Lecture 14 – others

- Outline for today
  - 回顾相关性分析和线性回归
  - 线性回归的变体和延伸
    - 非线性回归 nonlinear regression
    - 逻辑思谛回归 logistic regression
  - 其它常用方法
  - Summary
  - R Lab & Discussion

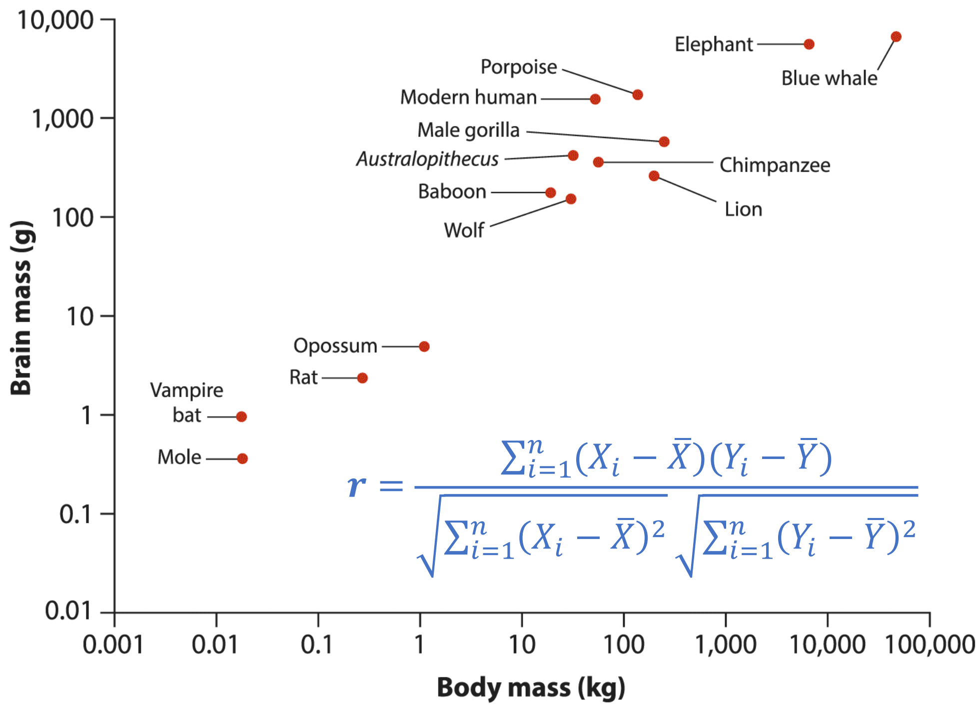
生物统计学

李 勤

生态与环境科学学院

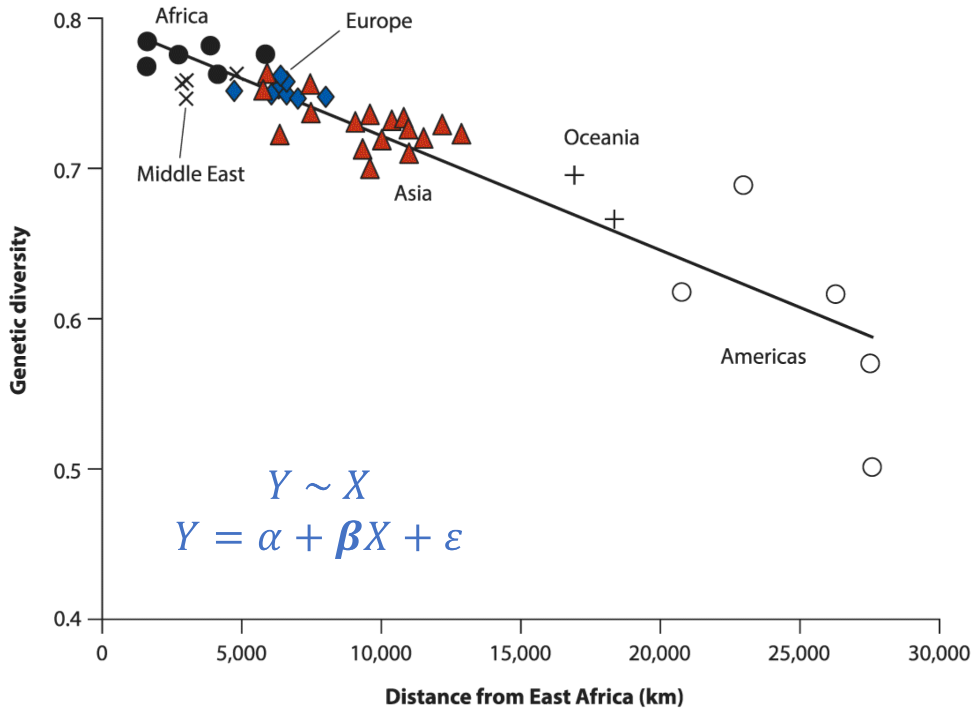
1. 回归——相关性和简单线性回归

- 相关性 correlation
  - 量化两个数值变量间的关联程度；
    - 反应了数据的分散程度；
- 线性回归 linear regression
  - 拟合回归线，可用以做预测；
    - 斜率反应了Y随X变化的变化率；



$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company



$$Y \sim X$$
$$Y = \alpha + \boldsymbol{\beta} X + \varepsilon$$

Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

1. 回归——相关性和简单线性回归：课堂练习

- Ch17-1 Does face shape predict aggression?
  - 男性的面部宽高比 (width-to-height ratio) 平均高于女性——这反映了在青春期期间雄性激素表达的差异。
  - 已知雄性激素能预测攻击性行为；那么，面部形状是否能预测攻击性呢?
  - 为此，Carré和McCormick（2008）比较了21名大学曲棍球 (hockey) 运动员的面部宽高比与每场比赛中因激烈违规行为（如打架或横扫）而获得的平均罚分。数据见"chap17q01FacesAndPenalties.csv"。
- 请尝试
  - 计算面部宽高比与平均罚分的相关系数?
  - 通过面部宽高比来预测罚分的回归线模型是?

https://whitlockschluter3e.zoology.ubc.ca/RLabs/R_tutorial_Correlation_Regression.html#r_commands_summary

R commands summary

Correlation coefficient

First numerical variable

Second numerical variable

```
cor(guppyData$fatherOrnamentation, guppyData$sonAttractiveness)
```

Testing correlation

First numerical variable

Second numerical variable

```
cor.test(guppyData$fatherOrnamentation, guppyData$sonAttractiveness)
```

Spearman's rank correlation

```
cor.test(guppyData$fatherOrnamentation, guppyData$sonAttractiveness, method = "spearman")
```

First numerical variable

Second numerical variable

Linear regression

Response numerical variable

Data frame

```
guppyRegression <- lm(sonAttractiveness ~ fatherOrnamentation, data = guppyData)
```

Explanatory numerical variable

Output of lm( )

```
summary(guppyRegression)
```

Residual plot

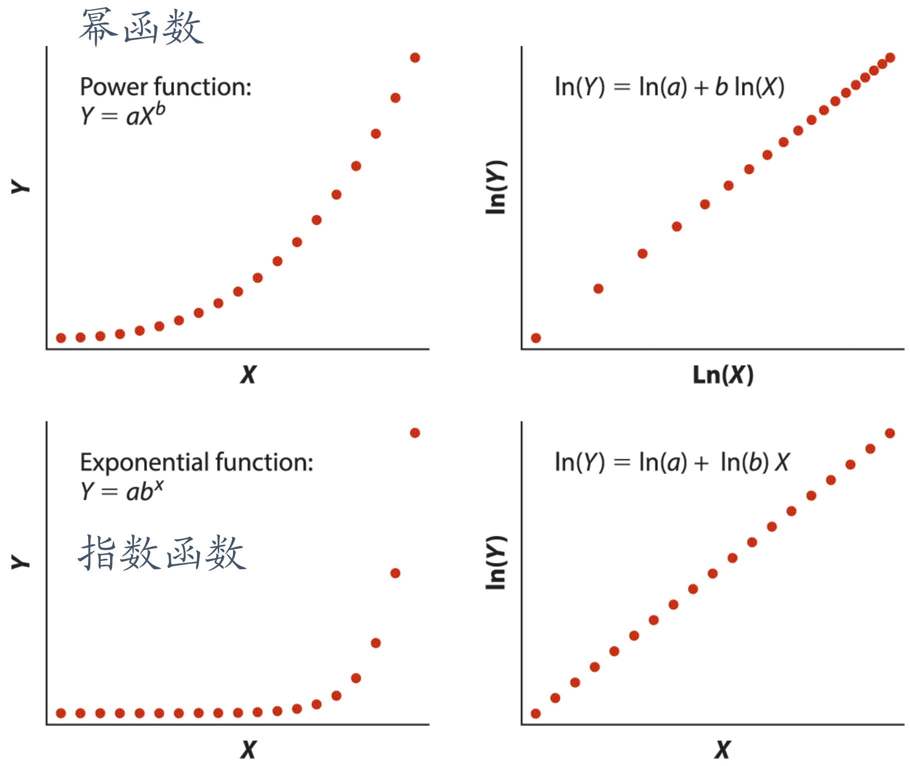
Output of lm( )

Explanatory numerical variable

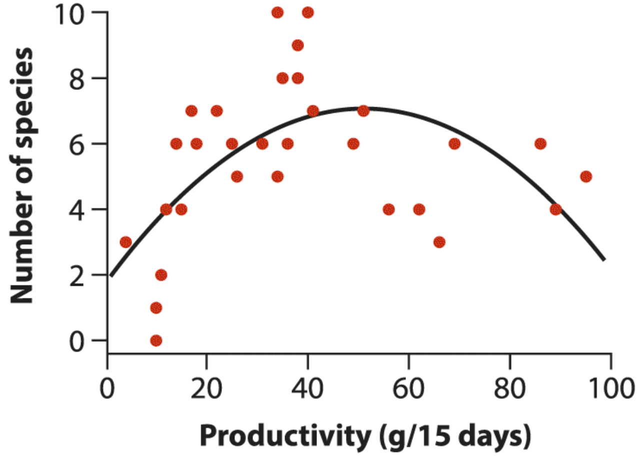
Data frame

```
plot(residuals(guppyRegression) ~ fatherOrnamentation, data = guppyData)
```

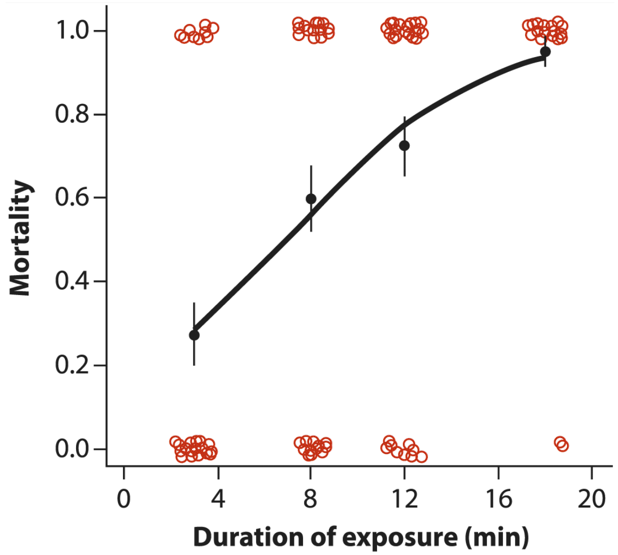
2. 简单线性回归的变体和延伸



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company



Whitlock & Schluter, *The Analysis of Biological Data*, 3e
© 2020 W. H. Freeman and Company

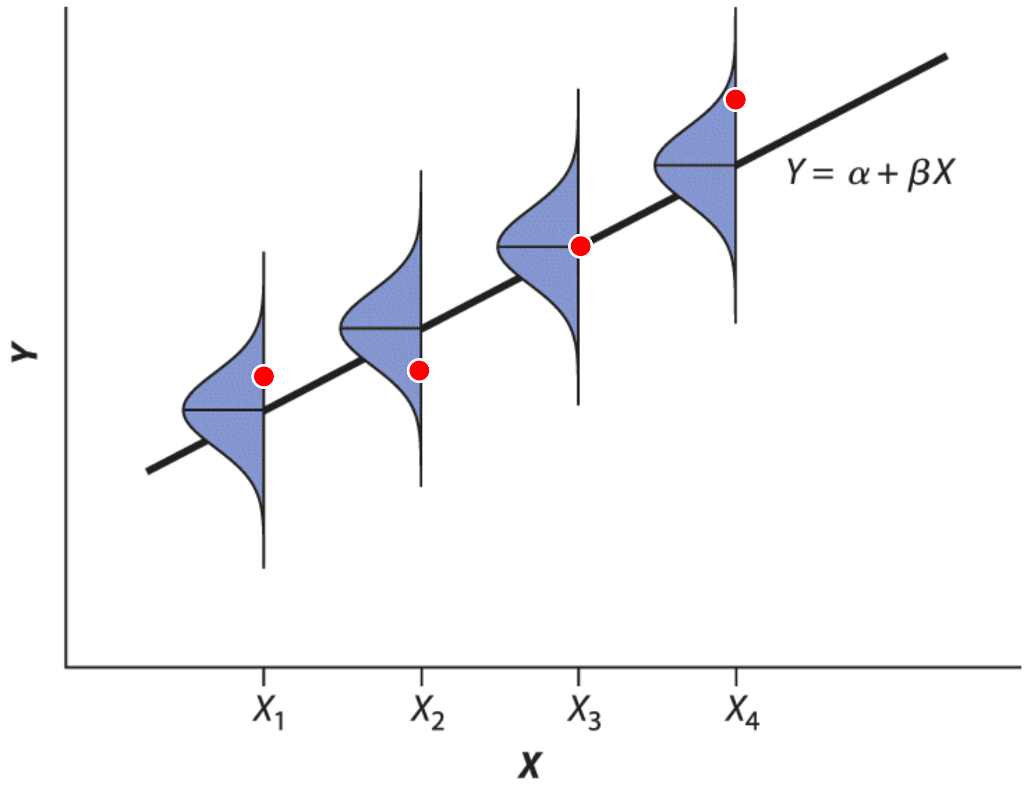


Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

2.1 简单线性回归的前提假设

- 前提假设 Assumptions
  - 每个X值对应的所有可能的Y值遵从正态分布，其均值位于真实回归线上;
  - 不同Y值的方差都相同;
  - 在每个X值处的Y的测量值/观察值代表了可能的Y值总体的一个随机样本;

  - 没有对X的正态性的要求;
  - 也不要求X是随机抽样而来;



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

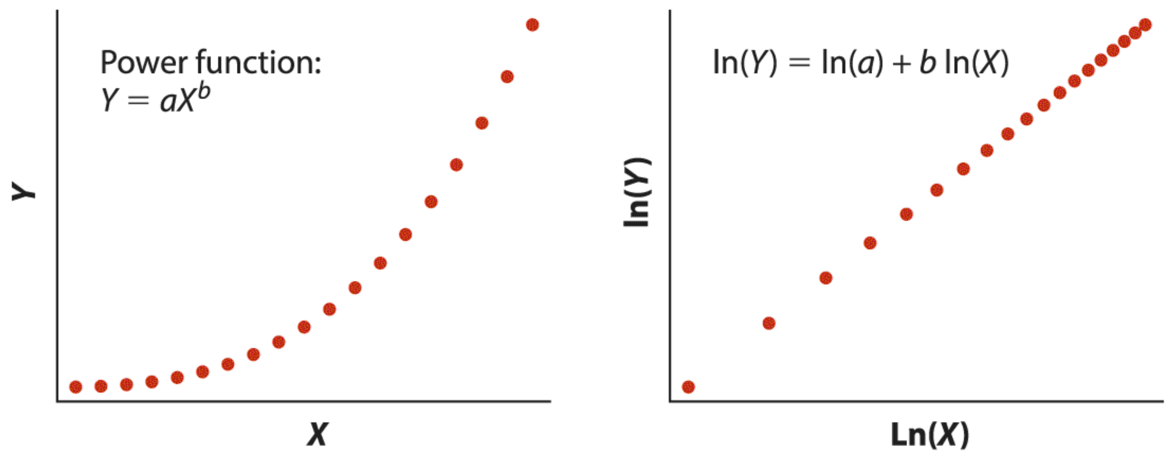
2.2 简单线性回归的测量误差

- 测量误差： 当变量的测量不完全准确时， 测量值≠真实值;
  - 测量误差对回归的影响与其对相关性的影响不同，取决于变量;
  - Y的测量误差会增加残差的方差，增大斜率估计的抽样误差（不确定性）；
  - X的测量误差会增加残差的方差，增大斜率的期望估计的偏差（bias）；

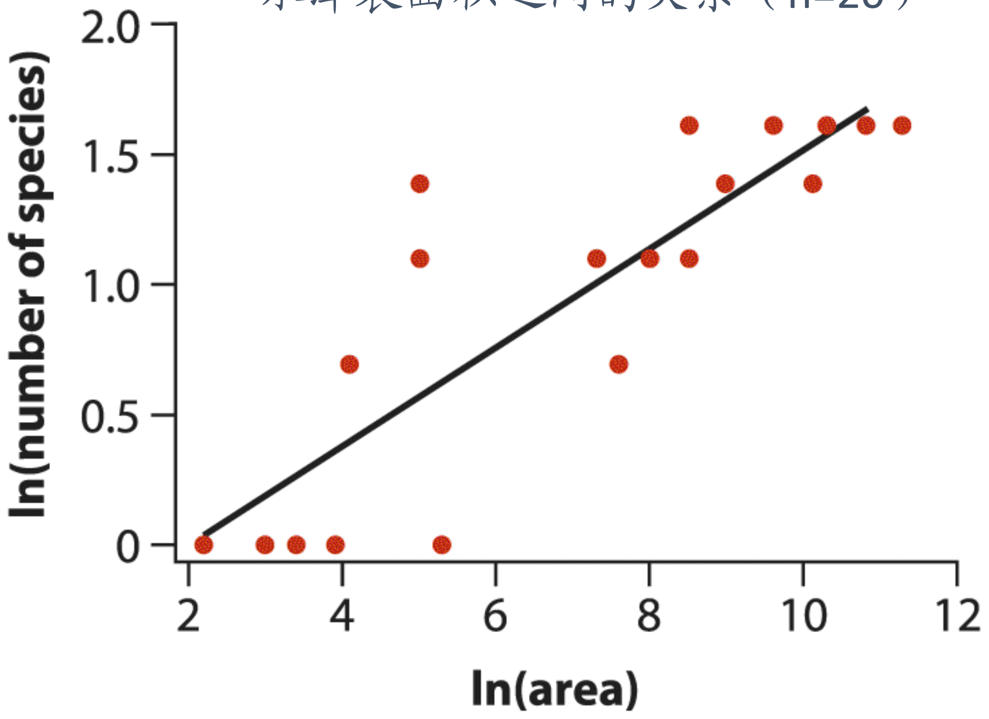
Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

2.3 简单线性回归的变体 Transformations

- 对数转换：对满足幂函数的X和Y进行转换从而满足线性关系；
  - $ln(Y) \sim ln(X)$



沙漠水潭中的鱼类物种数量与水潭表面积之间的关系（n=20）

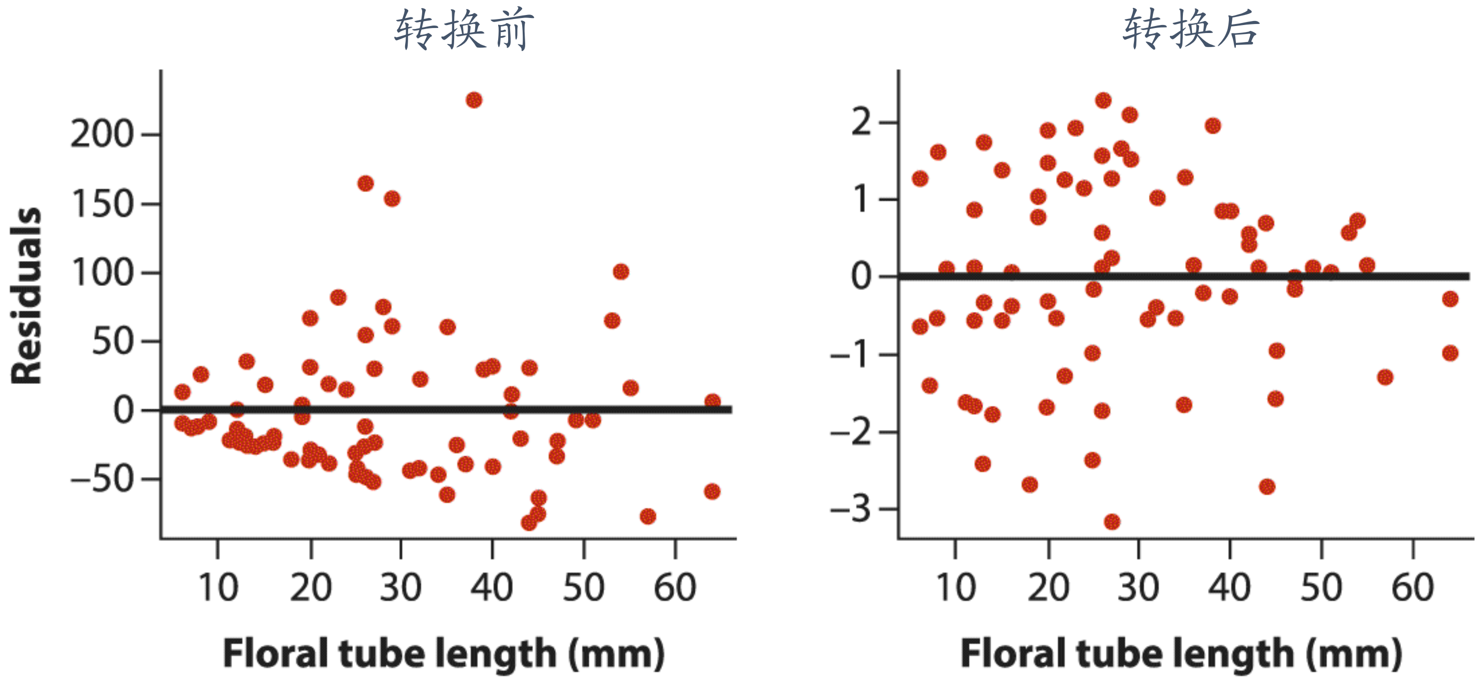


Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

2.3 简单线性回归的变体 Transformations

- 平方根转换：对Y进行转换从而满足残差方差相等的假设；
  - $\mathrm{sqrt}(\boldsymbol{Y}) \sim \boldsymbol{X}$

鸢尾草*Lapeirousia anceps* 的花朵接收的花粉数（由长吻蝇传粉）随着花筒长度的增加而增加;



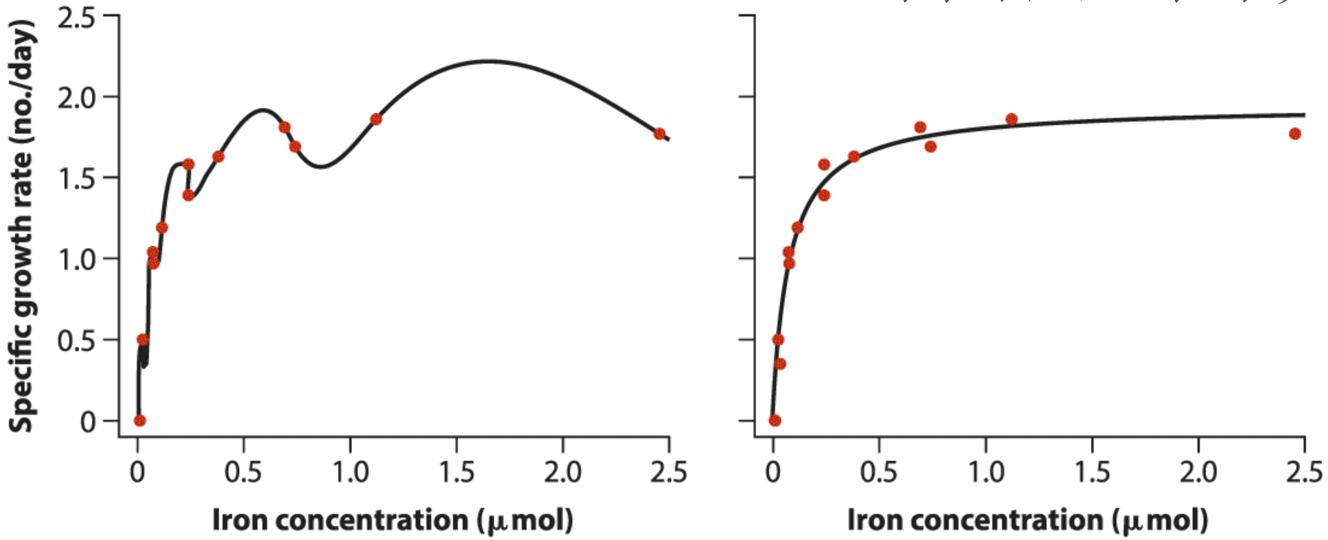
Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

2.3 非线性关系的回归 Regression with nonlinear relationships

- 基本原则
  - 除了不要求X和Y之间的线性关系，其余前提假设与线性回归一致；
    - 所以，很多数学函数可选 (so many mathematical functions to choose);
  - keep things simple ——越简单越好！

$min(\mathrm{MS}_{residual})$

铁对浮游植物生长的影响



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

2.3 非线性关系的回归 Regression with nonlinear relationships

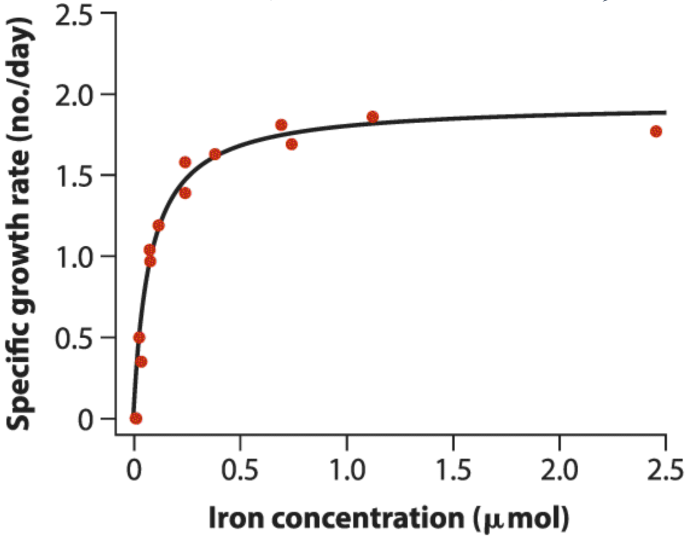
- （1）渐近线 （asymptote）

生物化学中常见的the Michaelis–Menten equation

- $Y = \frac{aX}{b+X}$

- a：渐近线常数；
- b：曲线上升速率；

铁对浮游植物生长的影响



2.3 非线性关系的回归 Regression with nonlinear relationships

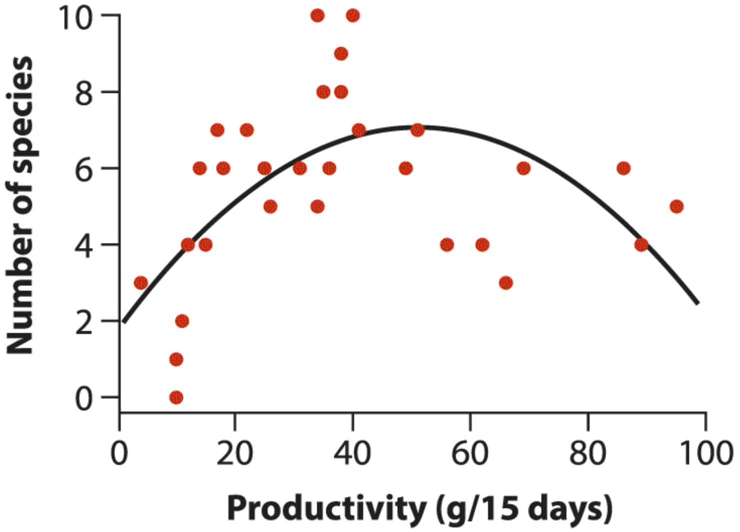
- （2）二次曲线 （quadratic curves）

生态学中常见的抛物线/单峰曲线

- $Y = a + bX + cX^2$

- a：截距;
- b：一次项的回归系数;
- c：二次项的回归系数;
  - 负值：驼峰;
  - 正值：U型;

植物物种数与池塘生产力之间的关系



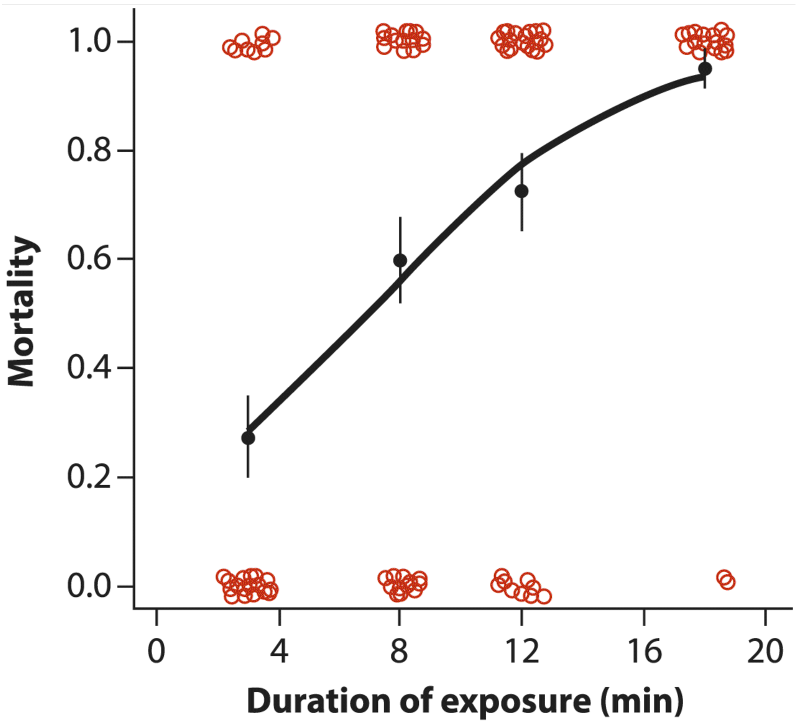
Whitlock & Schluter, *The Analysis of Biological Data*, 3e

2.3 逻辑斯谛回归 Logistic Regression

- A special type of nonlinear regression
- Y: a ***binary*** response variable
  - 1 vs 0
  - 存活 vs 死亡
  - 出现 vs 不出现
  - 萌发 vs 不萌发（种子）
  - 表现 vs 不表现（行为）
  - etc.

类型变量 ~ 数值变量

死亡率



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

暴露在低温的时间

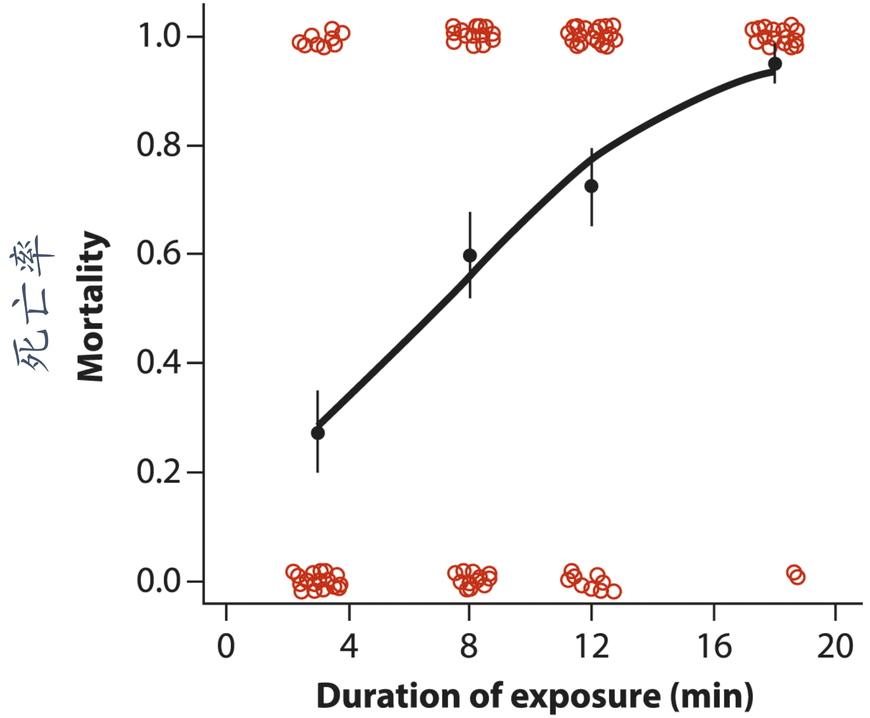
2.3 逻辑斯谛回归 Logistic Regression

- 前提假设
  - 每个X值处的结果为二项分布（1 vs 0）
- 模型构建
  - $\log\text{-odds}(Y) = \ln\left(\frac{Y}{1-Y}\right) = \alpha + \beta X$

出现“1”的比例/概率的比值比的自然对数值

普通的线性回归

类型变量 ~ 数值变量



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

暴露在低温的时间

2.3 逻辑斯谛回归 Logistic Regression

- $\text{log-odds}(Y) = \ln\left(\frac{Y}{1-Y}\right) = \alpha + \beta X$

- 逻辑斯谛回归属于一种广义线性回归 (generalized linear regression)
  - glm()
  - 对Y构建一个link function
    - link = logit
  - 参数估计:
    - α: 截距
    - β: 斜率

Fit a logistic regression

Use `summary()` to obtain estimates of regression coefficients with standard errors (Table 17.9-2).

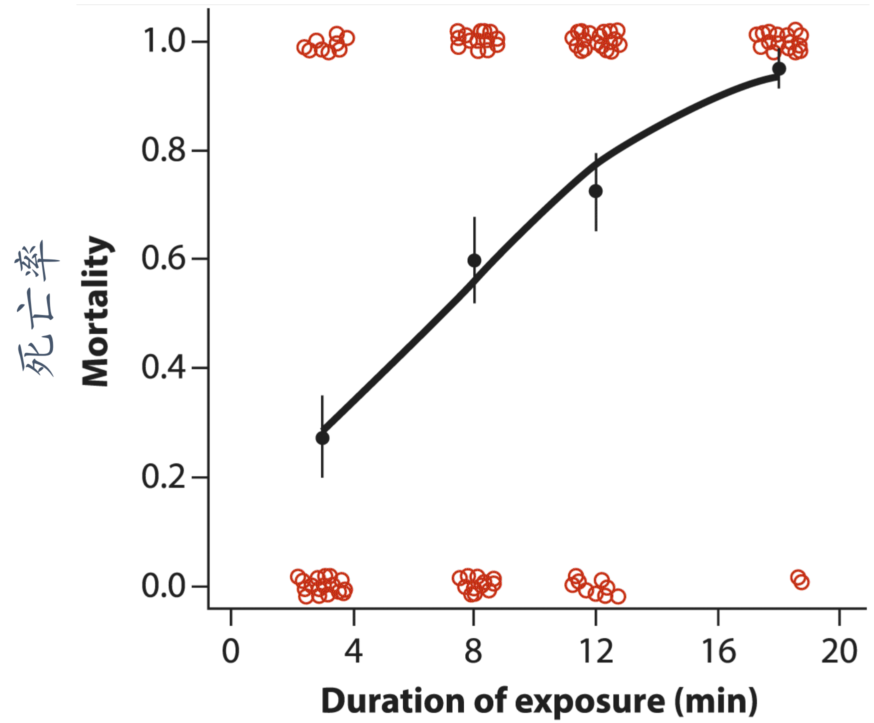
```
guppyGlm <- glm(mortality ~ exposureDurationMin, data = guppy,
    family = binomial(link = logit))
summary(guppyGlm)
```

```
##
## Call:
## glm(formula = mortality ~ exposureDurationMin, family = binomial(link = logit),
##     data = guppy)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3332  -0.8115   0.3688   0.7206   1.5943
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -1.66081    0.40651  -4.086 4.40e-05 ***
## exposureDurationMin  0.23971    0.04245   5.646 1.64e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 209.55  on 159  degrees of freedom
## Residual deviance: 164.69  on 158  degrees of freedom
## AIC: 168.69
##
## Number of Fisher Scoring iterations: 4
```

2.3 逻辑斯谛回归 Logistic Regression

- 拟合建模: $\text{log-odds}(Y) = \ln\left(\frac{Y}{1-Y}\right) = \alpha + \beta X$
  - $\text{logit}(Y) = -1.66 + 0.24X$
    - $\alpha = -1.66 \pm 0.41\text{SE}$
    - $\beta = 0.24 \pm 0.04\text{SE}$

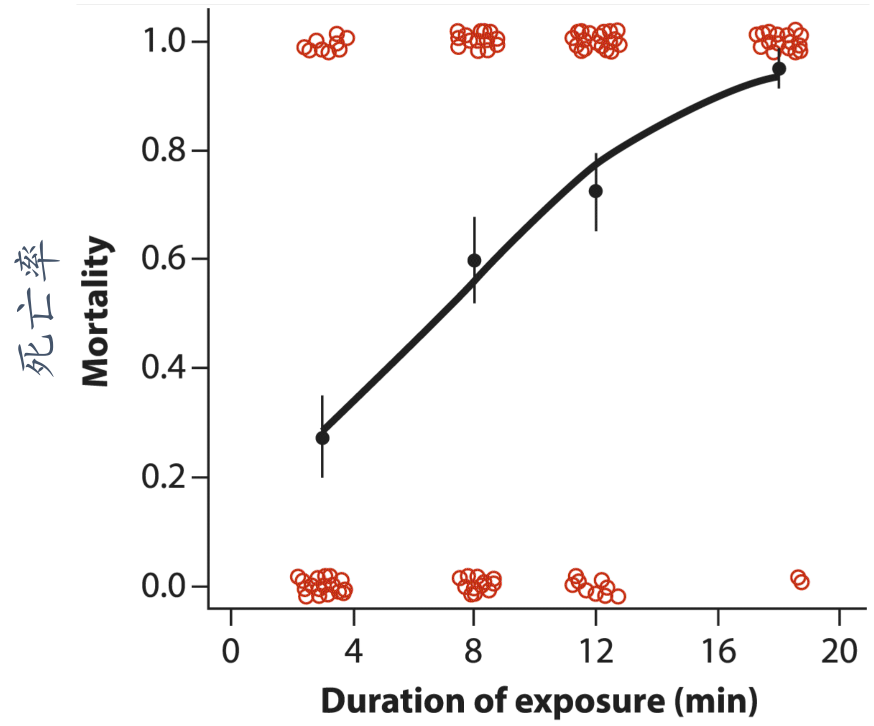
- 预测: $\hat{Y} = \dfrac{e^{\alpha+\beta X}}{1+e^{\alpha+\beta X}}$
    - $X = 10\text{min} \rightarrow \hat{Y} = \dfrac{e^{0.74}}{1+e^{0.74}} = 0.68$
    - 暴露在低温中10分钟会导致68%的死亡率。



暴露在低温的时间

2.3 逻辑斯谛回归 Logistic Regression

- 拟合建模: $\text{log-odds}(Y) = \ln\left(\frac{Y}{1-Y}\right) = \alpha + \beta X$
  - $\text{logit}(Y) = -1.66 + 0.24X$
    - $\alpha = -1.66 \pm 0.41\text{SE}$
    - $\beta = 0.24 \pm 0.04\text{SE}$
- 预测: $\hat{Y} = \frac{e^{\alpha+\beta X}}{1+e^{\alpha+\beta X}}$
    - 半数致死量: $\hat{Y} = 0.5$
    - $\ln\left(\frac{0.5}{1-0.5}\right) = \alpha + \beta X = 0 \rightarrow X = \frac{\alpha}{\beta}$



Whitlock & Schluter, *The Analysis of Biological Data*, 3e © 2020 W. H. Freeman and Company

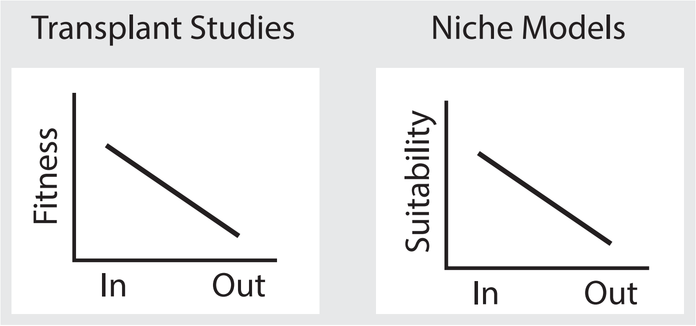
暴露在低温的时间

3. Others methods in Biology/Ecology/Evolution/Biogeography

- Multivariate linear regression
- Ordination (PCA, clustering, NMDS)
- Maximum Likelihood & Model Comparison (AIC/BIC)
- Bayesian methods
- SDM – spatial related
  - Statistical: GLM， GAM
  - Machine learning: MaxEnt, RandomForest
- Bioinformatics, genetics
- Macroevolution, Phylogenetics, Biogeographical models

通过移栽试验来探讨物种的地理分布边界与其生态位边界的重叠性

Meta-Analysis
荟萃分析



ECOLOGY LETTERS

Ecology Letters, (2016) **19**: 710–722 doi: 10.1111/ele.12604

REVIEW AND SYNTHESIS

A synthesis of transplant experiments and ecological niche models suggests that range limits are often niche limits

Julie A. Lee-Yaw,[1,2]* Heather M. Kharouba,[3] Megan Bontrager,[1] Colin Mahony,[4] Anna Mária Csergő,[5] Annika M.E. Noreen,[1] Qin Li,[1] Richard Schuster[4] and Amy L. Angert[1]

Abstract

Global change has made it important to understand the factors that shape species' distributions. Central to this area of research is the question of whether species' range limits primarily reflect the distribution of suitable habitat (i.e. niche limits) or arise as a result of dispersal limitation. Over-the-edge transplant experiments and ecological niche models are commonly used to address this question, yet few studies have taken advantage of a combined approach for inferring the causes of range limits. Here, we synthesise results from existing transplant experiments with new information on the predicted suitability of sites based on niche models. We found that individual performance and habitat suitability independently decline beyond range limits across multiple species. Furthermore, inferences from transplant experiments and niche models were generally concordant within species, with 31 out of 40 cases fully supporting the hypothesis that range limits are niche limits. These results suggest that range limits are often niche limits and that the factors constraining species' ranges operate at scales detectable by both transplant experiments and niche models. In light of these findings, we outline an integrative framework for addressing the causes of range limits in individual species.

Keywords

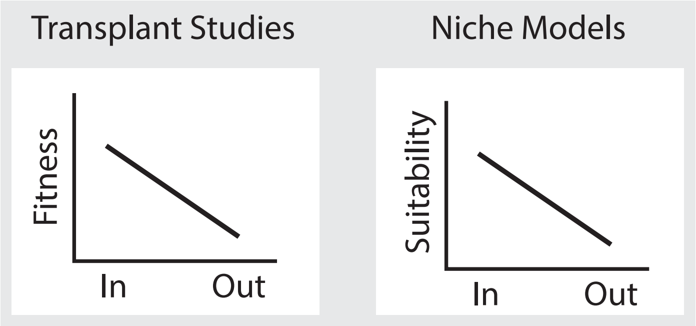
Abiotic constraints, climate, dispersal limitation, fitness, geographical distribution, over the edge transplant, species distribution modelling.

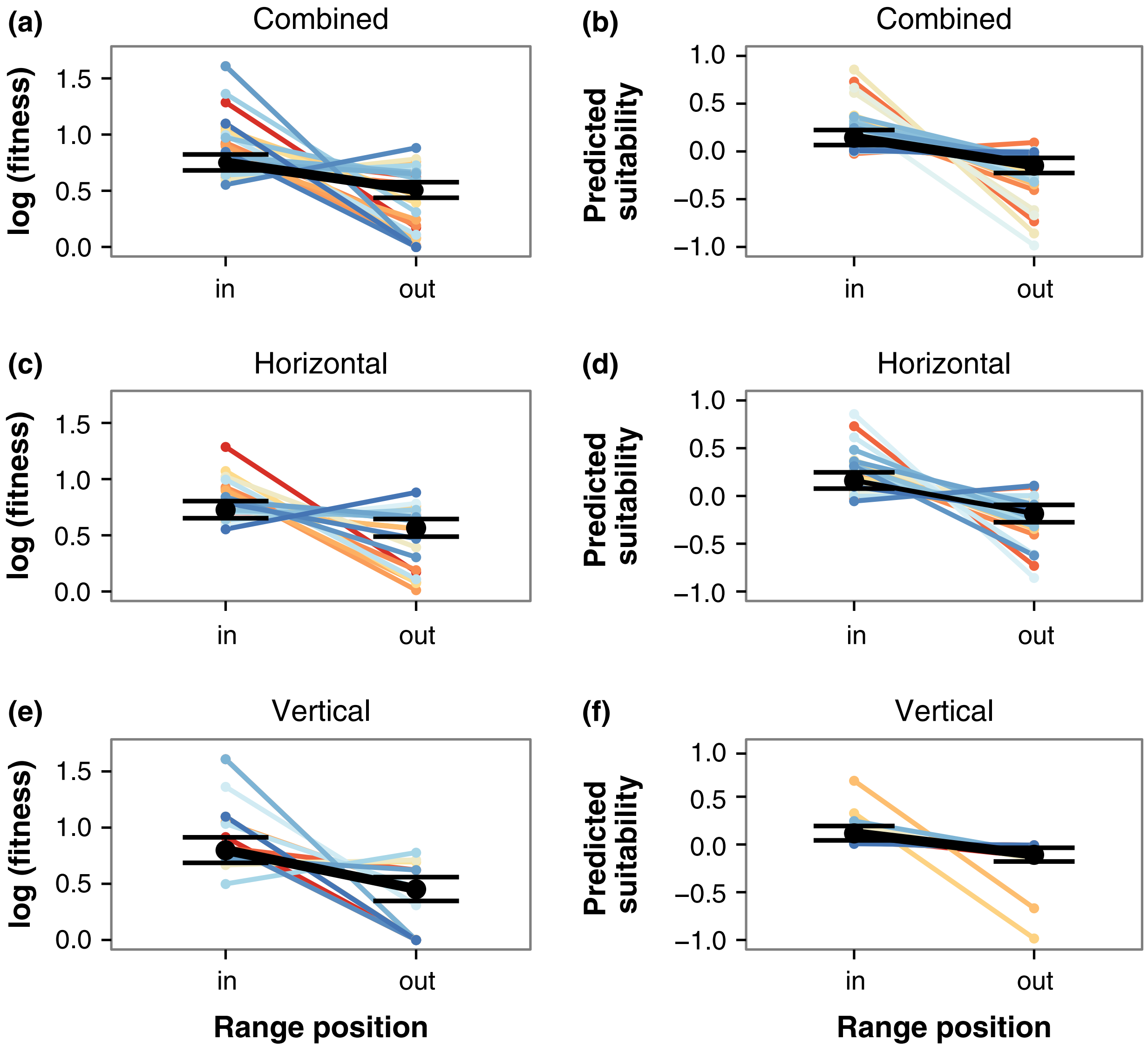
Ecology Letters (2016) **19**: 710–722

Sign-test 符号检验

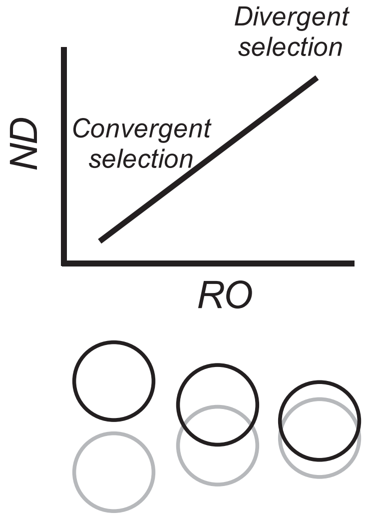
通过移栽试验来探讨物种的地理分布边界与其生态位边界的重叠性

Meta-Analysis
荟萃分析





探讨近缘物种（同一个属，closely related species）之间生态位分化和地理分布重叠程度的相关性



doi:10.1111/evo.13

The effect of range overlap on ecological niche divergence depends on spatial scale in monkeyflowers

Qin Li,[1,2,3] Dena L. Grossenbacher,[4] and Amy L. Angert[5]

[1]*Department of Botany, University of British Columbia, Vancouver, British Columbia, Canada*

[2]*Integrative Research Center, The Field Museum, Chicago, Illinois, USA*

[3]*E-mail: qin.li@biodiversity.ubc.ca*

[4]*Department of Biology, California Polytechnic State University, San Luis Obispo, California, USA*

[5]*Department of Botany and Zoology, University of British Columbia, Vancouver, British Columbia, Canada*

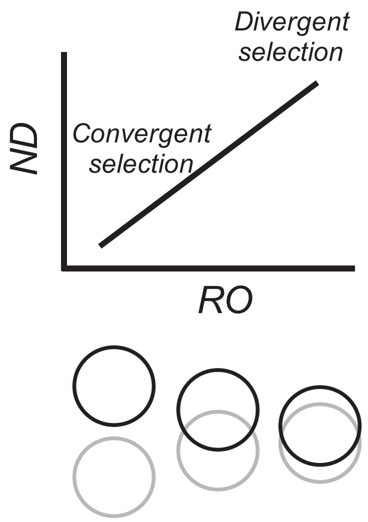
Received January 10, 2017
Accepted July 20, 2018

Patterns of niche divergence and geographical range overlap of closely related species provide insights into the evolutionary dynamics of ecological niches. When ranges overlap, shared selective pressures may preserve niche similarity along coarse-scale macrohabitat axes (e.g., bioclimates). Alternatively, competitive interactions may drive greater divergence along local-scale microhabitat axes (e.g., micro-topographical features). We tested these hypotheses in 16 species pairs of western North American monkeyflowers (*Erythranthe* and *Diplacus*, formerly *Mimulus*) with estimations of species' niches, geographic ranges, and a robust phylogeny. We found that macrohabitat niche divergence decreased with increasing range overlap, consistent with convergent selection operating at a coarse scale. No significant relationship was detected for microhabitat niches. Additionally, niche divergence was greater for recently diverged pairs along all macrohabitat niche axes, but greater for distantly diverged pairs along one microhabitat axis related to vegetation cover. For species pairs with partially overlapping ranges, greater microhabitat divergence was detected in sympatry than in allopatry for at least one niche axis for three of four pairs, consistent with character displacement in sympatry. Thus, coarse- and local-scale niche divergence show dissimilar patterns in relation to range overlap and divergence time, perhaps because the relative importance of convergent versus divergent selection depends on spatial scale.

KEY WORDS: *Mimulus*, monkeyflower, niche divergence, range overlap, phylogenetic distance.

Evolution 2018

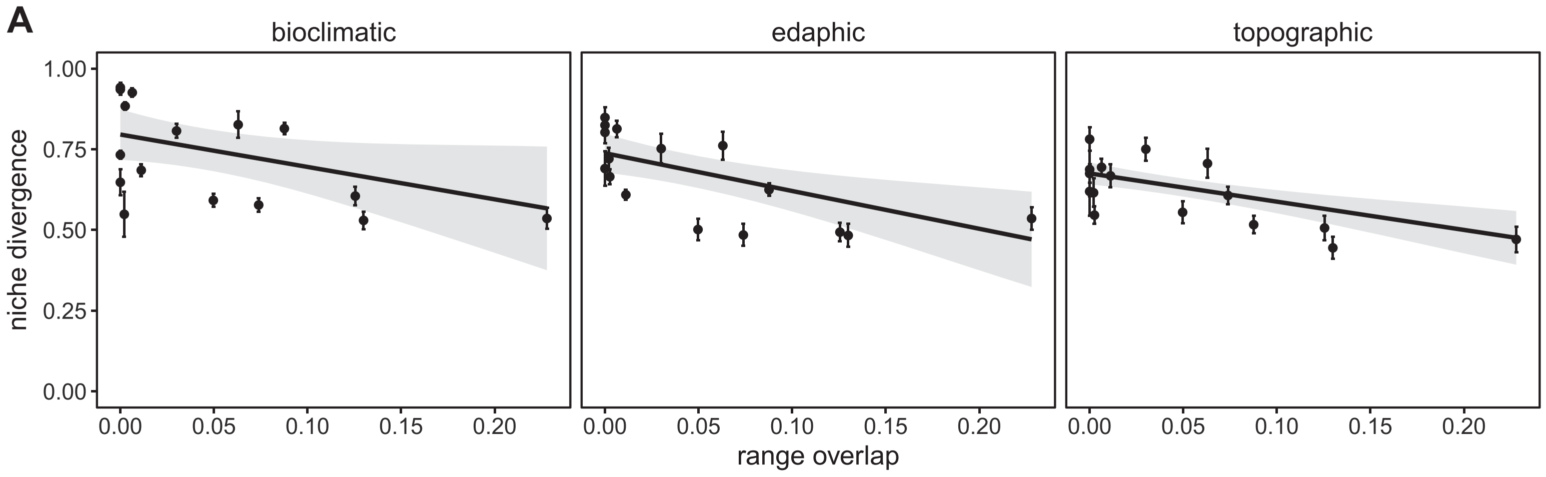
探讨近缘物种（同一个属，closely related species）之间生态位分化和地理分布重叠程度的相关性

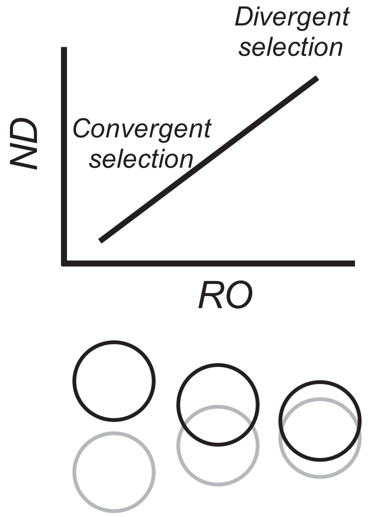


Mimulus, Monkeyflower (猴面花，沟酸浆属)

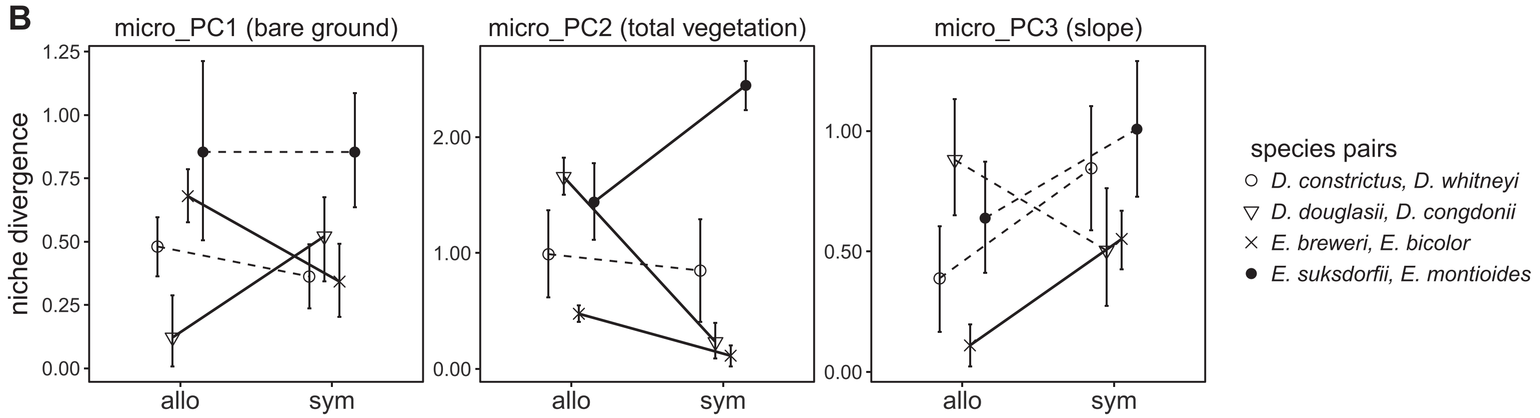
Linear Regression 线性回归

探讨近缘物种（同一个属，closely related species）之间生态位分化和地理分布重叠程度的相关性





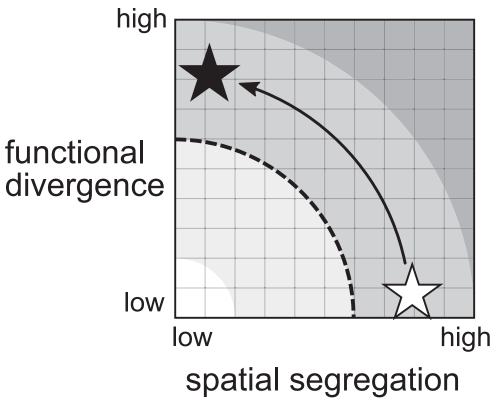
Confidence Interval 置信区间比较



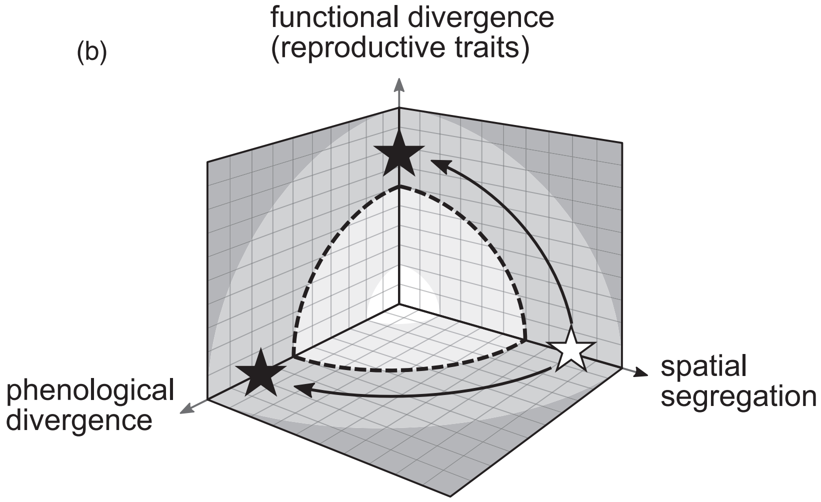
生物多样性格局

- 空间尺度上的共存
- 时间尺度上的分化

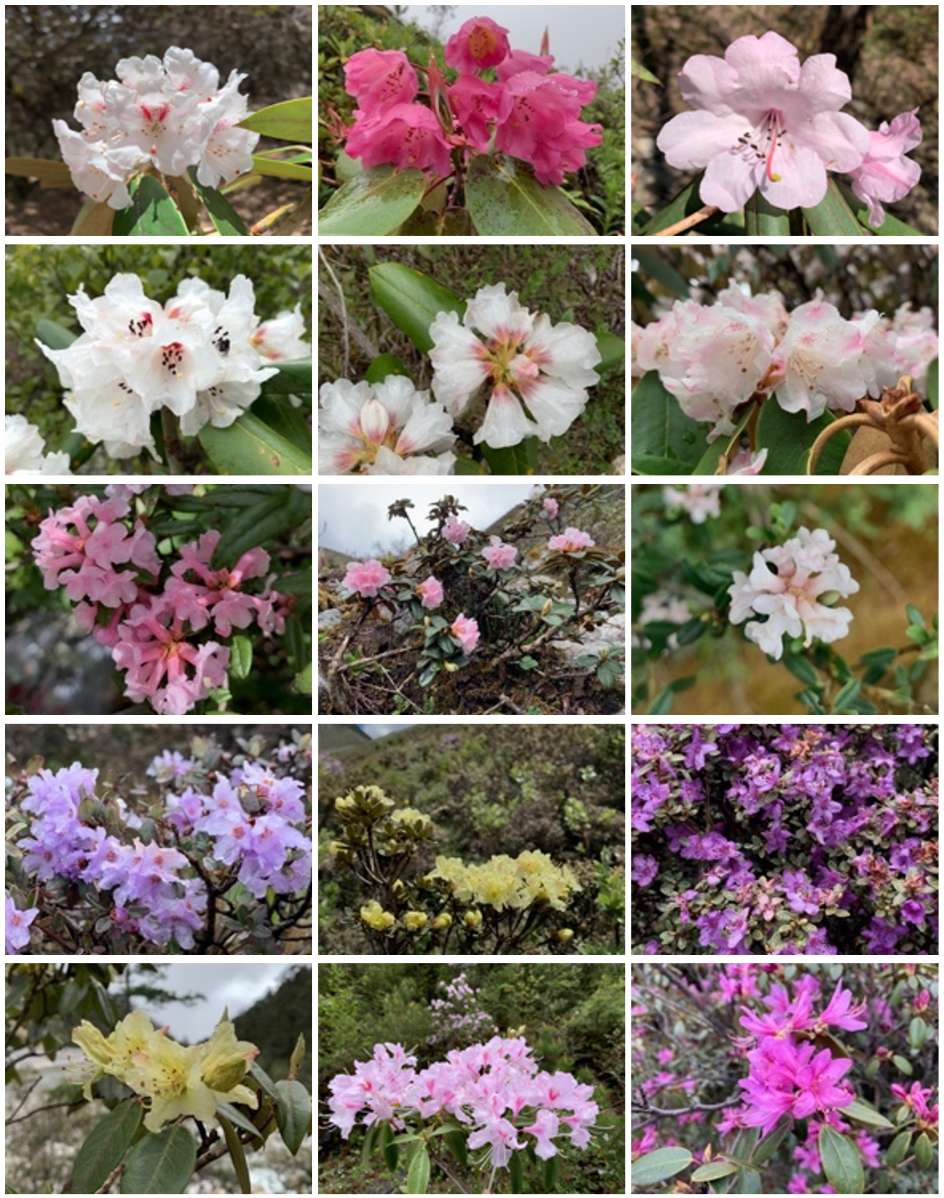
(a)



(b)



Rhododendron (杜鹃花属)

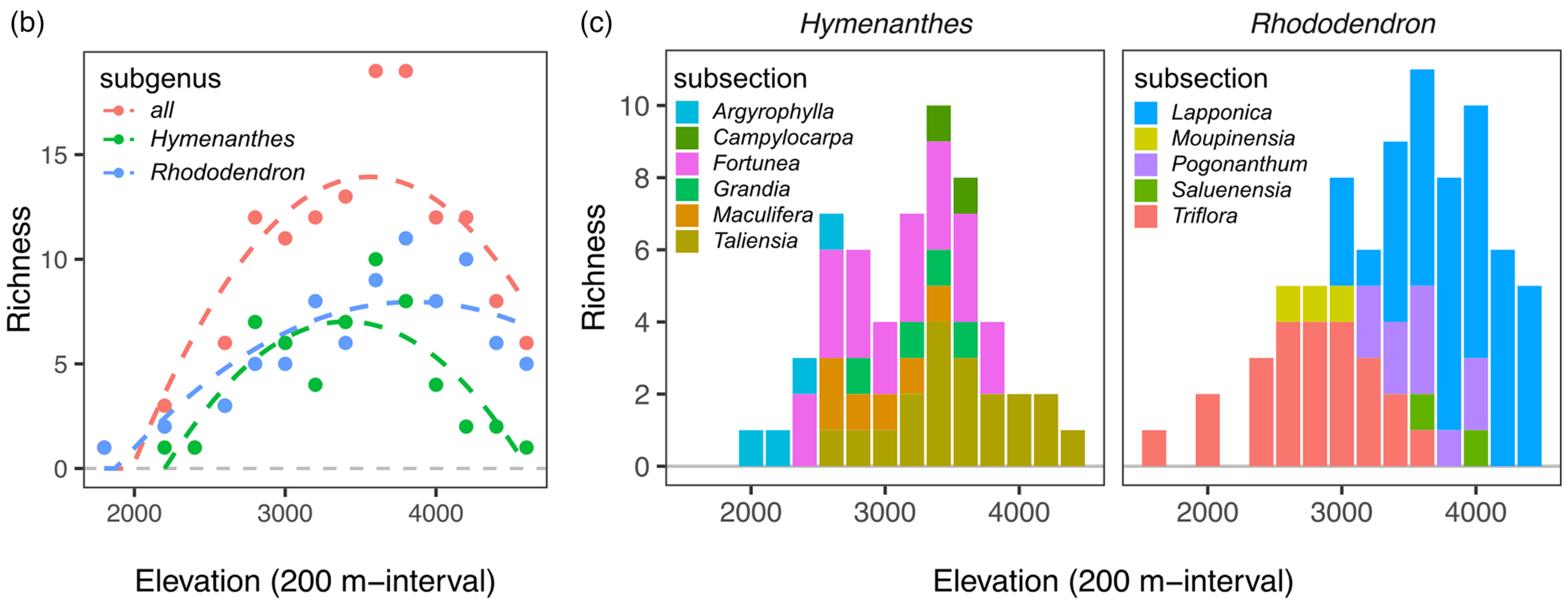


© Qin Li

生物多样性格局

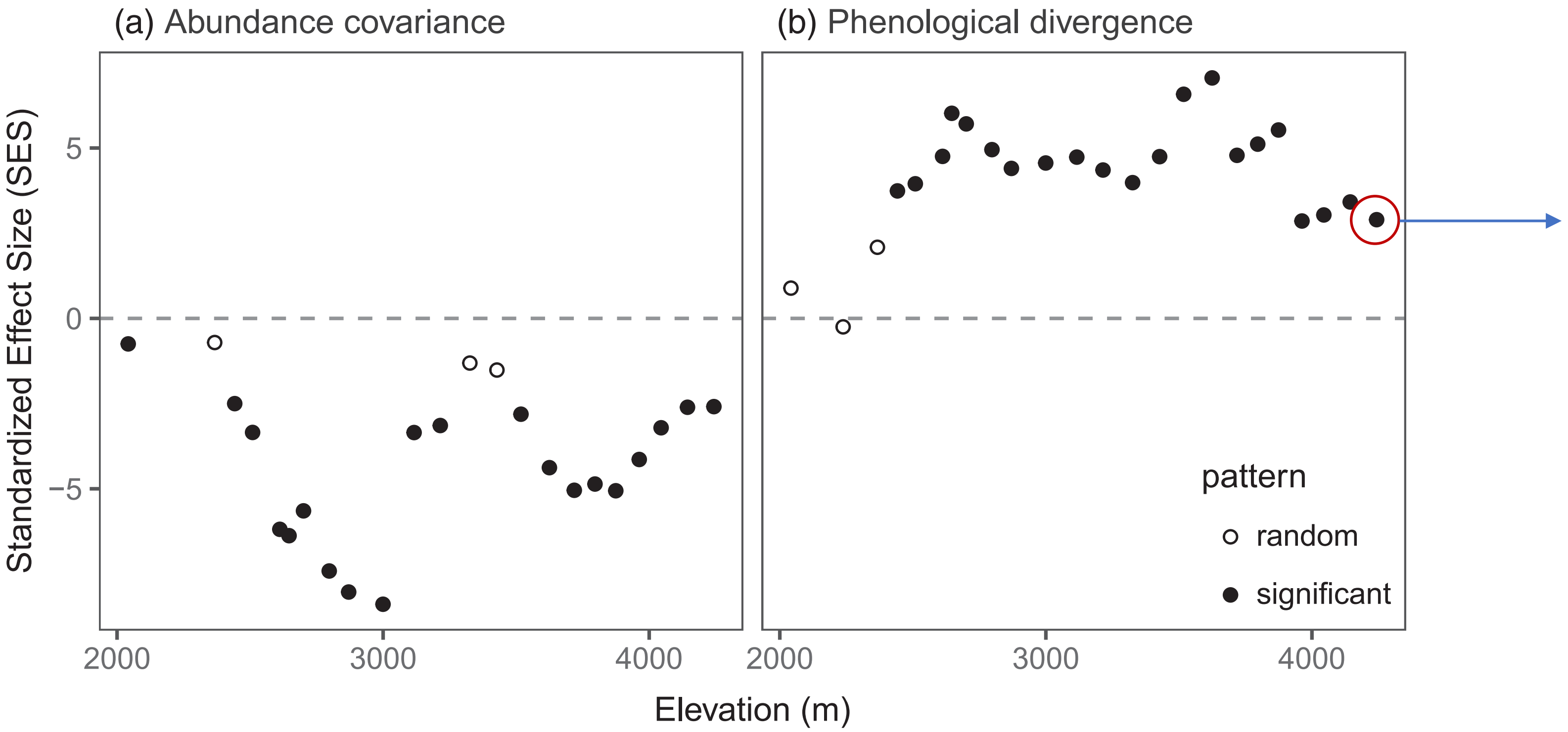
- 空间尺度上的共存
- 时间尺度上的分化

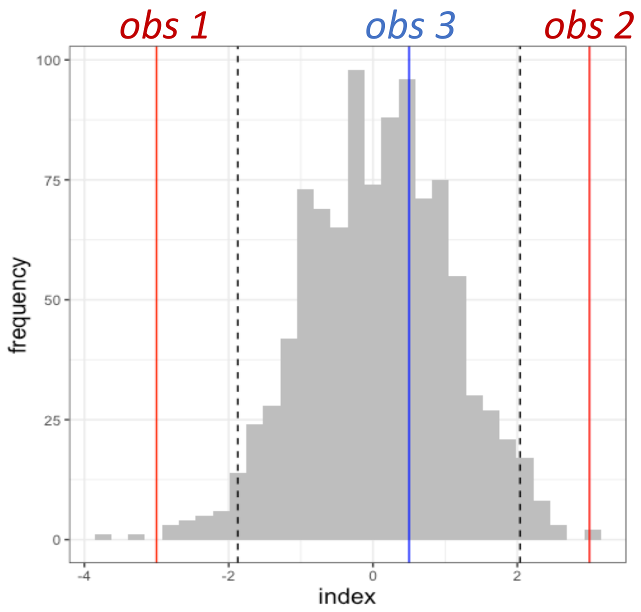
沿海拔梯度的多样性



生物多样性格局

- 空间尺度上的共存
- 时间尺度上的分化





Permutation test

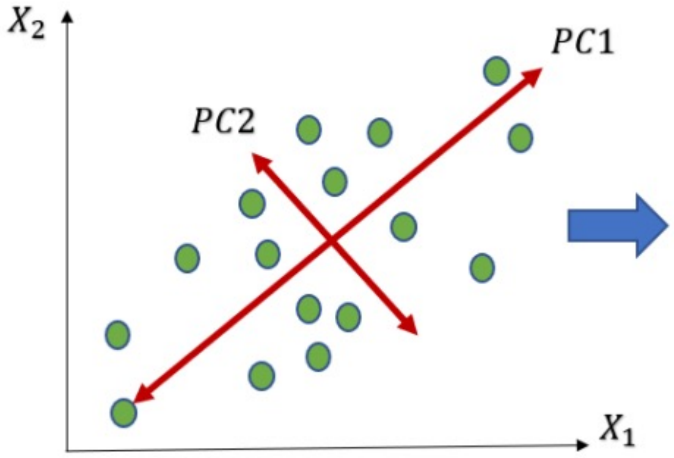
置换检验

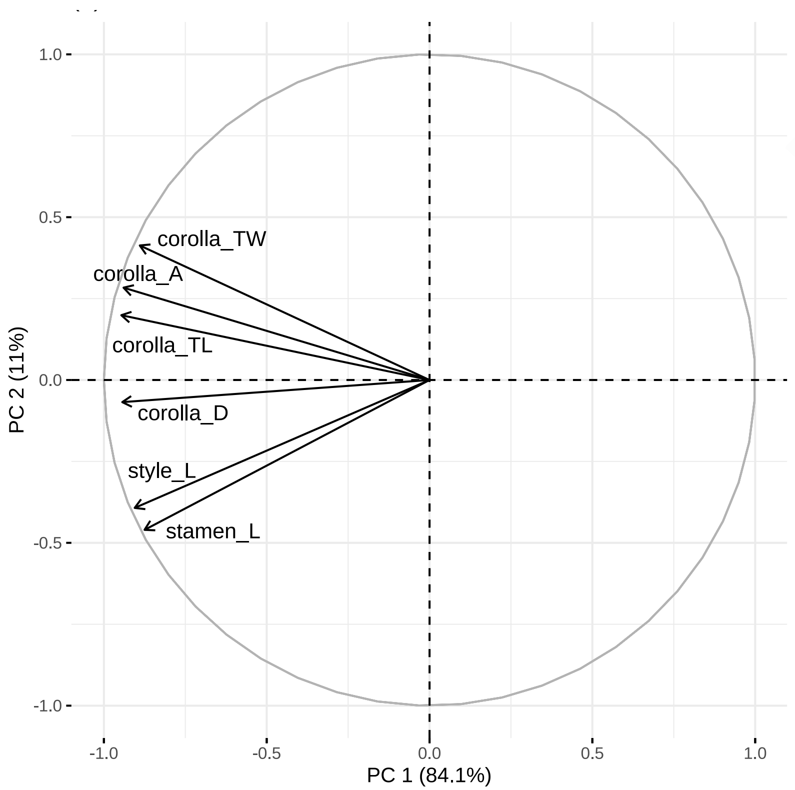
生物多样性格局

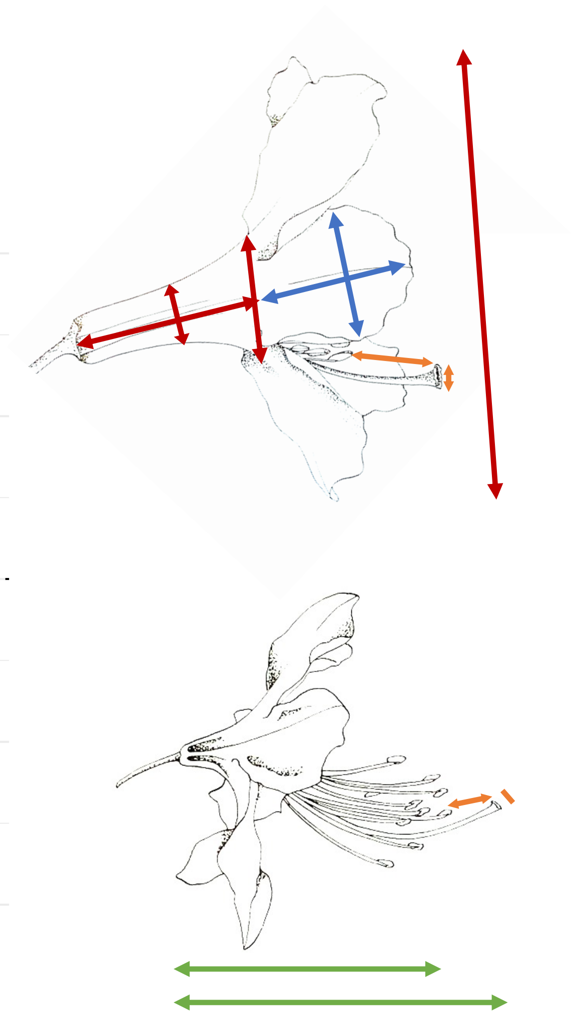
- 空间尺度上的共存
- 时间尺度上的分化
- 性状分化

PCA 主成分分析

Principal Component Analysis



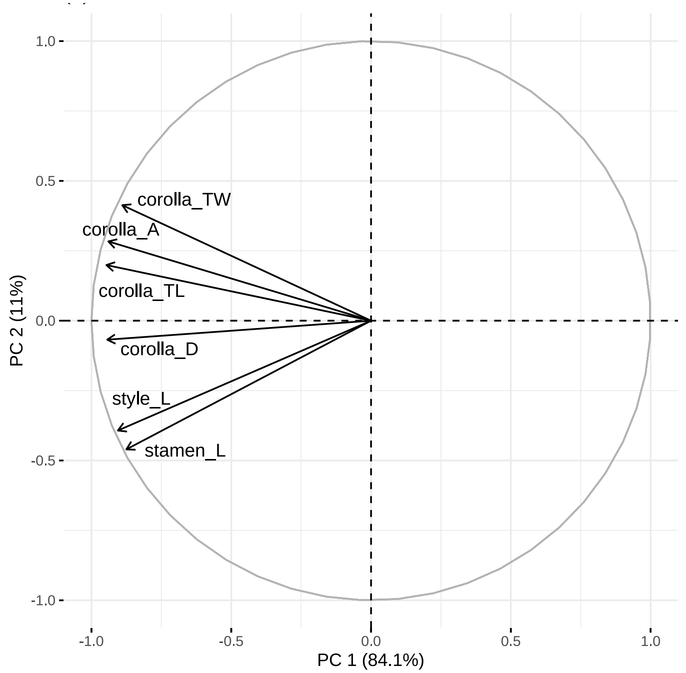


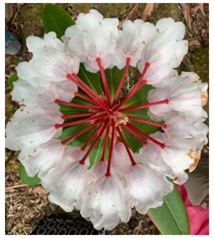


生物多样性格局

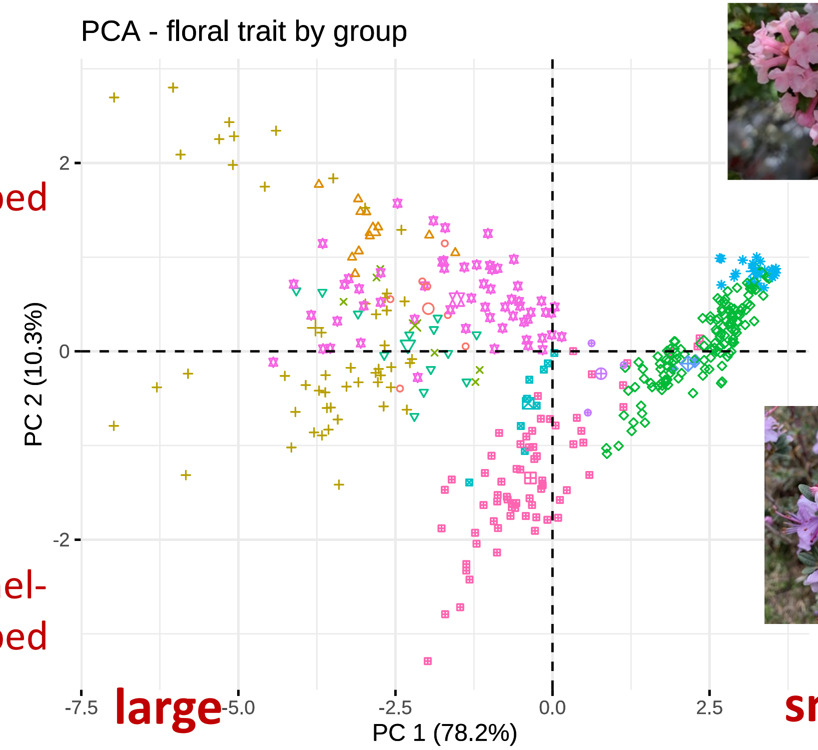
- 空间尺度上的共存
- 时间尺度上的分化
- 性状分化

PCA 主成分分析
Principal Component Analysis





bell-shaped

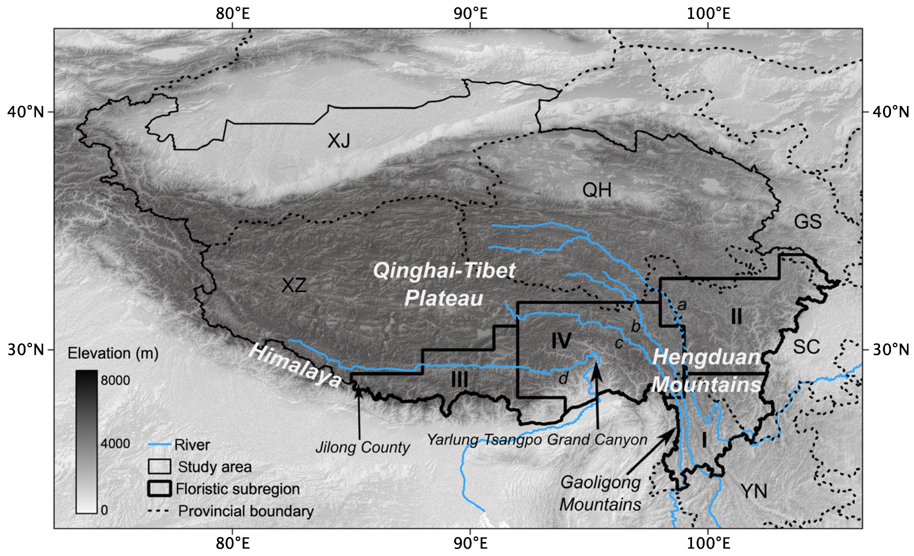


small

funnel-shaped

生物多样性格局

- 量化横断山区植物区系



New Phytologist

Research

Grade of Membership models reveal geographical and environmental correlates of floristic structure in a temperate biodiversity hotspot

Qin Li[1], Hang Sun[2], David E. Boufford[3], Bruce Bartholomew[4], Peter W. Fritsch[5], Jiahui Chen[2], Tao Deng[2] and Richard H. Ree[1]

[1]Department of Science and Education, Field Museum, Chicago, IL 60605, USA; [2]Key Laboratory for Plant Diversity and Biogeography of East Asia, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming, Yunnan 650201, China; [3]Harvard University Herbaria, Cambridge, MA 02138, USA; [4]Department of Botany, California Academy of Sciences, Golden Gate Park, San Francisco, CA 94118, USA; [5]Botanical Research Institute of Texas, Fort Worth, TX 76017, USA

Authors for correspondence:
Qin Li
Email: qli@fieldmuseum.org

Richard H. Ree
Email: rree@fieldmuseum.org

Received: *3 November 2020*
Accepted: *28 March 2021*

New Phytologist (2021)
doi: 10.1111/nph.17443

Key words: biogeographical regionalization, biotic turnover, floristic structure, Hengduan Mountains, seed plants, species motif, tropical-temperate divide.

Summary

- Identifying the contours and correlates of species turnover is central to understanding the nature of biogeographical regions. The Hengduan Mountains region of south-central China (HMR) is well known for its high diversity of plants, but its boundaries and internal floristic structure are poorly understood, especially in relation to geographical and environmental factors.
- With data on occurrences and elevational ranges of seed plants across the HMR and adjacent areas of the greater Qinghai-Tibet Plateau, we identified motifs (distinct species assemblages) by Grade of Membership models, and characterized relative contributions of geography, elevation, and climate to their spatial patterns.
- Motifs segregate primarily by latitude, elevation, and correlated environmental variables, most sharply across the tropical-temperate divide. Secondarily, they segregate by longitude and geographical features, and reveal a novel divide across the Jinsha River. A core set of motifs corresponds to previous delineations of the HMR.
- The HMR biodiversity hotspot is more a mosaic of floristic elements than a cohesive entity. Grade of Membership models effectively reveal the geographical contours of biotic structure, and are a valuable new tool for biogeographical analysis.

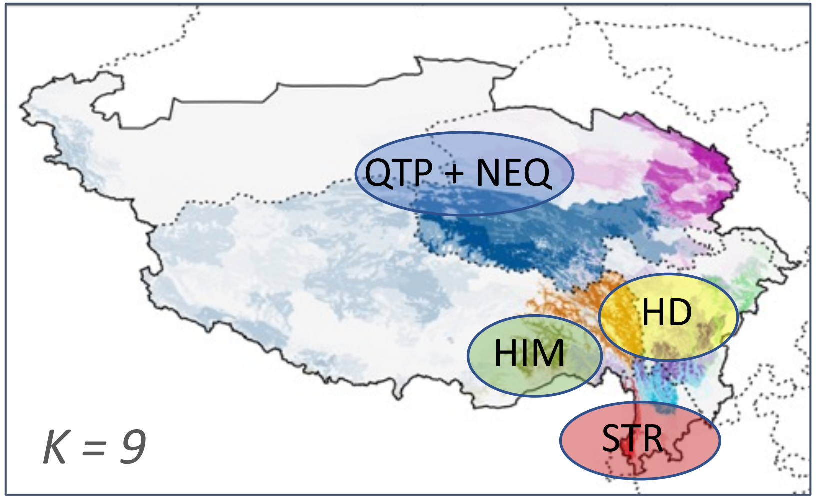
生物多样性格局

- 量化横断山区植物区系

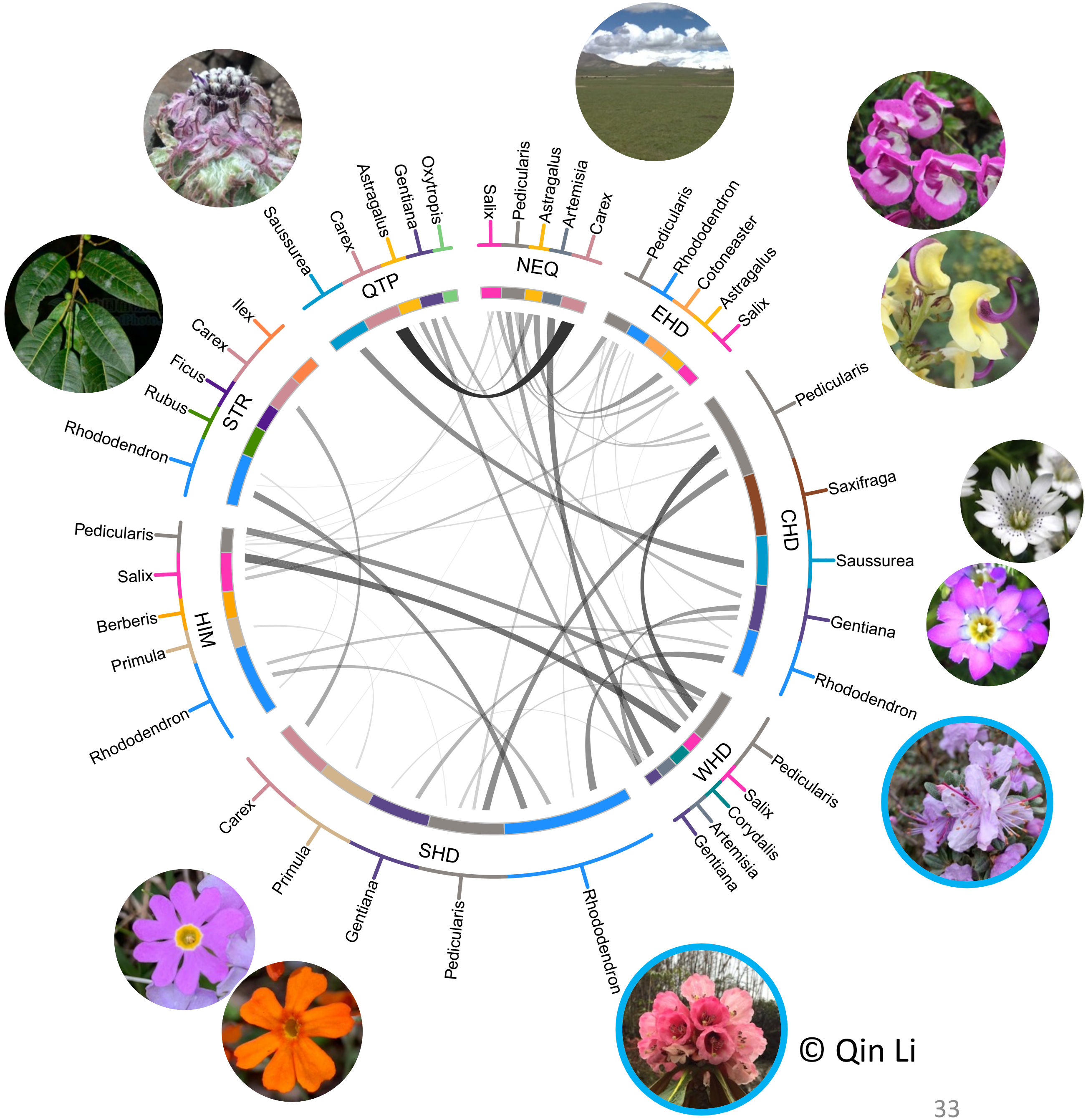
贝叶斯模型

Bayesian Models

$$p_{ng} = \sum_{k=1}^{K} \omega_{nk}\theta_{kg}$$



(Li et al. New Phytologist, 2021)

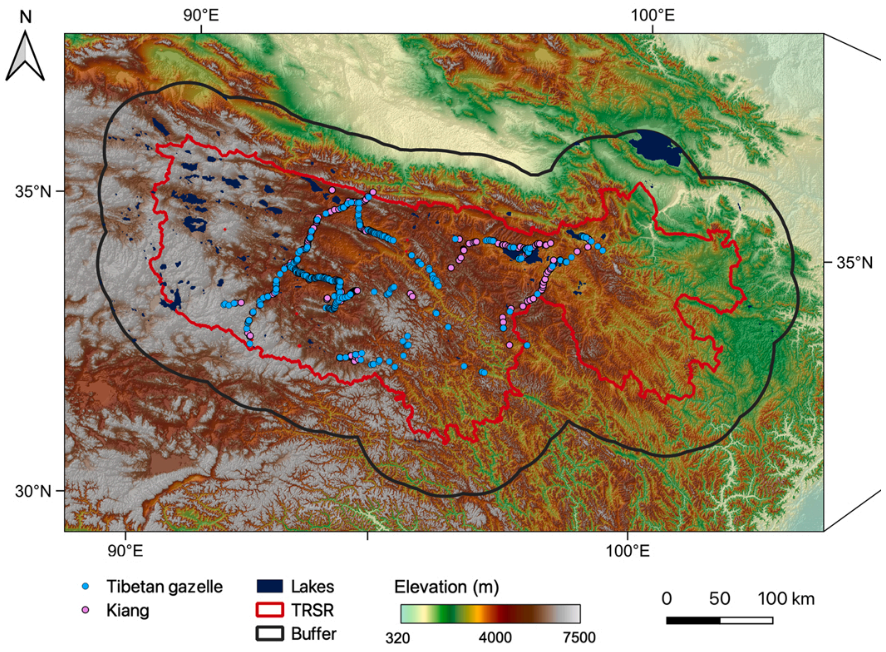


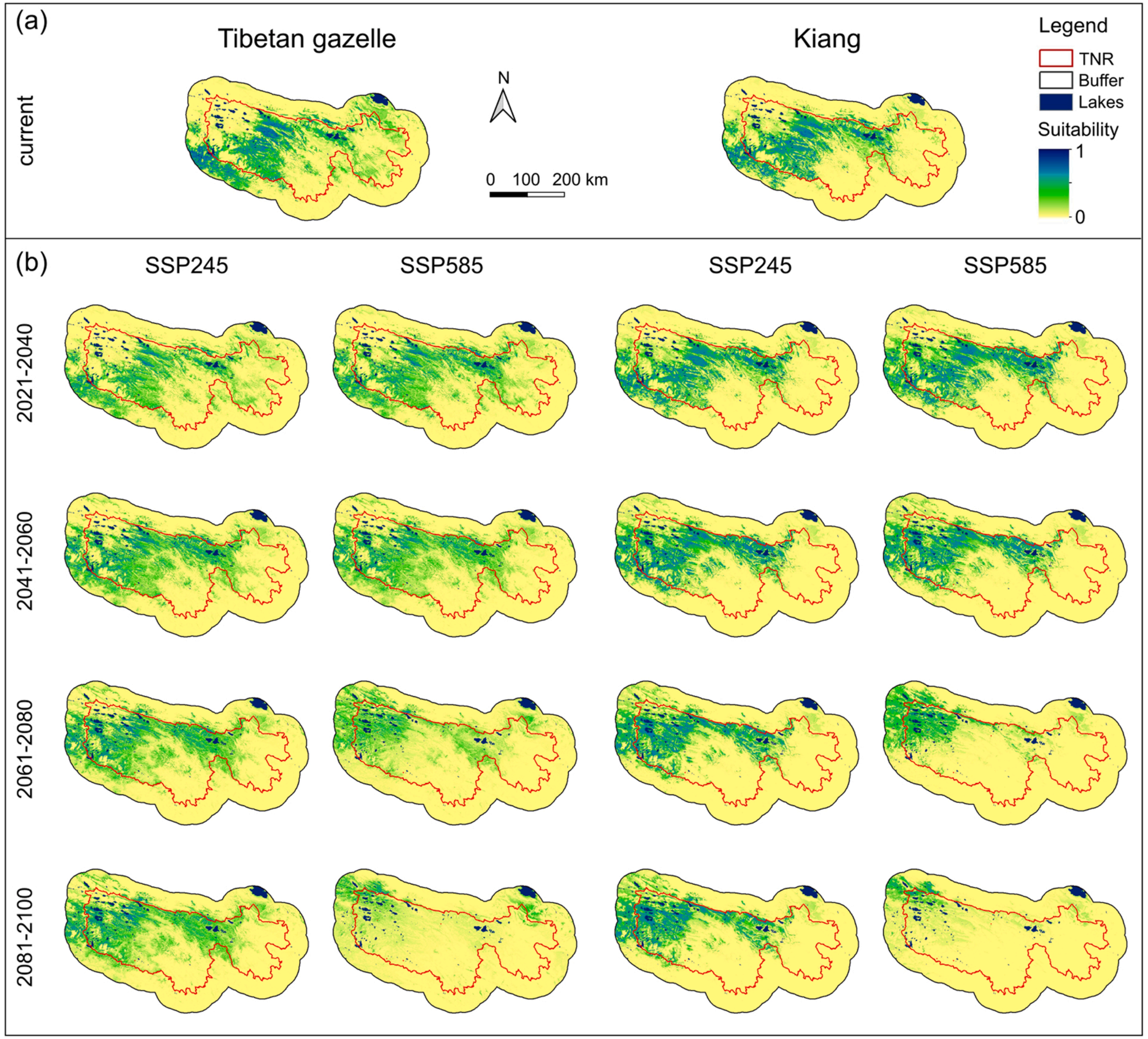
© Qin Li

全球变化和物种响应

- 物种地理分布的变化
  - 物种分布模型
  - Species Distribution Model

MaxEnt：一种机器学习方法
（物种出现点 + 环境数据）

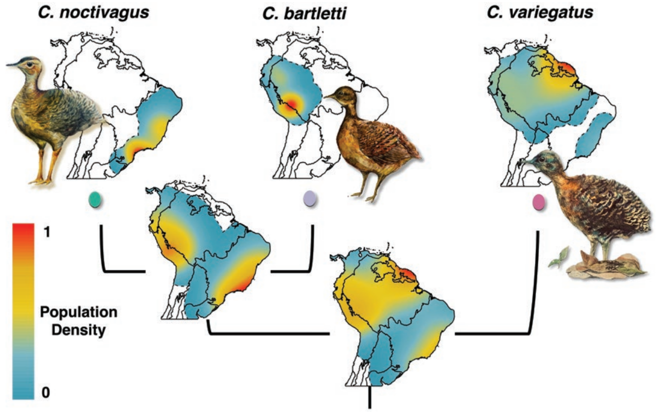




(Zhang, Li, et al. Global Ecology and Conservation, 2022)

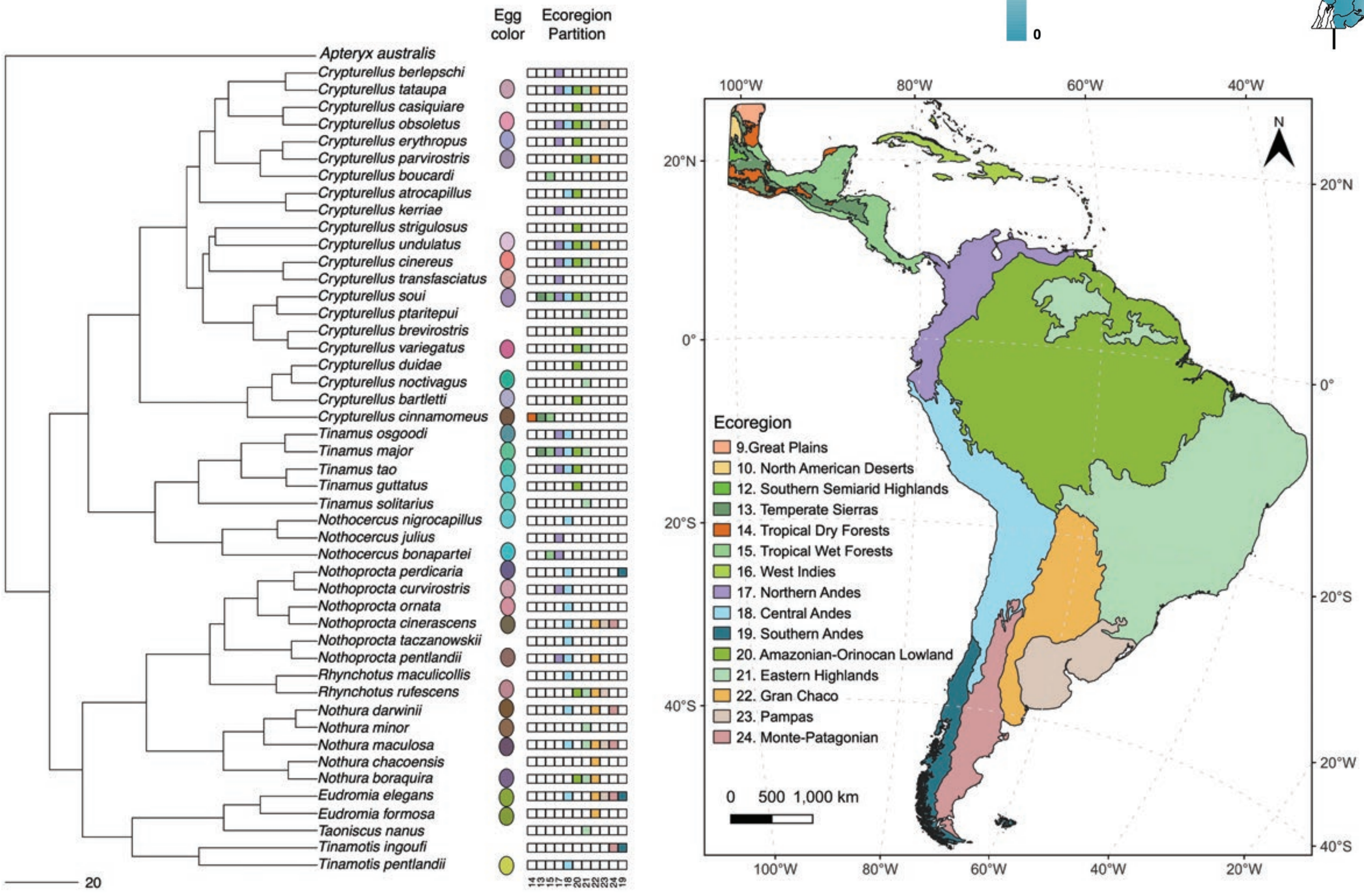
物种谱系关系及性状演化

- Phylogenetic relationship
- Character evolution



最大似然法
Maximum likelihood

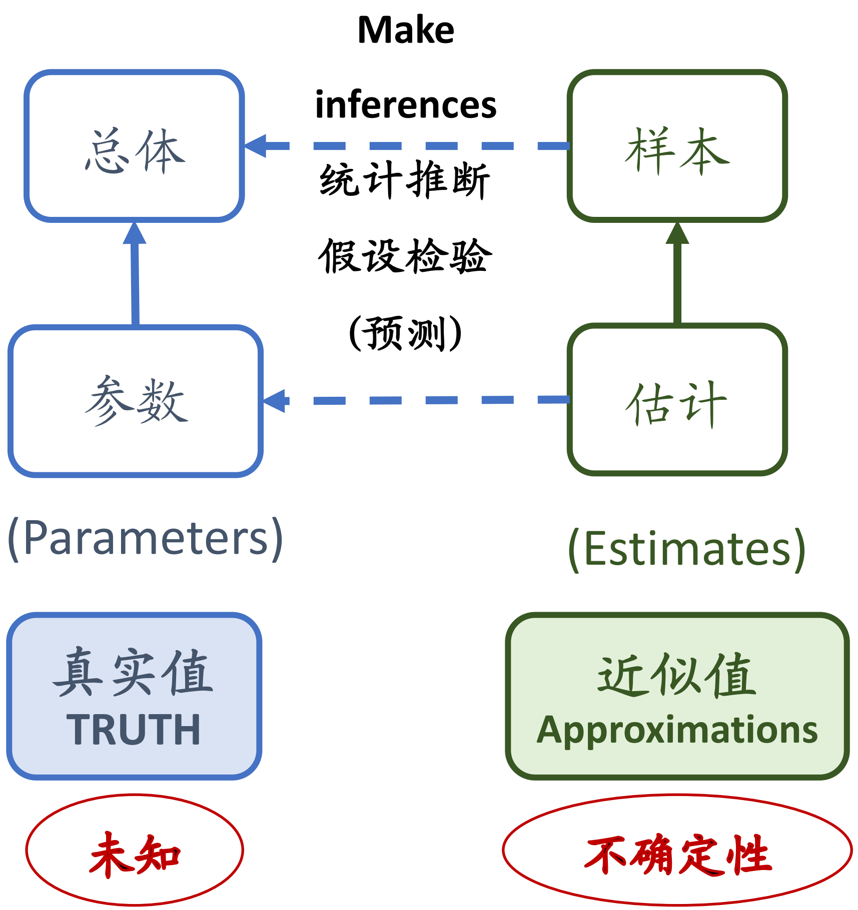
模型比较
Model comparison



(Li, et al. Evolution, 2023)

4. 课程总结

描述性统计量	Descriptive Statistics	总体 Population (parameter)	样本 Sample (estimate)
均值	Mean	μ	$\bar{Y}$
方差	Variance	σ^2	s^2
标准差	Standard deviation	σ	s
标准误	Standard error	$\sigma_{\bar{Y}}$	$\mathrm{SE}_{\bar{Y}}$
相关系数	Correlation coefficient	ρ	r
截距	Intercept	α	a
回归斜率	Regression coefficient	β	b

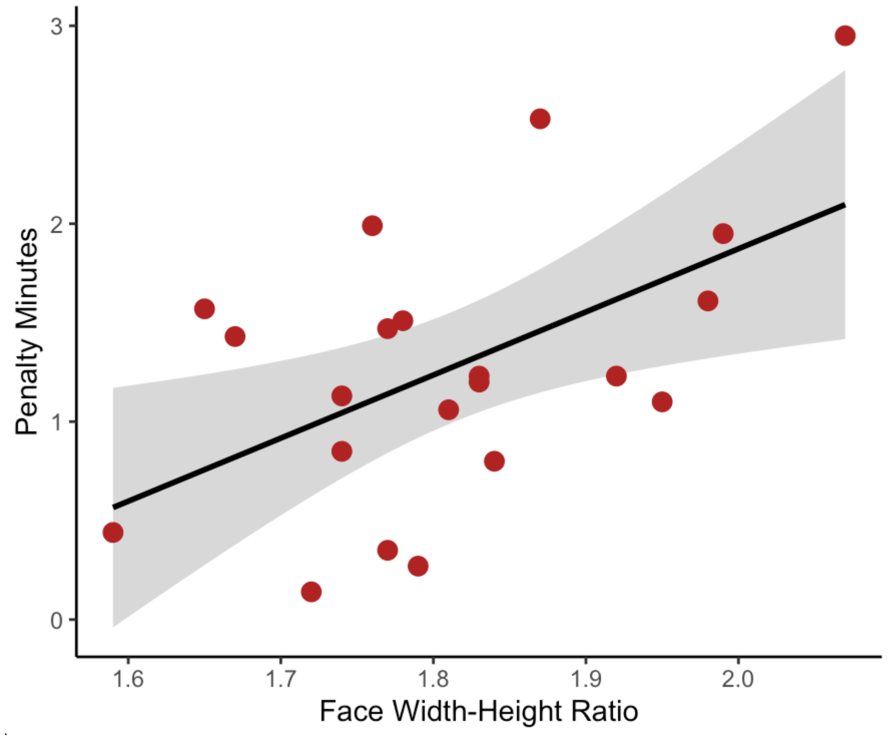


1. 回归——相关性和简单线性回归：课堂练习

- Ch17-1 Does face shape predict aggression?
  - 数据见"chap17q01FacesAndPenalties.csv"。
  - 计算面部宽高比与平均罚分的相关系数？ $r = 0.537$
  - 通过面部宽高比来预测罚分的回归线模型是？
    - $Y = -4.505 + 3.189X$
    - $0.783 < \beta < 5.595$

```r
face <- read.csv("R-Labs/chap17q01FacesAndPenalties.csv")

cor.test(~ PenaltyMinutes + FaceWidthHeightRatio,
         data = face)
faceRegression <- lm(PenaltyMinutes ~ FaceWidthHeightRatio,
                     data = face)
summary(faceRegression)
```



4. 课程总结

- 学无止境
- 成长型思维