


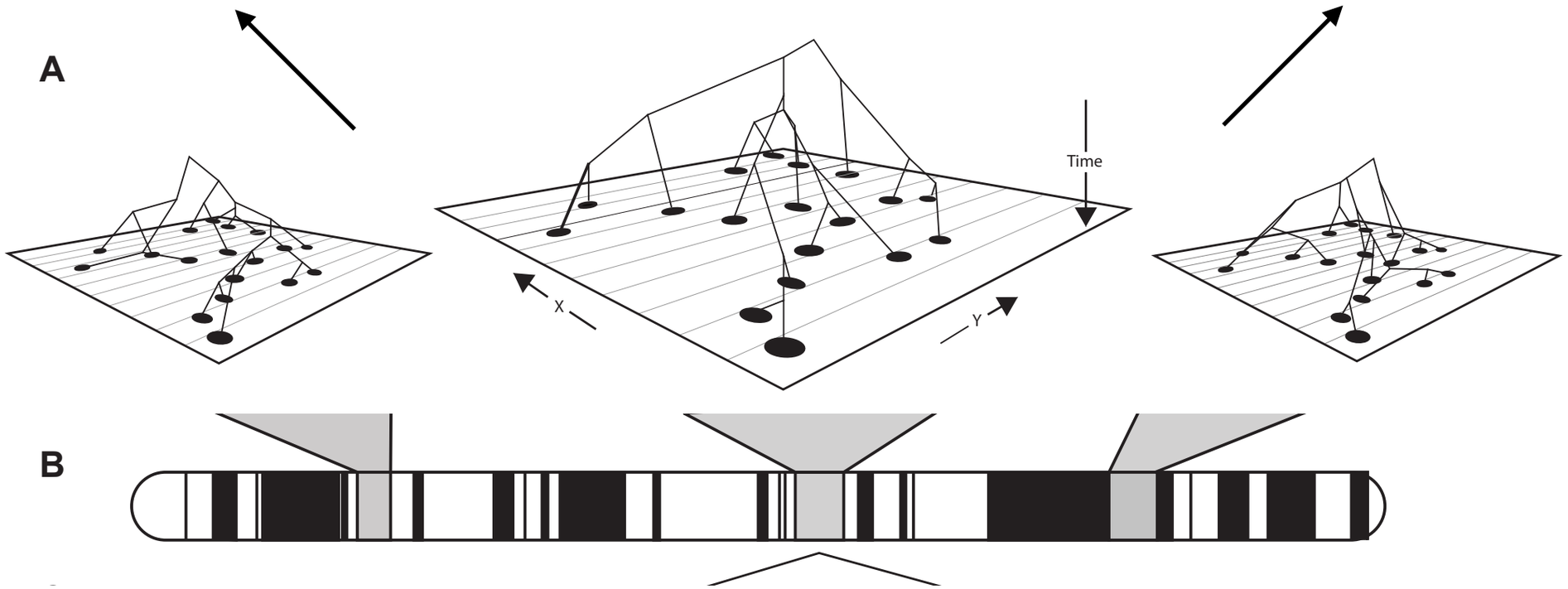
MODEL FREE

Locator— (deep) learning space



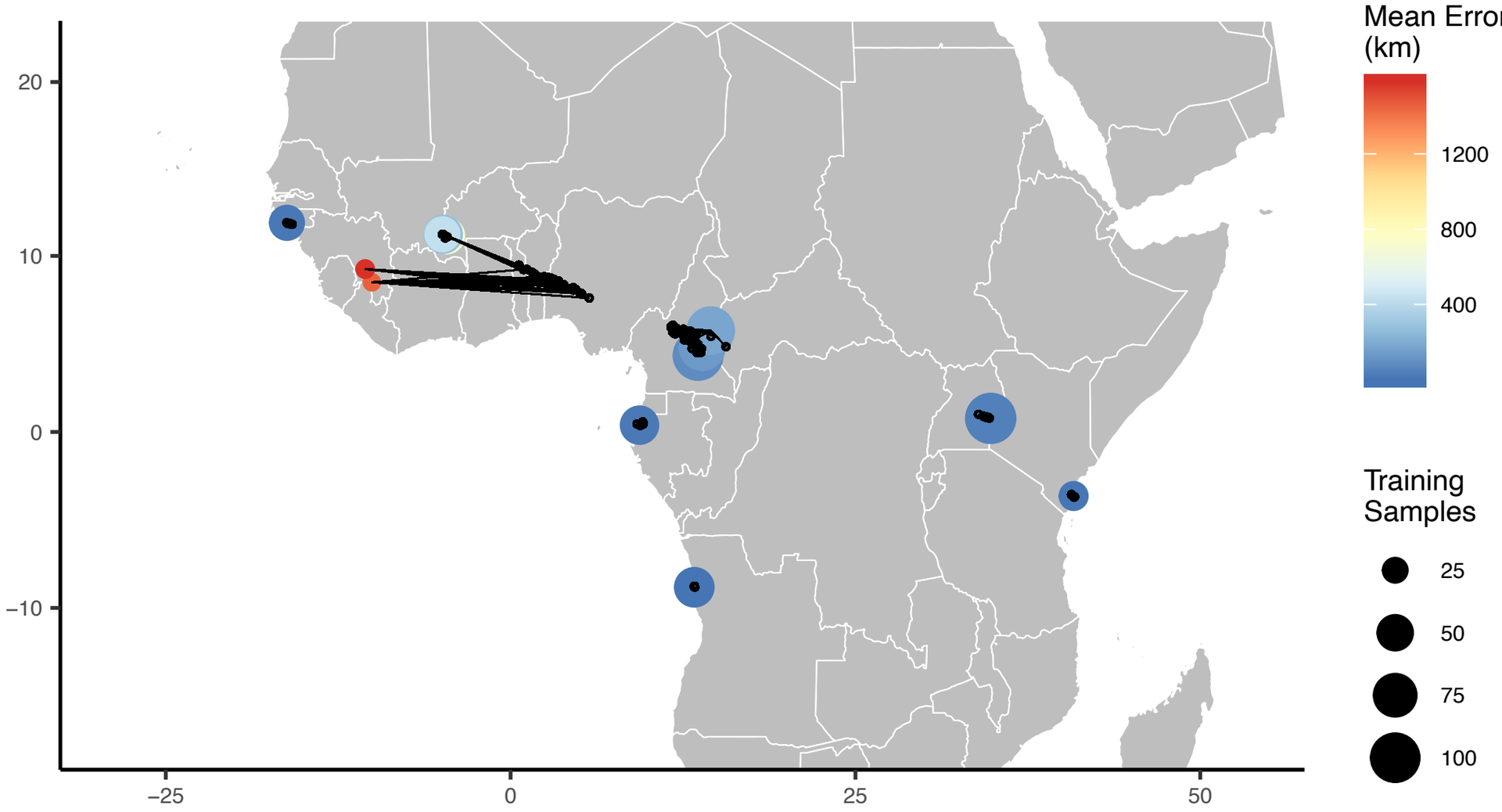
Locator— (deep) learning space





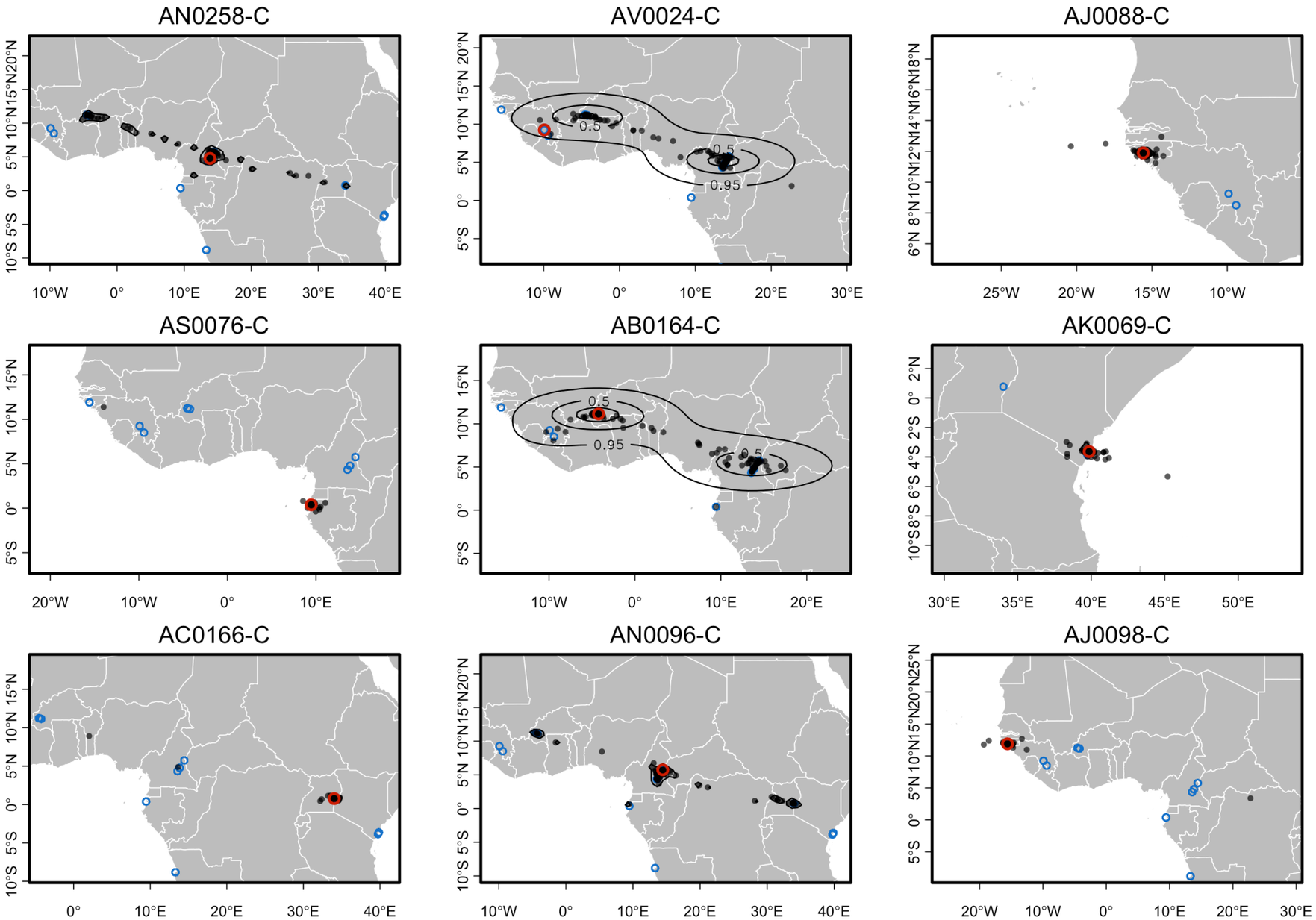
How best to assess uncertainty?
Block bootstrap?

Locator— (deep) learning space



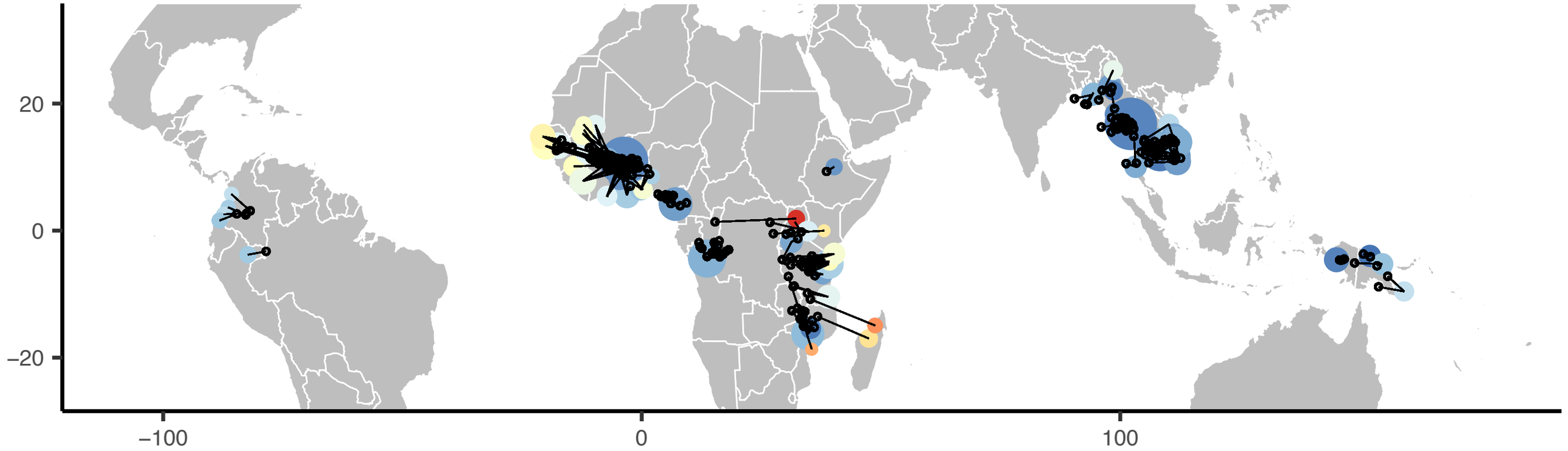
Anopheles gambiae - ag1000g data
median error = 5.7km

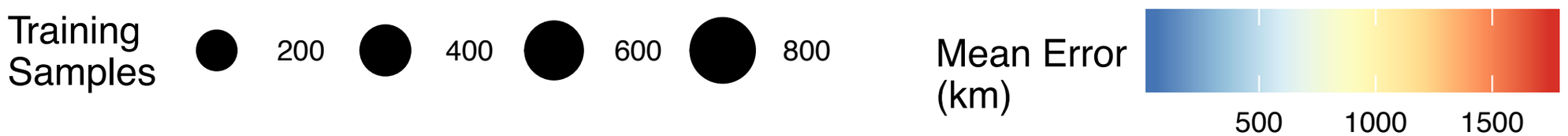
Locator— (deep) learning space



Anopheles gambiae - ag1000g data

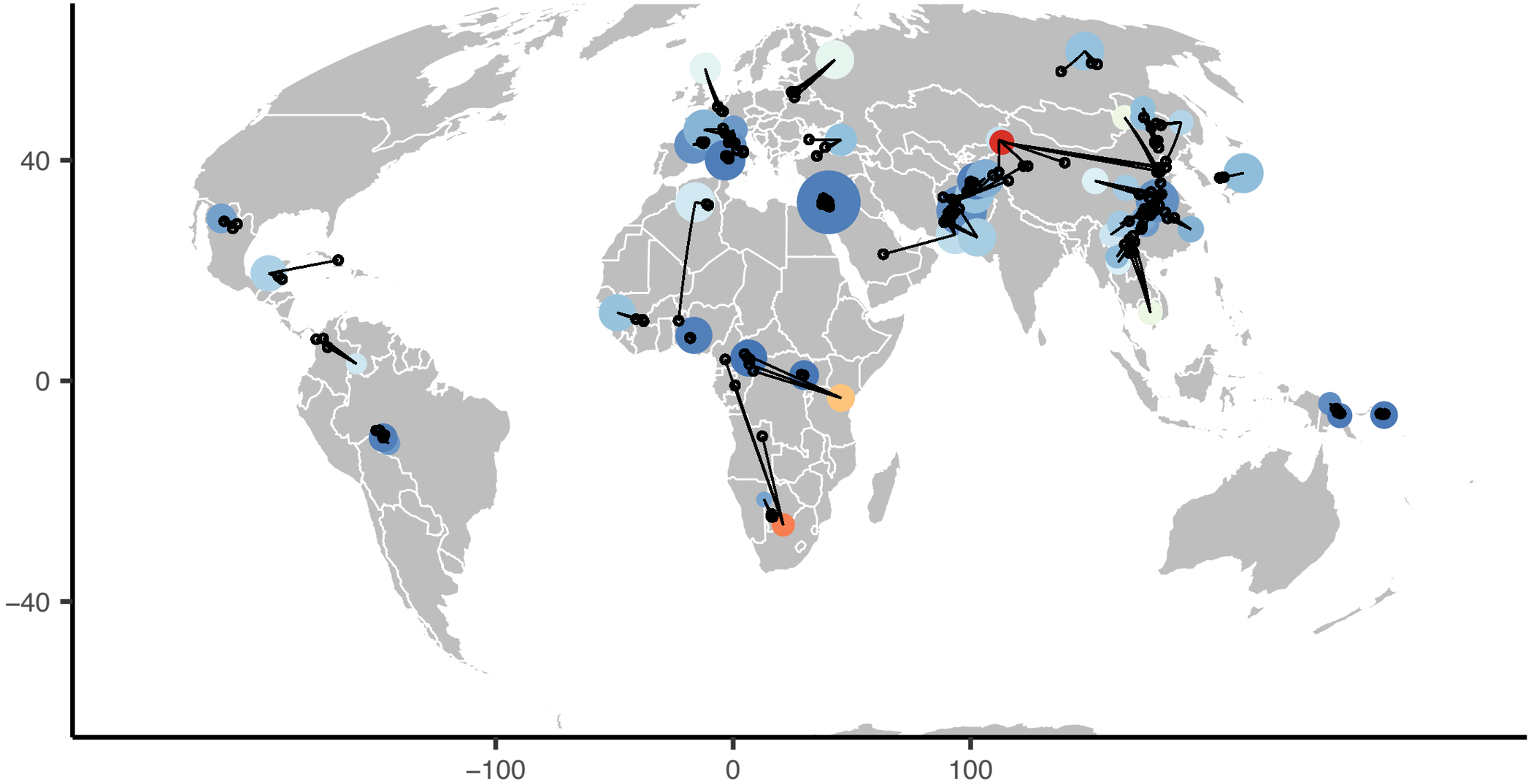
Locator— (deep) learning space





Plasmodium falciparum- Pf7K dataset
median error = 16.9 km

Locator— (deep) learning space



Training Samples 20 40 60 80

Mean Error (km) 1000 2000 3000

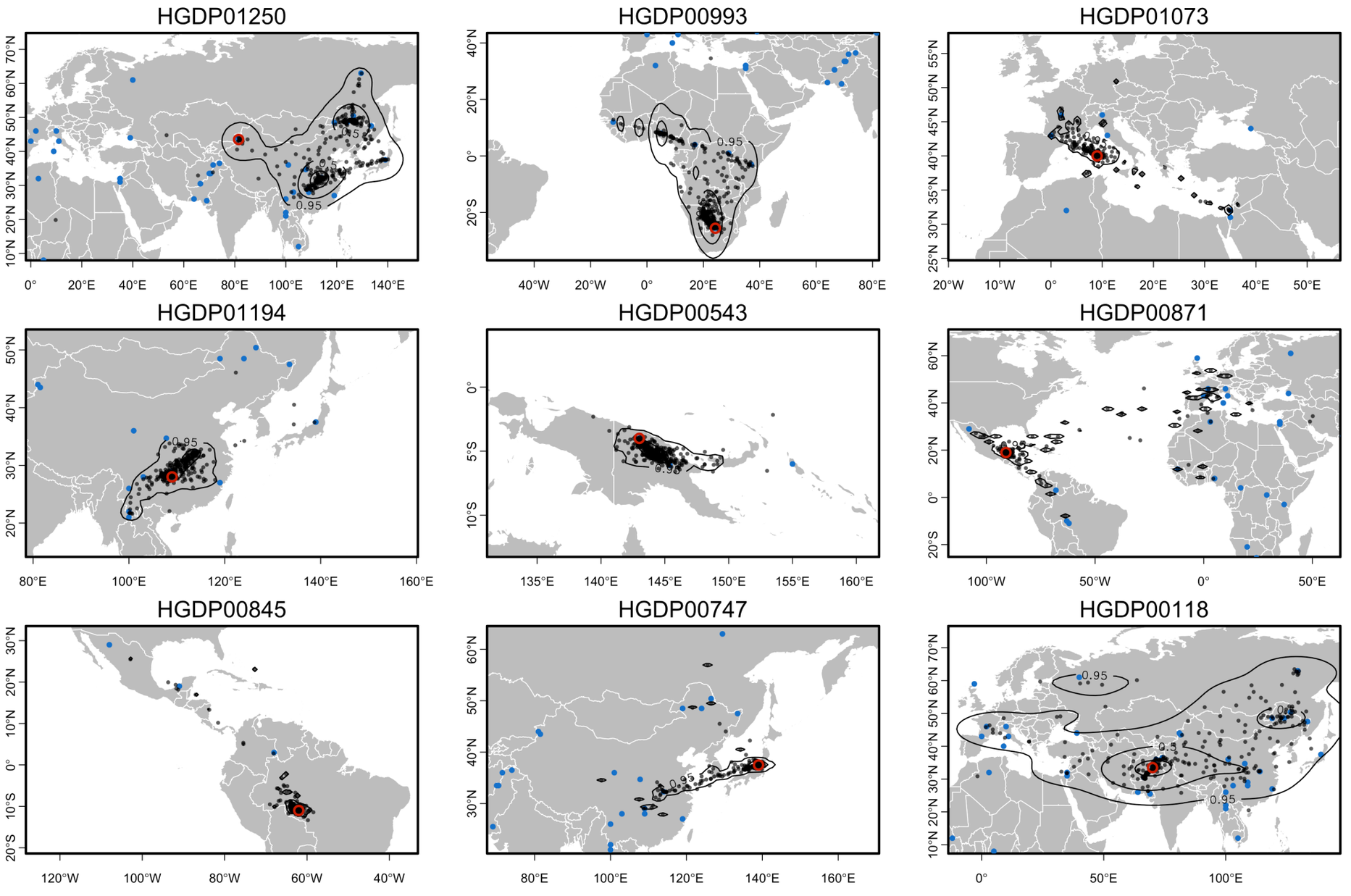
eLife
TOOLS AND RESOURCES

Predicting geographic location from genetic variation with deep neural networks

CJ Battey*, Peter L Ralph, Andrew D Kern

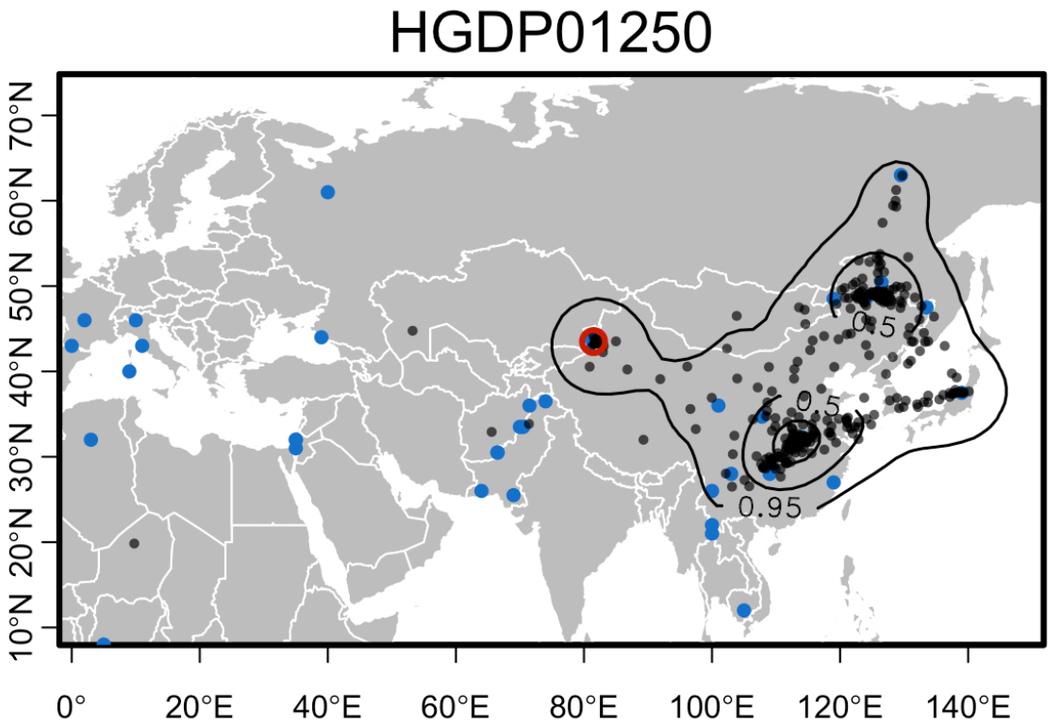
Humans - HGDP
median error = 85km

Locator— (deep) learning space



Humans - HGDP

Locator— (deep) learning space



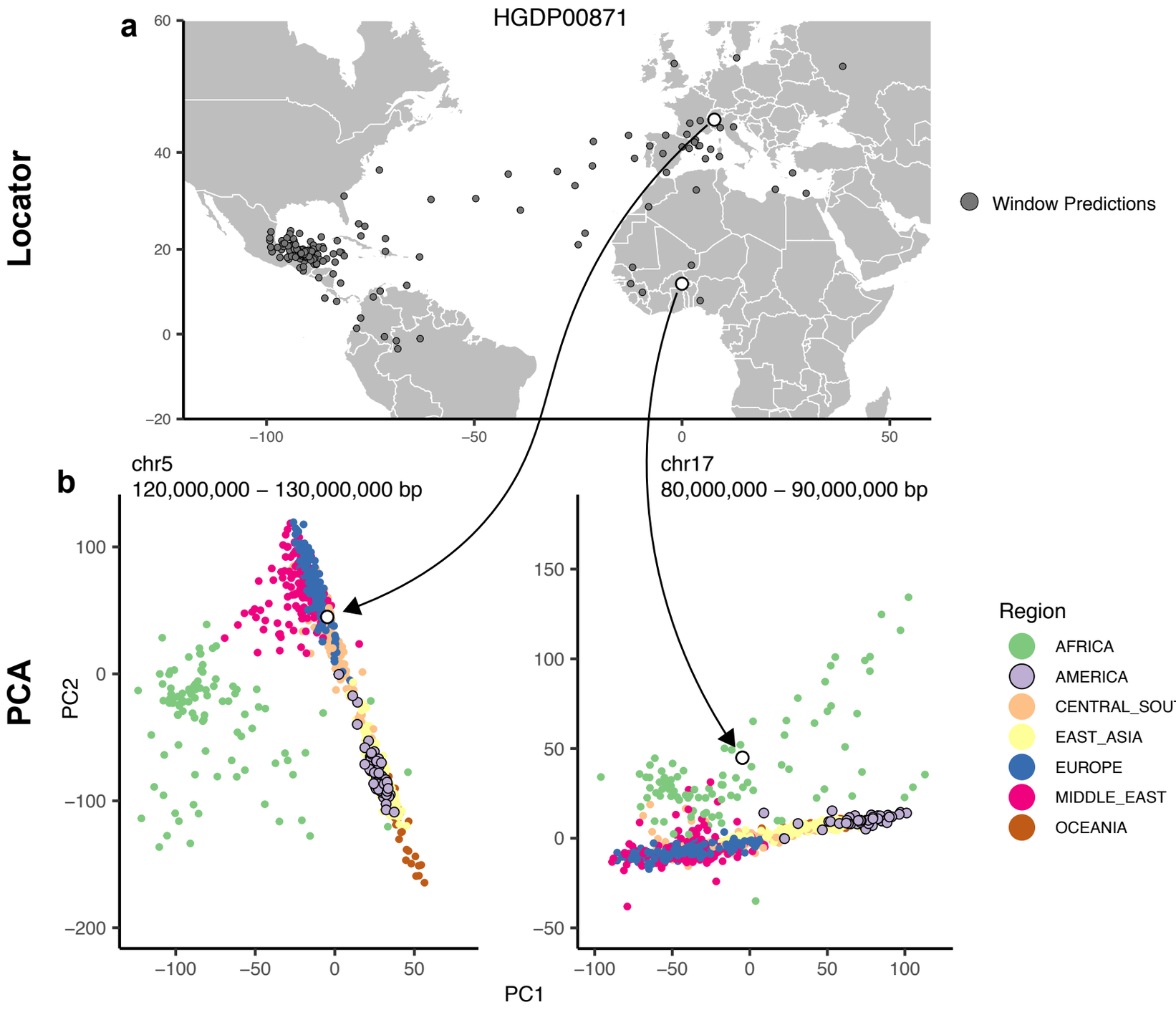
Xibe individual

Xibe

Xibe is a Tungusic language spoken in Xinjiang Uyghur Autonomous Region in north west China by about 30,000 people. It is closely related to Manchu, though the Xibe people consider themselves a separte ethnic group. The Xibe were moved to the region in 1764 by the Ch'ing emperor Qianlong. The language is also known as Sibe, Xibo or Sibo.

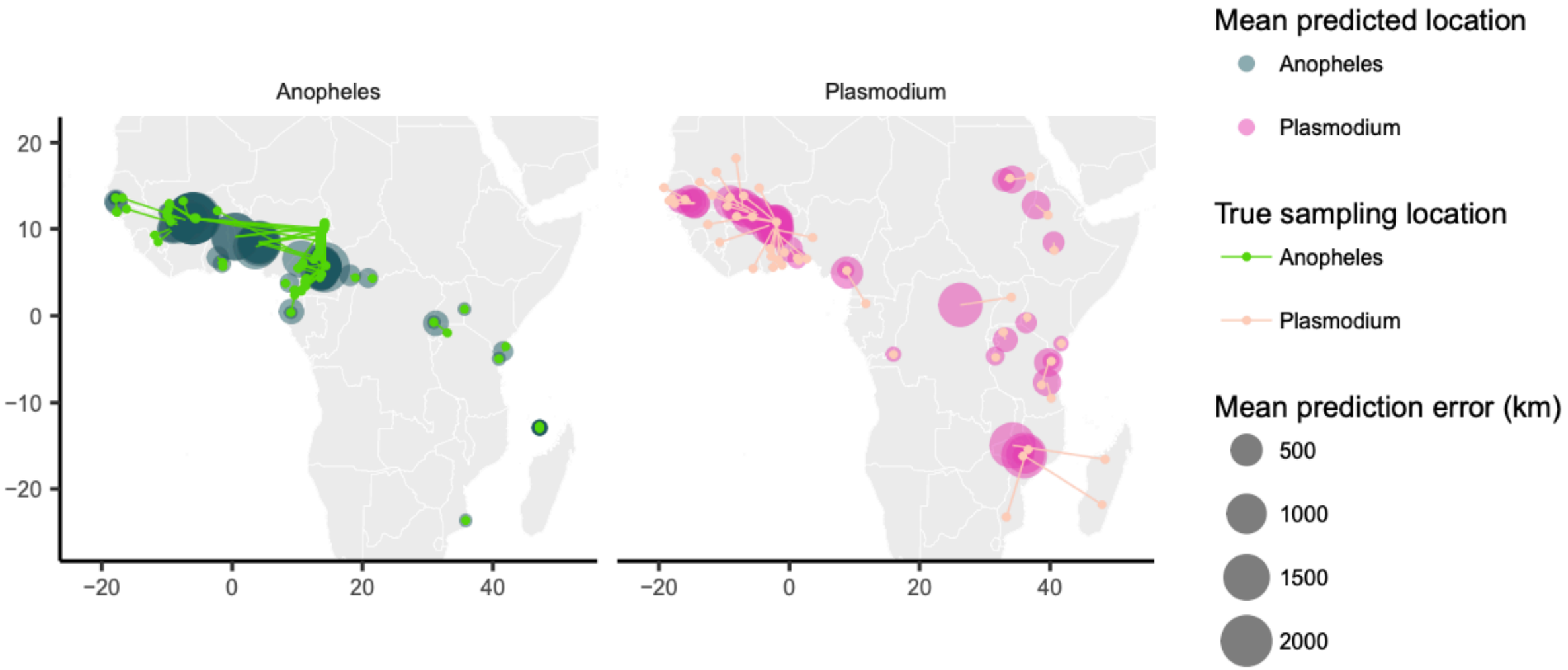
Humans - HGDP

Locator— interpretation

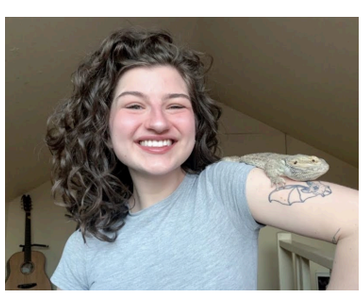


Mayan outlier windows?

Co-geography in a host-parasite system

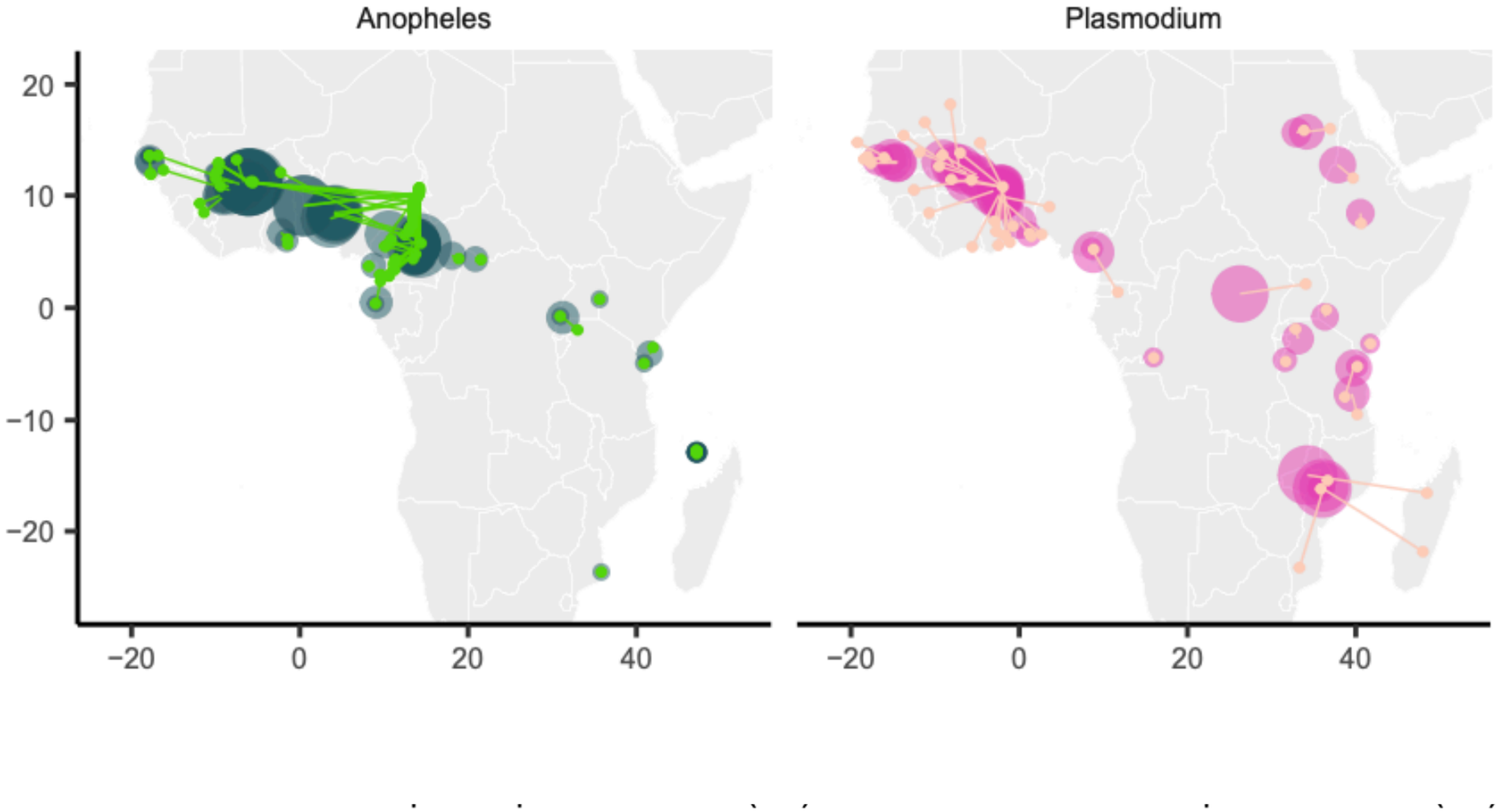


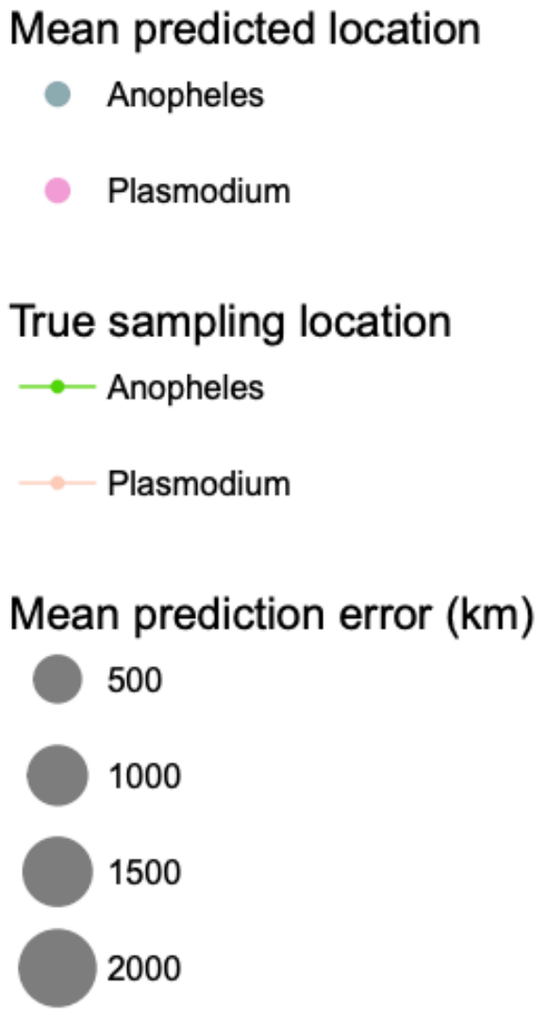
are geographies of Plasmodium & Anopheles linked?



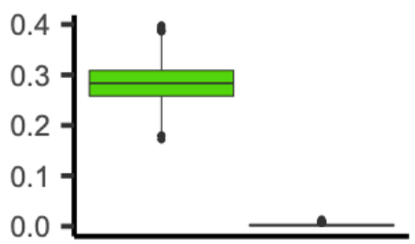
Clara Rehmann

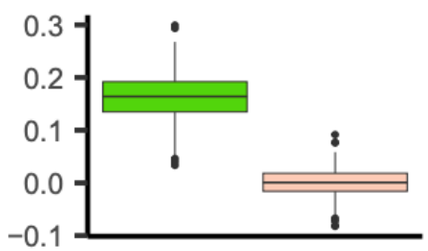
Interpreting Locator residuals

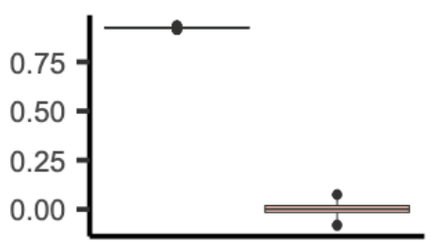


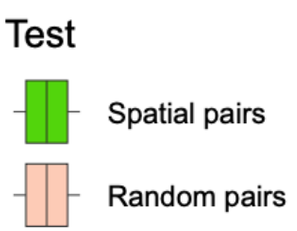


Clara Rehmann









so correlated— but is this real?

Interpreting Locator residuals

Facial recognition is accurate if you're a white dude

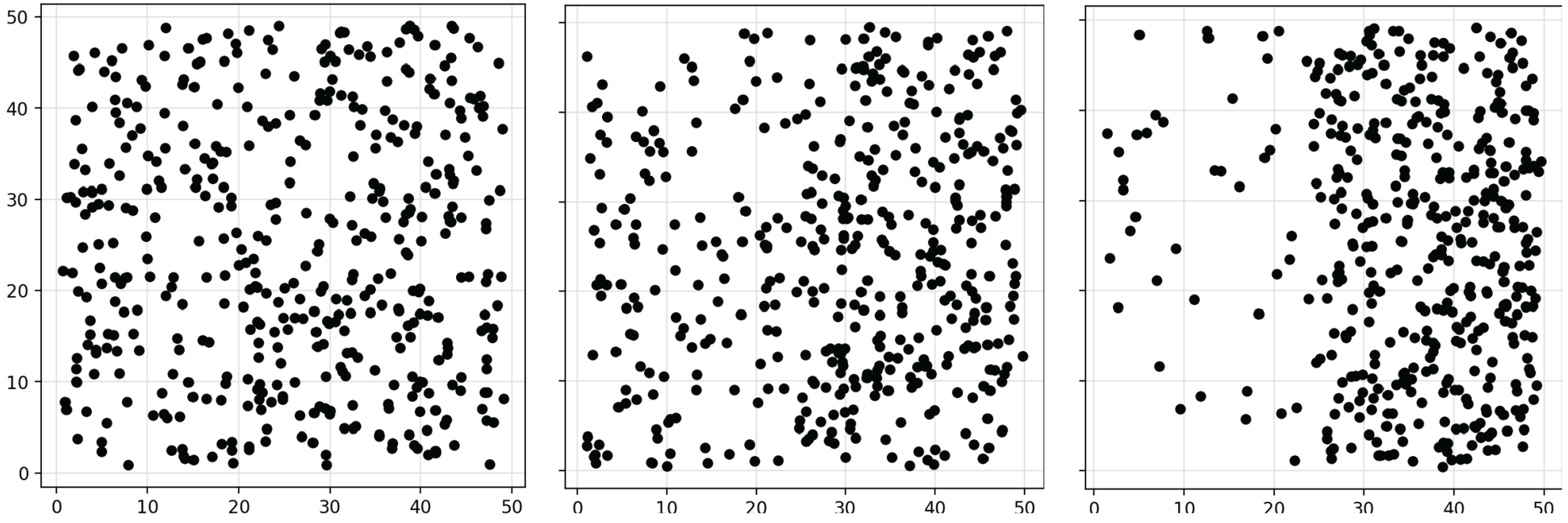
Gender was misidentified in **up to 1 percent of lighter-skinned males** in a set of 385 photos.

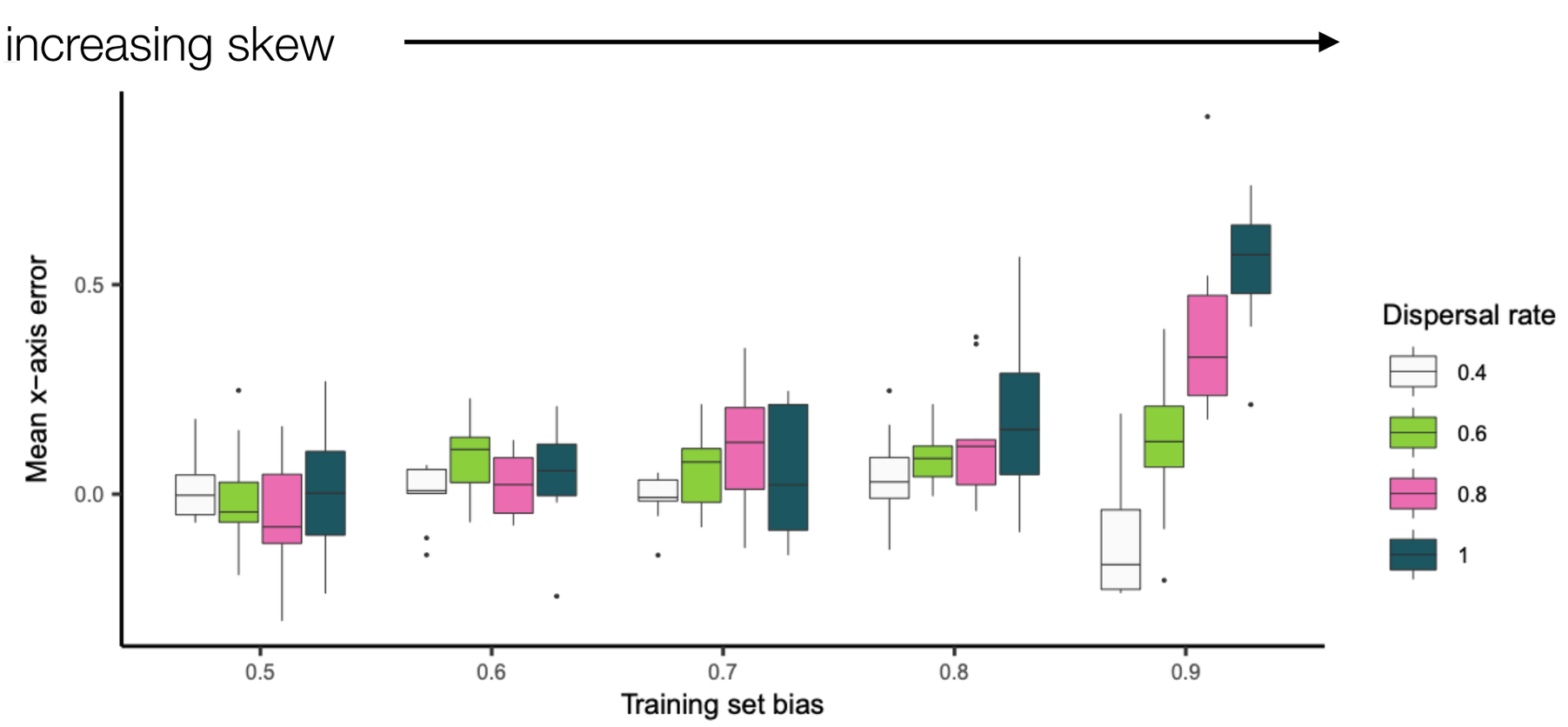
Gender was misidentified in **35 percent of darker-skinned females** in a set of 271 photos.

---

biased training set leads to biased predictions

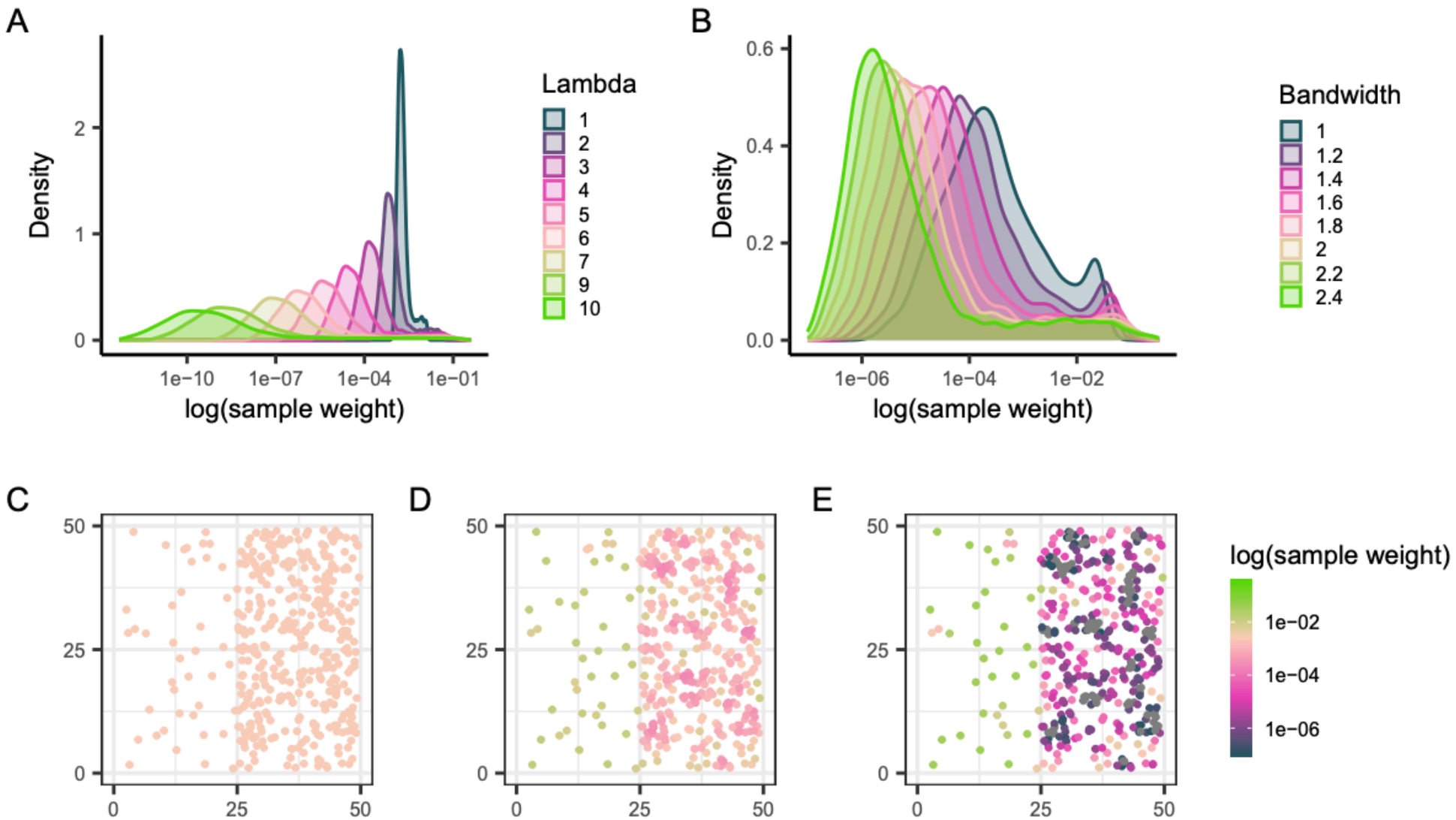
Interpreting Locator residuals





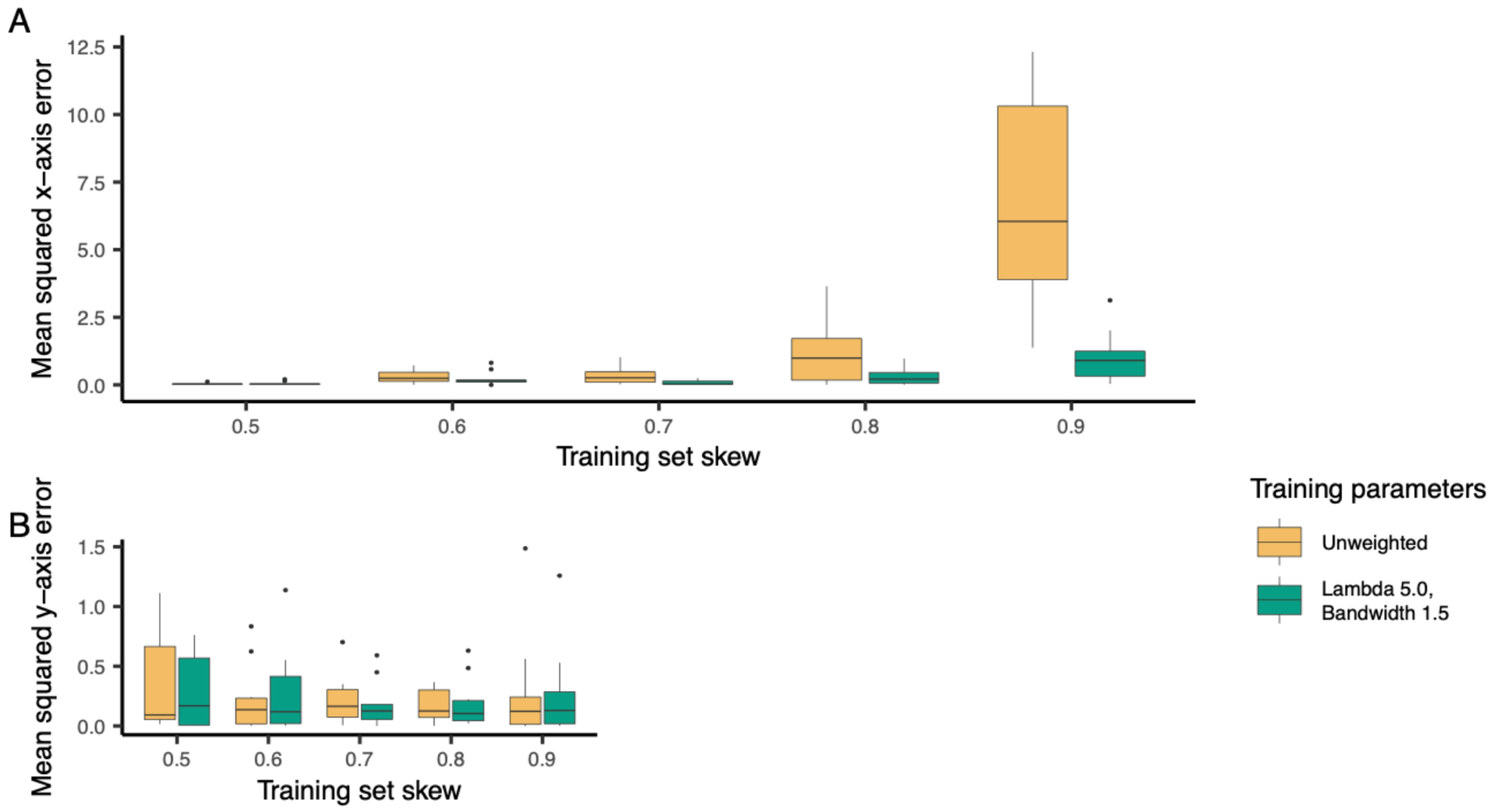
biased density of samples in training set leads to biased predictions

Interpreting Locator residuals



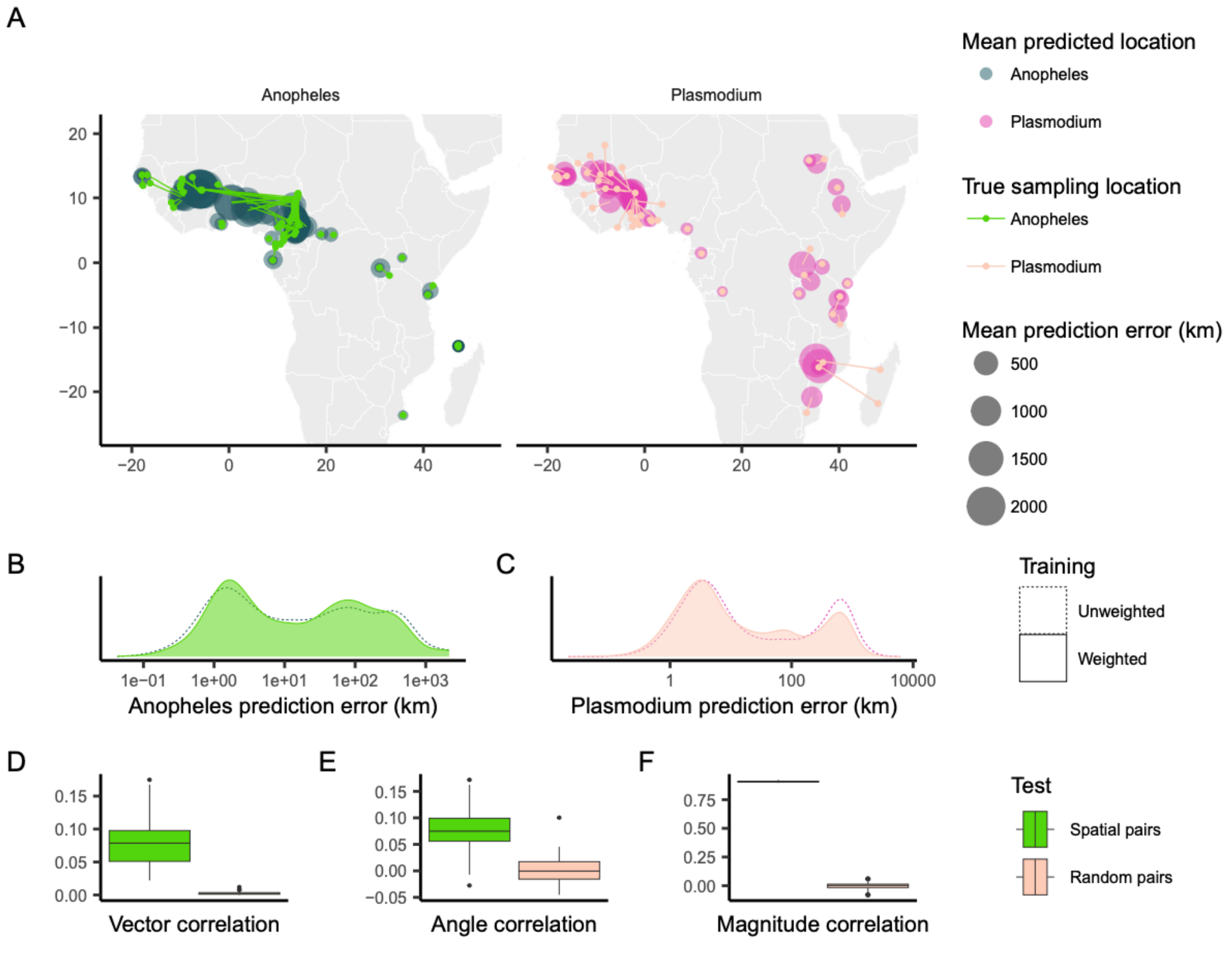
Can correct bias by weighting training loss function

Interpreting Locator residuals



Can correct bias by weighting training loss function

Co-geography in a host-parasite system



Clara Rehmann

After correction, correlation remains!

Co-geography in a host-parasite system

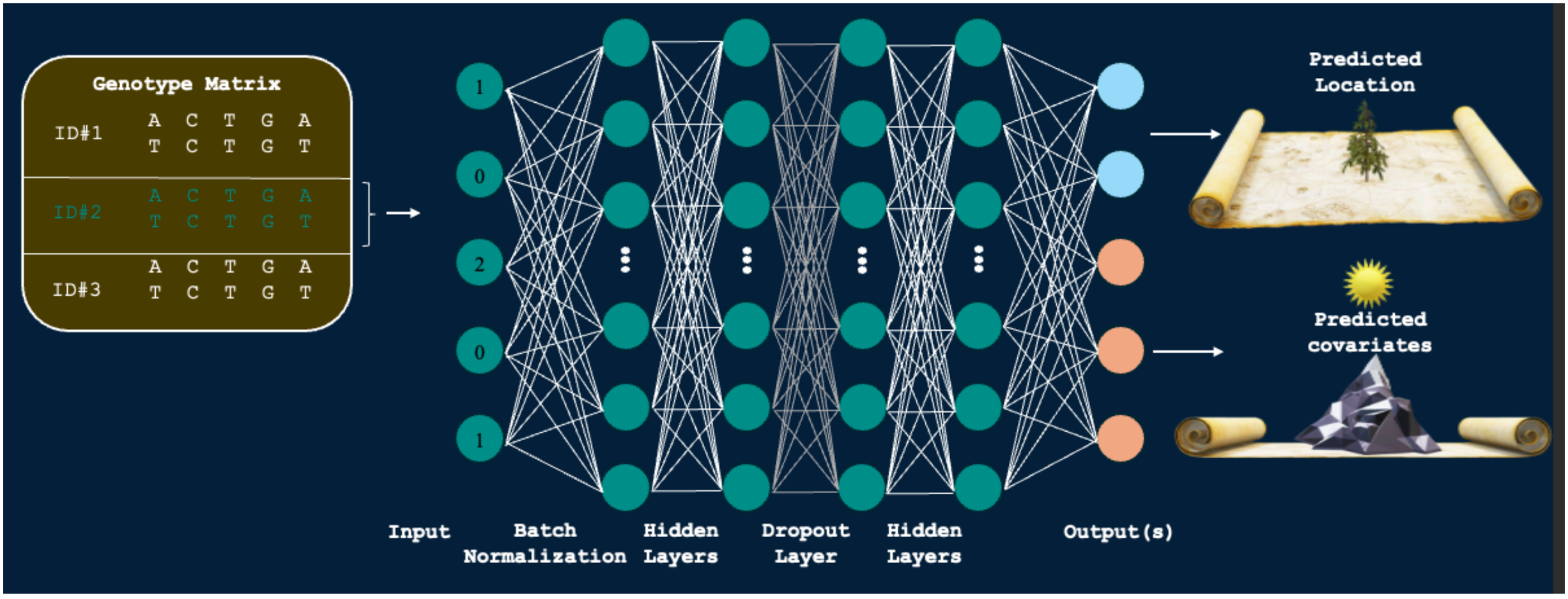
Evaluating evidence for co-geography in the *Anopheles–Plasmodium* host–parasite system

Clara T Rehmann ✉, Peter L Ralph, Andrew D Kern Author Notes

G3 Genes|Genomes|Genetics, Volume 14, Issue 3, March 2024, jkae008,
https://doi.org/10.1093/g3journal/jkae008

Clara Rehmann

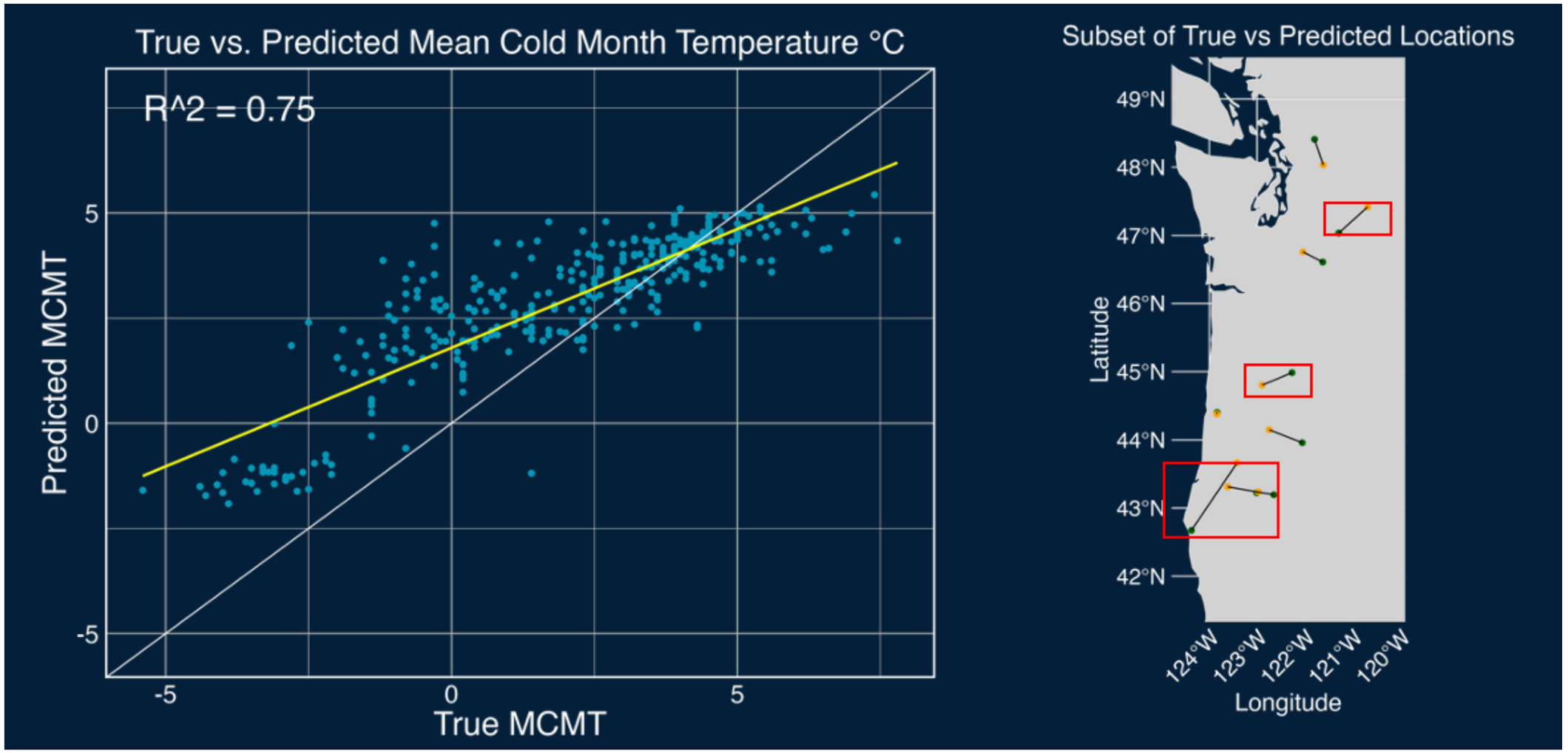
Ecolocator



Jordan Rodriguez

Extending the locator model to local adaptation

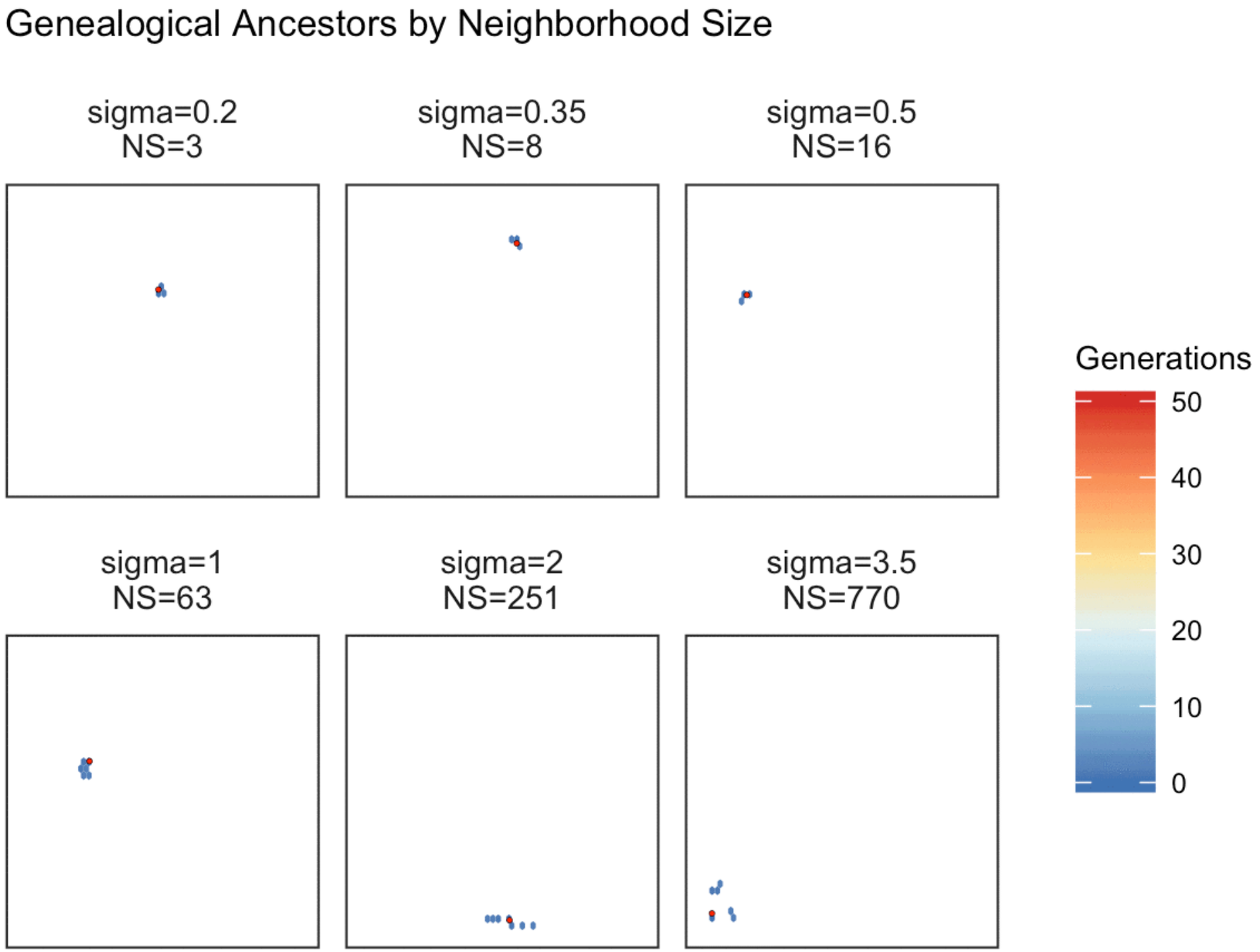
Ecolocator



Jordan Rodriguez

Extending the locator model to local adaptation

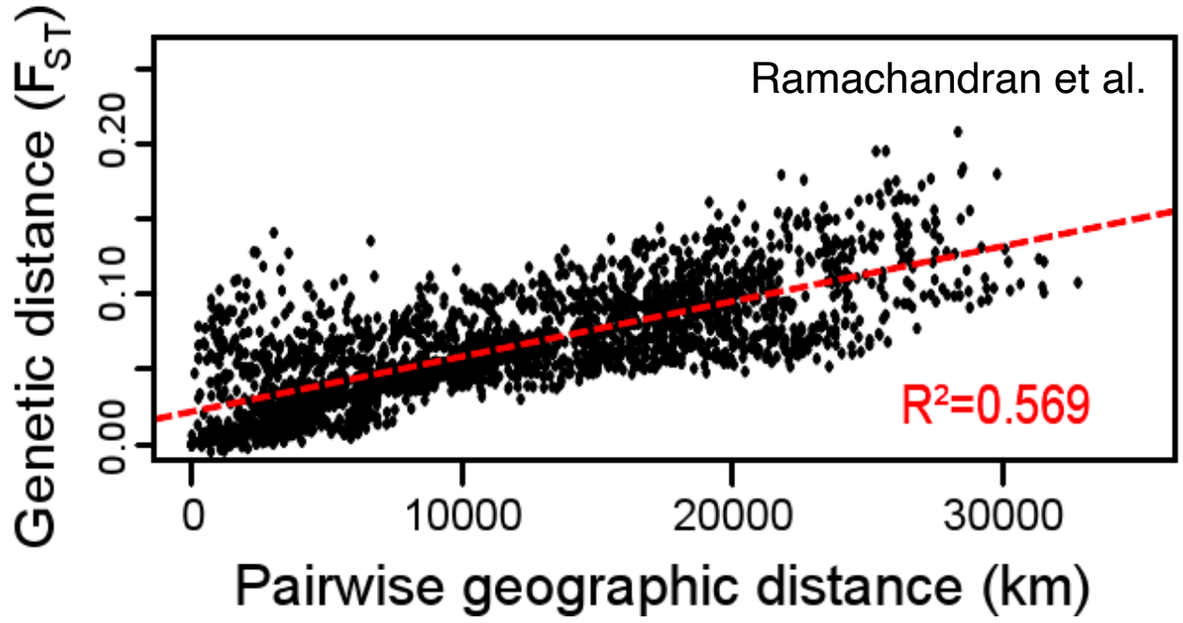
Estimating σ



Chris Smith

Estimating σ

Isolation by distance



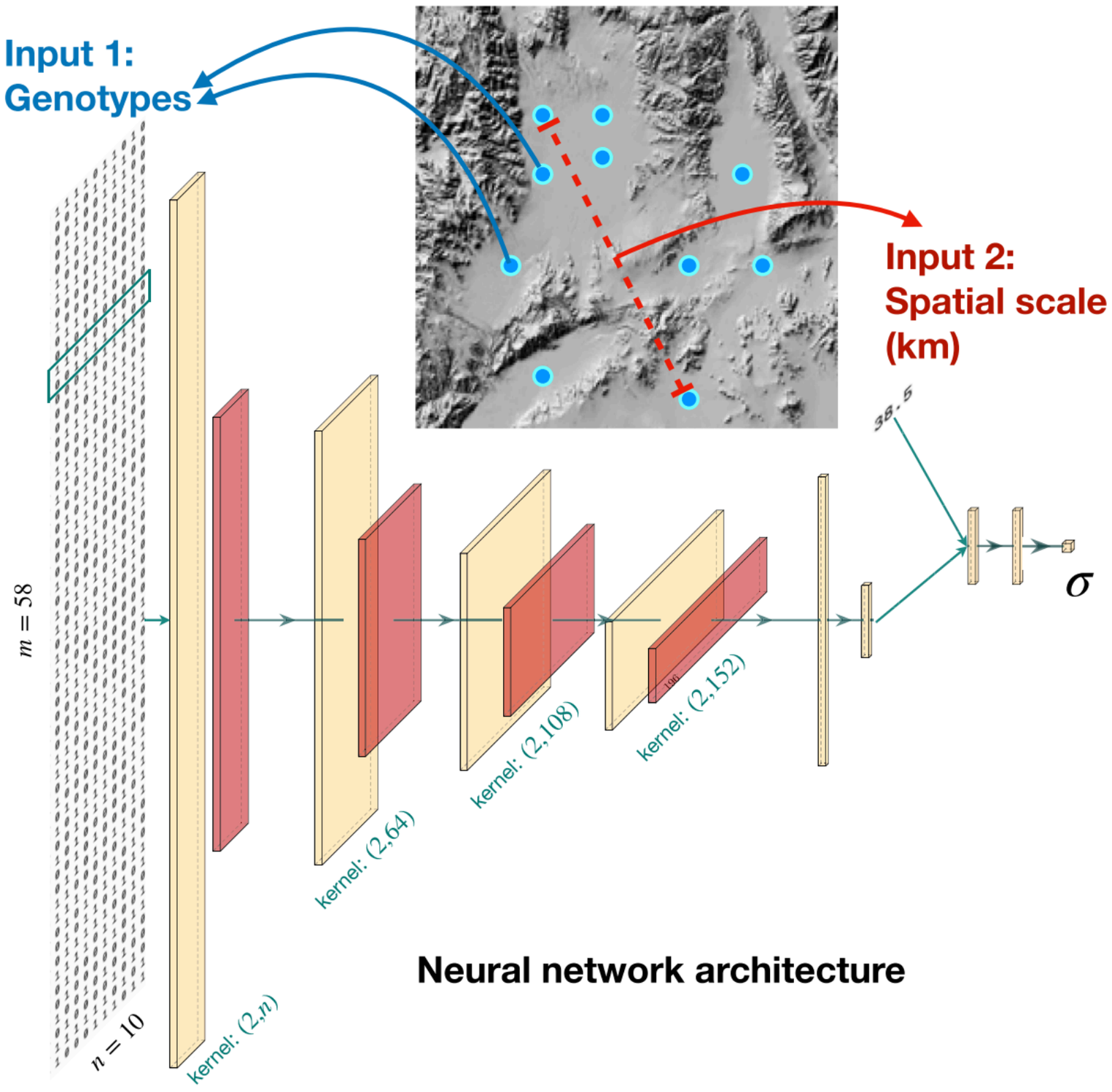
Classic result

Rousset's method— fit this regression, slope is approx $\frac{1}{4N\pi\sigma^2}$

BUT

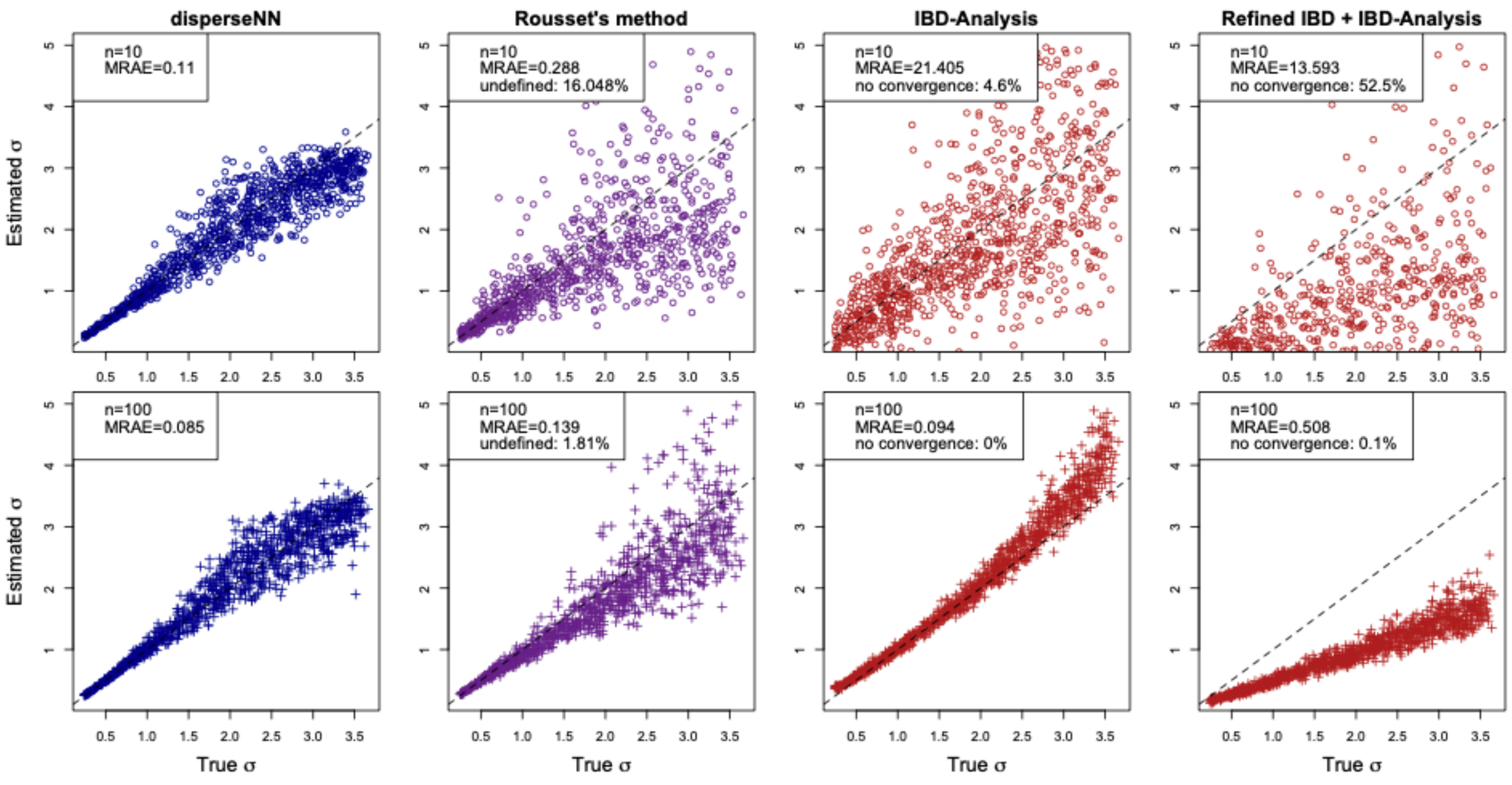
need to know local N!

disperseNN



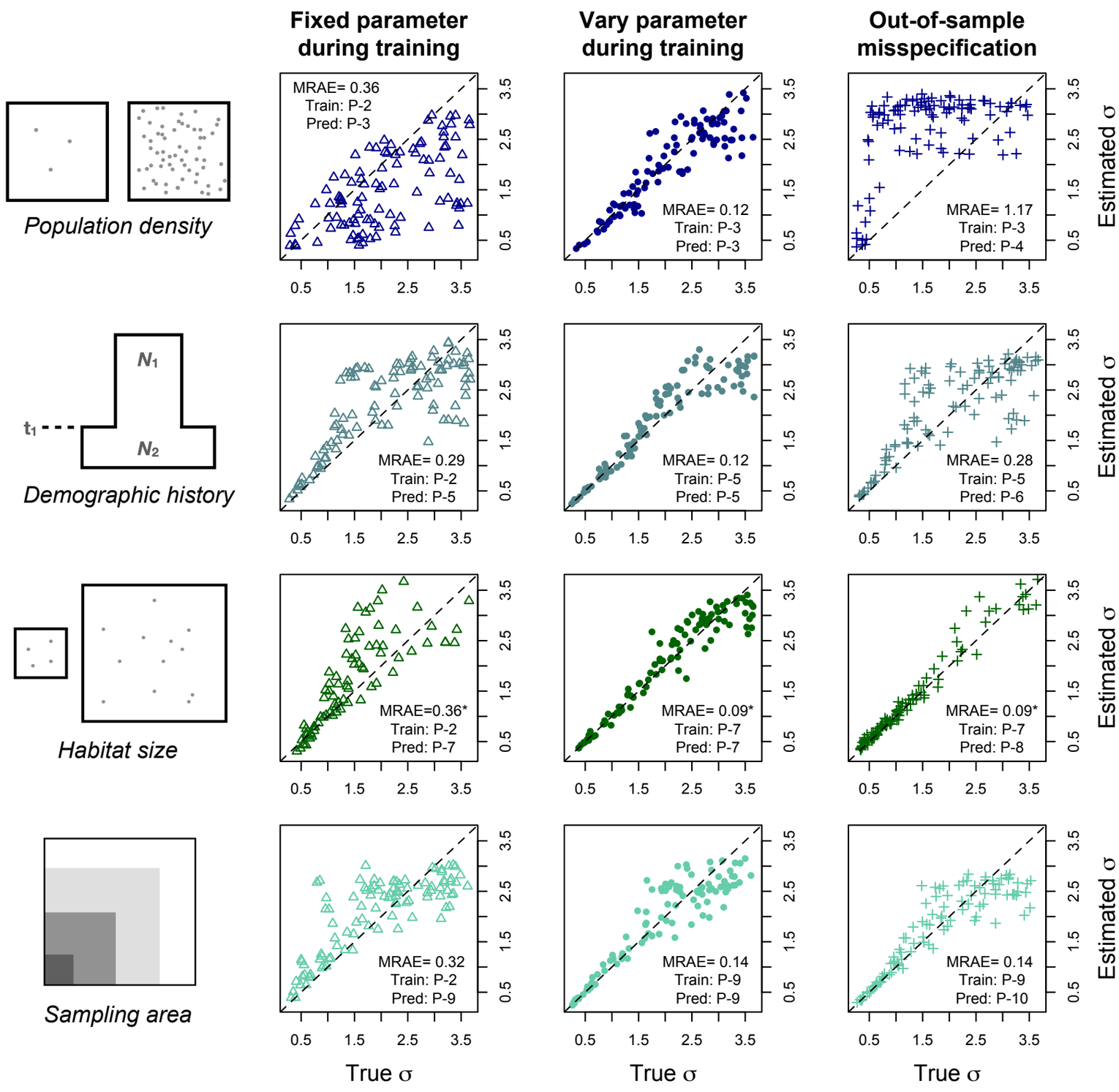
Neural network architecture

disperseNN



disperseNN works really well, particularly at small sample size
(assume perfect knowledge of N or perfect IBD tract for other me)

disperseNN



disperseNN sensitive to misspecification but can train our way out of it (mostly)

disperseNN

Species	Common name	Region	σ (km)	95% CI (km)	Previous (km)	N_{loc}	n	S (km)	M. dist.
Zosterops borbonicus	Réunion grey white-eye	Réunion	4.97	(1.76, 13.83)	NA	295	41	62	4.59
Peromyscus leucopus	white-footed mouse	New York	0.77	(0.32, 1.67)	0.03-0.11	-231	12	38	8.15
Anopheles gambiae	African malaria mosquito	Cameroon	10.29	(2.00, 48.03)	0.04-0.5	52	29	278	9.62
Bombus bifarius	two-form bumble bee	Washington	14.75	(5.60, 37.28)	1.2-5	1,147	14	273	10.47
Bombus vosnesenskii	yellow-faced bumble bee	California	7.70	(1.21, 38.11)	1.2-5	3,944	18	169	11.83
Hippoglossus hippoglossus	Atlantic halibut	Canada	4.29	(0.71, 33.85)	NA	-5,546	11	193	14.59
Crassostrea virginica	eastern oyster	Canada	1.52	(0.72, 4.31)	21.9	1,435	13	187	19.69
Canis lupus	grey wolf	N. America	15.68	(2.36, 107.3)	98-147	35	13	721	25.42
Helianthus petiolaris	prairie sunflower	Kansas	1.00	(0.39, 3.52)	0.156	9	11	204	45.28
Zosterops olivaceus	Réunion olive white-eye	Réunion	1.05	(0.27, 4.36)	NA	2,392	10	50	45.97
Helianthus argophyllus	silverleaf sunflower	Texas	1.04	(0.38, 4.08)	0.156	57	30	307	86.49
Arabidopsis thaliana	thale cress	Spain	1.36	(0.28, 5.05)	0.001	35	35	80	198.25
Arabidopsis thaliana	thale cress	Sweden	0.44	(0.20, 0.93)	0.001	84	84	325	428.17

Empirical estimates from diverse set of organisms

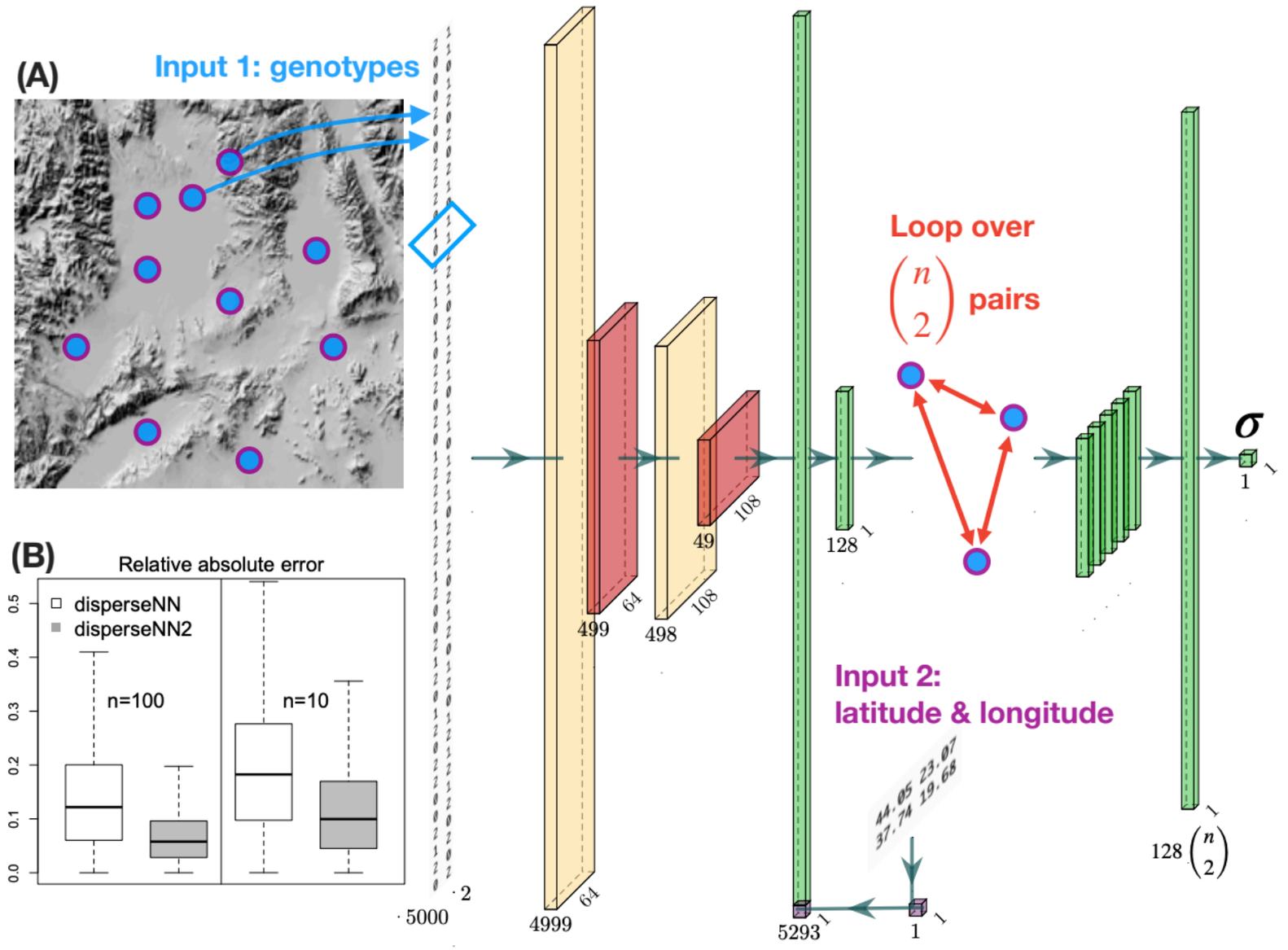
Dispersal inference from population genetic variation using a convolutional neural network

Chris C. R. Smith ,* Silas Tittes , Peter L. Ralph , Andrew D. Kern

Institute of Ecology and Evolution, University of Oregon, Eugene, OR 97403, USA

*Corresponding author: Institute of Ecology and Evolution, University of Oregon, Eugene, OR 97403, USA. Email: chriscs@uoregon.edu

disperseNN2: electric boogaloo



disperseNN2: a neural network for estimating dispersal distance from georeferenced polymorphism data

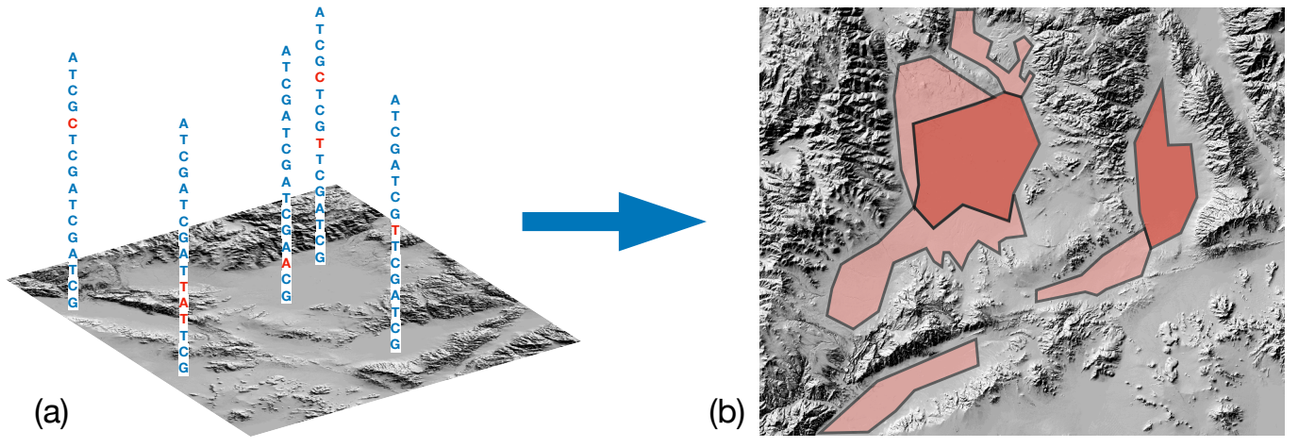
Chris C. R. Smith & Andrew D. Kern

BMC Bioinformatics **24**, Article number: 385 (2023) | Cite this article

mapNN

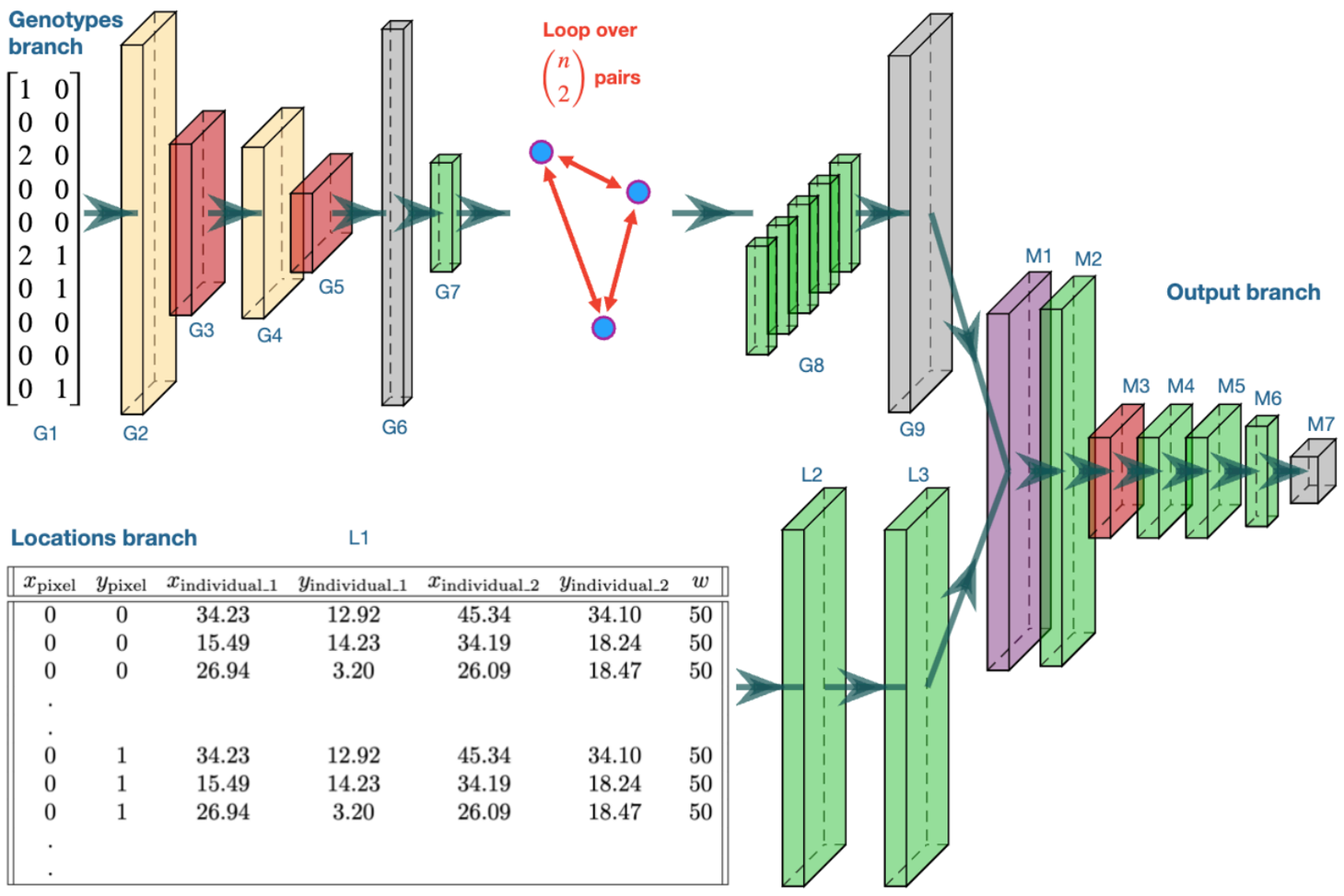
But dispersal need not be homogenous across space!

mapNN



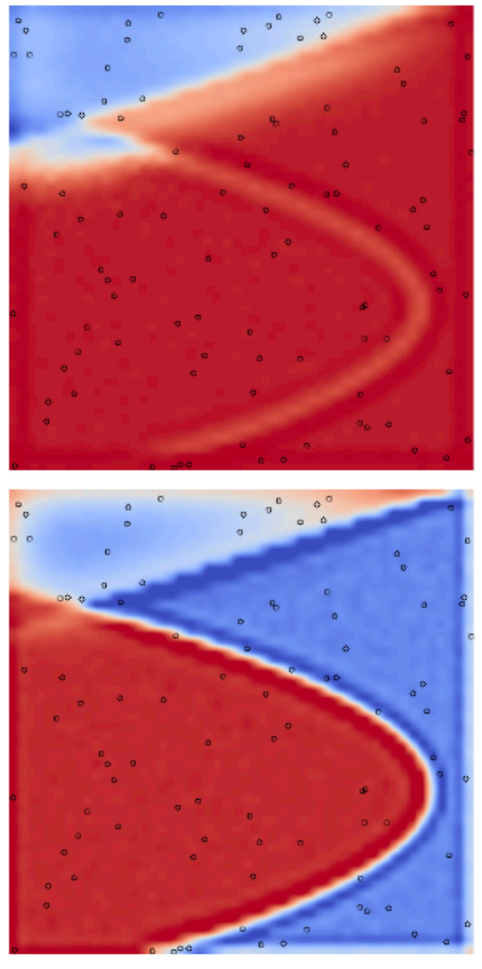
Predicting maps of dispersal and density with a segmentation network

mapNN



x_{pixel}	y_{pixel}	$x_{\text{individual\_1}}$	$y_{\text{individual\_1}}$	$x_{\text{individual\_2}}$	$y_{\text{individual\_2}}$	w
0	0	34.23	12.92	45.34	34.10	50
0	0	15.49	14.23	34.19	18.24	50
0	0	26.94	3.20	26.09	18.47	50
.						
.						
0	1	34.23	12.92	45.34	34.10	50
0	1	15.49	14.23	34.19	18.24	50
0	1	26.94	3.20	26.09	18.47	50
.						
.						

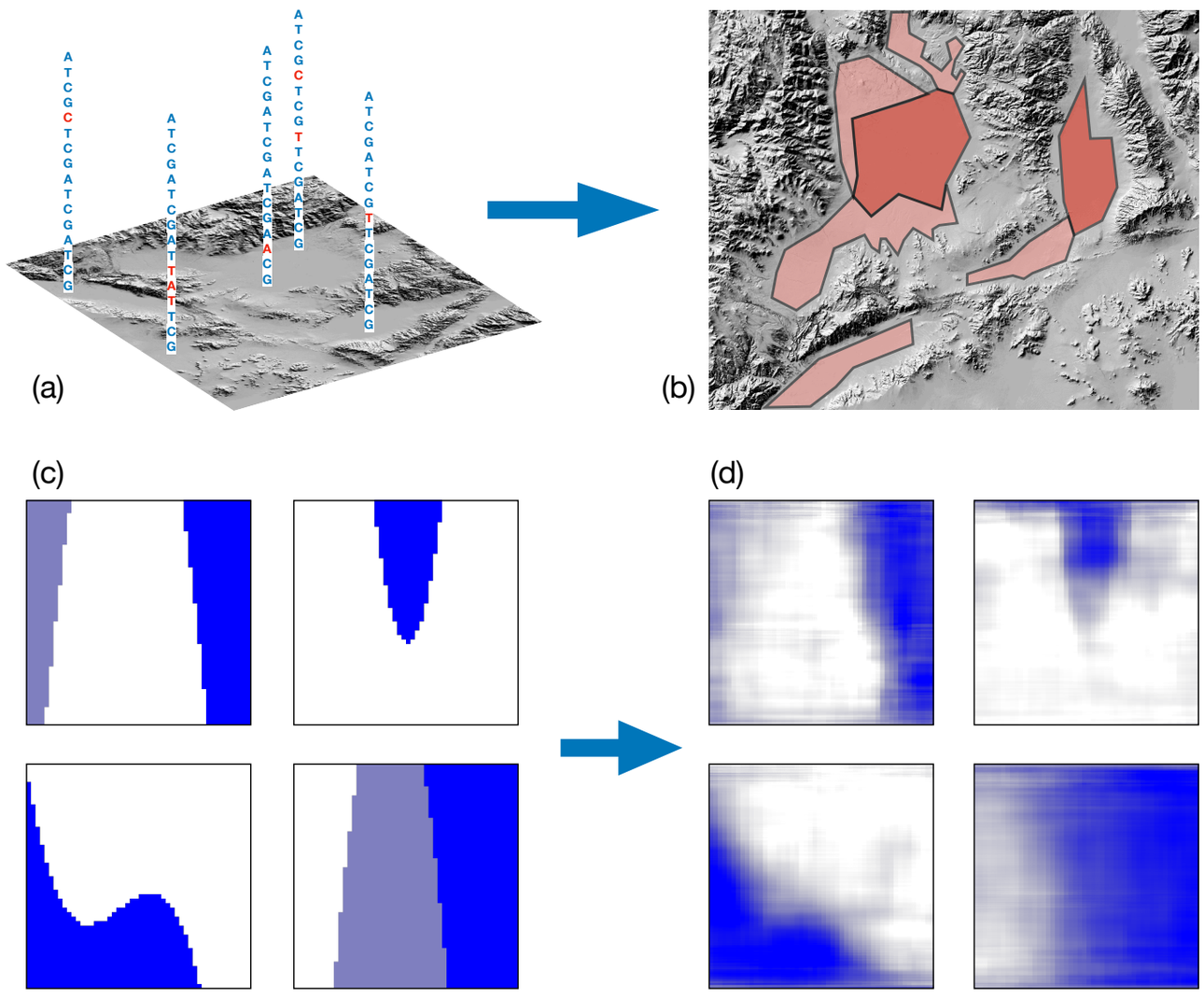
Dispersal



Density

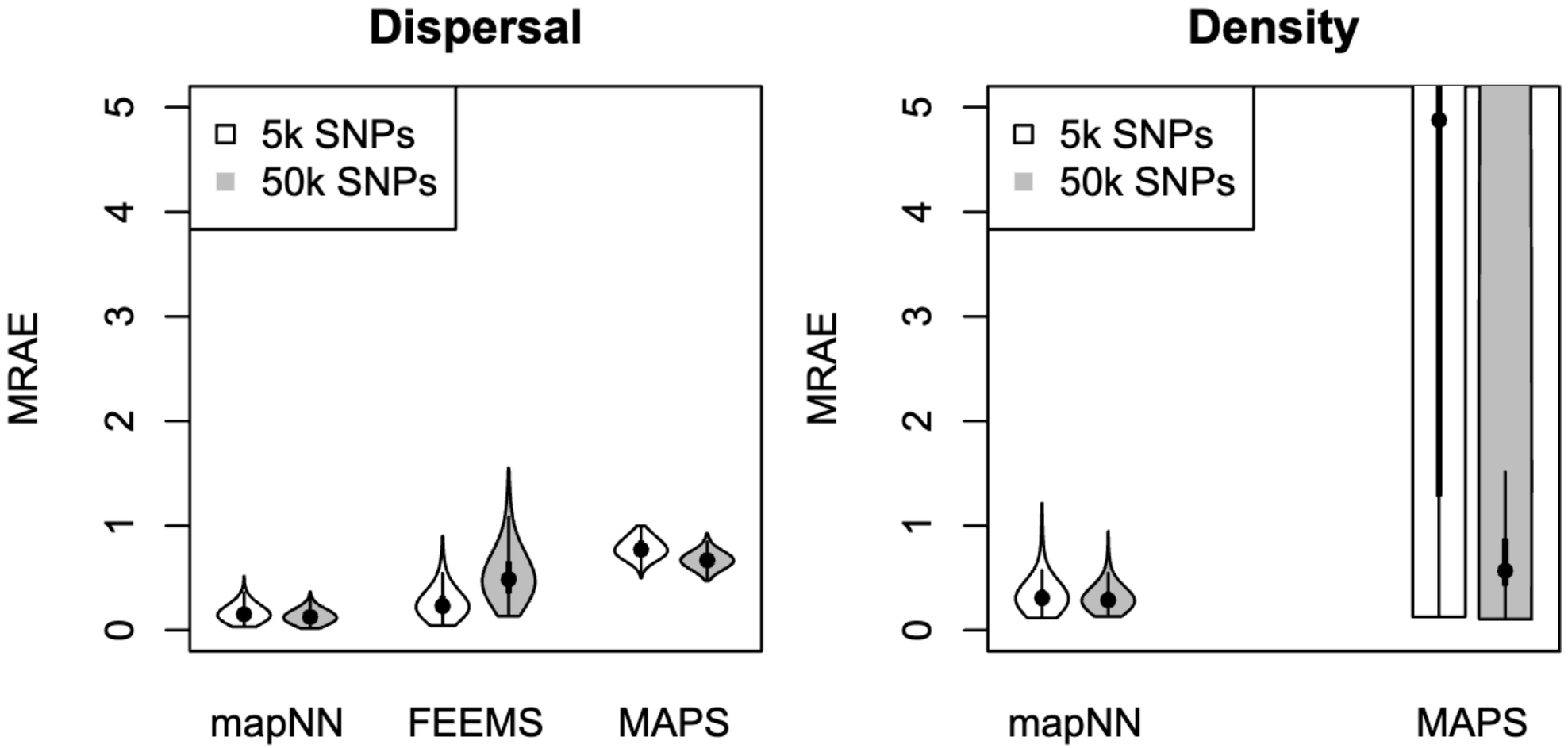
Predicting maps of dispersal and density with a segmentation network

mapNN



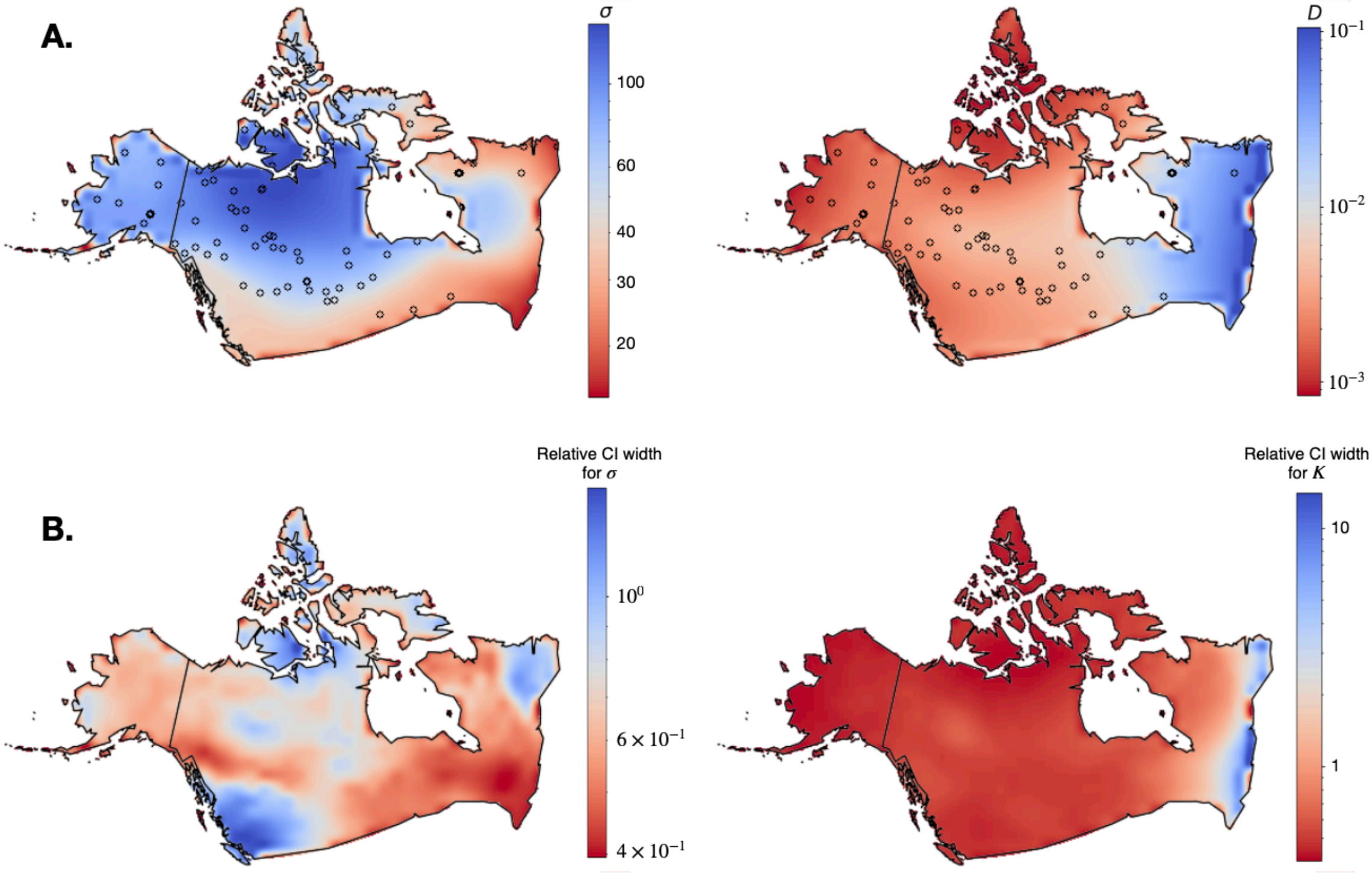
Predicting maps of dispersal with an segmentation network

mapNN



Outcompetes only other existing methods that are based on likelihood

mapNN



Applied to 97 wolf genotypes collected from North America

Space summary

- Now can perform very realistic, continuous space simulations (Battey et al. 2020)

- Space Matters! Summary stats, GWAS

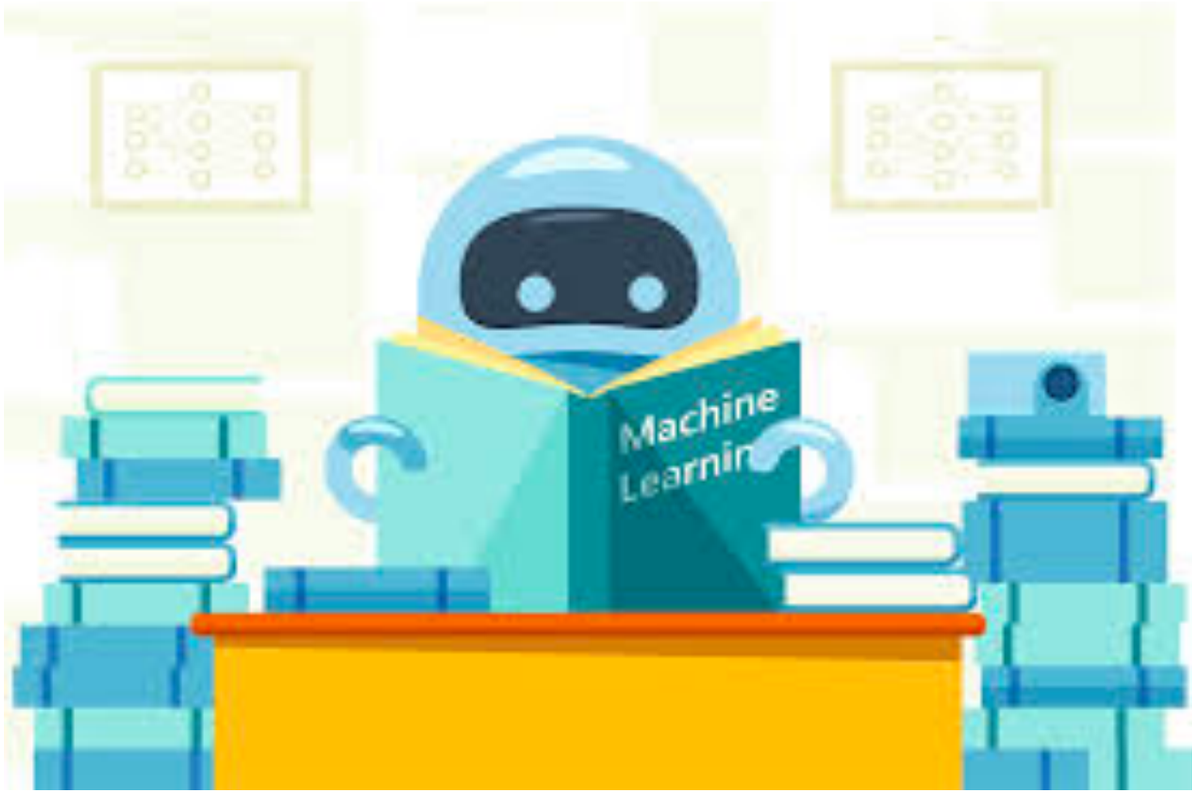
- Can use DNN with georeferenced genomes for model free inference (Locator / Ecolocator)

- biased training an issue!

- Can estimate dispersal independent of density using a CNNs (disperseNN / disperseNN2)

- mapNN provides SOTA estimates of heterogeneous maps of dispersal/ density

take aways



- lots of ways that machine learning can help evolutionary biology / genetics

- huge growth potential— few people working in the area

- many opens questions still:
  - best architectures? prob depends
  - way to deal with uncertainty?
  - ways to build DL models that are robust?

Thanks!

Jeff Adrion
CJ Battey
Gabby Coffing
Clara Rehmann
Chris Smith
Jared Galloway
Dan Schrider
Chuck Langley
Peter Ralph

all members of the Kern-Ralph lab