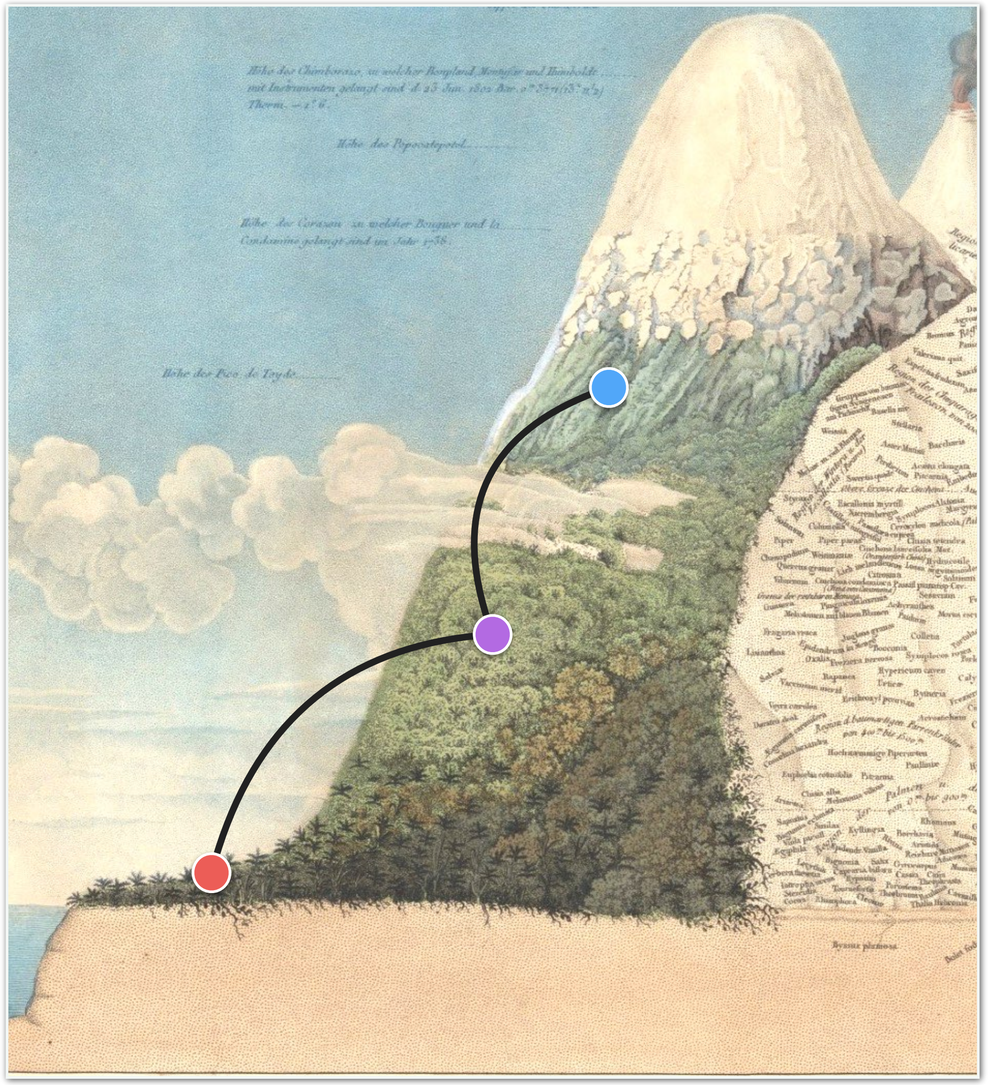
Using Gradient Forest and Generalized Dissimilarity Modeling to analyze and map current and future patterns of adaptive genomic variation

Matt Fitzpatrick
University of Maryland Center for Environmental Science
Appalachian Lab

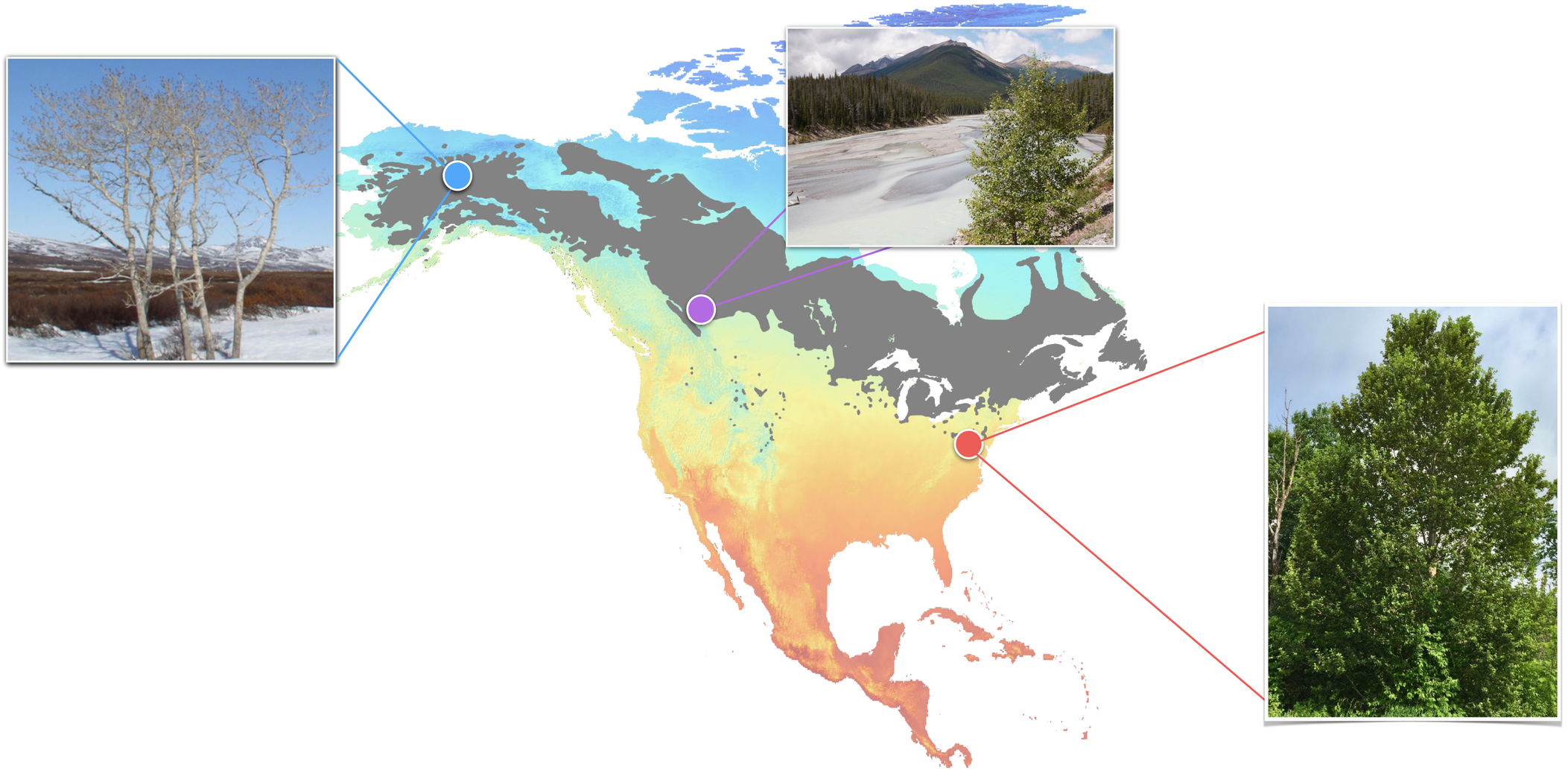
mfitzpatrick@umces.edu

University of Maryland
CENTER FOR ENVIRONMENTAL SCIENCE
APPALACHIAN LABORATORY

β-diversity: change (turnover) in community structure along spatial, temporal or environmental gradients



Change in genomic composition associated with local adaptation along environmental gradients



“Community-level” biodiversity models based on **non-linear** compositional turnover functions

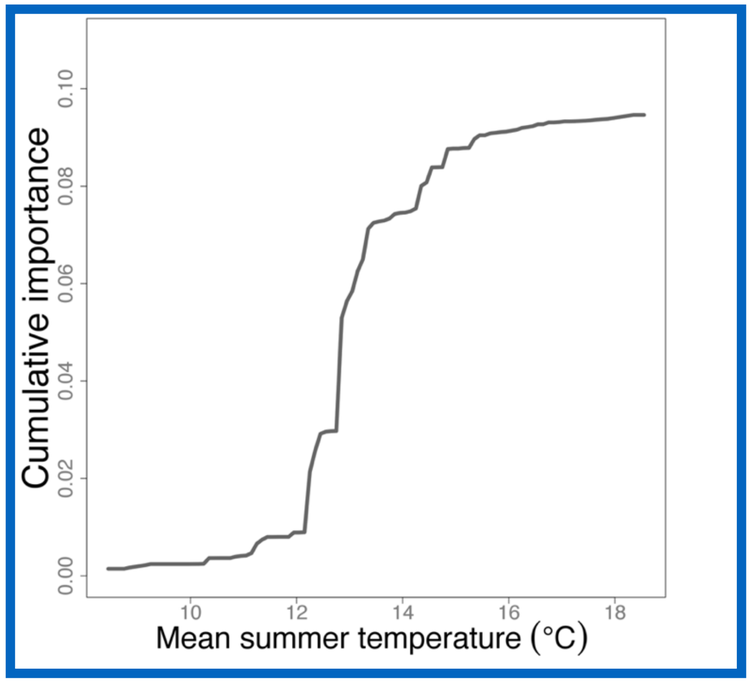
1. Generalized Dissimilarity Modeling (GDM; Ferrier et al. 2007, Mokany et al. 2022)

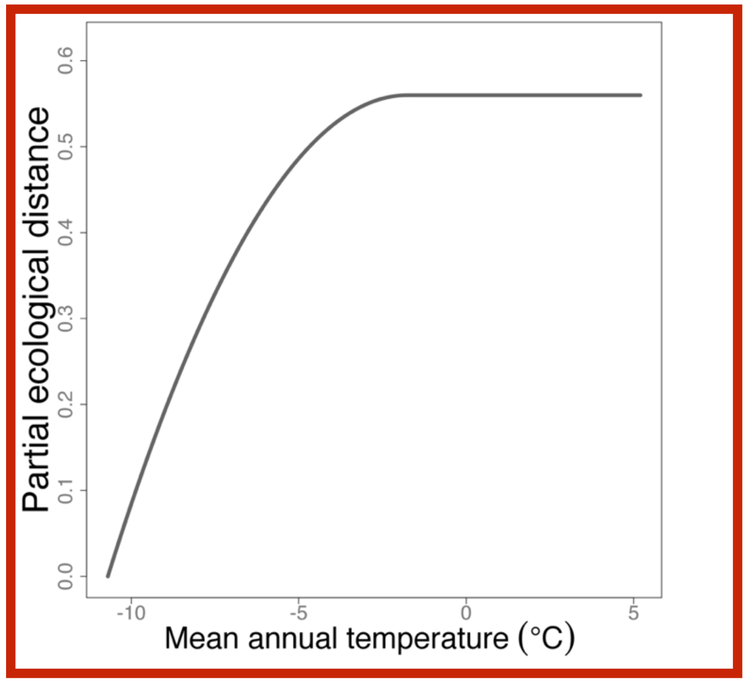
2. Gradient Forests (GF; Ellis et al. 2012)

“Community-level” biodiversity models based on **non-linear** compositional turnover functions

1. Generalized Dissimilarity Modeling (GDM; Ferrier et al. 2007, Mokany et al. 2022)

2. Gradient Forests (GF; Ellis et al. 2012)

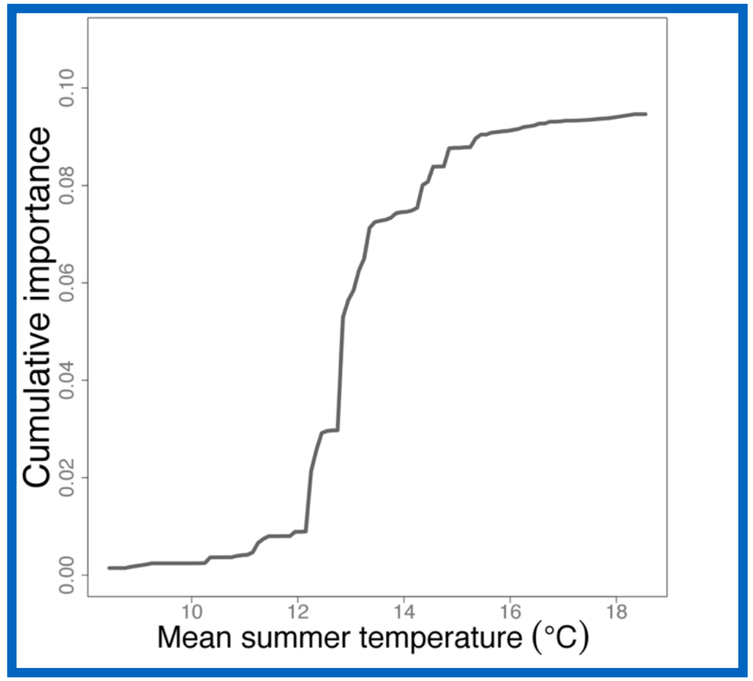


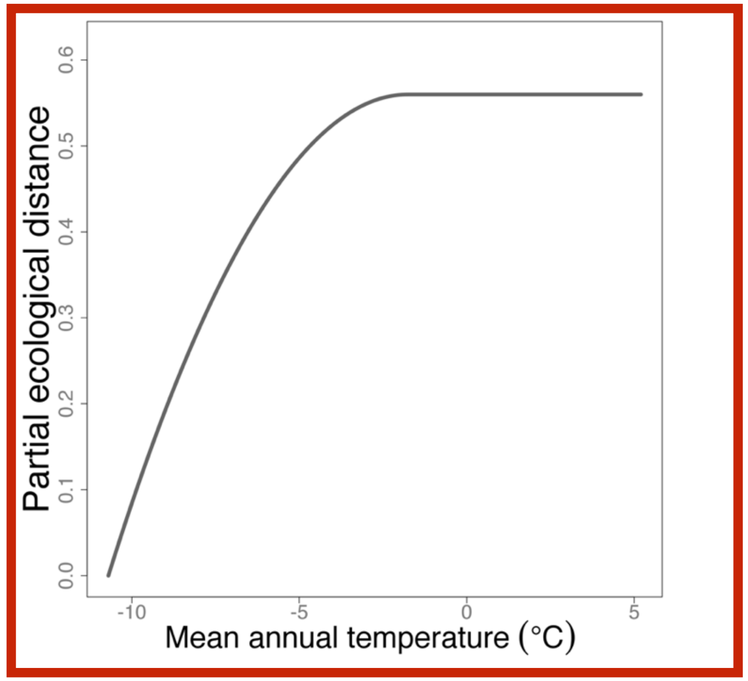


“Community-level” biodiversity models based on **non-linear** compositional turnover functions

Shape = rate of change in allele frequencies along each gradient

Height = relative importance of that gradient in explaining genomic patterns

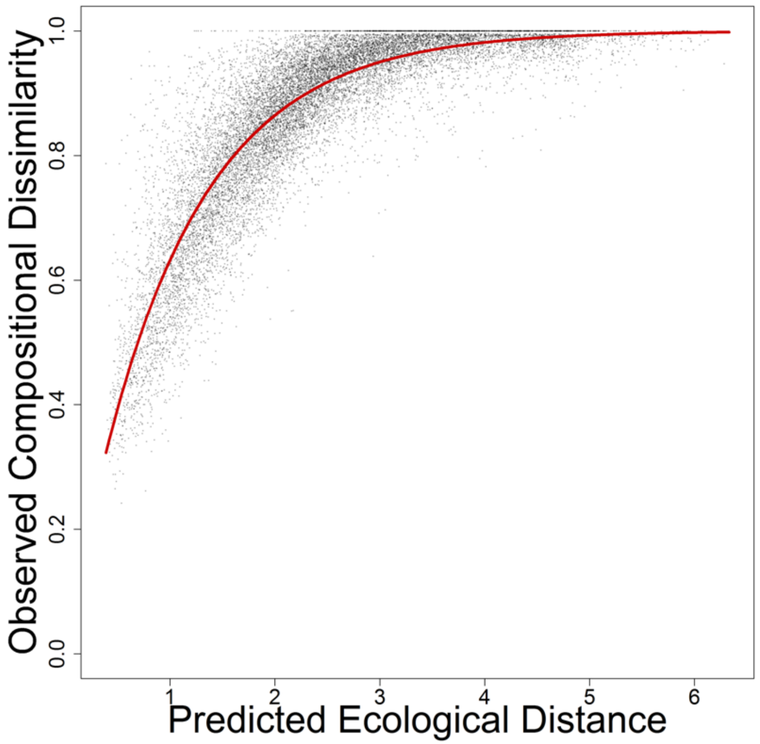


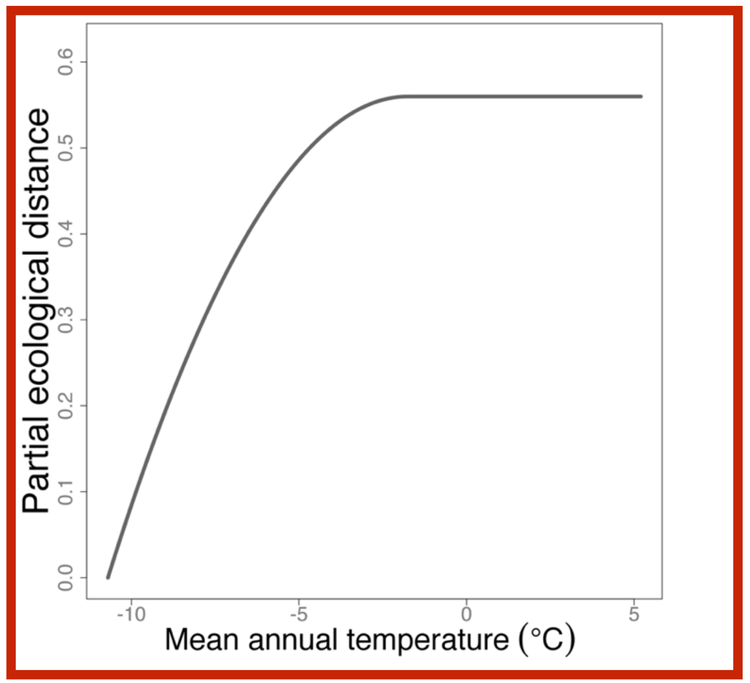


GDM: Spatial variation in genome composition as a function of geographic and environmental distance

Generalized Dissimilarity Modeling (GDM; Ferrier et al. 2007, Mokany et al. 2022)

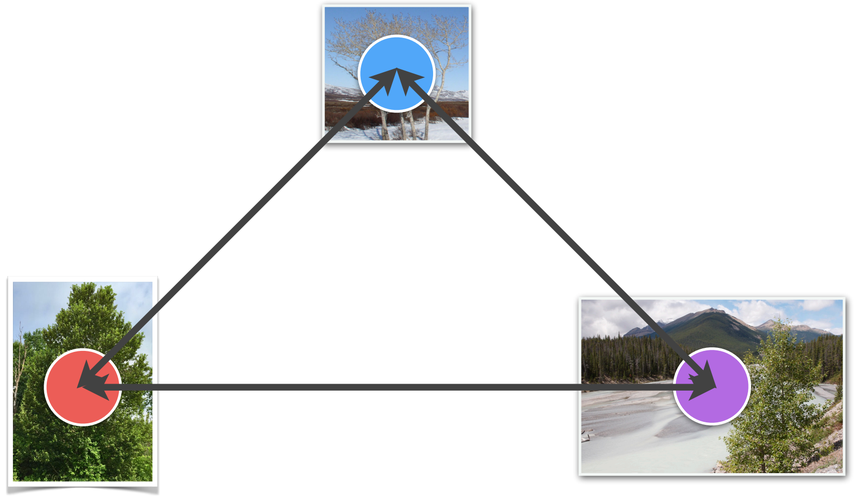
- GLM-like framework that models pairwise biological dissimilarity as a function of environmental & geographic distance

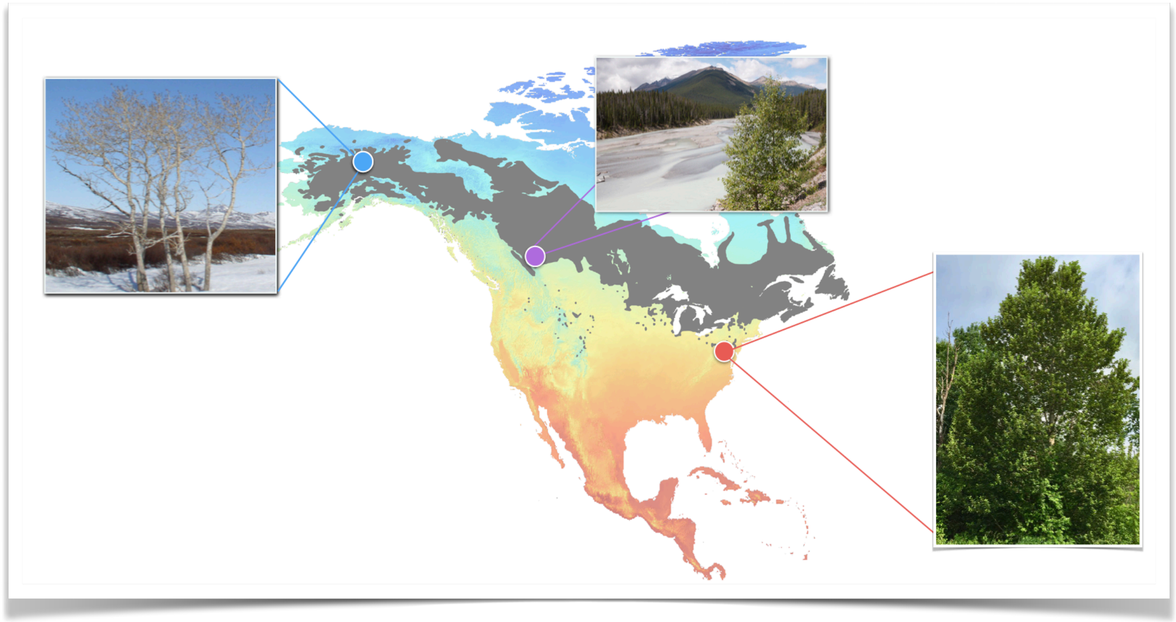




GDM: Spatial variation in genome composition as a function of geographic and environmental distance

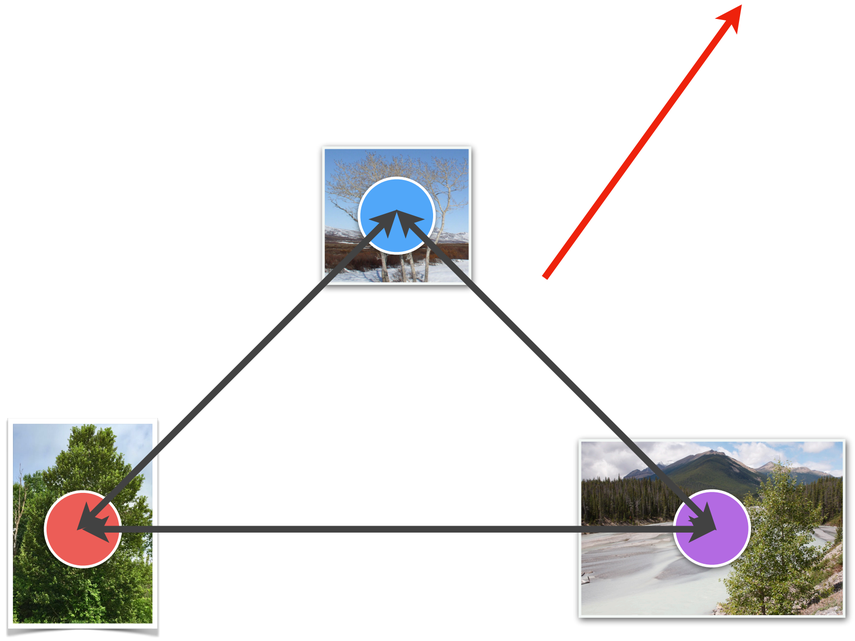
$$-\ln\left(1 - d_{ij}\right) = a_0 + \sum_{p=1}^{n} \left| f_p(x_{pi}) - f_p(x_{pj}) \right|$$

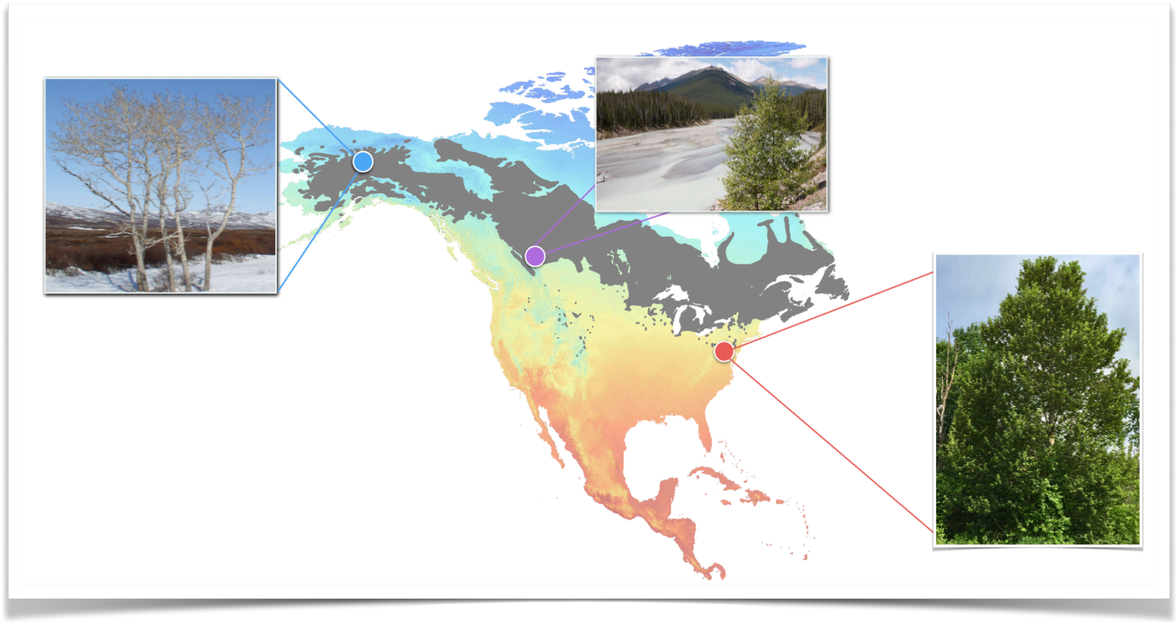




GDM: Spatial variation in genome composition as a function of geographic and environmental distance

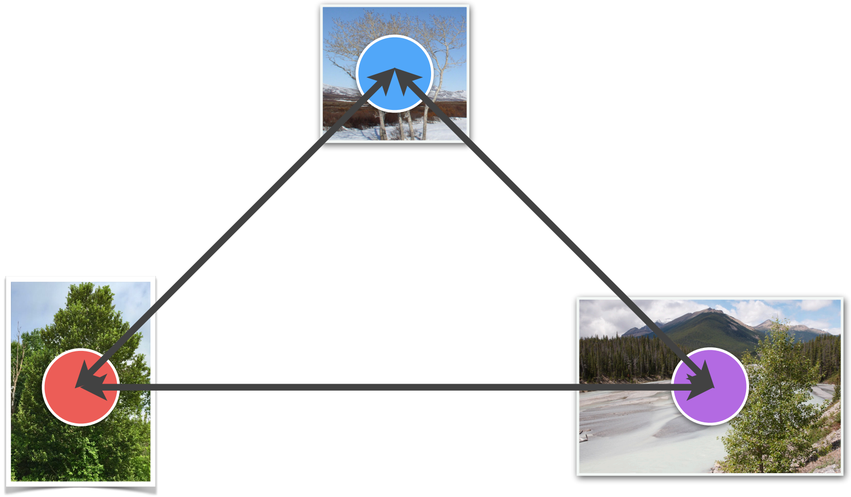
$$-\ln\left(1 - d_{ij}\right) = a_0 + \sum_{p=1}^{n} \left| f_p(x_{pi}) - f_p(x_{pj}) \right|$$

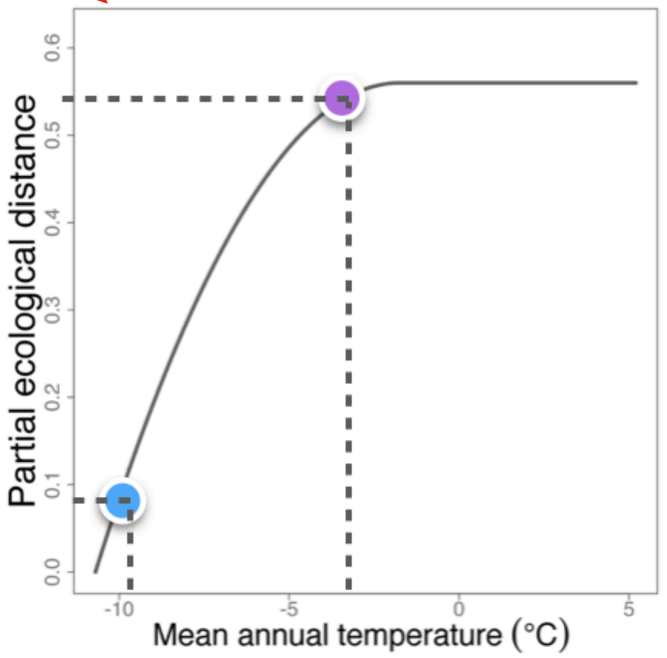




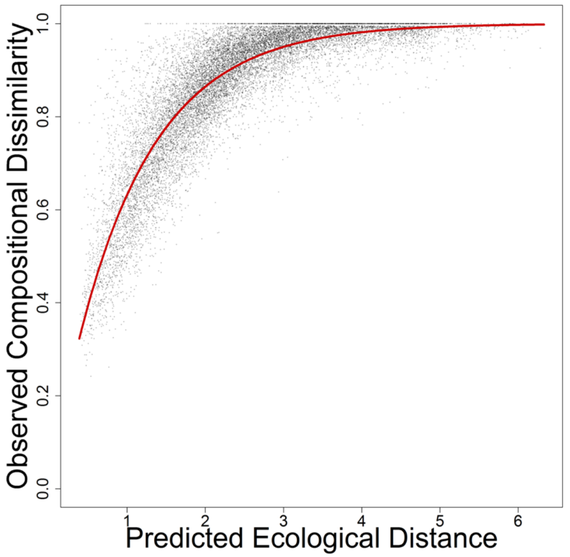
GDM: Spatial variation in genome composition as a function of geographic and environmental distance

$$-\ln\left(1 - d_{ij}\right) = a_0 + \sum_{p=1}^{n} \left| f_p(x_{pi}) - f_p(x_{pj}) \right|$$





GDM



Biological distance metric

$$-\ln\left(1 - d_{ij}\right) = a_0 + \sum_{p=1}^{n} \left| f_p(x_{pi}) - f_p(x_{pj}) \right|$$

Model intercept

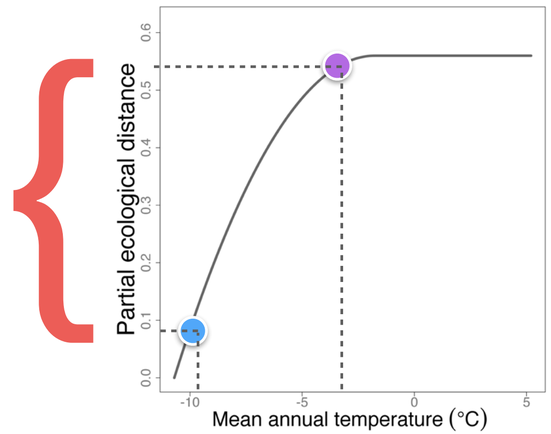
GDM

Biological distance metric

Scaled Environmental Distance

$$-\ln\left(1 - d_{ij}\right) = a_0 + \sum_{p=1}^{n} \left| f_p(x_{pi}) - f_p(x_{pj}) \right|$$

Model intercept



Predicting genetic differentiation between any pair of locations (and/or times)

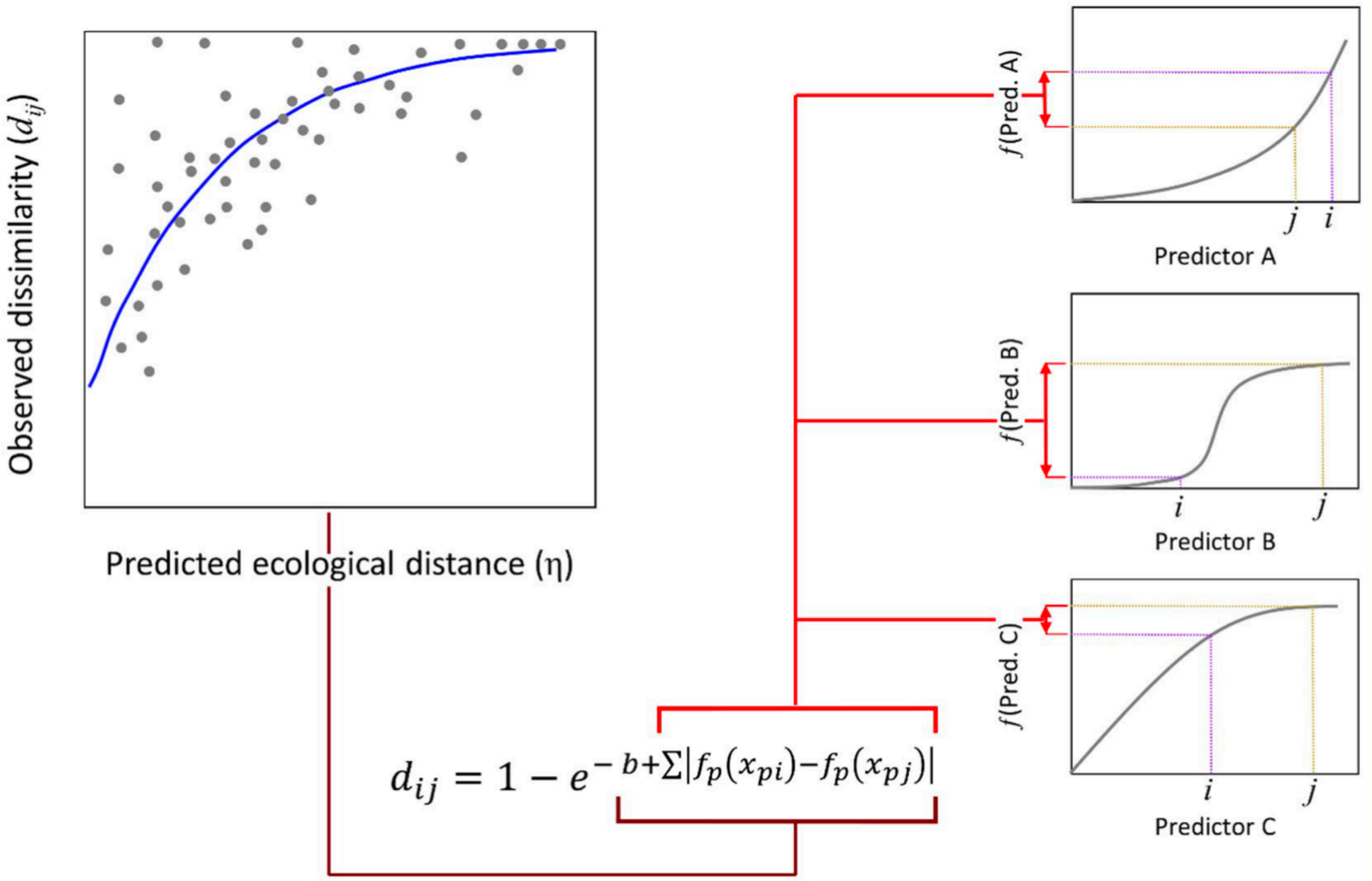
Biological distance metric

Model intercept

Scaled Environmental Distance

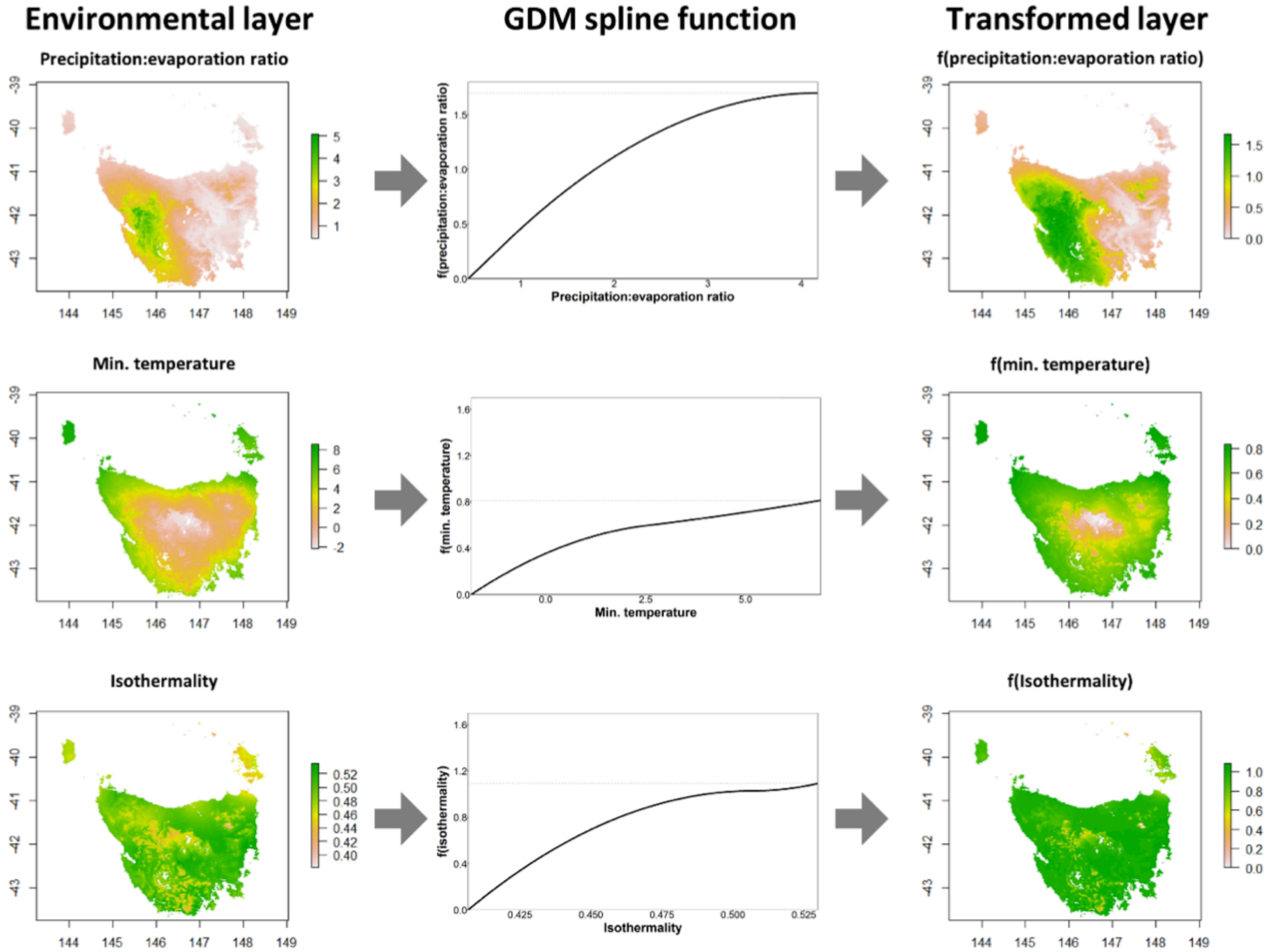
$$d_{ij} = 1 - e^{-(a+E_{ij})}$$

$$E_{ij} = \sum_{p=1}^{n} \left| f_p(x_{pi}) - f_p(x_{pj}) \right|$$



Mokany et al. (2022)

gdm.transform()



Mokany et al. (2022)

GDM: Input data formats - biological

(1) Site by species matrix

Site name	X	Y	Species 1	Species 2	Species 3	Species 4	Species 5	Species 6	Species 7
Site A	140.3	-39.2	1	1	0	0	0	0	1
Site B	141.2	-38.3	0	1	0	1	1	0	0
Site C	139.6	-37.5	0	1	0	0	1	1	0
Site D	143.8	-39.1	0	1	1	0	1	0	0
Site E	140.3	-39.2	1	0	0	0	0	1	1
Site F	141.2	-39.1	1	1	1	1	0	1	0
Site G	138.1	-37.3	0	1	0	0	1	1	0
Site H	143.8	-37.5	1	1	0	1	0	0	1
Site I	140.3	-36.2	0	0	0	1	0	1	1

(2) X, Y, species list

X	Y	Species
140.3	-39.2	*E.dab*
141.2	-38.3	*E.dab*
139.6	-37.5	*D.bop*
143.8	-39.1	*D.bop*
140.3	-39.2	*D.bop*
141.2	-39.1	*D.bop*
138.1	-37.3	*Z.bam*
143.8	-37.5	*Z.bam*
140.3	-36.2	*Z.bam*

(3) site by site distance matrix

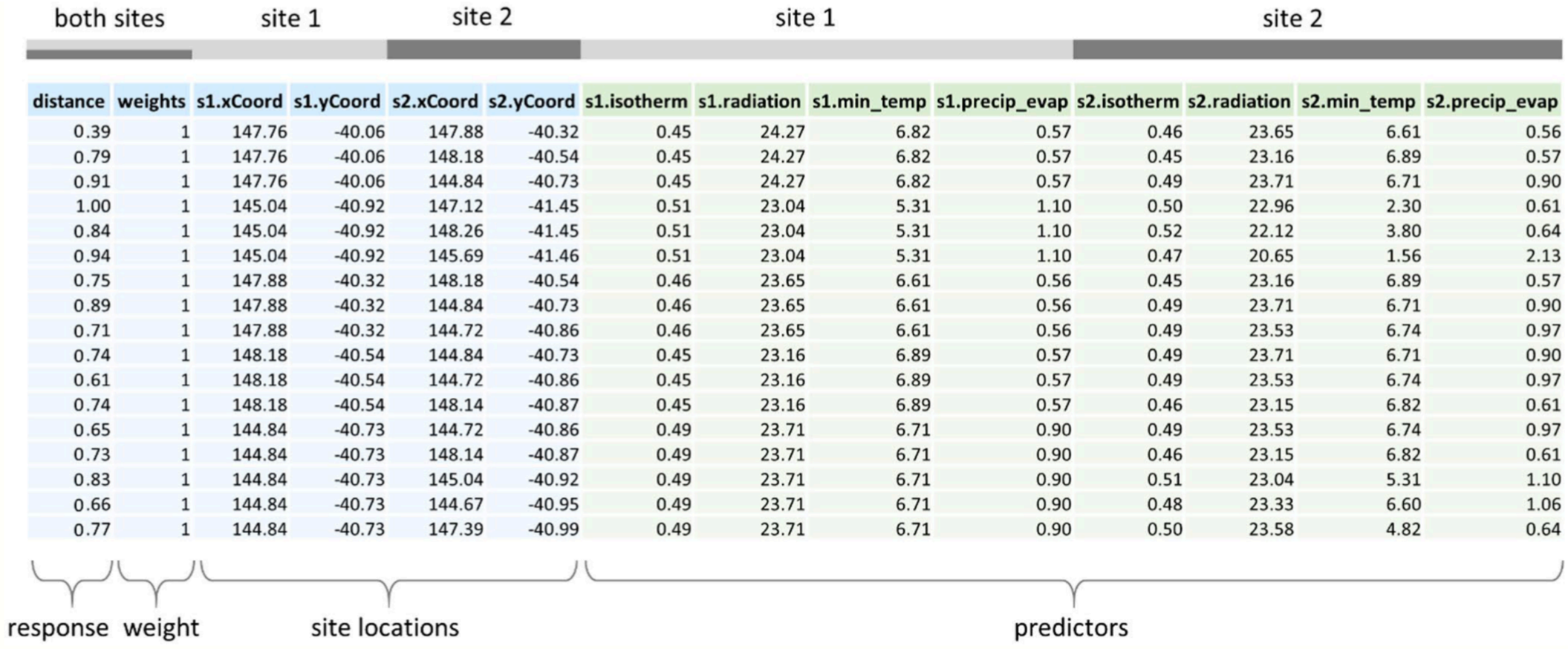
Site name	Site A	Site B	Site C	Site D	Site E	Site F	Site G
Site A	0	0.56	0.94	1.00	0.22	1.00	0.78
Site B	0.56	0	1.00	0.31	0.89	1.00	1.00
Site C	0.94	1.00	0	1.00	0.65	1.00	0.26
Site D	1.00	0.31	1.00	0	1.00	1.00	0.55
Site E	0.22	0.89	0.65	1.00	0	0.84	1.00
Site F	1.00	1.00	1.00	1.00	0.84	0	1.00
Site G	0.78	1.00	0.26	0.55	1.00	1.00	0

Genetic distance
Phylogenetic distance
Trait distance

Mokany et al. (2022)

GDM: Site-pair formatted data table

gdm::formatsitepair()



distance	weights	s1.xCoord	s1.yCoord	s2.xCoord	s2.yCoord	s1.isotherm	s1.radiation	s1.min_temp	s1.precip_evap	s2.isotherm	s2.radiation	s2.min_temp	s2.precip_evap
0.39	1	147.76	-40.06	147.88	-40.32	0.45	24.27	6.82	0.57	0.46	23.65	6.61	0.56
0.79	1	147.76	-40.06	148.18	-40.54	0.45	24.27	6.82	0.57	0.45	23.16	6.89	0.57
0.91	1	147.76	-40.06	144.84	-40.73	0.45	24.27	6.82	0.57	0.49	23.71	6.71	0.90
1.00	1	145.04	-40.92	147.12	-41.45	0.51	23.04	5.31	1.10	0.50	22.96	2.30	0.61
0.84	1	145.04	-40.92	148.26	-41.45	0.51	23.04	5.31	1.10	0.52	22.12	3.80	0.64
0.94	1	145.04	-40.92	145.69	-41.46	0.51	23.04	5.31	1.10	0.47	20.65	1.56	2.13
0.75	1	147.88	-40.32	148.18	-40.54	0.46	23.65	6.61	0.56	0.45	23.16	6.89	0.57
0.89	1	147.88	-40.32	144.84	-40.73	0.46	23.65	6.61	0.56	0.49	23.71	6.71	0.90
0.71	1	147.88	-40.32	144.72	-40.86	0.46	23.65	6.61	0.56	0.49	23.53	6.74	0.97
0.74	1	148.18	-40.54	144.84	-40.73	0.45	23.16	6.89	0.57	0.49	23.71	6.71	0.90
0.61	1	148.18	-40.54	144.72	-40.86	0.45	23.16	6.89	0.57	0.49	23.53	6.74	0.97
0.74	1	148.18	-40.54	148.14	-40.87	0.45	23.16	6.89	0.57	0.46	23.15	6.82	0.61
0.65	1	144.84	-40.73	144.72	-40.86	0.49	23.71	6.71	0.90	0.49	23.53	6.74	0.97
0.73	1	144.84	-40.73	148.14	-40.87	0.49	23.71	6.71	0.90	0.46	23.15	6.82	0.61
0.83	1	144.84	-40.73	145.04	-40.92	0.49	23.71	6.71	0.90	0.51	23.04	5.31	1.10
0.66	1	144.84	-40.73	144.67	-40.95	0.49	23.71	6.71	0.90	0.48	23.33	6.60	1.06
0.77	1	144.84	-40.73	147.39	-40.99	0.49	23.71	6.71	0.90	0.50	23.58	4.82	0.64

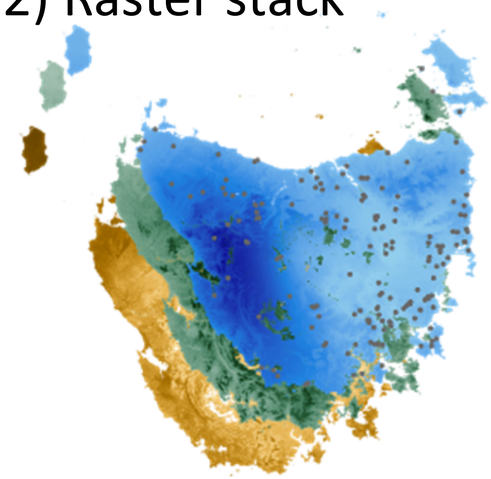
Mokany et al. (2022)

GDM: Input data formats - Predictor variables

(1) Site by predictor matrix

Site name	X	Y	Predictor 1	Predictor 2	Predictor 3	Predictor 4
Site A	139.6	-37.5	0	2003	8.3	1.00
Site B	143.8	-39.1	0.56	1945	7.8	0.31
Site C	140.3	-39.2	0.94	632	6.2	1.00
Site D	141.2	-39.1	1.00	1284	3.7	0
Site E	138.1	-37.3	0.22	789	3.2	1.00
Site F	143.8	-37.5	1.00	321	4.9	1.00
Site G	140.3	-36.2	0.78	2236	6.7	0.55

(2) Raster stack



(3) site by site distance matrix

Site name	Site A	Site B	Site C	Site D	Site E	Site F	Site G
Site A	0	0.56	0.94	1.00	0.22	1.00	0.78
Site B	0.56	0	1.00	0.31	0.89	1.00	1.00
Site C	0.94	1.00	0	1.00	0.65	1.00	0.26
Site D	1.00	0.31	1.00	0	1.00	1.00	0.55
Site E	0.22	0.89	0.65	1.00	0	0.84	1.00
Site F	1.00	1.00	1.00	1.00	0.84	0	1.00
Site G	0.78	1.00	0.26	0.55	1.00	1.00	0

Mokany et al. (2022)

GDM: Environmental Data

- Continuous variables
- Categorial (risky) and only if ordinal
- Geographic / Spatial predictors
  - Least cost paths
  - Resistance distance
- Biological predictors
  - Dissimilarity matrix for another taxonomic group?

GDM: Key outputs

- Percent deviance explained
- I-Spline turnover function for each predictor with sum of coefficients > 0
- Fitted model relating observed differentiation to scaled ecological distance (can predict pairwise distances across space and/or through time = genetic offset)

GDM: ‘predictions’

- Transformed environmental data (from the fitted I-splines): ***gdm.transform()***
- Predicted pairwise biological distance between all locations: ***predict.gdm()***
  - In space:
    - Expected genetic differentiation between populations / locations (spatial offset)
  - In time:
    - Expected degree of maladaptation in units of the response variable (e.g. F_{ST}) = genetic offset

Generalized Dissimilarity Modeling - Further reading

- Ferrier, S., Manion, G., Elith, J., & Richardson, K. (2007). Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. Diversity and distributions, 13(3), 252-264.

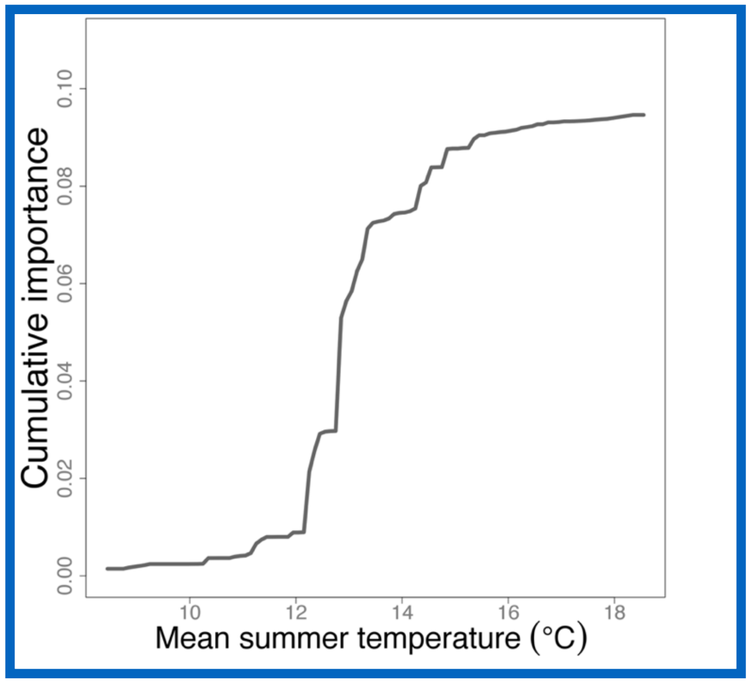
- Mokany, K., Ware, C., Woolley, S. N., Ferrier, S., & Fitzpatrick, M. C. (2022). A working guide to harnessing generalized dissimilarity modelling for biodiversity analysis and conservation assessment. Global Ecology and Biogeography, 31(4), 802-821.

- GDM website: https://mfitzpatrick.al.umces.edu/gdm/
- Github: https://github.com/fitzLab-AL/gdm

GF: Spatial variation in genome composition as a function of “split importance”

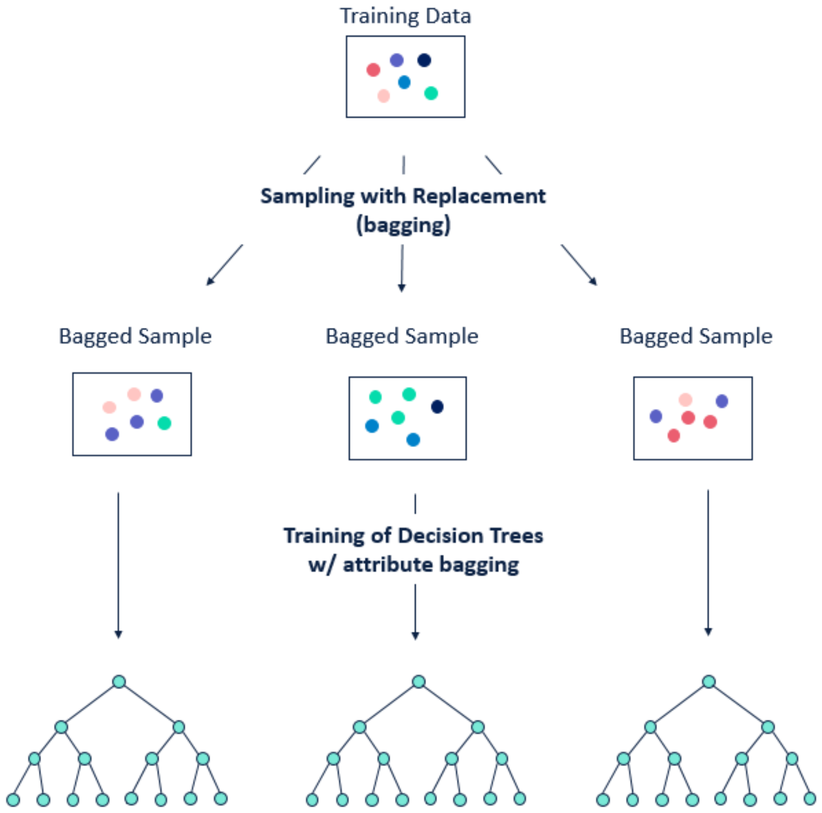
Gradient Forests (GF; Ellis *et al.* 2012)

- Nonparametric, machine learning, regression tree method (randomForests)
- Turnover functions for individual SNPs and for all SNPs combined



Random Forests

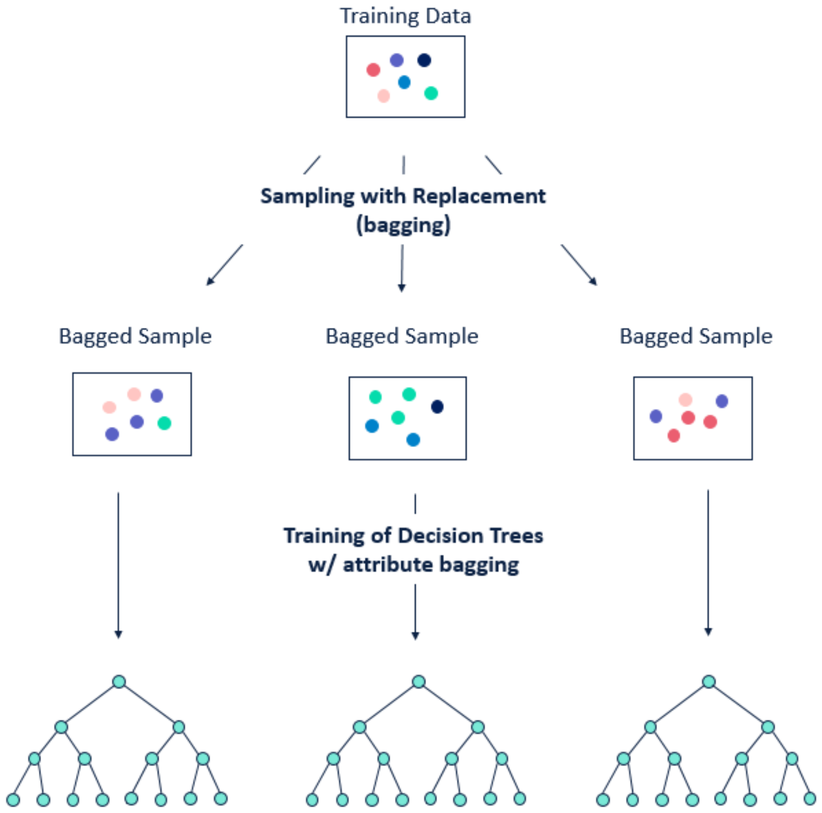
- Exploit strengths of decisions trees while dealing with weaknesses
- Build a large number of trees and aggregate



Random Forests

- Steps

1. Subsample data many times with replacement
2. Retain out-of-bag (OOB) sample (~30%) for model evaluation
3. Build a large number of trees (~500), selecting variables at random for each split
4. Evaluate with OOB
5. Avoid overfitting by averaging forest of trees



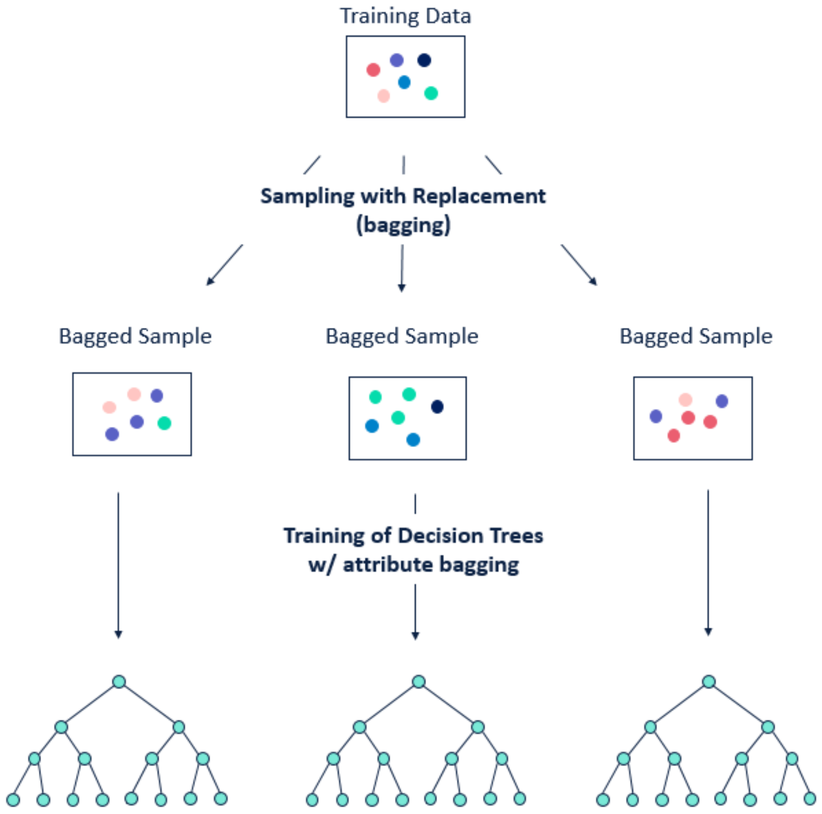
Random Forests

- Variable importance:

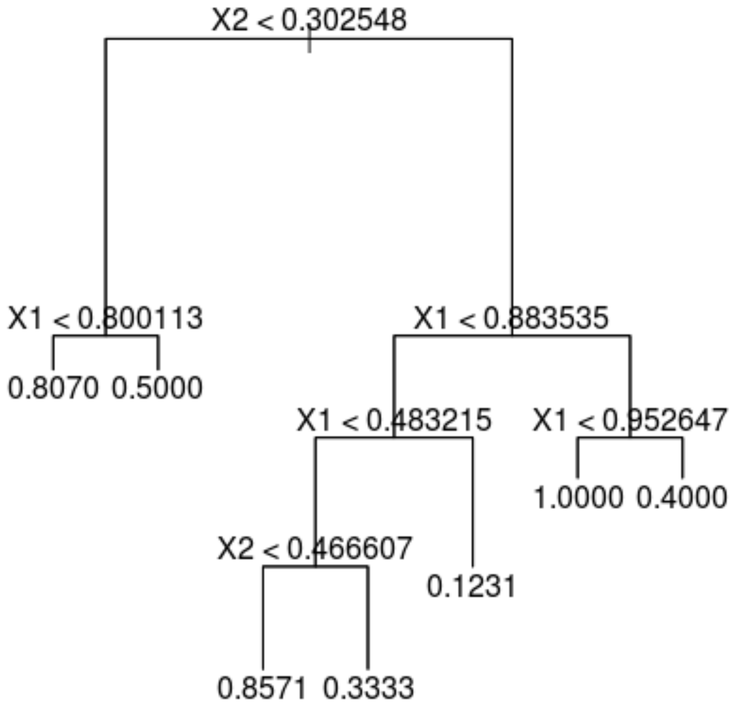
  1. Calculate error using OOB

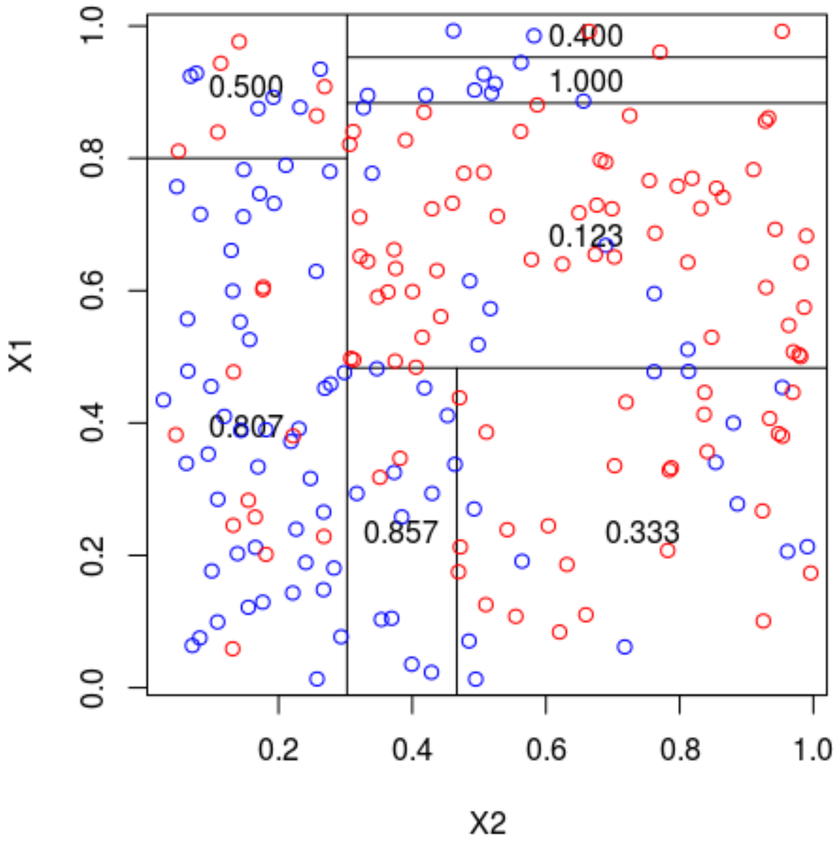
  2. Randomize each variable in turn and calculate how much error rates change

     - Important variables = large increase in error when randomized

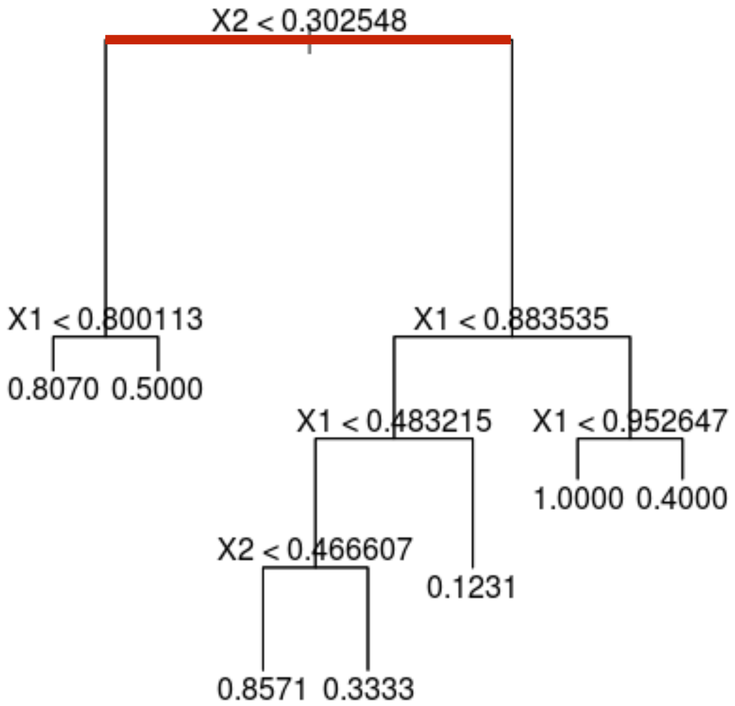


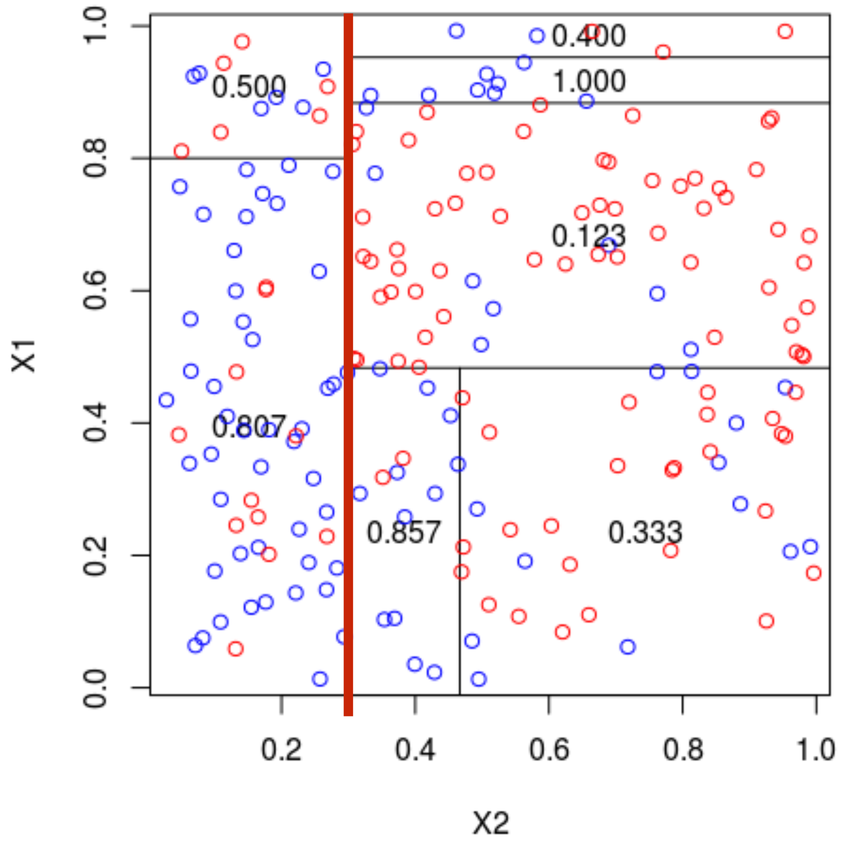
Random Forests



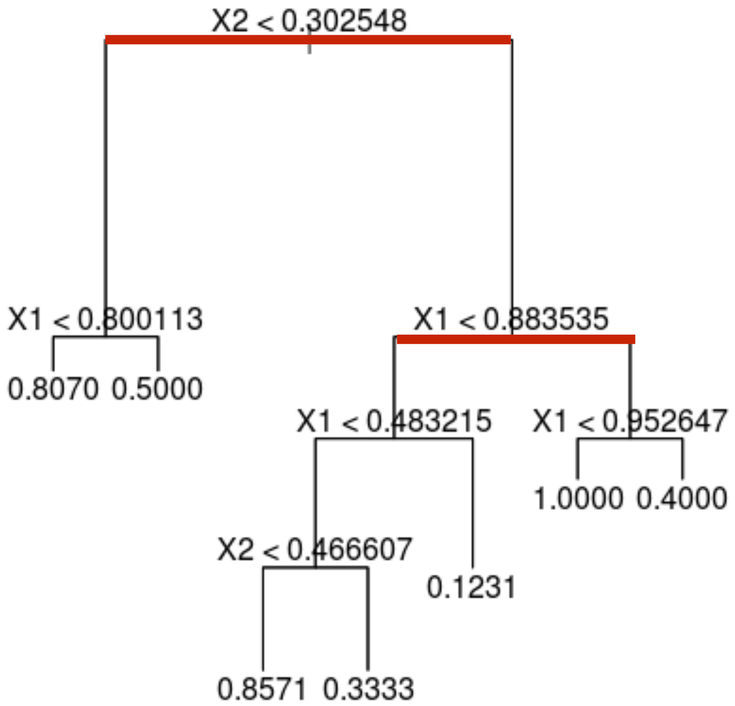


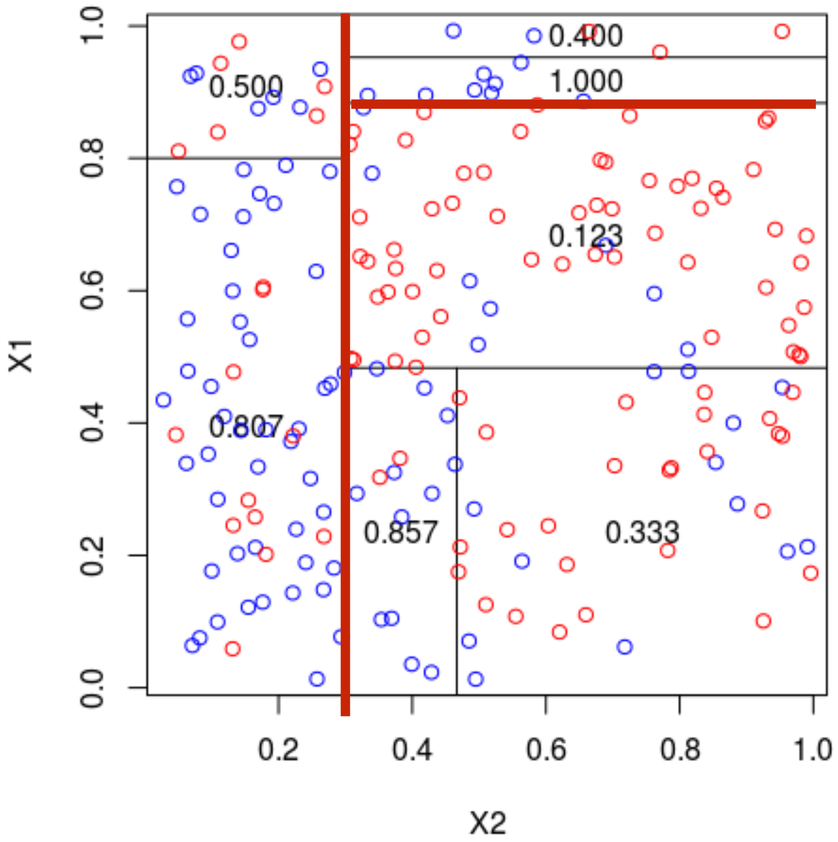
Random Forests



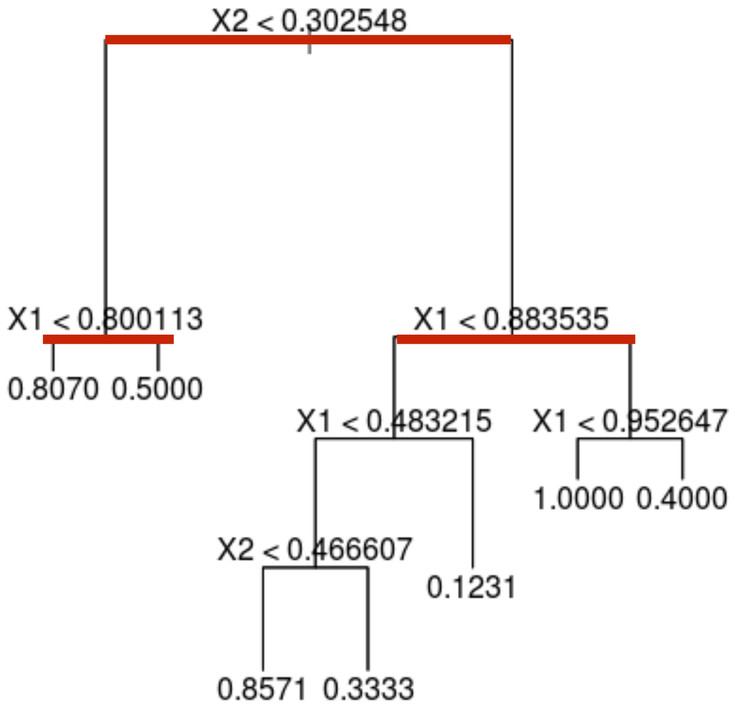


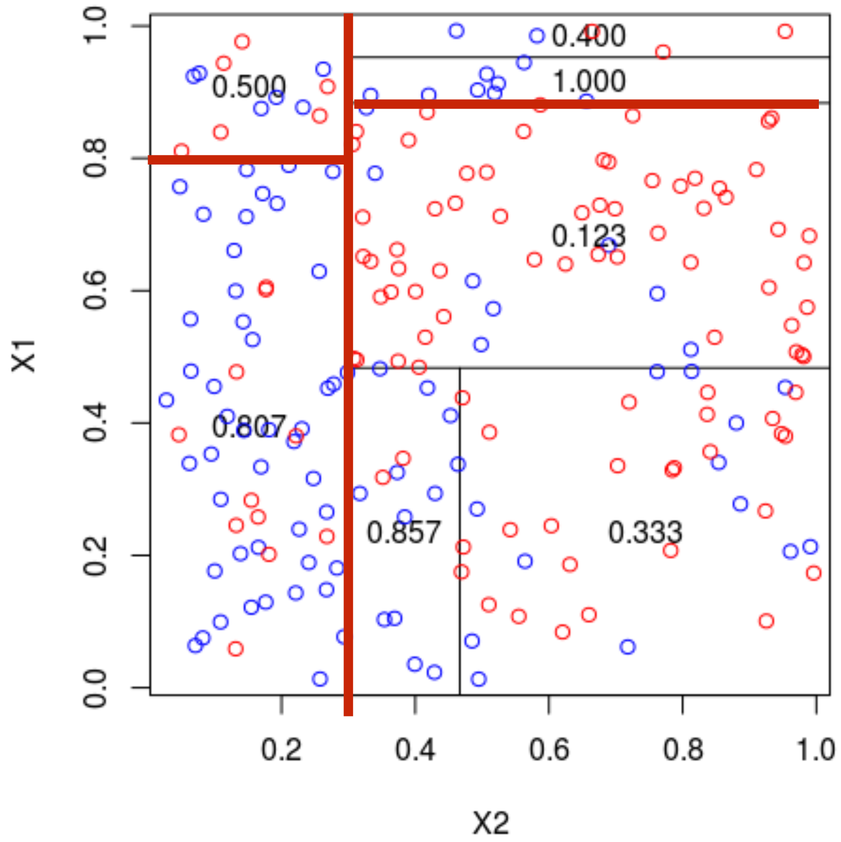
Random Forests



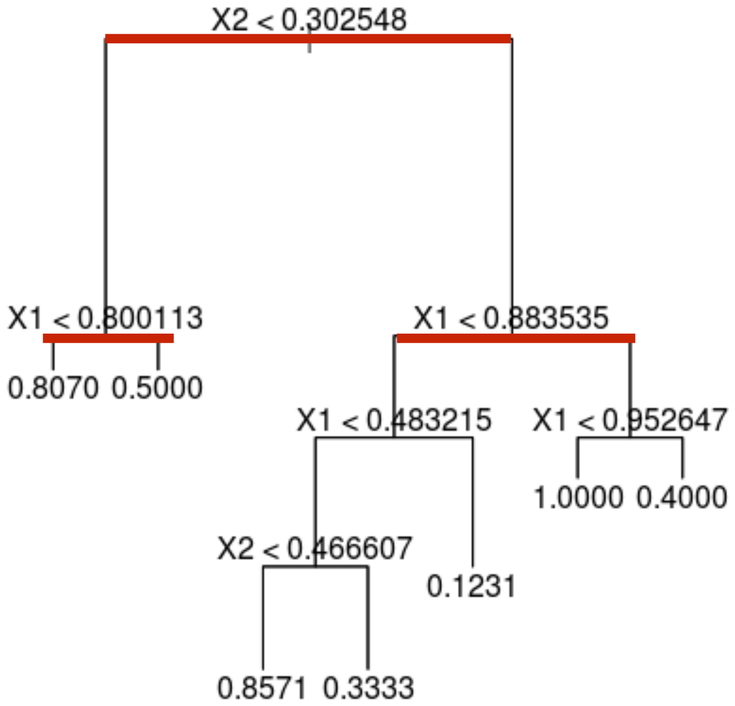


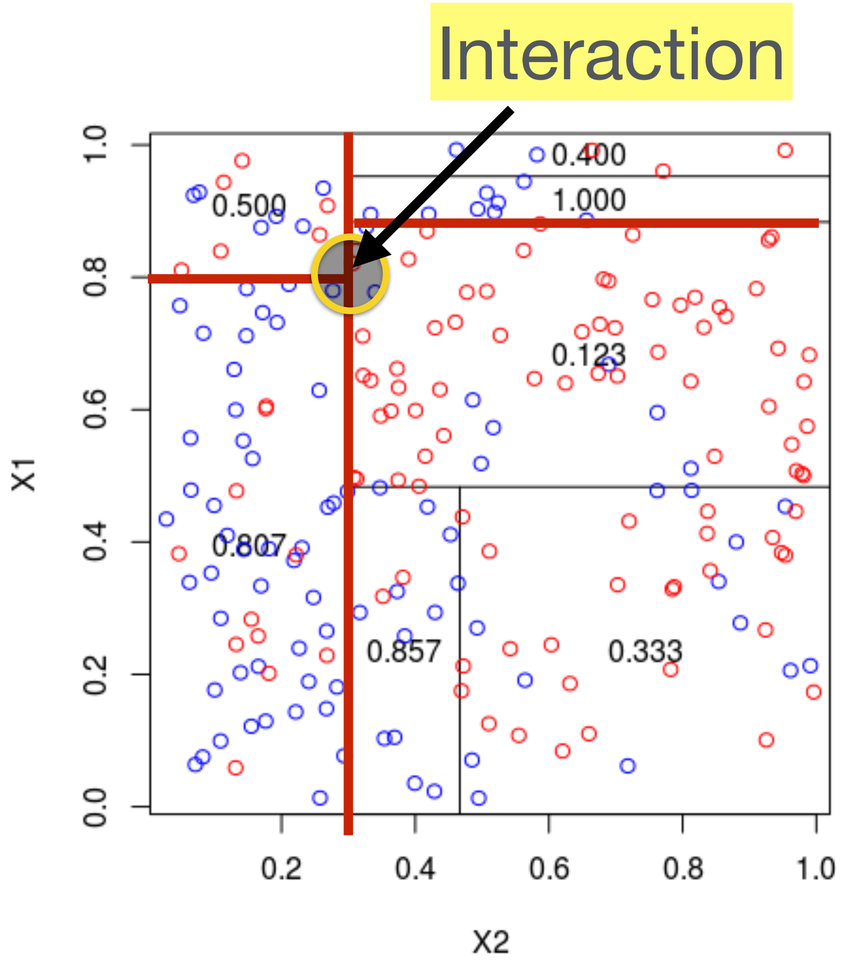
Random Forests



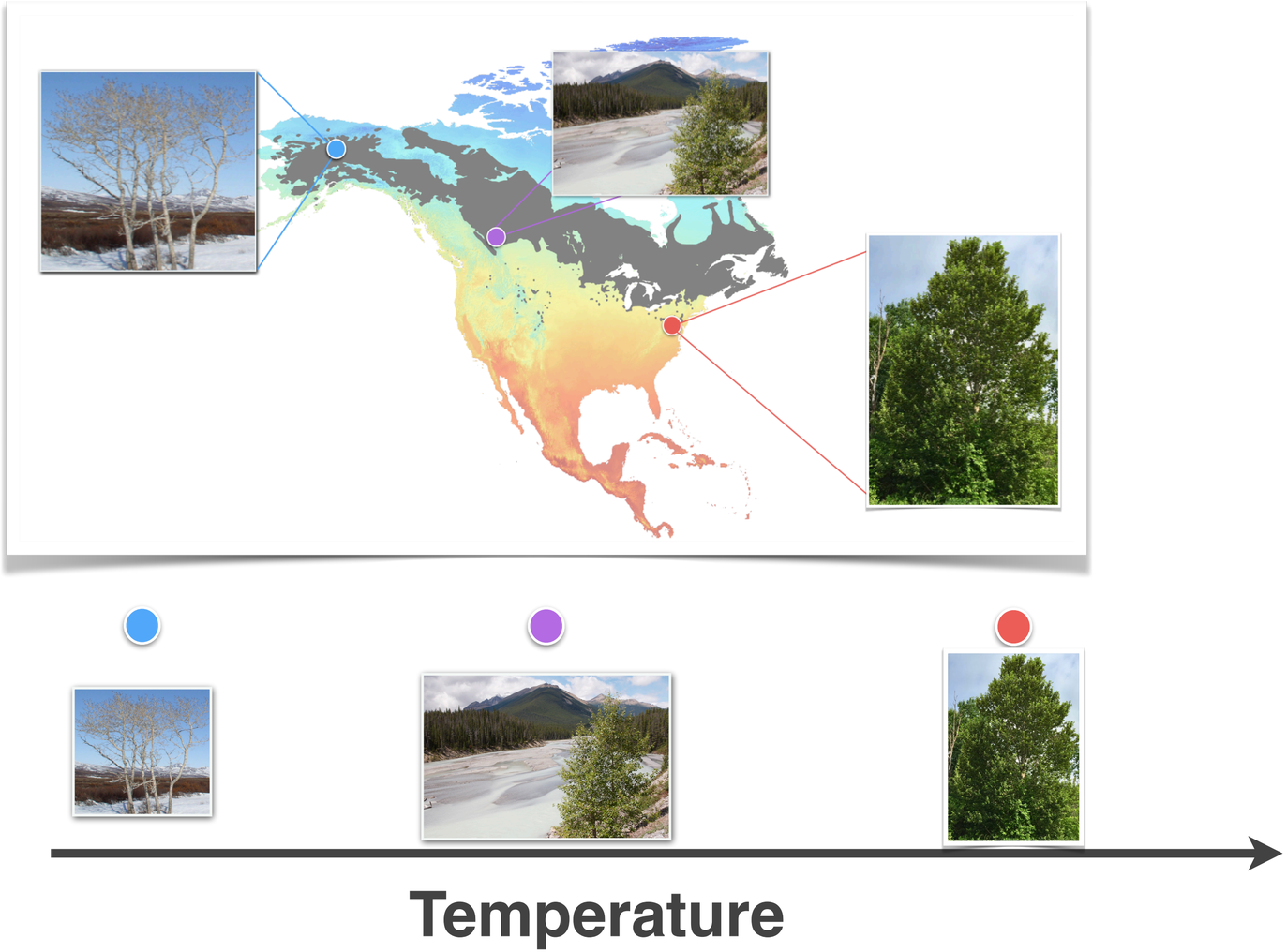


Random Forests

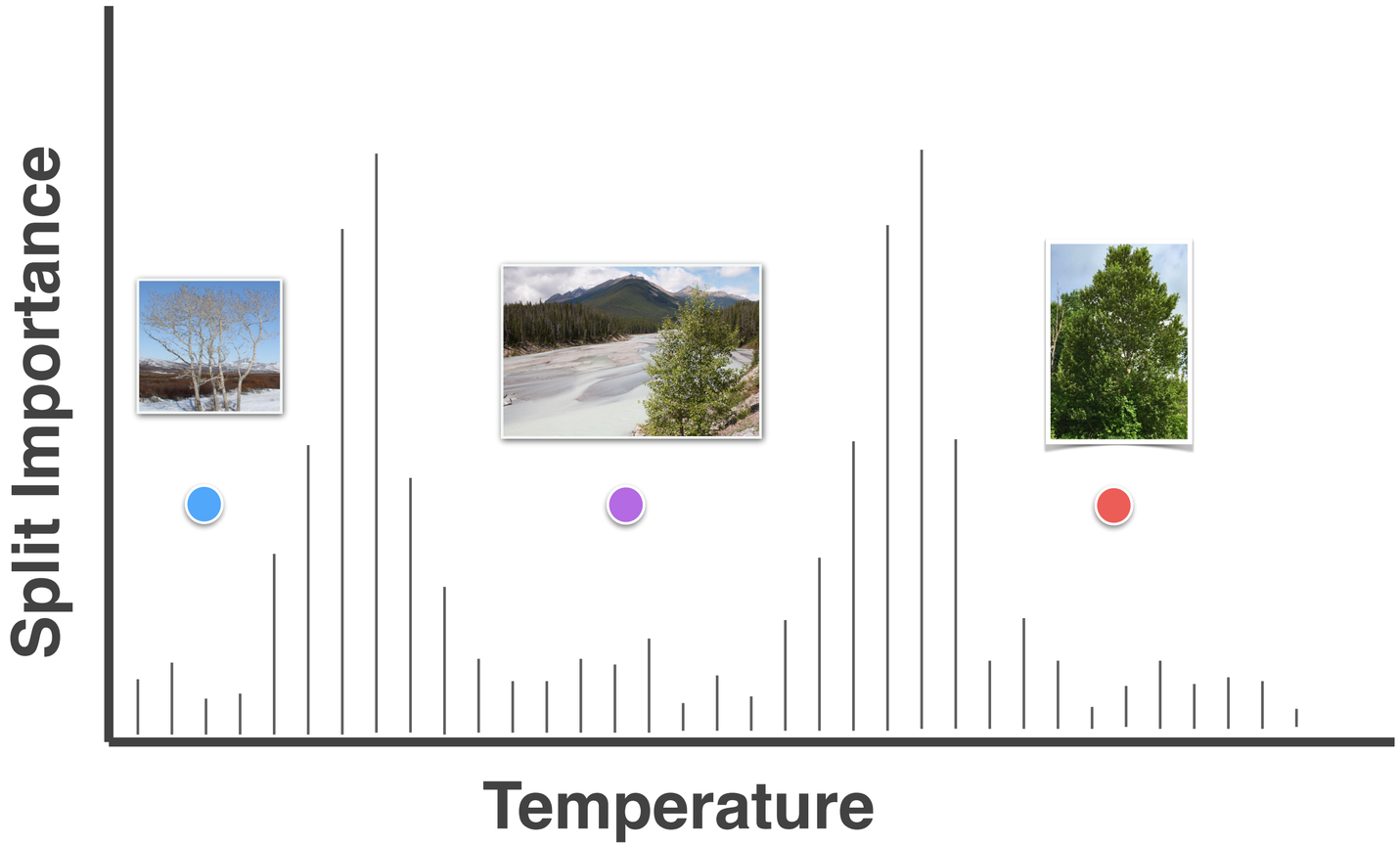




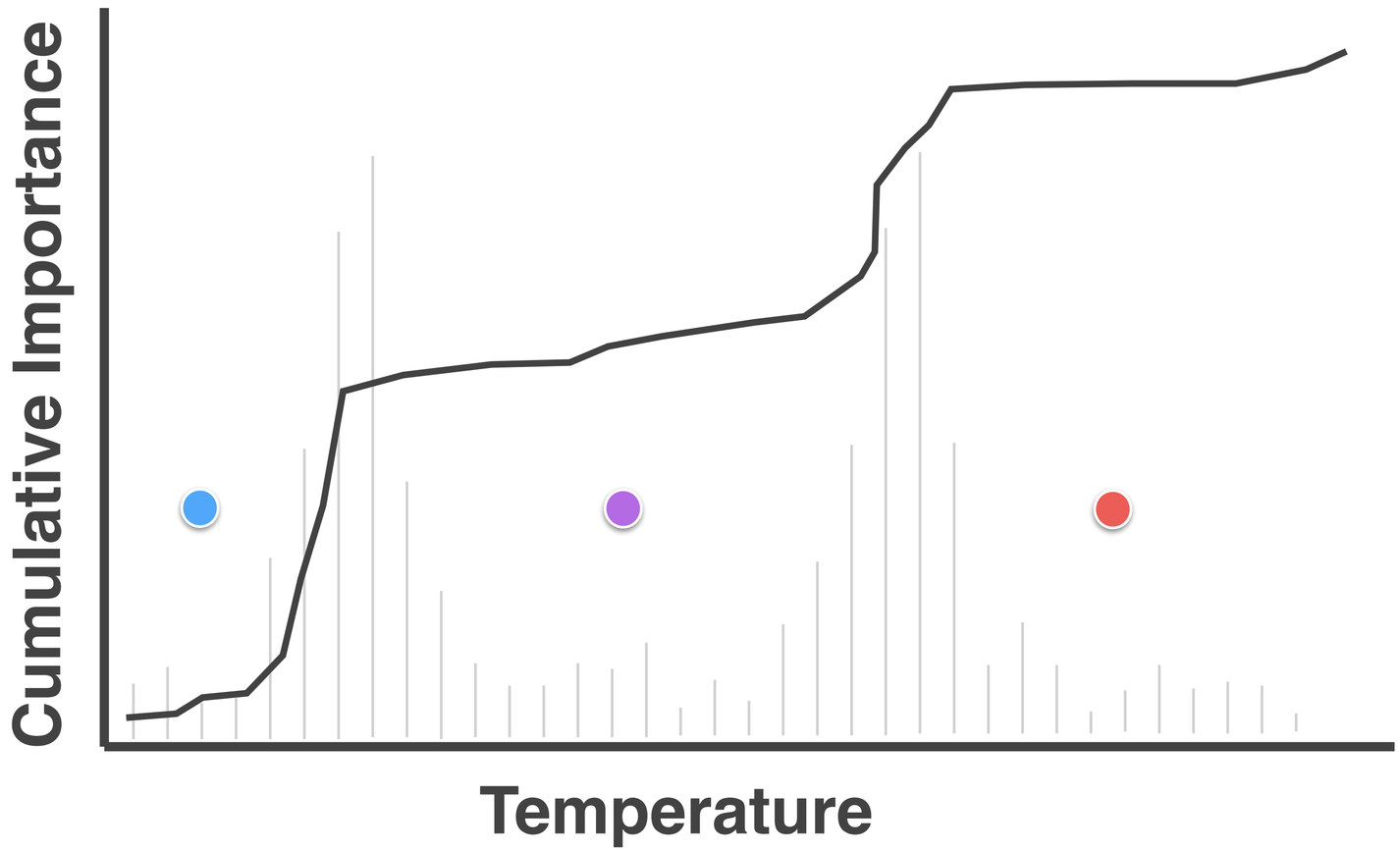
GF: Spatial variation in genome composition as a function of “split importance”



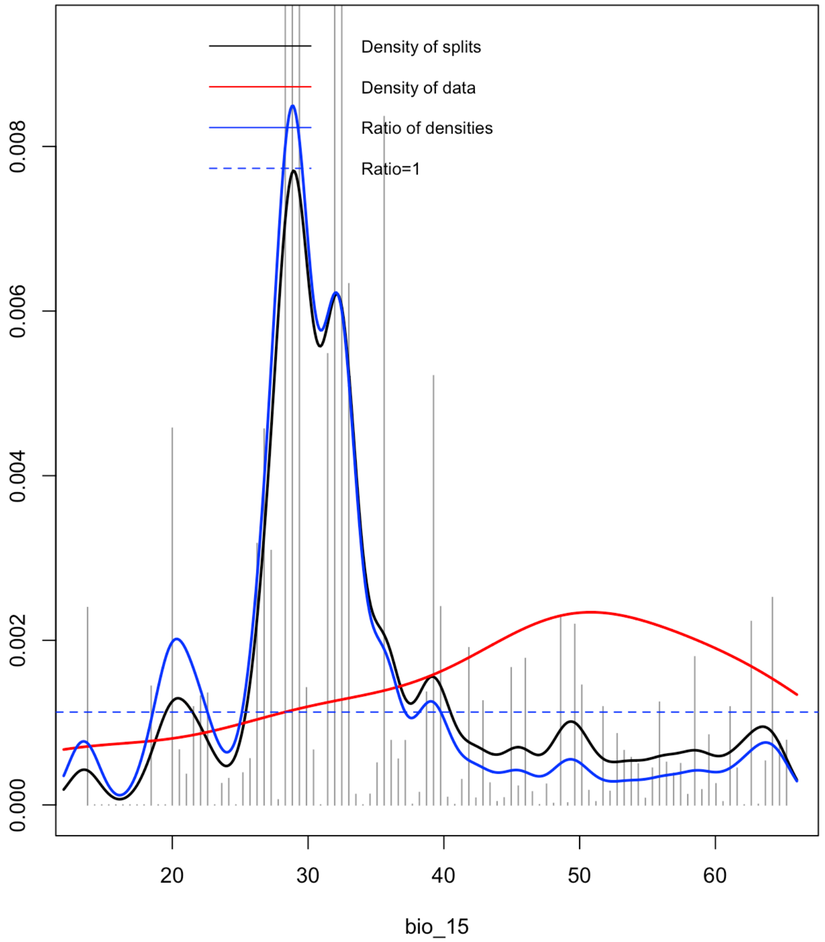
GF: Turnover as a function of “split importance”

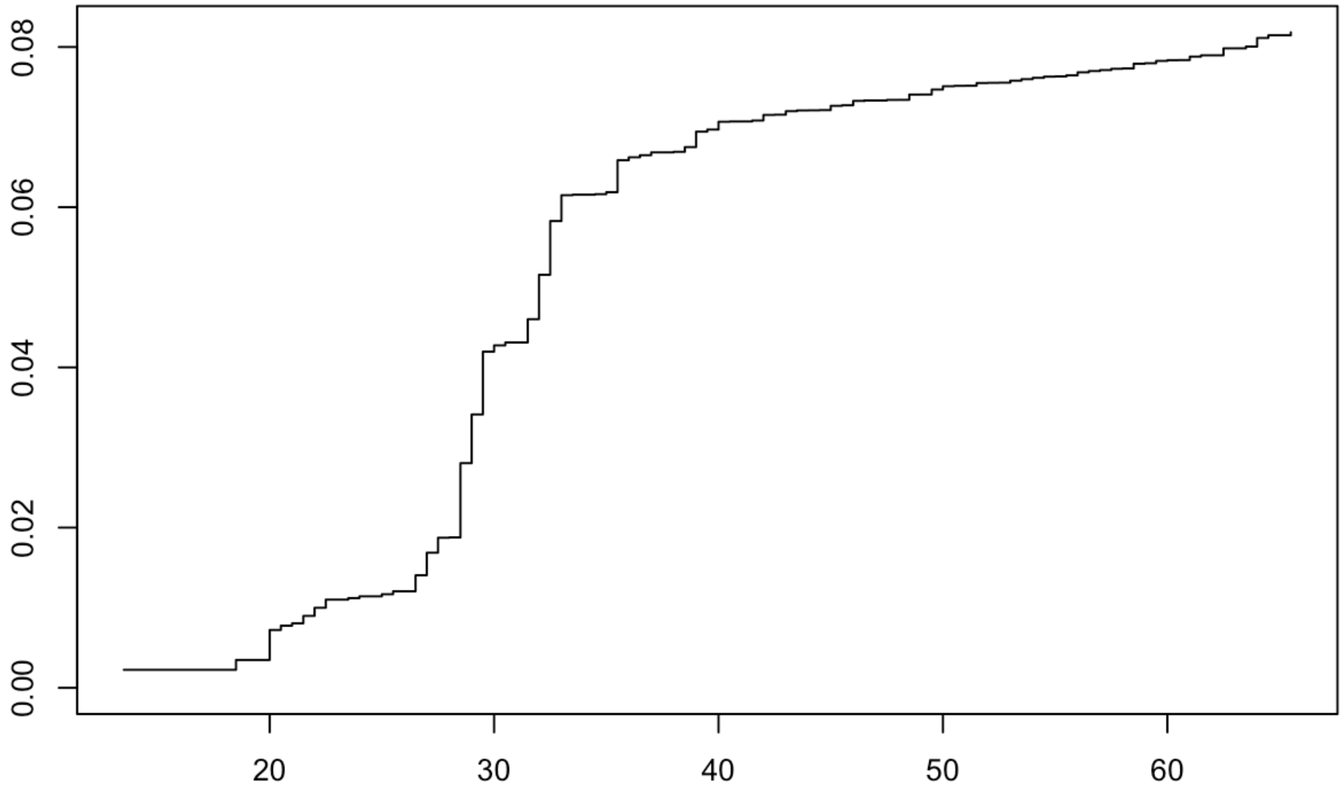


GF: Turnover as a function of “split importance”



GF: Turnover as a function of “split importance”





GF Inputs

Genomic data
(Allele frequencies)

Environmental covariates

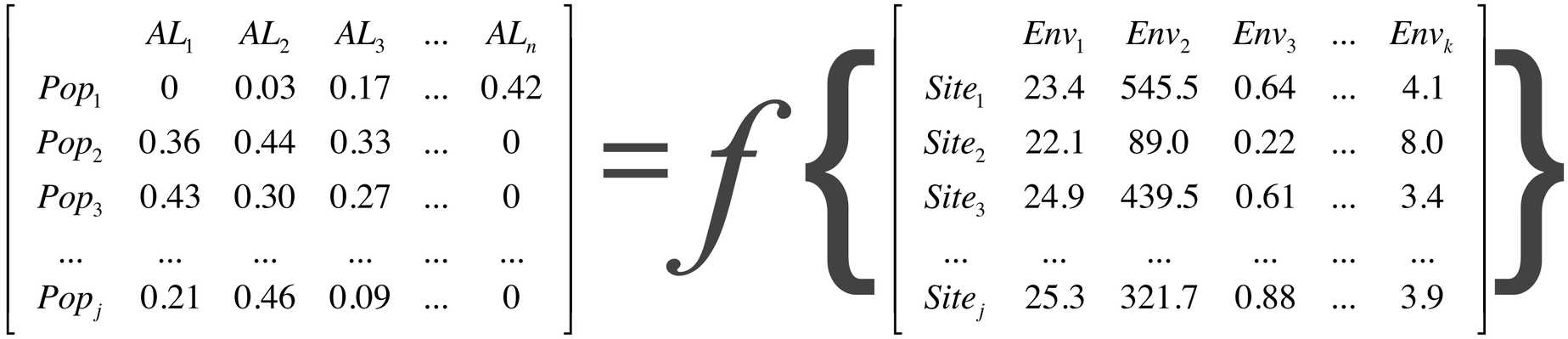
$$\begin{bmatrix} & AL_1 & AL_2 & AL_3 & \dots & AL_n \\ Pop_1 & 0 & 0.03 & 0.17 & \dots & 0.42 \\ Pop_2 & 0.36 & 0.44 & 0.33 & \dots & 0 \\ Pop_3 & 0.43 & 0.30 & 0.27 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ Pop_j & 0.21 & 0.46 & 0.09 & \dots & 0 \end{bmatrix} = f\left\{ \begin{bmatrix} & Env_1 & Env_2 & Env_3 & \dots & Env_k \\ Site_1 & 23.4 & 545.5 & 0.64 & \dots & 4.1 \\ Site_2 & 22.1 & 89.0 & 0.22 & \dots & 8.0 \\ Site_3 & 24.9 & 439.5 & 0.61 & \dots & 3.4 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ Site_j & 25.3 & 321.7 & 0.88 & \dots & 3.9 \end{bmatrix} \right\}$$

- Biological data
  - site (rows) x allele frequencies (columns) table
    - Allele frequencies - basically any continuous response of interest or binomial
    - Transforming response data suggested by Ellis et al. (2012) -my experience has been that this is not necessarily straightforward

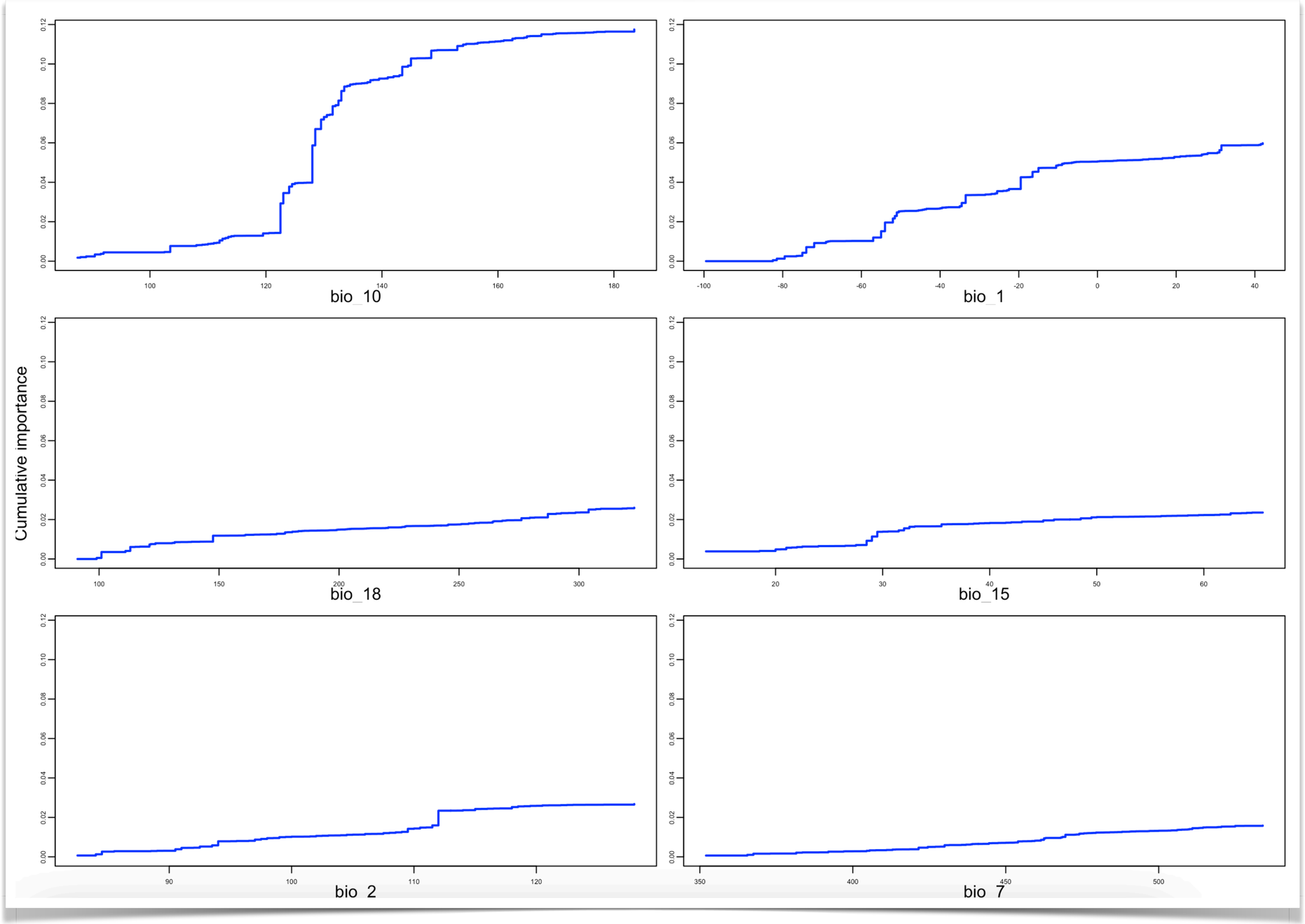
GF Inputs

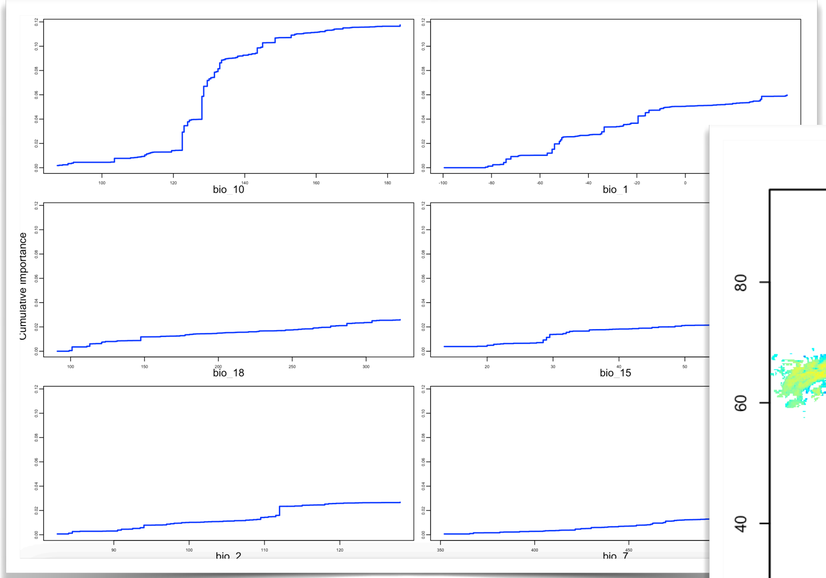
Genomic data
(Allele frequencies)

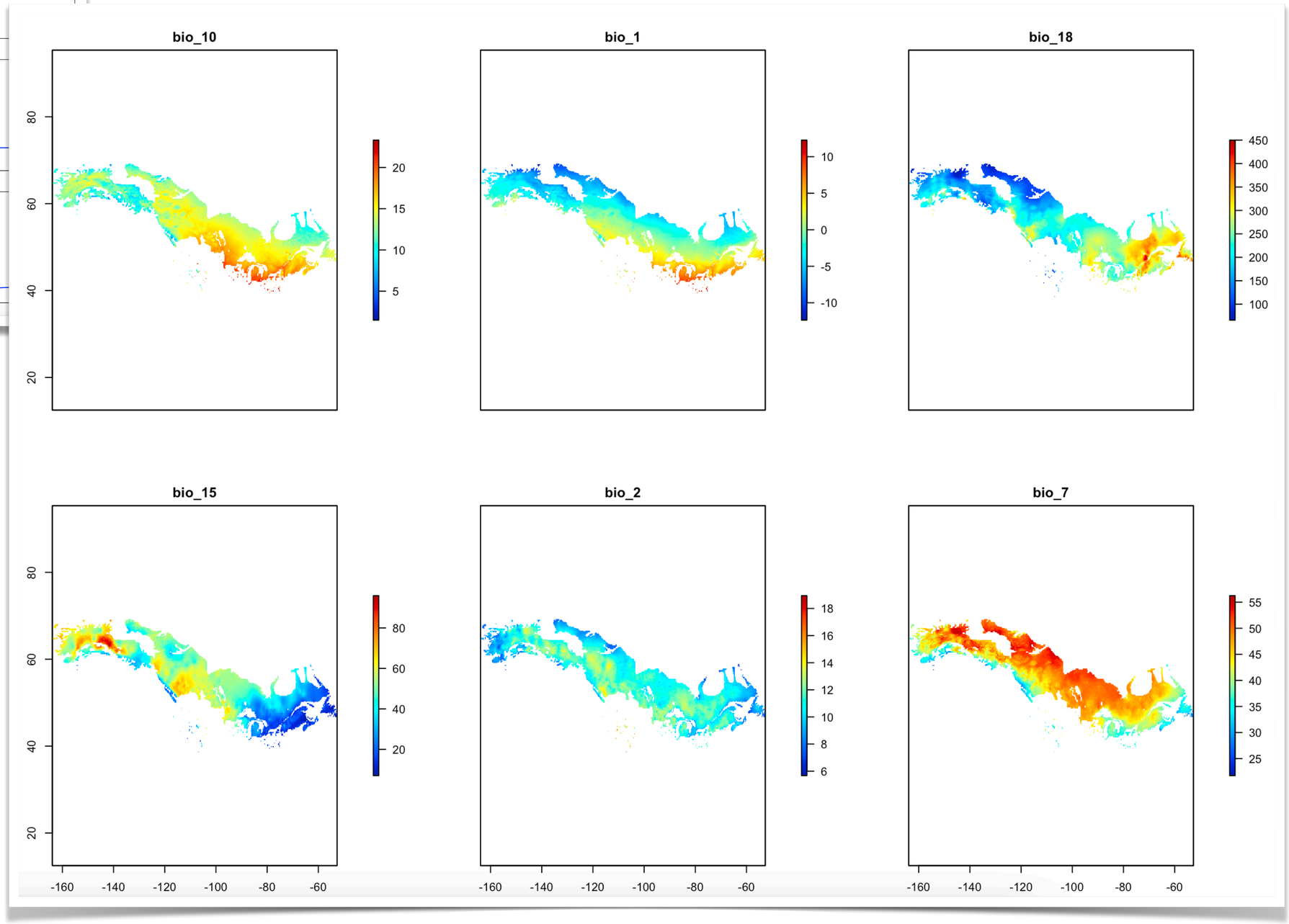
Environmental covariates

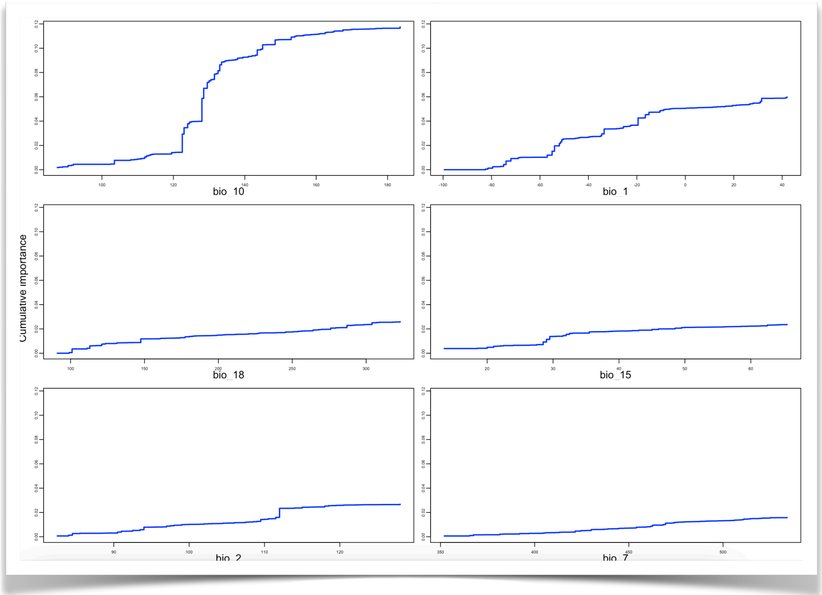


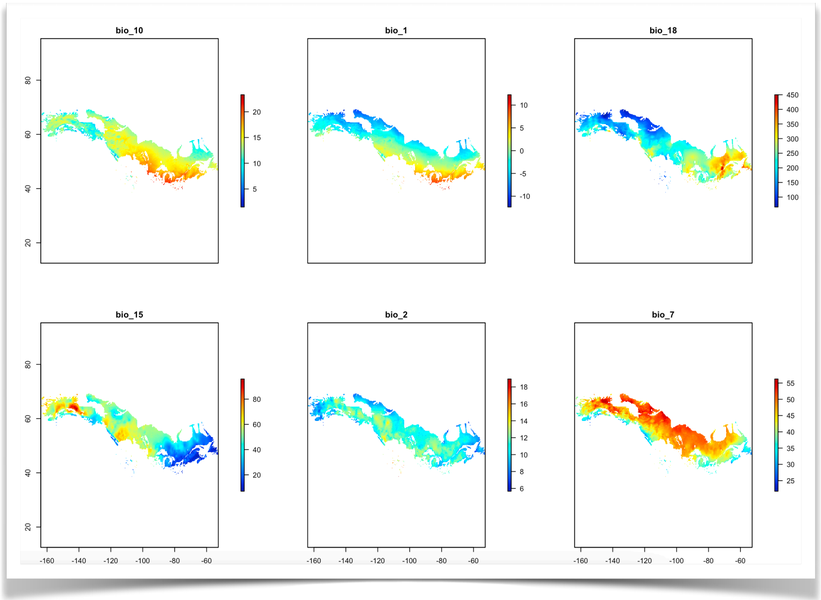
- Environmental data
  - Continuous & categorical
  - Covariates to capture spatial structure (e.g., Moran's Eigenvector Maps)

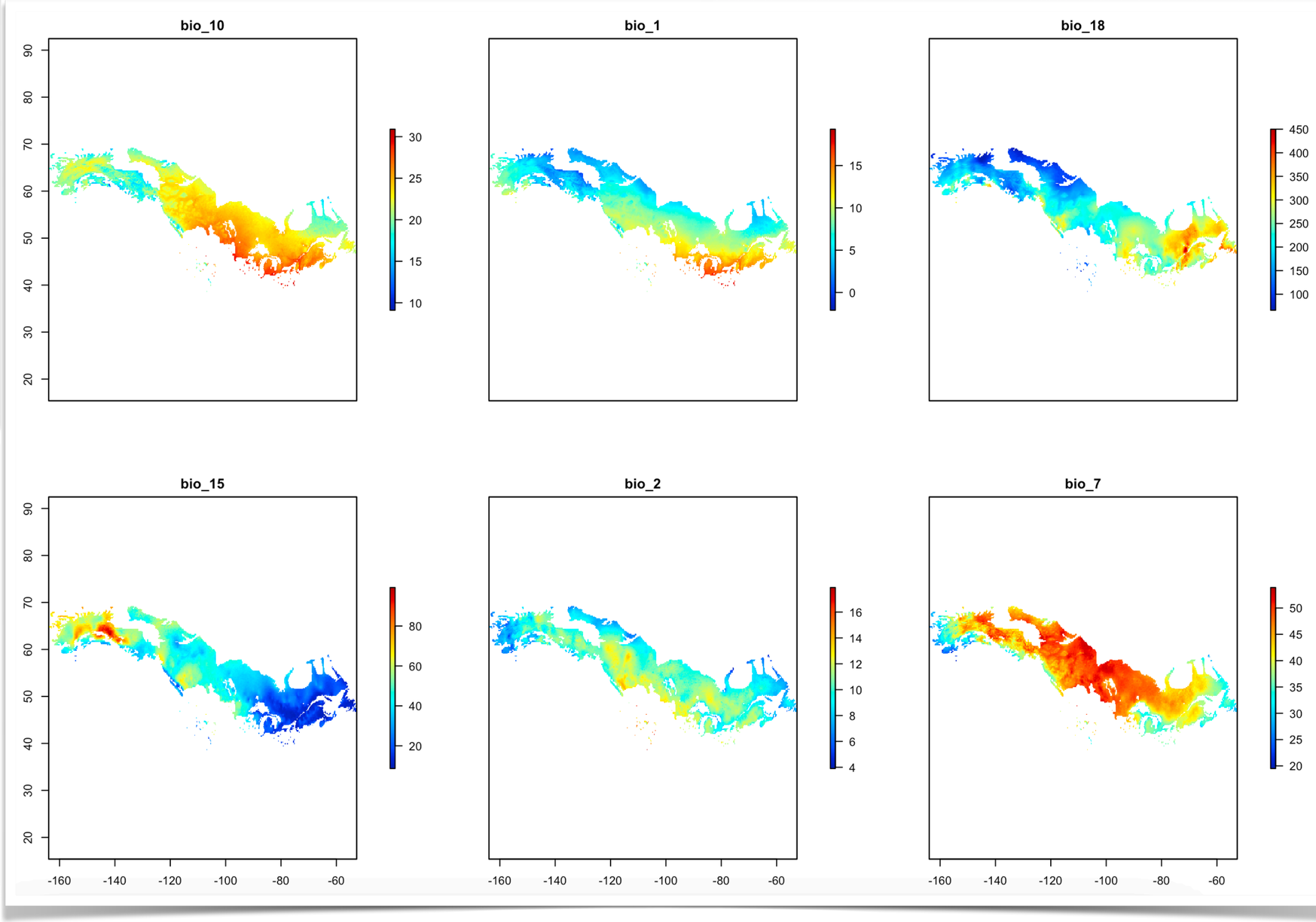

Cumulative importance
0.00
0.02
0.04
0.06
0.08
0.10
0.12
100
120
140
160
180
bio_10
-100
-80
-60
-40
-20
0
20
40
bio_1
100
150
200
250
300
bio_18
20
30
40
50
60
bio_15
90
100
110
120
bio_2
350
400
450
500
bio_7


Cumulative importance
bio_10
bio_1
bio_18
bio_15
bio_2
bio_7

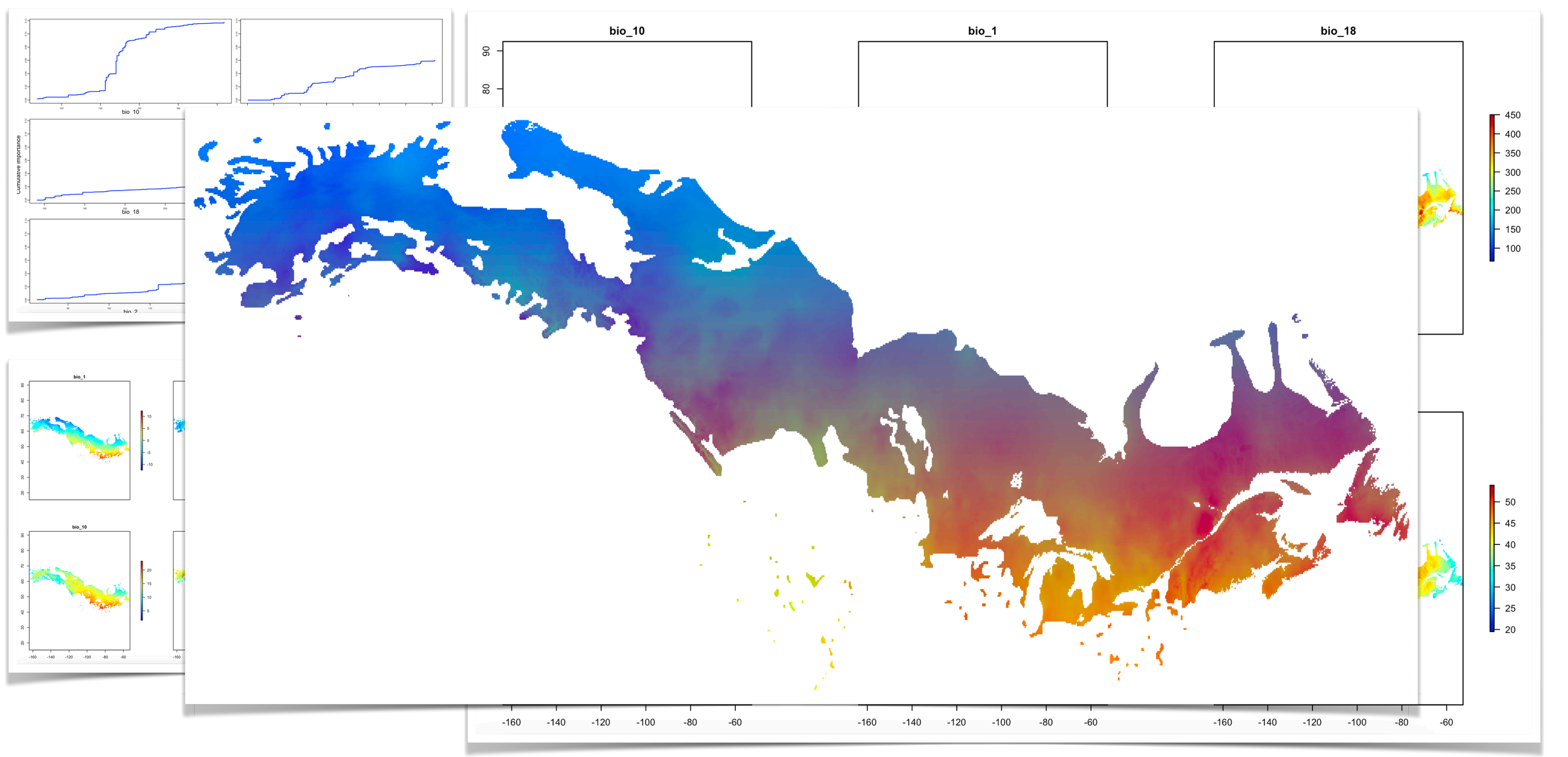

bio_10
bio_1
bio_18
bio_15
bio_2
bio_7
-160
-140
-120
-100
-80
-60
20
40
60
80





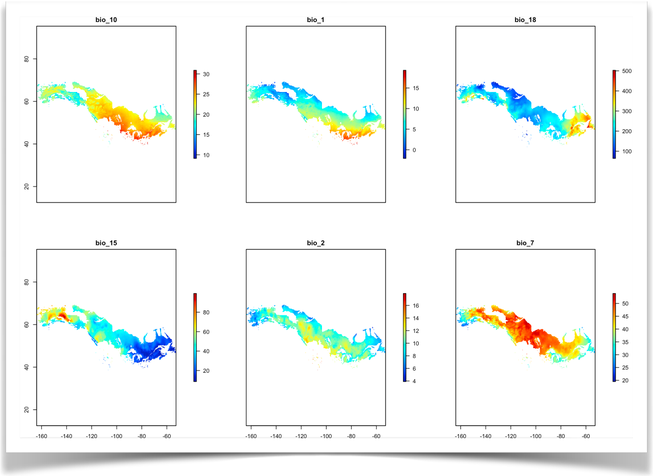


predict.gradientForest()

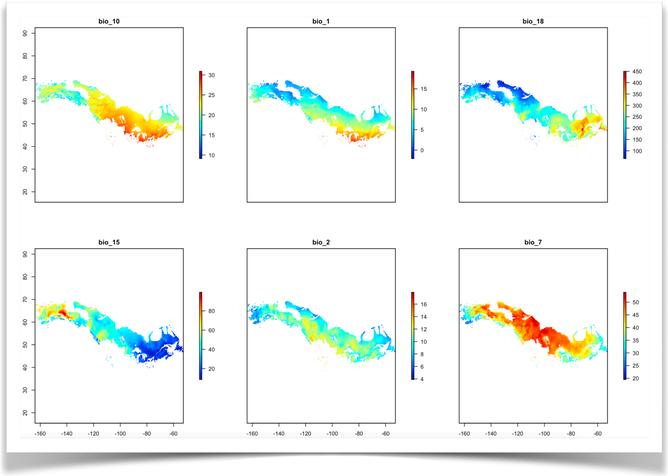

bio_10
bio_1
bio_18
Cumulative importance
bio_10
bio_18
bio_1
bio_10
90
80
-160
-140
-120
-100
-80
-60
450
400
350
300
250
200
150
100
50
45
40
35
30
25
20

Genetic offset = Euclidean distance between GF-transformed env. spaces

Transformed future climate



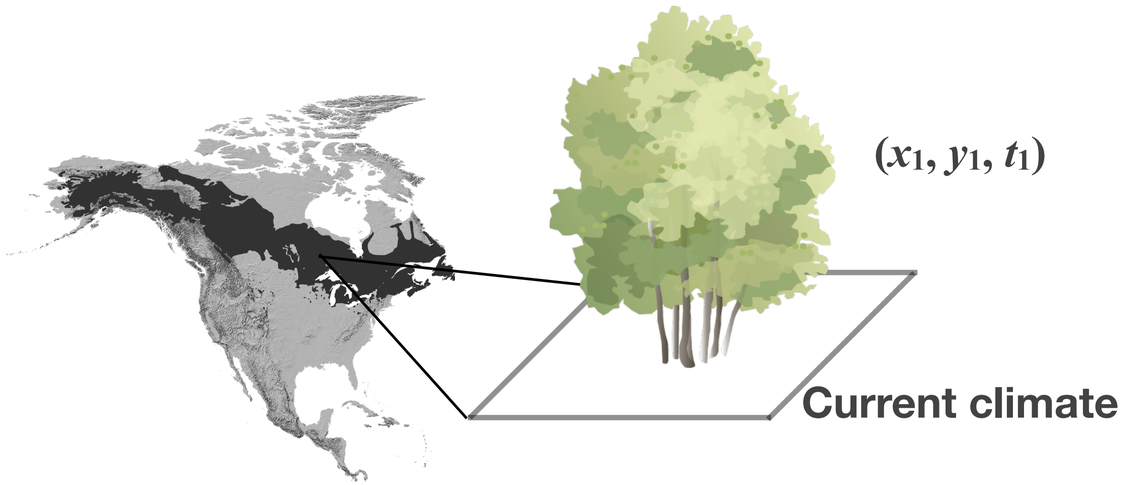
Transformed current climate



GF genetic offset increases with the number of predictor variables in the model

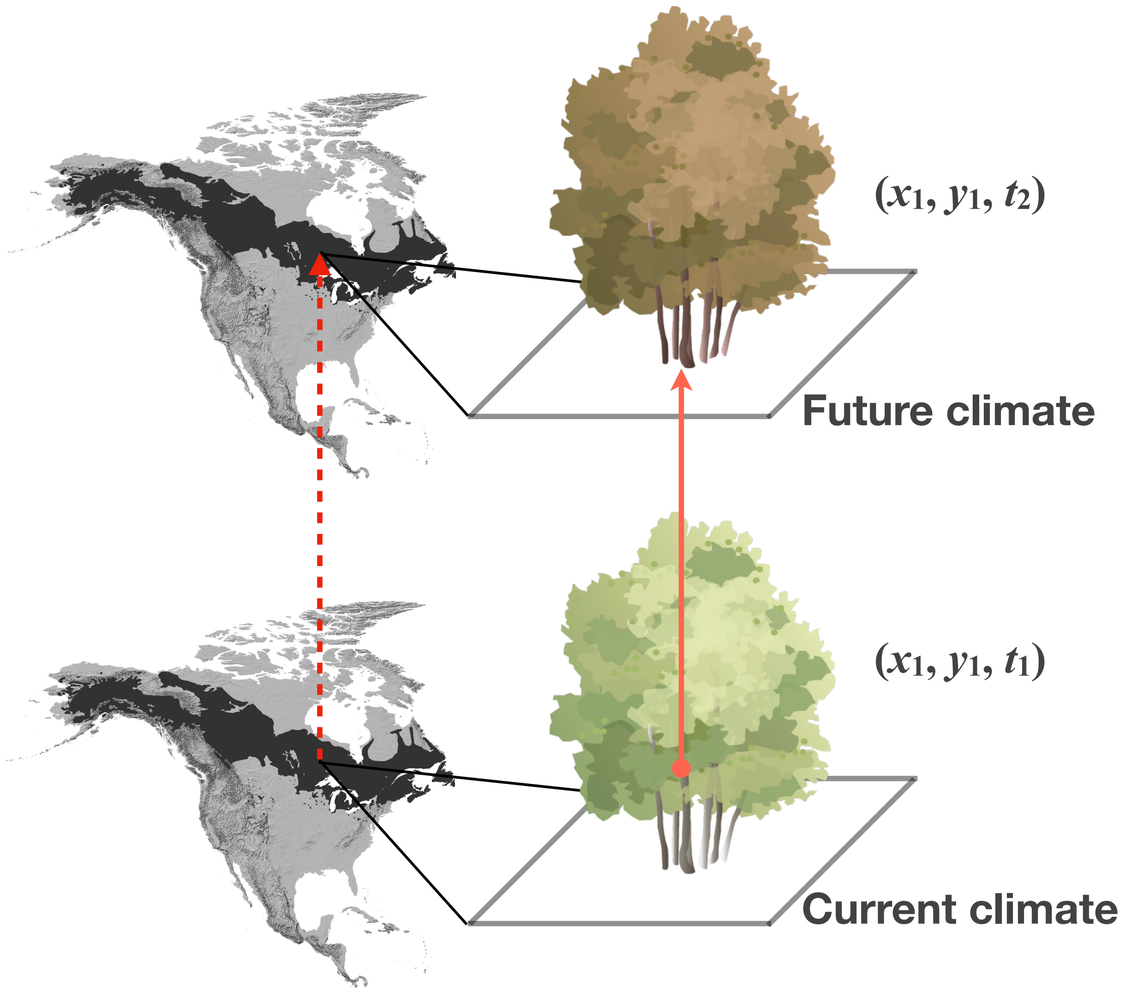
Gougherty et al. (2021) Nature Climate Change

“Local offset”: *in situ* disruption of gene-climate associations



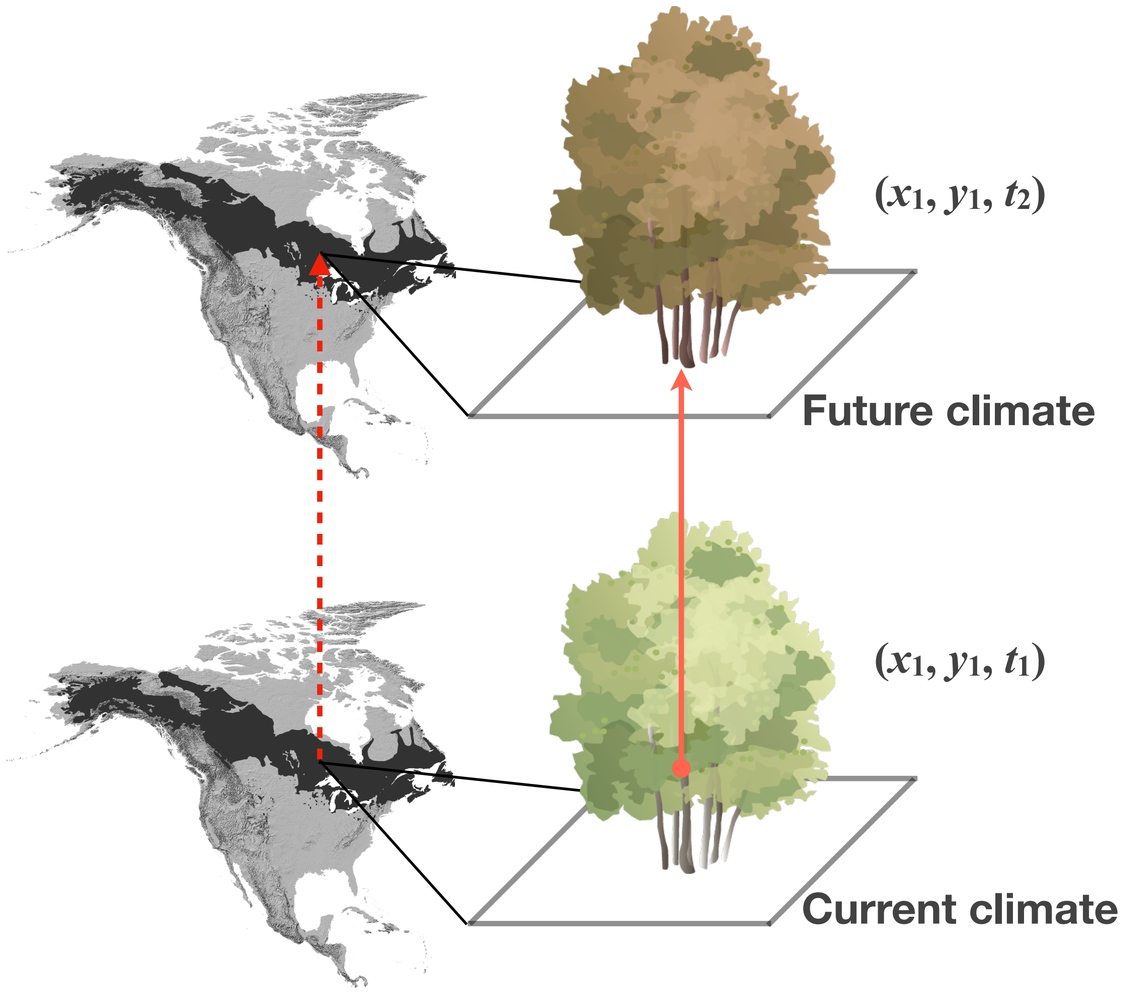
Gougherty et al. (2021) Nature Climate Change

“Local offset”: *in situ* disruption of gene-climate associations



Gougherty et al. (2021) Nature Climate Change

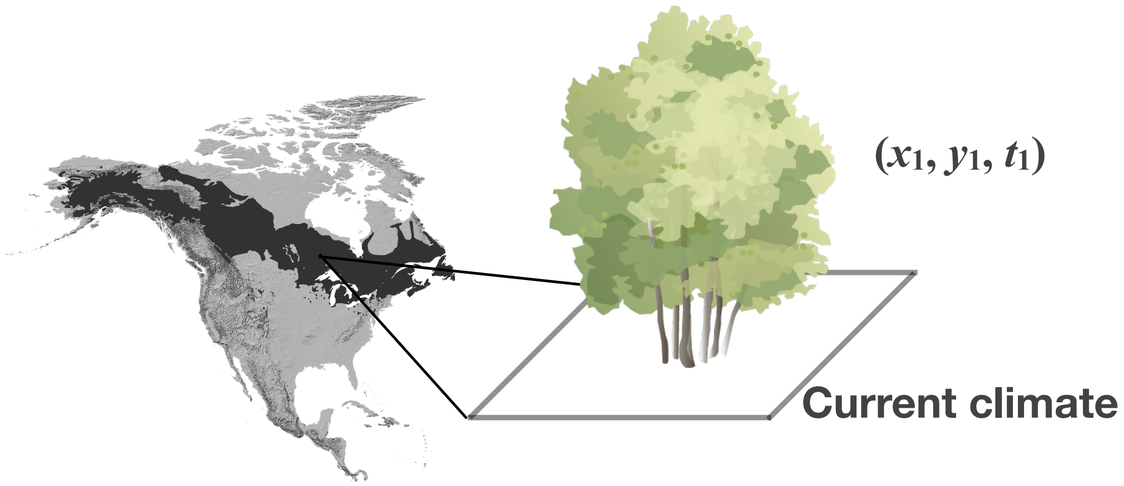
“Local offset”: *in situ* disruption of gene-climate associations



“Local offset”: How maladapted will this population be given its exposure to climate change in its current location?

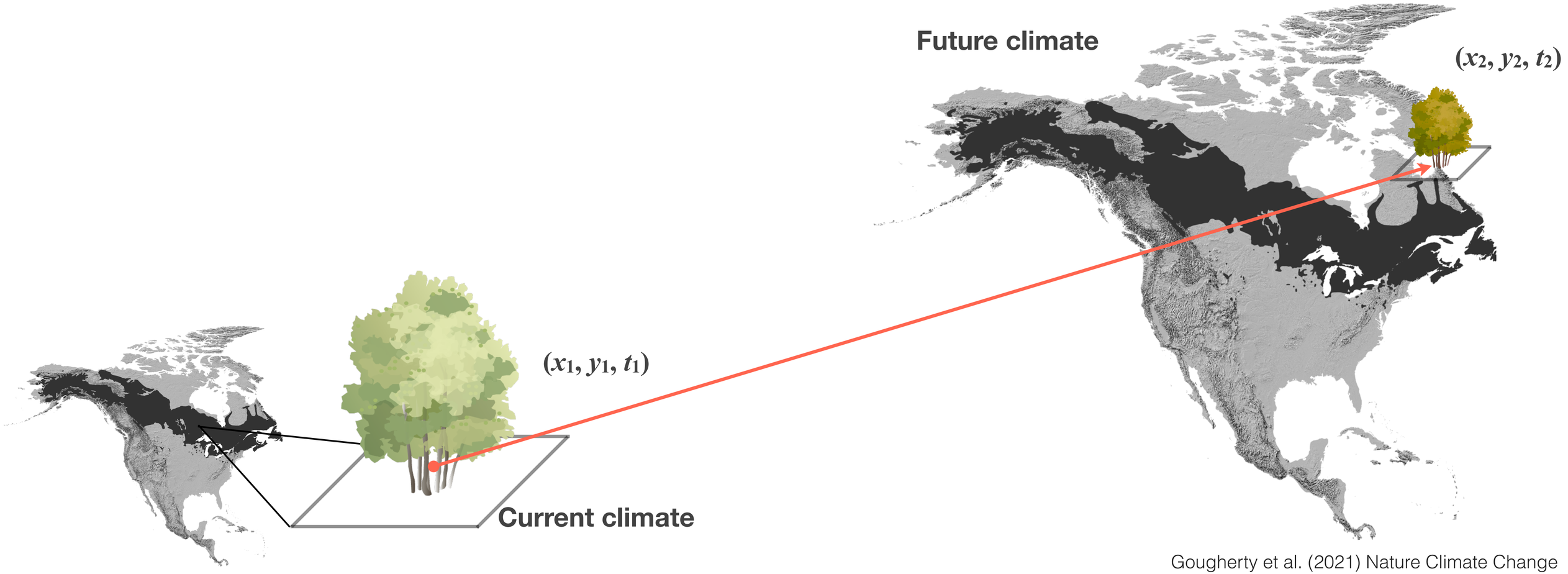
Gougherty et al. (2021) Nature Climate Change

“Forward offset”: disruption of gene-climate associations assuming migration



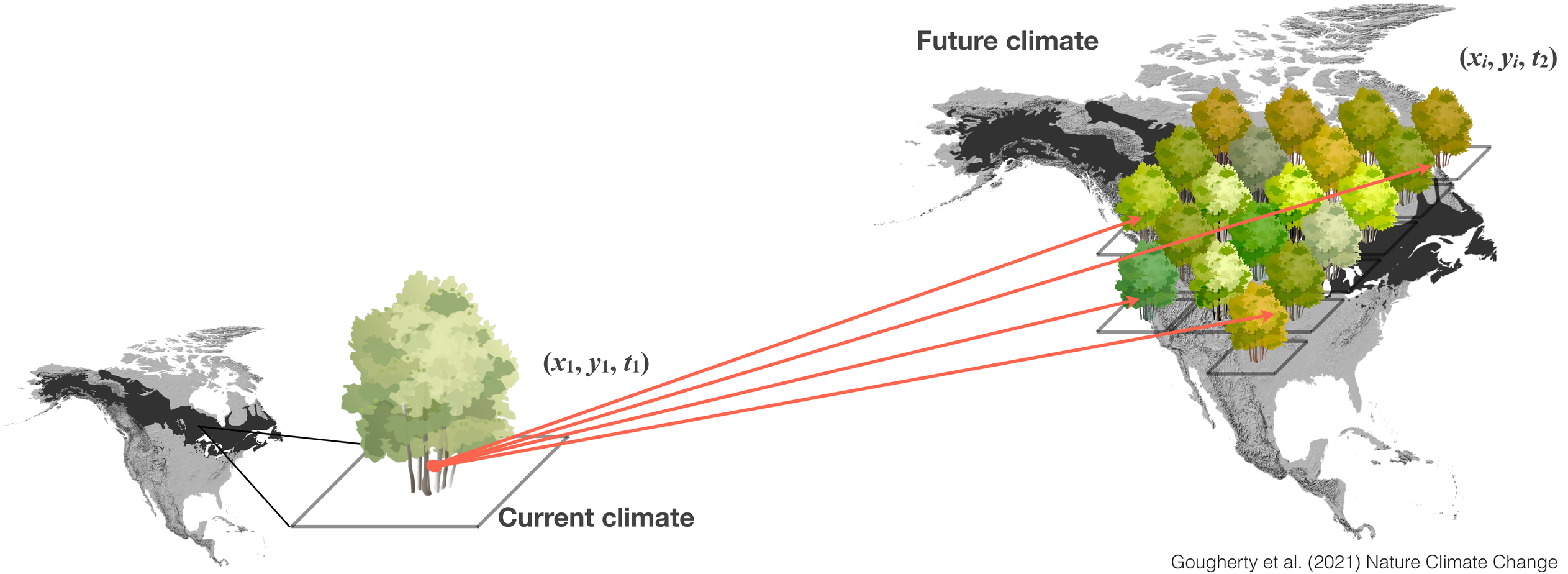
Gougherty et al. (2021) Nature Climate Change

“Forward offset”: disruption of gene-climate associations assuming migration



Gougherty et al. (2021) Nature Climate Change

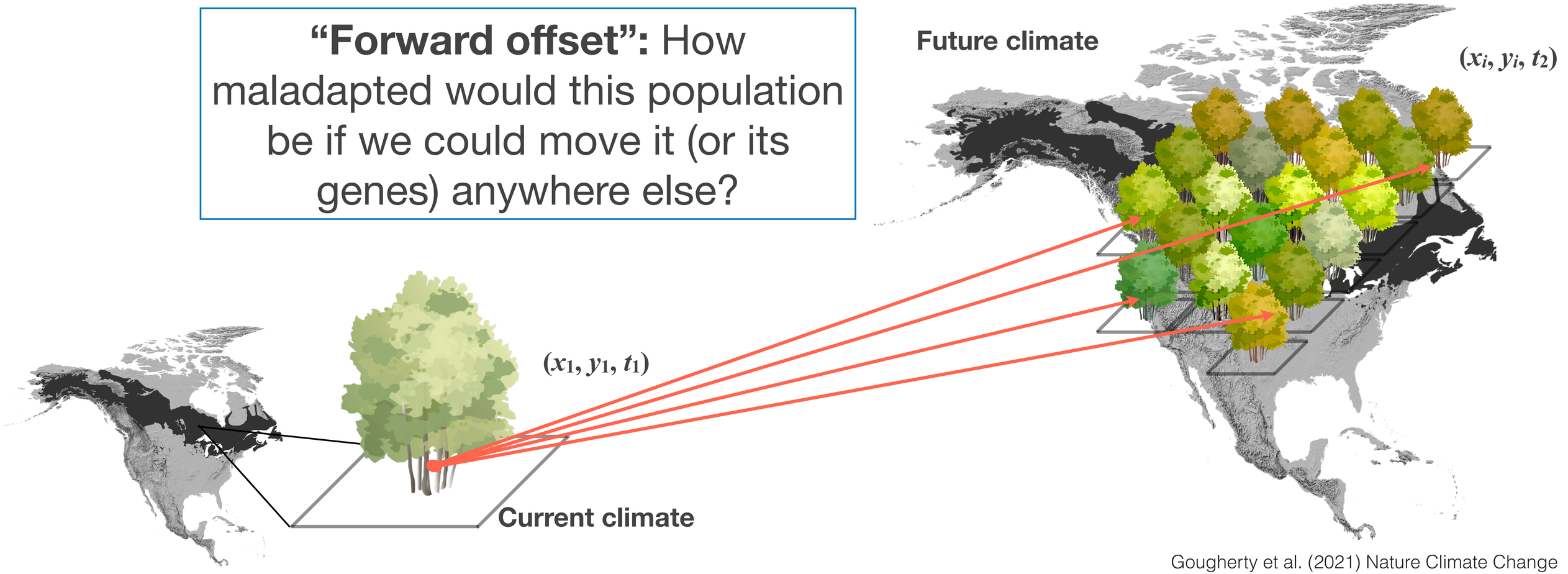
“Forward offset”: disruption of gene-climate associations assuming migration



Gougherty et al. (2021) Nature Climate Change

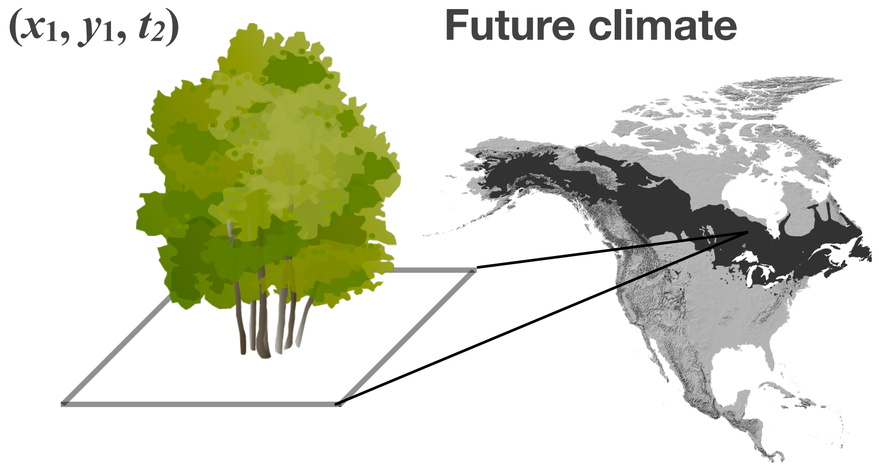
“Forward offset”: disruption of gene-climate associations assuming migration

“Forward offset”: How maladapted would this population be if we could move it (or its genes) anywhere else?



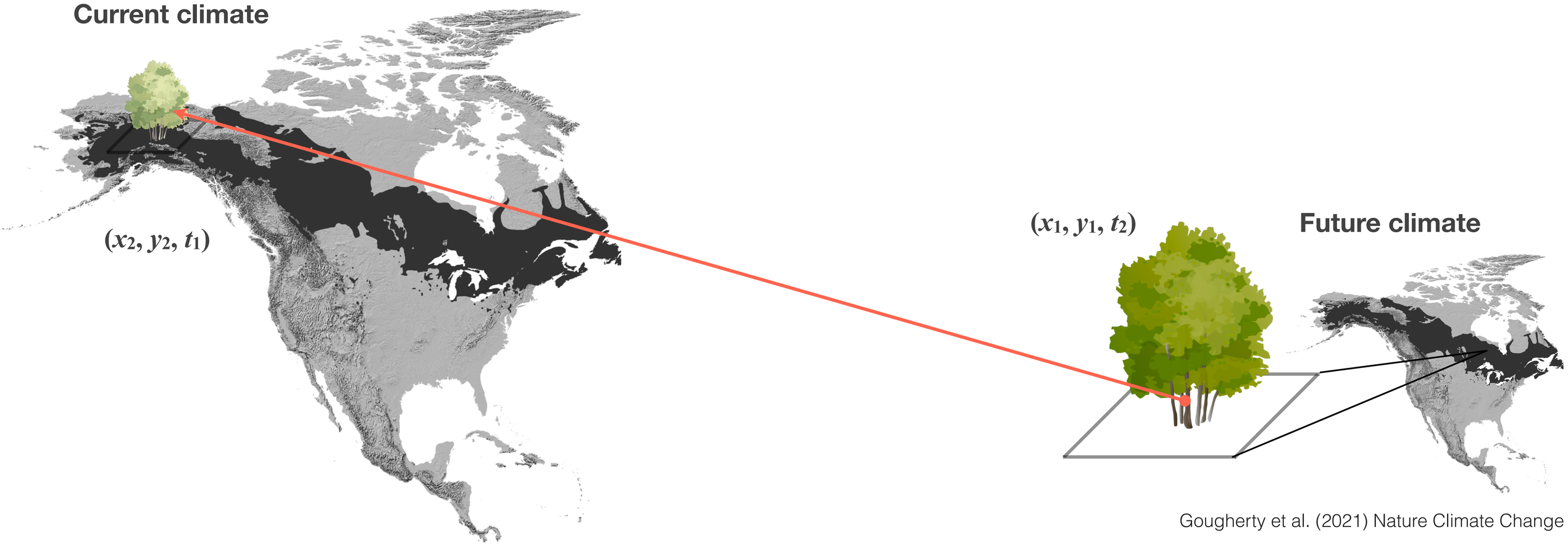
Gougherty et al. (2021) Nature Climate Change

“Reverse offset”: preadaptation of existing populations to future climate



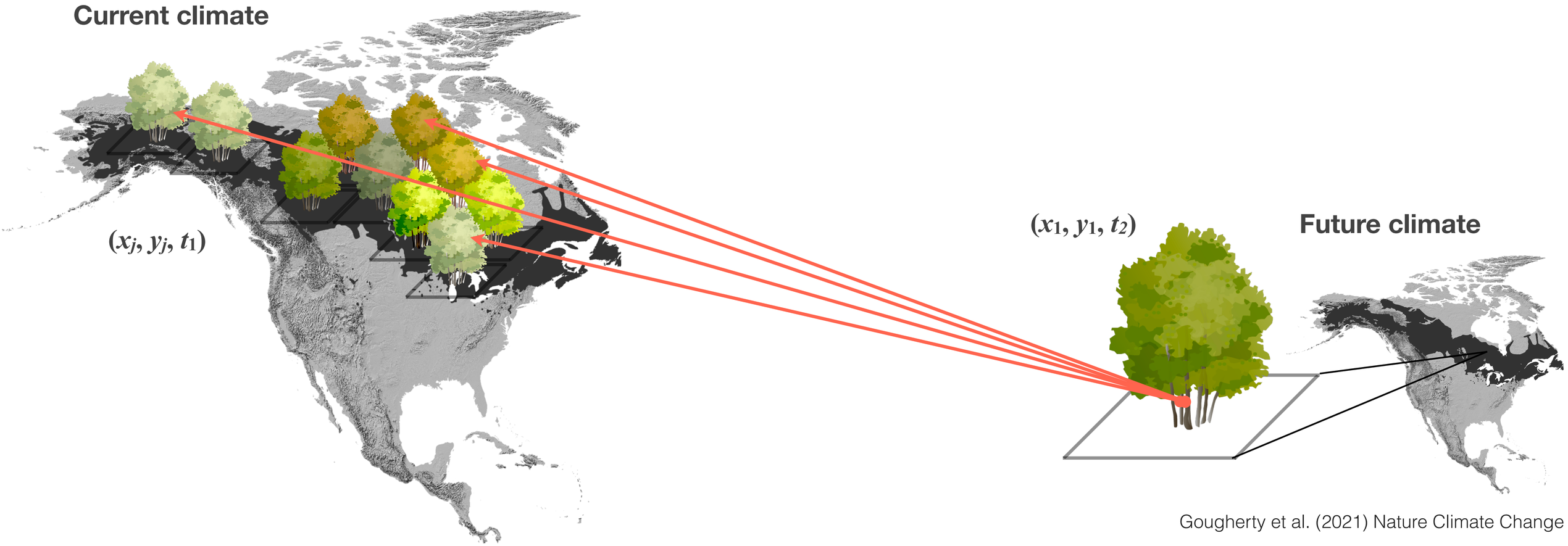
Gougherty et al. (2021) Nature Climate Change

“Reverse offset”: preadaptation of existing populations to future climate



Gougherty et al. (2021) Nature Climate Change

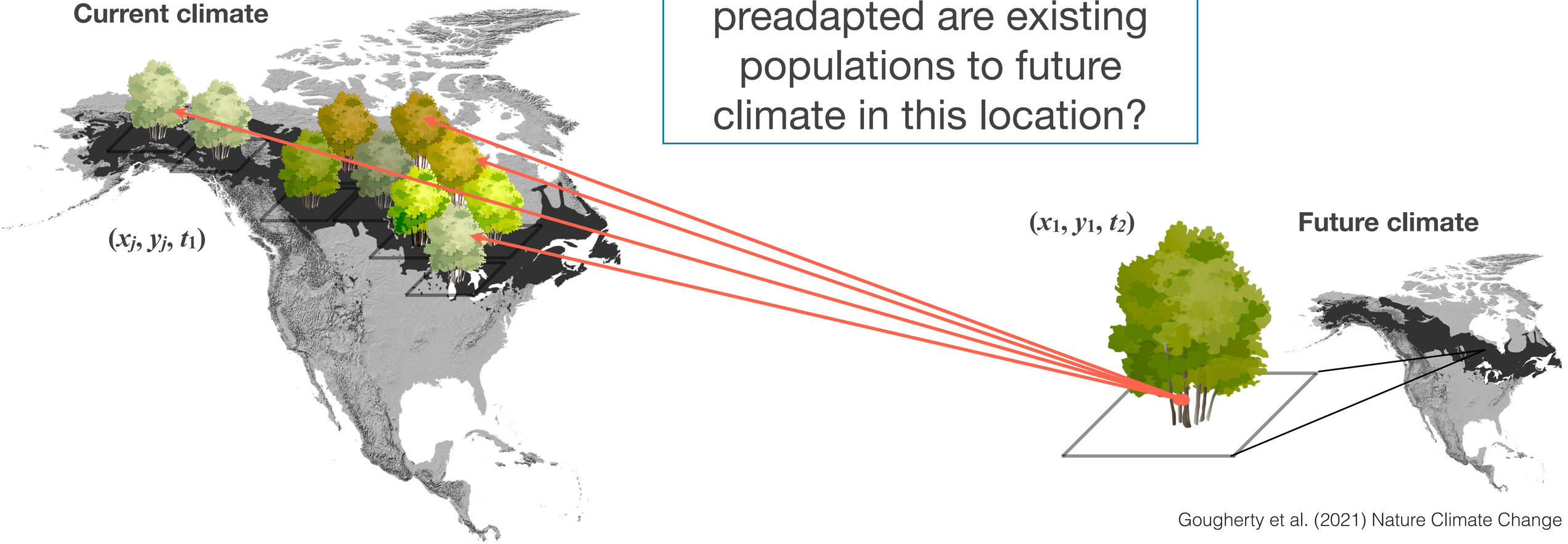
“Reverse offset”: preadaptation of existing populations to future climate



Gougherty et al. (2021) Nature Climate Change

“Reverse offset”: preadaptation of existing populations to future climate

“Reverse offset”: How preadapted are existing populations to future climate in this location?



Gougherty et al. (2021) Nature Climate Change

Climate change risk =

$$f(\textit{local offset}, \textit{forward offset}, \textit{reverse offset})$$

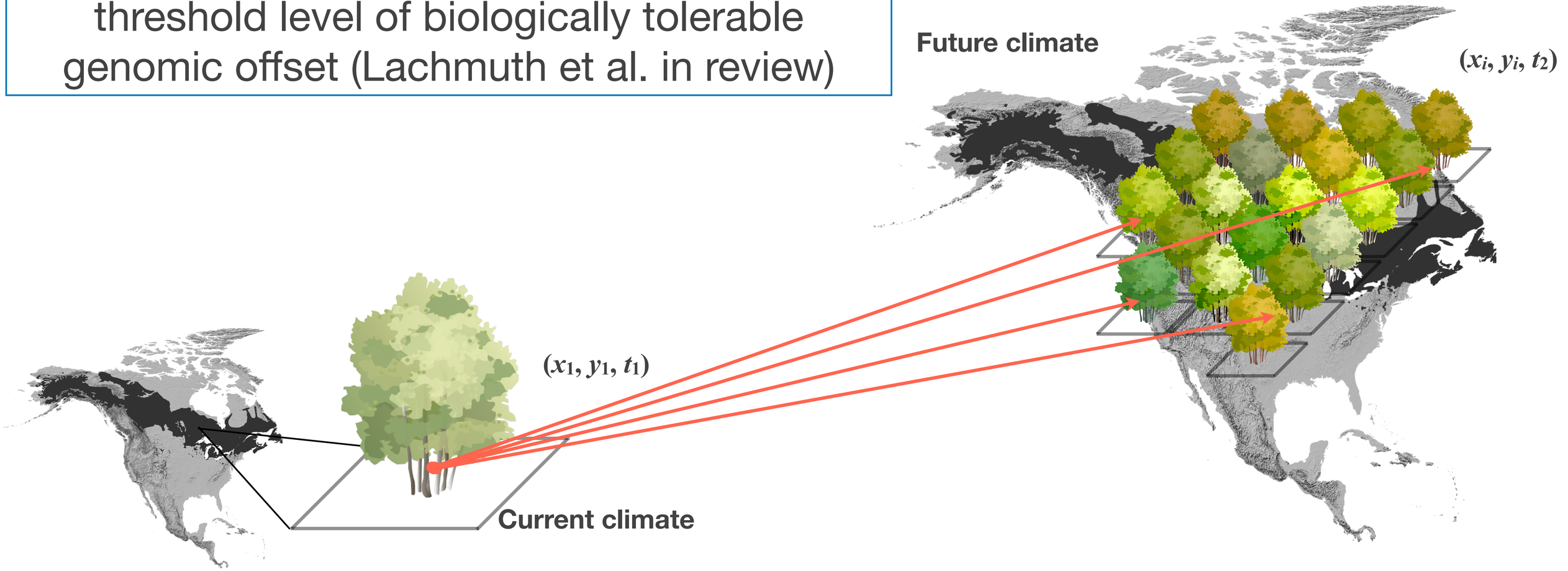
in situ maladaptation
assuming no migration

Possibility of extirpation
assuming migration

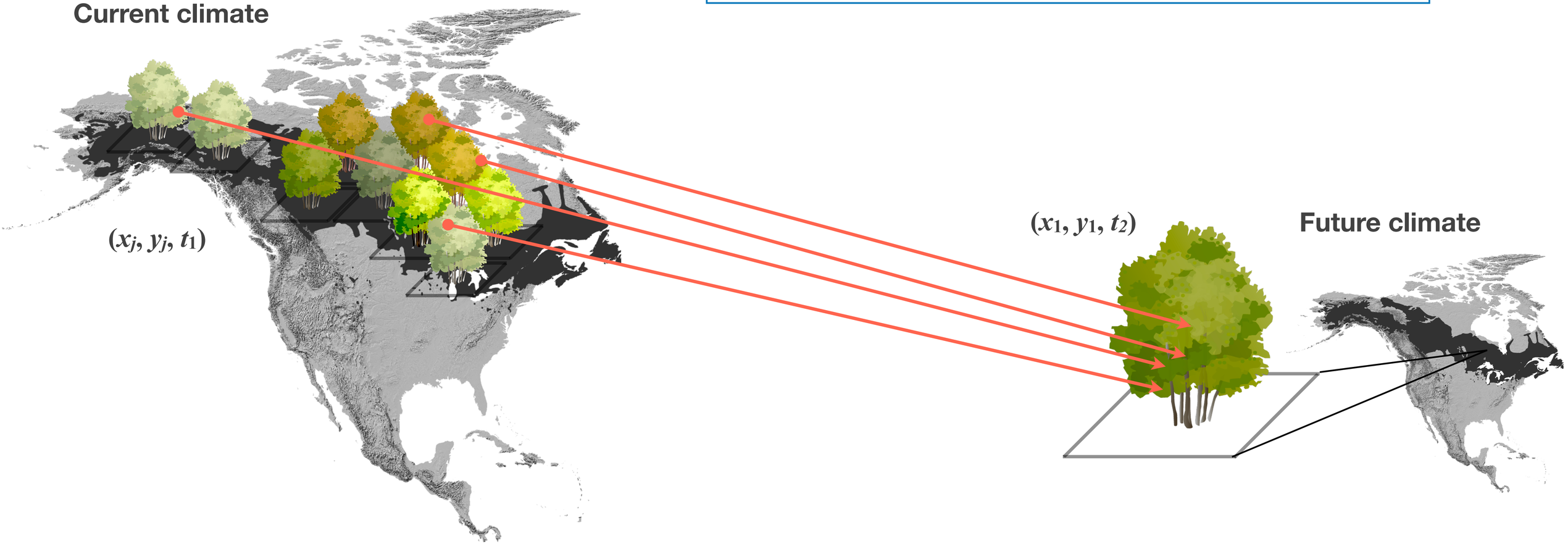
Preadaptation to future climates

Gougherty et al. (2021) Nature Climate Change

“Donor importance”: the percentage of recipient locations to which a given population could be transferred without exceeding a threshold level of biologically tolerable genomic offset (Lachmuth et al. in review)



“Recipient importance”: the % of donor populations that could be transferred to a given location, each without exceeding the tolerable offset threshold (Lachmuth et al. in review).



nature climate change

ARTICLES

https://doi.org/10.1038/s41558-020-00968-6

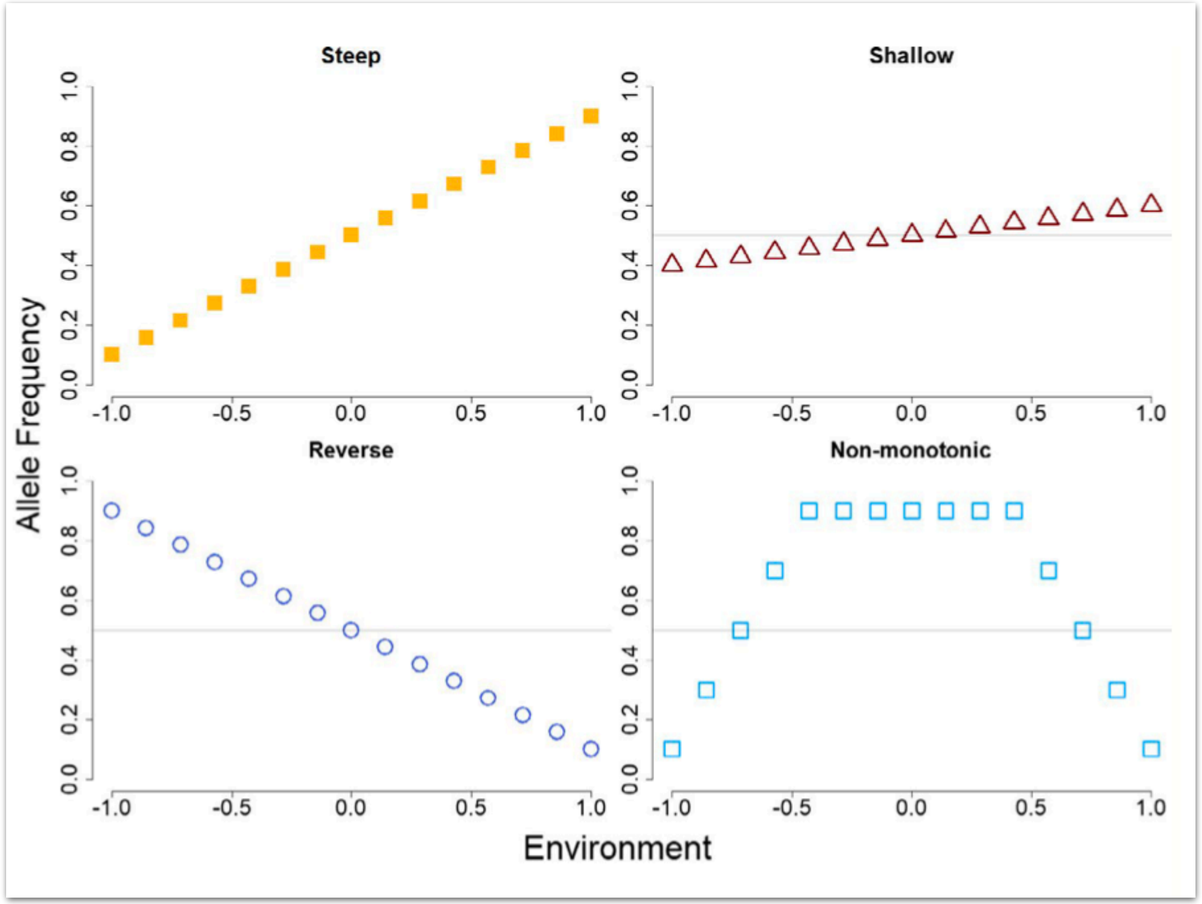
Check for updates

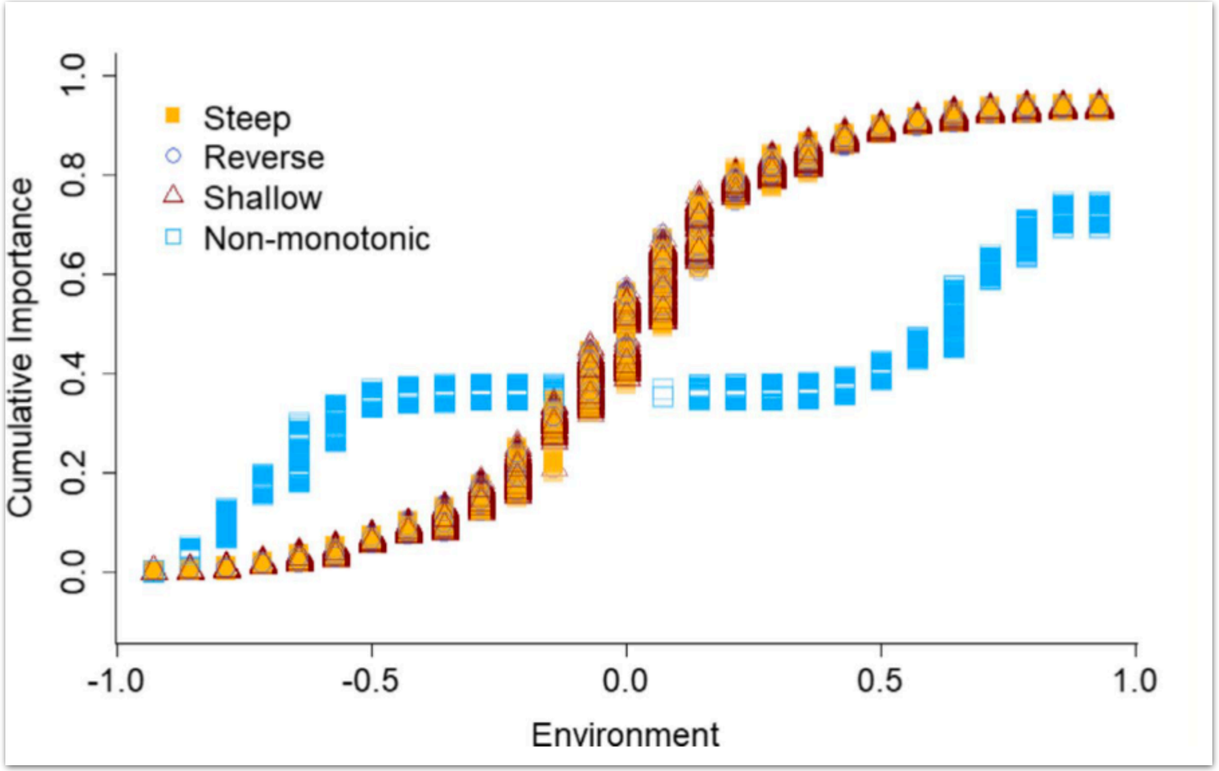
Maladaptation, migration and extirpation fuel climate change risk in a forest tree species

Andrew V. Gougherty[1,3]✉, Stephen R. Keller[2] and Matthew C. Fitzpatrick[1]

Lachmuth et al. (in review) Novel genomic offset metrics account for local adaptation in climate suitability forecasts and inform assisted migration.

Cautions


Steep
Shallow
Reverse
Non-monotonic
Allele Frequency
Environment
0.0
0.2
0.4
0.6
0.8
1.0
-1.0
-0.5
0.0
0.5
1.0


Steep
Reverse
Shallow
Non-monotonic
Cumulative Importance
Environment
0.0
0.2
0.4
0.6
0.8
1.0
-1.0
-0.5
0.0
0.5
1.0

Gradient Forests: Further reading

- Ellis, N., Smith, S. J., & Pitcher, C. R. (2012). Gradient forests: calculating importance gradients on physical predictors. Ecology, 93(1), 156-168.

- Fitzpatrick, M. C., & Keller, S. R. (2015). Ecological genomics meets community-level modelling of biodiversity: Mapping the genomic landscape of current and future environmental adaptation. Ecology letters, 18(1), 1-16.

- Gougherty, A. V., Keller, S. R., & Fitzpatrick, M. C. (2021). Maladaptation, migration and extirpation fuel climate change risk in a forest tree species. Nature Climate Change, 11(2), 166-171.

GF website: https://r-forge.r-project.org/R/?group_id=973

Pros / Cons - GDM

- Simple, relatively easy to understand model
- Computationally efficient
- Can easily incorporate space
- Magnitude of genetic offsets have clear(er) interpretation

- No variable interactions
- Model assessment / comparison can be challenging
- Can require rescaling of genetic distances to facilitate model convergence (which would alter predicted offsets)

Pros / Cons - GF

- All aspects of model performance reported as R^2 values
- Can model complex patterns, including interactions
- Functions for each allele and all alleles collectively
- Can combine multiple fitted models
- Categorical predictors

- No easy way to incorporate / deal with spatial structure
- Computationally intense
- “Predictions” are simply transformations of the covariates - no inference on actual changes in allele frequencies
- Magnitude of genetic offsets only comparative in a relative sense, but see Lachmuth et al. (in review)

Thank you!

People

Thibaut Capblancq

Vikram Chhatre

Erica Duda

Simon Ferrier

Andy Gougherty

Natalie Haydt

Susanne Lachmuth

Katie Lotterhos

Steve Keller

Aki Laruson

Karel Mokany

Diego Neito-Lugilde

Raju Soolanayakanally

SSMPG Summer School

Thibaut Capblancq

Olivier Francois

Funding

Matt Fitzpatrick

mfitzpatrick@umces.edu

University of Maryland Center for Environmental Science

Appalachian Lab