Biodiversity data wrangling: Linking large phylogenies with species traits and ecologies

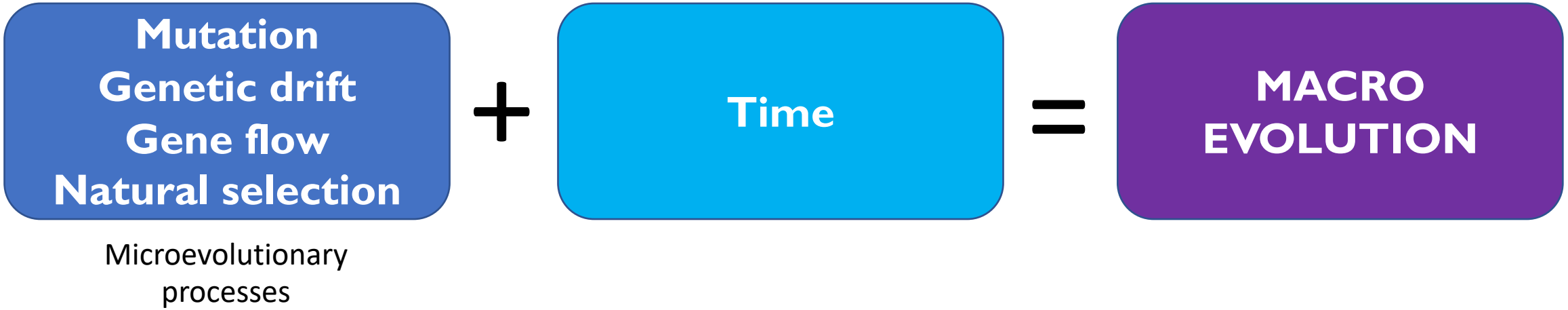
Module: Macroevolution

Carolina Siniscalchi

Botany 2022

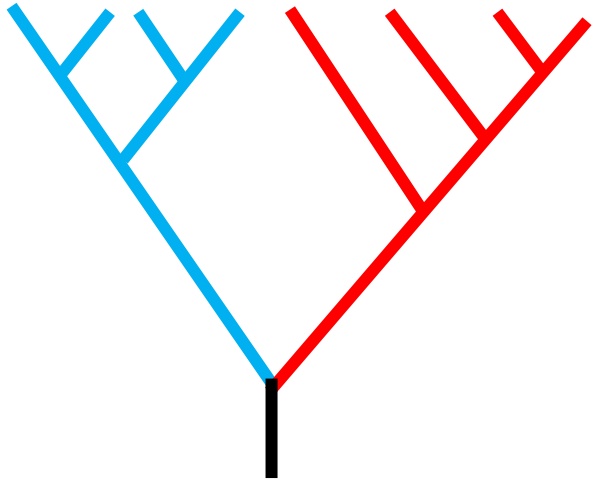
Macroevolution

- Evolution above the species level (e.g., clade level)
- Large trends and transformations in evolution:
    - Radiation of flowering plants
    - Appearance of specific traits, e.g., nitrogen fixation

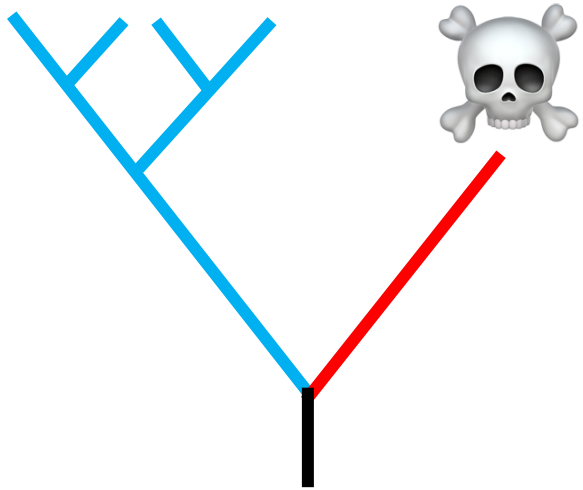


Macroevolution

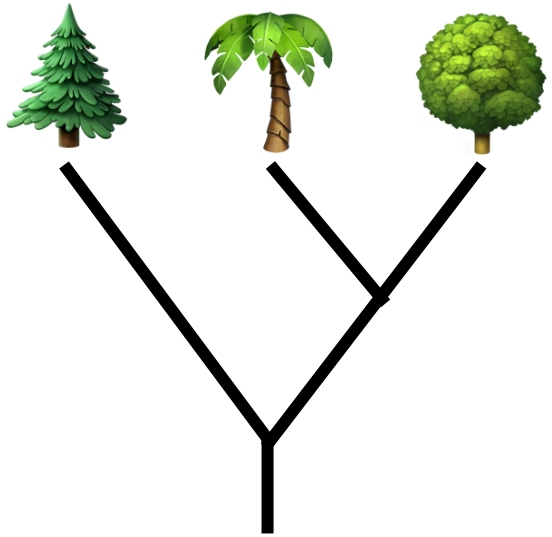
Some processes that are commonly studied in macroevolutionary studies:



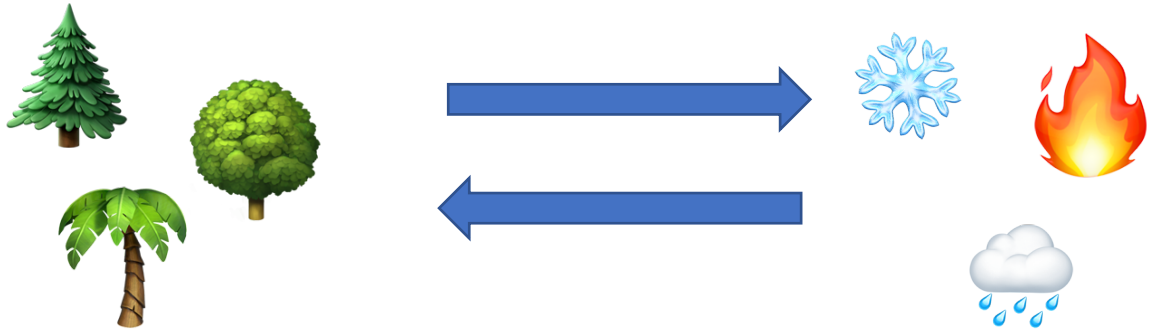
SPECIATION



EXTINCTION



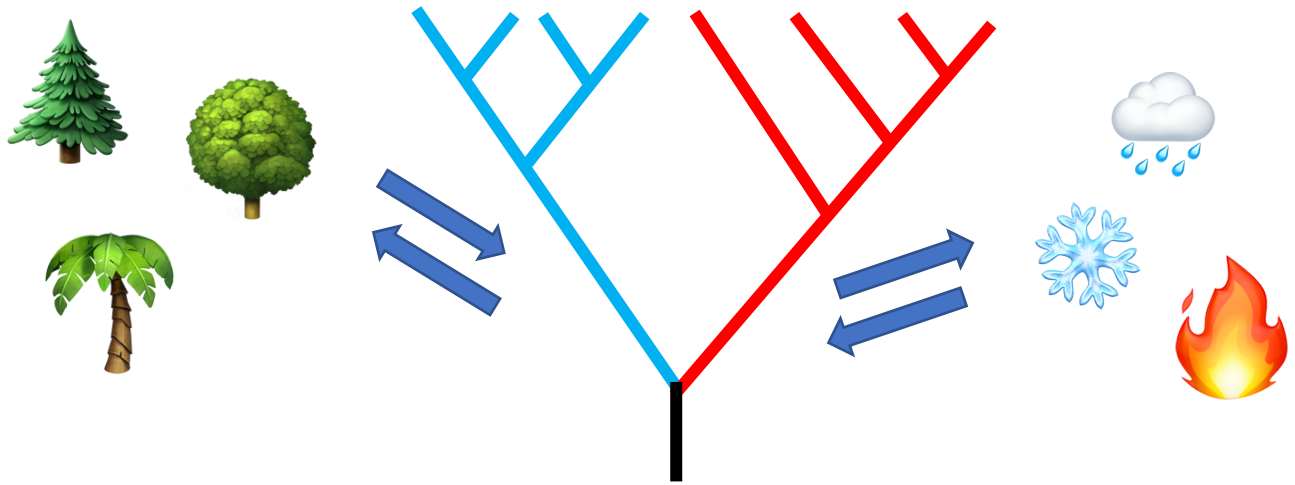
CHARACTER CHANGE



TRAIT/ENVIRONMENT INTERACTIONS

Macroevolution

- Phylogenies are fundamental parts of macroevolutionary studies!
- They provide a comparative framework in which we organize the information about the study organism
- We can extraction valuable information from the phylogenetic trees themselves (e.g., diversification rates, divergence times)



Macroevolution

- What this talk ISN'T about:
  - How to assemble sequence data
  - How to get a phylogenetic tree, comparison of methods, orthologs and paralogs, gene tree discordance, etc
  - How to date your tree or do biogeographical analysis
  - Philosophical discussions about macroevolution and criticism of methods
  - How to write code

Macroevolution

- What this talk IS about:
    - Give a user perspective of basic steps on macroevolutionary analysis, including:
        - Types of data that are frequently used in macroevolutionary studies
        - How to organize your data and get it ready for analysis
        - Some examples of analysis that combine phylogenetic, environmental and trait data

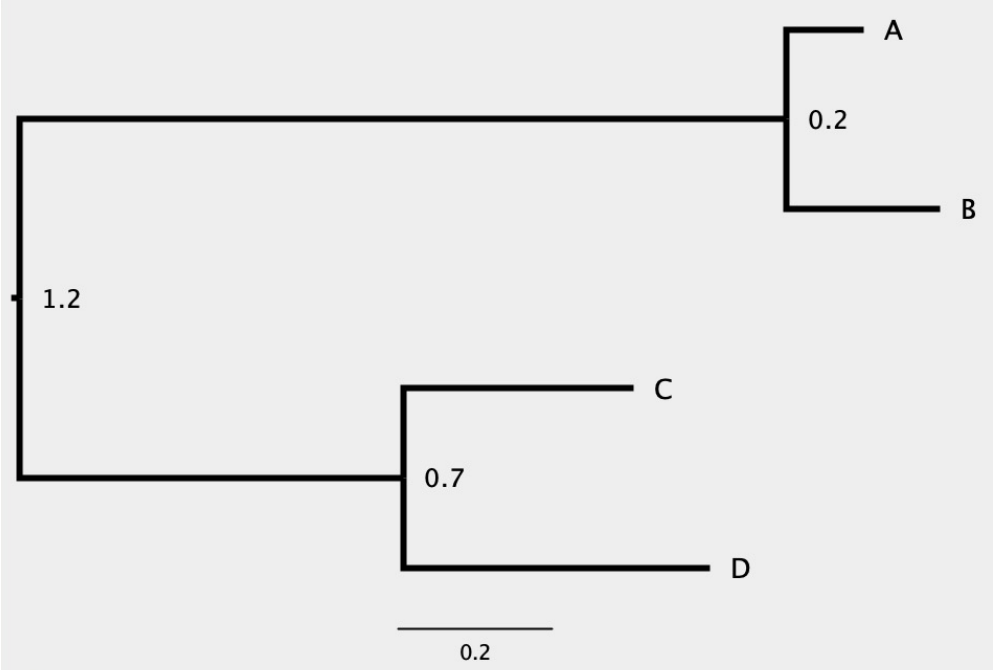
Step One: The Tree

Step One: The Tree

- Not a pdf with a tree figure on it
- We need a text-style file that R (and other programs/languages) can understand
- Newick (.tre, .tree, .nwk) or nexus (.nex)

Step One: The Tree

((A:0.1,B:0.2),(C:0.3,D:0.4):0.5);



Step One: The Tree

- Methods that involve diversification rates need trees that are:
  - dated, ultrametric
  - fully bifurcating (no polytomies)
- Some methods are sensitive to sampling, so the completeness of your tree should be considered

Step Two: Environmental Data

Step Two: Environmental Data

- First things first: we need to know where our species of interest are found
- Geographical records!
- Where to find them?

Step Two: Environmental Data

How to obtain geographical records:

1) herbarium specimens

- coordinates on the specimen

- georeferencing based on location

(manually or with tools like GeoLocate)

Step Two: Environmental Data

How to obtain geographical records:

2) literature review

- species descriptions (if recent)
- monographs
- floras

***Additional Specimens Examined (Paratypes)*—Brazil.** Bahia, Gentio do Ouro, Distrito de Santo Inácio, on rocky hillside called Pedra da Mulher just south of town, [-42.733333°, -11.116667°], ca. 500–600 m, 25 Feb 1977, *R. M. Harley 19029* (CEPEC, NY, RB, US); ibid., área muito seca, última chuva em dezembro de 1989, [-42.733333°, -11.116667°], ca. 500 m, 5 Oct 1990, *A. Freire-Fierro et al. 1782* (SPF); ibid., área muito seca, última chuva em dezembro de 1989, [-42.733333°, -11.116667°], ca. 500 m, 5 Oct 1990, *A. Freire-Fierro et al. 1787* (SPF); ibid., Caminho para Santo Inácio, [-42.710278°, -11.057778°], 680 m, 24 Jun 1996, *M. L. Guedes et al. 2998* (ALCB, CEPEC, HUEFS, SPF); ibid., ca. 24 km S de Xique-Xique, na Estrada para Santo Inácio, 16 Jun 1994, *L. P. de Queiroz & N. S. Nascimento 3958* (HUEFS); ibid., [-42.666667°, -11.100000°], 19 Jun 1998, *J. Santino de Assis 210* (RB); ibid., Vale das Pedras (CASF), [-42.721389°, -11.111111°], 14 Apr 2000, *S. S. Lima s.n.* (ALCB); ibid., dunas vicariantes, [-42.721389°, -11.111111°], 2 Jun 2000, *S. S. Lima s.n.* (ALCB); ibid., ramal para a cachoeira, [-42.721667°, -11.096667°], 536 m, 20 Jul 2000, *M. M. da Silva et al. 466* (HUEFS); ibid., Serra de Sapé, [-42.718722°, -11.189722°], 624 m, 26 May 2009, *J. A. Siqueira-Filho et al. 2060* (HVASF, RB); ibid., Serra do Sapê, Estrada de Gentio do Ouro para Santo Inácio, [-42.717506°, -11.192153°], 622 m, 7 Nov 2015, *C. M. Siniscalchi & J. Vidal 630* (SPF); ibid., Estrada para a cachoeira, acesso pela Estrada que liga a rodovia BA-330 a Santo Inácio, [-42.721667°, -11.096667°], 532 m, 7 Nov 2015, *C. M. Siniscalchi & J. Vidal 631* (SPF).

Step Two: Environmental Data

How to obtain geographical records:

3) GBIF

4) iNaturalist (GBIF pulls records from there)

SPECIES | ACCEPTED

Chresta harleyi H.Rob.

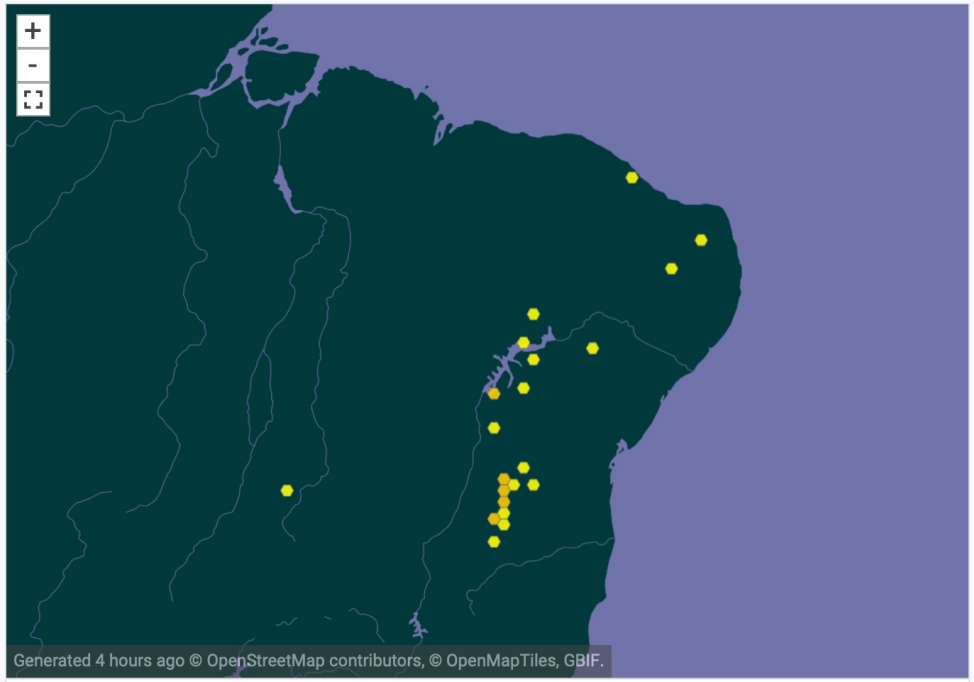
Published in: H. Rob. In: Phytologia, 53(6): 385. (1983).
source: Synonymic Checklists of the Vascular Plants of the World

168 OCCURRENCES

OVERVIEW METRICS REFERENCE TAXON

36 OCCURRENCES WITH IMAGES

126 GEOREFERENCED RECORDS



SEARCH OCCURRENCES | 168 RESULTS

TABLE GALLERY MAP TAXONOMY METRICS DOWNLOAD

Scientific name	Country or area	Coordinates	Month & year	Occurrence status
Chresta harleyi H.Rob.	Brazil	14.3S, 42.5W	2022 March	Present
Chresta harleyi H.Rob.	Brazil	8.7S, 41.6W	2020 July	Present
Chresta harleyi H.Rob.	Brazil	14.3S, 42.5W	2017 March	Present

Step Two: Environmental Data

- Always in decimal degrees
- Record clean-up: remove duplicate values, outliers, records that fall in the wrong place (e.g., in the ocean, in the middle of a city), around botanical gardens/herbaria, etc
- Always save as a .csv or .tsv file

ID	long	lat
Chresta_harleyi	-42.48333333	-14.15
Chresta_harleyi	-42.52116667	-14.25616667
Chresta_harleyi	-42.52133333	-14.256
Chresta_harleyi	-42.52	-14.25366667
Chresta_harleyi	-42.52233333	-14.26666667
Chresta_harleyi	-42.52283333	-14.26
Chresta_harleyi	-42.42143333	-14.83376667
Chresta_harleyi	-42.52516667	-14.53583333
Chresta_harleyi	-42.57116667	-14.74166667
Chresta_harleyi	-42.77116667	-15.38383333
Chresta_harleyi	-42.61666667	-15.33333333
Chresta_harleyi	-42.466667	-14.066667
Chresta_harleyi	-41.958611	-13.523056
Chresta_harleyi	-42.538333	-14.350556
Chresta_harleyi	-42.546111	-14.27
Chresta_harleyi	-42.515833	-14.897778
Chresta_harleyi	-42.524167	-14.571389
Chresta_harleyi	-42.510833	-14.628333
Chresta_harleyi	-43.572778	-14.745
Chresta_harleyi	-42.54	-14.586667

Step Two: Environmental Data

- OK, I have my geographical records, what now?

Step Two: Environmental Data

- High-resolution spatial data is increasingly becoming available for different variables
- BIOCLIM: 19 climatic variables – data from weather stations interpolated with satellite images, with 1 km^2 resolution
- Format: geotiff (image file [tiff] that contains georeferencing information)

INTERNATIONAL JOURNAL OF CLIMATOLOGY
Int. J. Climatol. (2017)
Published online in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/joc.5086

RMetS
Royal Meteorological Society

WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas

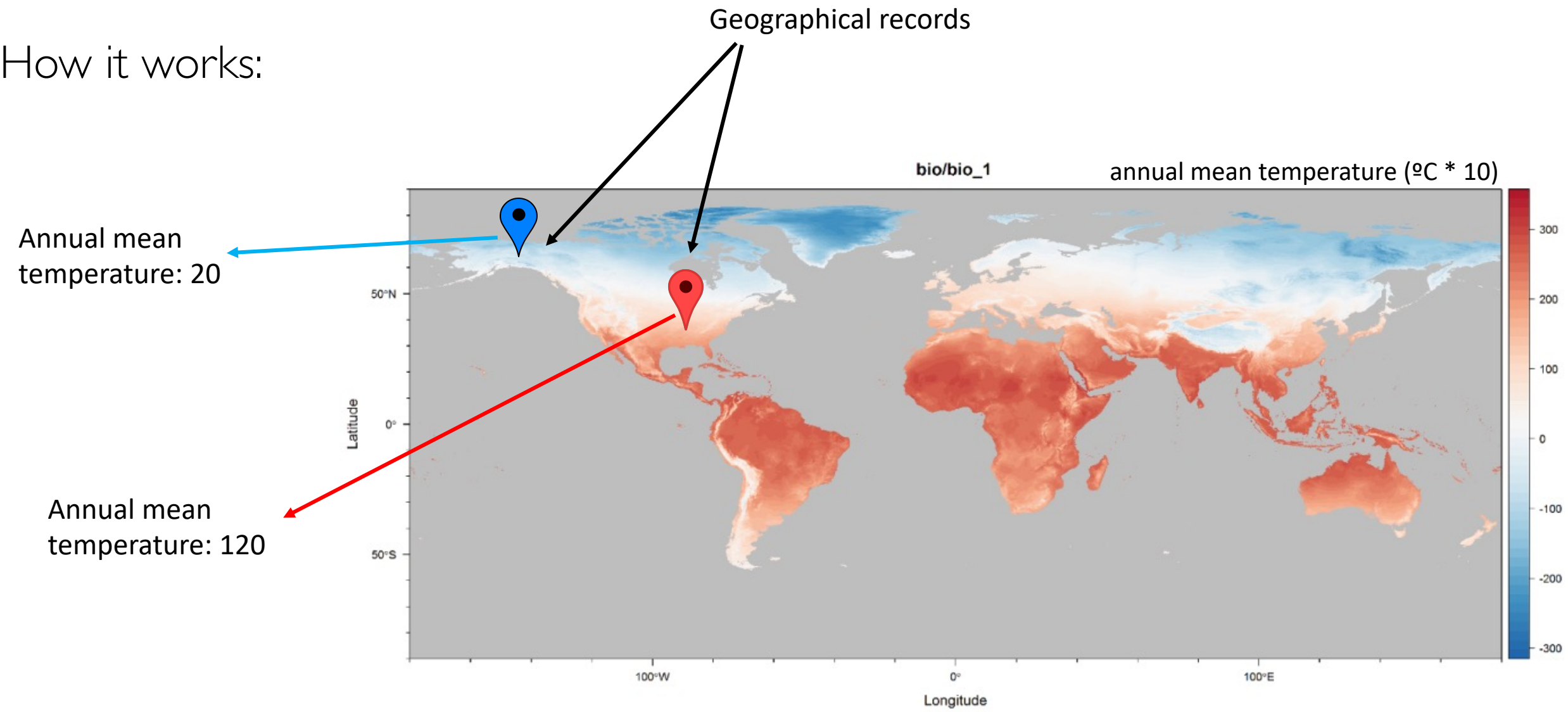
Stephen E. Fick[a*] and Robert J. Hijmans[b]

Step Two: Environmental Data

- Other sources of spatial data:
  - GTOPO30: global digital elevation model (elevation, aspect, slope)
  - SoilGrids: global digital mapping of soil characteristics (e.g., pH, carbon, nitrogen, sand, silt and clay content)
  - MODIS LandCover: global land cover classification (type of vegetation)

Step Two: Environmental Data

How it works:



Step Two: Environmental Data

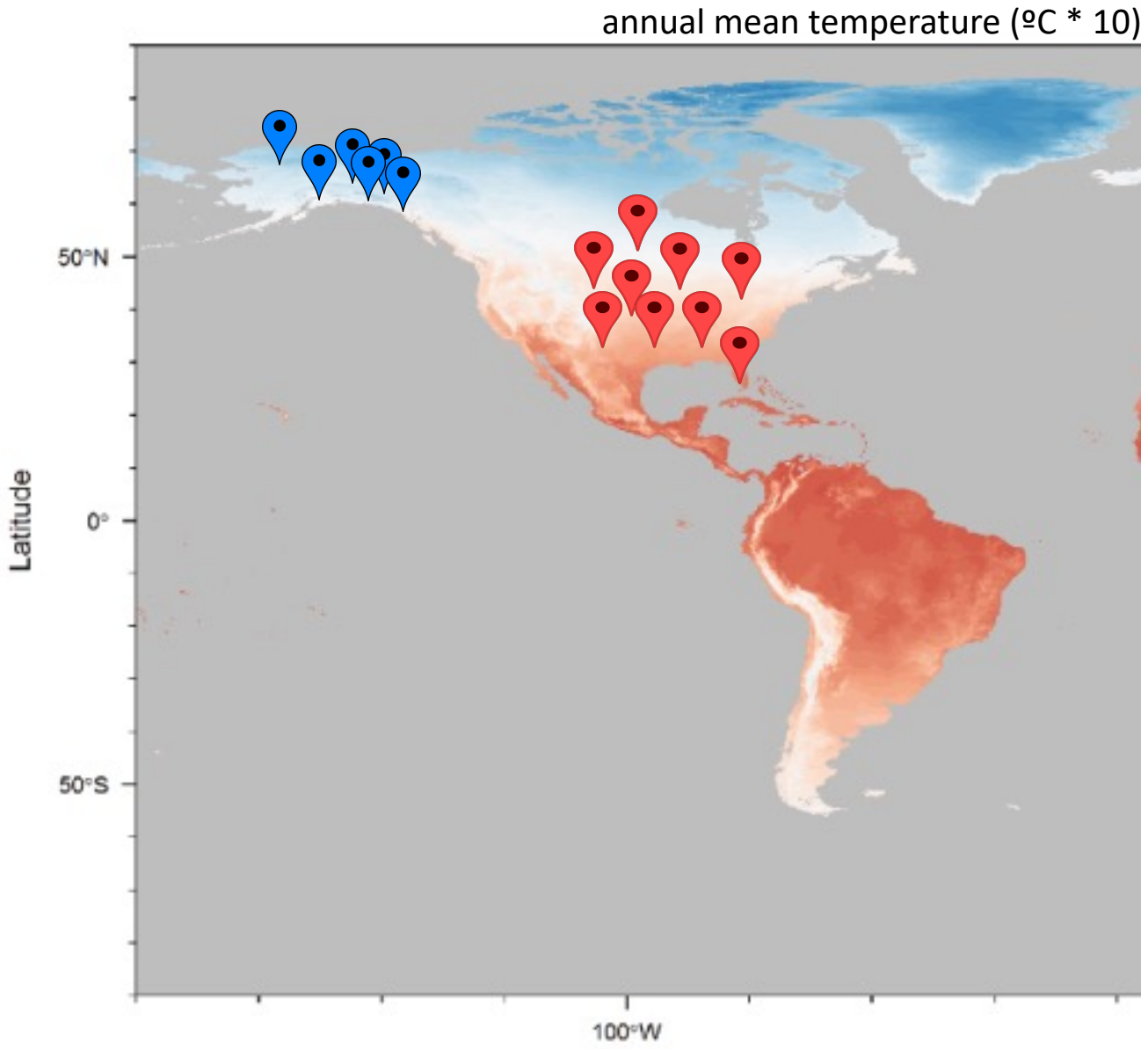
How it works:

- Look up the value of the variable in each geographical record of the species of interest
- Calculate the average of the variable for that species

Average species A: -5

Average species B: 80

- Rinse and repeat for all species/variables



Step Two: Environmental Data

– Also possible to calculate the range (max – min) of each species

– The final result will be a comma- or tab-delimited file with the average values of all variables for all species

– This process can be automatized using custom python or R scripts

Species	BIO1	BIO2	BIO3	BIO4	BIO5
A	8.85714	153.171	39.2571	7678.51	205.029
B	53.6434	139.91	44.006	6113.26	234.767
C	116.882	98.2353	47.7647	3461.71	223.824
D	91.0497	152.025	42.7516	7090.15	281.944
E	109.907	156.953	45.6512	6448.67	286.581
F	78.4754	134.754	37.623	7239.16	282
G	64.3875	158.546	38.3083	8676.86	281.806
H	20.837	153.266	40.0345	7827.42	226.505
I	87.2	147.72	44.2	5966.2	278.44
J	25.5769	149.154	39.1154	7253.92	221.385
K	52.0417	126.542	39.2917	6396.08	225.833
L	34.2329	155.959	39.8037	8010.14	242.648
M	46.6	144.971	42.4	6334.54	232.057
N	156.734	131.596	65.5963	1771.97	253.716
O	160.308	144.231	41.1538	6267.62	336.308
P	-12.3636	125.545	37.5455	6103.55	145

Step Two: Environmental Data

–comma-delimited file (.csv)

– each line is a row of the table

– columns separated by commas

– if you need to look at the values in a .csv file, it's easier to open them in excel

```
Species,BIO1,BIO2,BIO3,BIO4,BIO5
A,8.85714,153.171,39.2571,7678.51,205.029
B,53.6434,139.91,44.006,6113.26,234.767
C,116.882,98.2353,47.7647,3461.71,223.824
D,91.0497,152.025,42.7516,7090.15,281.944
E,109.907,156.953,45.6512,6448.67,286.581
F,78.4754,134.754,37.623,7239.16,282
G,64.3875,158.546,38.3083,8676.86,281.806
H,20.837,153.266,40.0345,7827.42,226.505
I,87.2,147.72,44.2,5966.2,278.44
J,25.5769,149.154,39.1154,7253.92,221.385
K,52.0417,126.542,39.2917,6396.08,225.833
L,34.2329,155.959,39.8037,8010.14,242.648
M,46.6,144.971,42.4,6334.54,232.057
N,156.734,131.596,65.5963,1771.97,253.716
O,160.308,144.231,41.1538,6267.62,336.308
P,-12.3636,125.545,37.5455,6103.55,145
```

Step Three: Trait Data

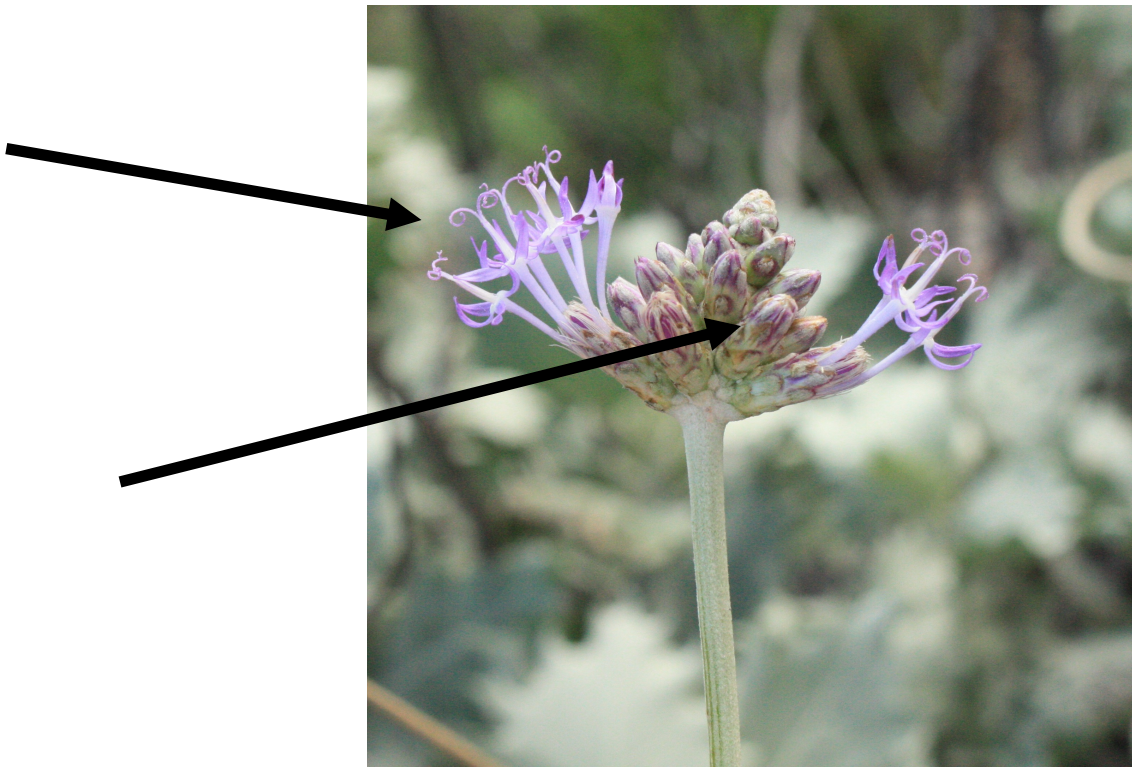
Step Three: Trait Data

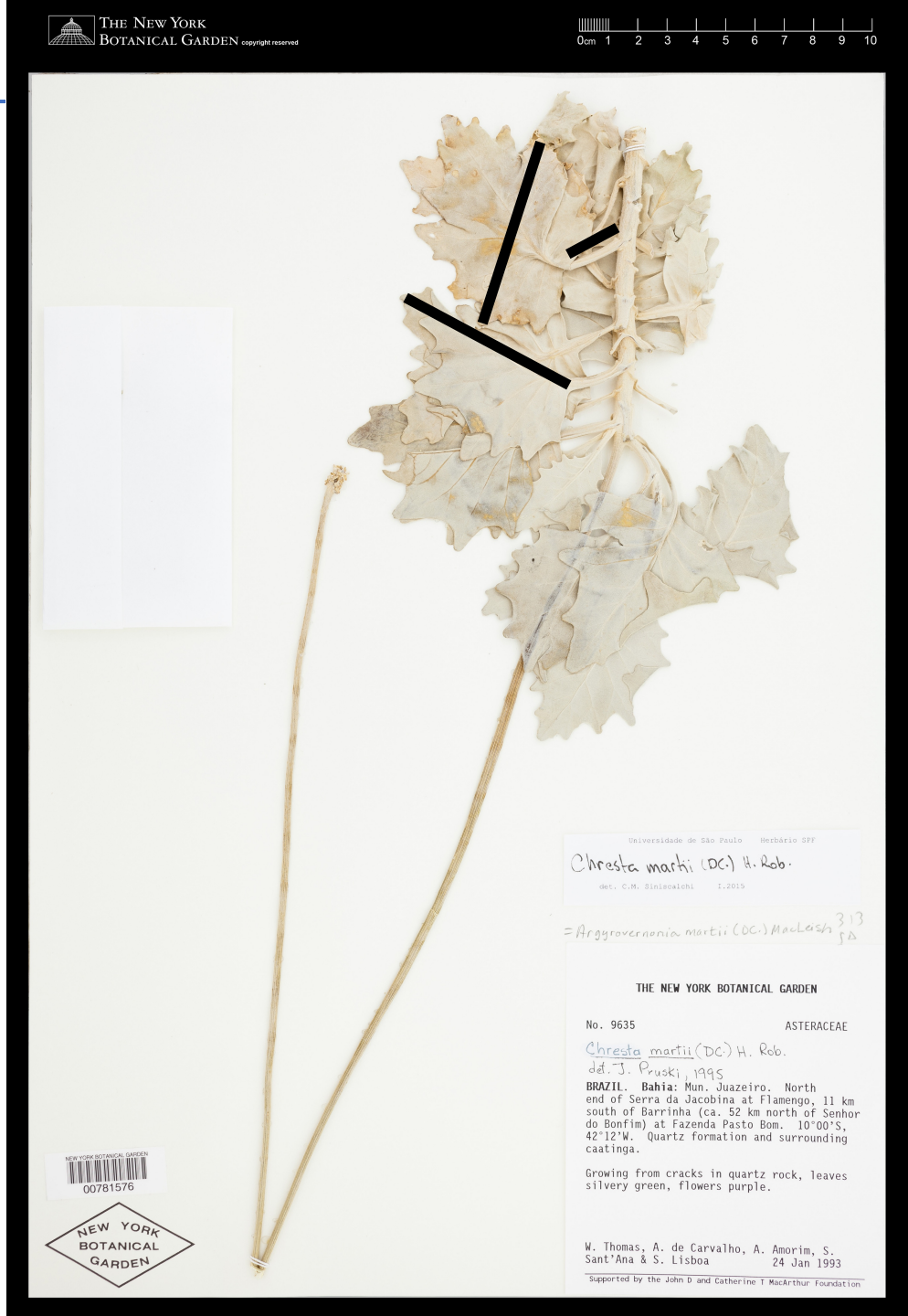
- In macroevolutionary studies, we select traits that could potentially explain evolutionary processes in our group of interest, such as increased diversification and occupation of new habitats.
- Some examples:
    - Ancient and current polyploidy
    - Morphological changes that impact pollination and pollinator type
    - Changes in photosynthesis metabolism
    - Occurrence of symbiosis, like nodulation

Step Three: Trait Data

- Where to obtain traits?

  1) Live and preserved specimens





Step Three: Trait Data

- Where to obtain traits?

  2) Literature (monographs, floras)

TABLE 1. Comparison of key characters of *Chresta artemisiifolia* and related species.

	C. artemisiifolia	***C. harleyi***	***C. hatschbachii***	***C. subverticillata***	***C. martii***
Leaves	non-aromatic	non-aromatic	non-aromatic	non-aromatic	aromatic
Leaf blade	constricted, following the venation	expanded	expanded	expanded	expanded
Syncephalium growth	determinate	determinate	determinate	sometimes indeterminate, resembling spikes	indeterminate
Scape length (cm)	5–11.5	12–38	1–5	6–25	30–60
Phyllary apex	acute	acute	acute	acute	rounded
Twin-hairs	adjoined up to the end, both cells with similar length	adjoined up to the end, both cells with similar length	adjoined up to the end, both cells with similar length	adjoined up to the end, both cells with similar length	separate at the base, unequal length
Pollen type	*Chresta*-type II	*Chresta*-type II	*Chresta*-type II	*Chresta*-type II	*Chresta*-type I

Step Three: Trait Data

- Where to obtain traits?

  3) Specialized databases (e.g., CCDB, genome size database)

1

Summary Statistics				
	Mean	Min	Max	Std Dev
1C (pg)	1.48	0.79	2.64	0.70

Genus	Species	Subspecies	DNA Amount 1C (pg)	Original Reference
Vernonia	incana		0.79	Vega & Dematteis,2017
Vernonia	echioides		1.23	Vega & Dematteis,2017
Vernonia	fasciculata	subsp. fasciculata	1.25	Sonnier,2016
Vernonia	echioides		2.64	Vega & Dematteis,2017

CCDB → *Angiosperms* → *Compositae* → *Chresta*

2 species in *"Chresta"* :

Show Statistics *Export to CSV* *Export F*

Taxon name	Status	Median (n)
Chresta angustifolia Gardner	Accepted	17
Chresta angustifolia	Synonym of Chresta angustifolia Gardner	17

Step Three: Trait Data

- Where to obtain traits?

  4) Specialized literature (e.g., carbon isotope reviews, books about nodulation, etc)

Table 1.4 Known nodulation of genera (as in Lewis et al., 2005), in tribe Ingeae

Genus	Species	Nod.
Abarema Pittier	46	13
Albizia Durazz	120–140	46
Archidendron F. Muell.	94	6
Archidendropsis F. Muell.	14	2
Blanchetiodendron Barneby & Grimes	1	?
Calliandra Benth.	~135	25
Cathormium (Benth.) Hasske.	1	?
Cedrelinga Ducke	1	1
Chloroleucon (Benth.) Britton & Rose	10	4
Cojoba Britton & Rose	12	1
Ebenopsis Britton & Rose	3	2
Enterolobium Mart.	11	8
Faidherbia A. Chev.	1	1
Falcataria (Nielsen) Barneby & Grimes	3	1
Guinetia L. Rico & M. Sousa	1	?
Havardia Small	5	2
Hesperalbizia Barneby & Grimes	1	1
Hydrochorea Barneby & Grimes	3	3
Inga Mill.	~300	63
Leucochloron Barneby & Grimes	4–5	?
Lysiloma Benth.	8–9	4
Macrosamanea Britton & Rose	11	3
Marmaroxylon Killip	9–13	?
Painteria Britton & Rose	3	?
Pararchidendron I.C. Nielsen	1	?
Paraserianthes I.C. Nielsen[1]	1	1
Pithecellobium Mart.	18	6
Pseudosamanea Harms	2	1
Samanea Merr.	3	2
Serianthes Benth.	~18	2
Sphinga Barneby & Grimes	3	?
Thailentadopsis Kosterm	3	?
Viguieranthus Villiers	~23	?
Wallaceodendron Koord.	1	1
Zapoteca H.M. Hern.	20	0?
Zygia P. Browne	45–50	10

[1] The number of species listed in both ILDIS and GRIN is 3, but there is only one in Lewis et al. (2005).

Step Three: Trait Data

Different types of trait data:

- Discrete

  (binary or multi-state)

- Continuous

taxa	style_length	flower_color	sweeping_hair
Chresta_plantaginifolia	11	purple	acute
Chresta_angustifolia	12	purple	subulate
Chresta_souzae	10	purple	subulate
Chresta_scapigera	10	purple	subulate
Chresta_sphaerocephala	9	purple	clavate
Chresta_exsucca	9	purple	clavate
Chresta_pycnocephala	8	purple	subulate
Chresta_curumbensis	30	red	subulate
Chresta_speciosa	37	red	subulate
Chresta_martii	11	purple	subulate
Chresta_harleyi	11	purple	subulate
Chresta_pacourinoides	10	purple	subulate
Chresta_filicifolia	13	purple	lageniform
Chresta_subverticillata	12	purple	subulate
Chresta_heteropappa	5.5	purple	lageniform
Chresta_artemisiifolia	10.5	purple	lageniform
Chresta_hatschbachii	11	purple	subulate

Step Three: Trait Data

Different types of trait data:

- Discrete

  (binary or multi-state)
- Continuous
- .csv!
- No spaces between

  words!

```
taxa,style_length,flower_color,sweeping_hair
Chresta_plantaginifolia,11,purple,acute
Chresta_angustifolia,12,purple,subulate
Chresta_souzae,10,purple,subulate
Chresta_scapigera,10,purple,subulate
Chresta_sphaerocephala,9,purple,clavate
Chresta_exsucca,9,purple,clavate
Chresta_pycnocephala,8,purple,subulate
Chresta_curumbensis,30,red,subulate
Chresta_speciosa,37,red,subulate
Chresta_martii,11,purple,subulate
Chresta_harleyi,11,purple,subulate
Chresta_pacourinoides,10,purple,subulate
Chresta_filicifolia,13,purple,lageniform
Chresta_subverticillata,12,purple,subulate
Chresta_heteropappa,5.5,purple,lageniform
Chresta_artemisiifolia,10.5,purple,lageniform
Chresta_hatschbachii,11,purple,subulate
```

Step Three: Trait Data

Different types of trait data:

- Discrete (binary or multi-state)
- Continuous

- The type of trait data will influence what type of analysis you can run
- Some methods are developed thinking specifically about one type of data, and don't apply to the other (e.g., equal-rates state reconstruction only with discrete data, phylogenetic signal methods initially developed for continuous data)
- Some methods allow both types of data (e.g., PGLS)

Now let's get to it!

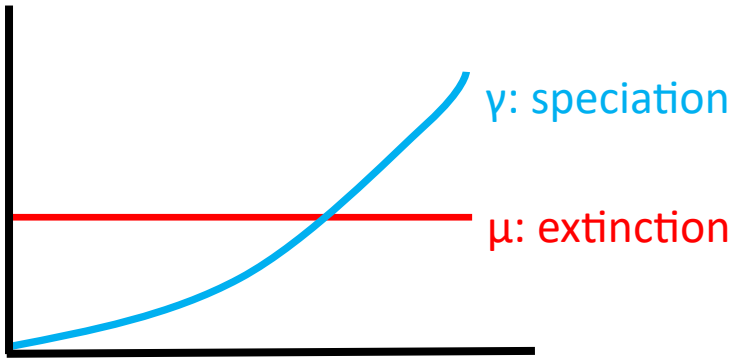
Macroevolution

Some questions we can ask:

- Is a certain trait related to changes in diversification rates?
- Are certain environmental conditions related to shifts in diversification?
- Is a certain trait related to shifts to different types of environment?
- Are environmental shifts related to morphological changes?
- Among many others….

Diversification rates

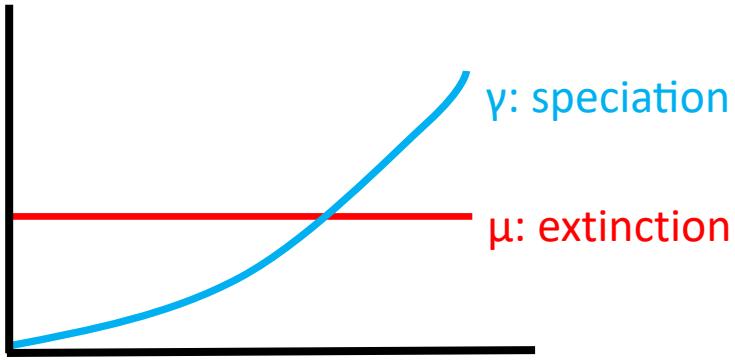
- What are diversification rates?
  - Rates at which new species form (speciation) and go extinct (extinction) – birth-death process



- There are many methods available to estimate diversification rates
  - Fossil time series
  - Phylogenetic trees

Diversification rates

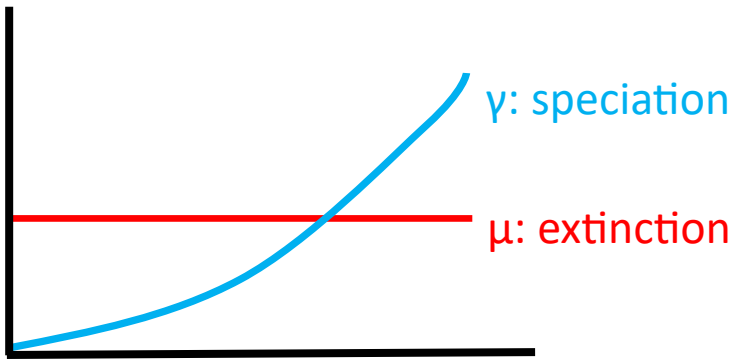
- Diversification rates are dependent both on:
  - Sampling (completeness of the tree)
  - Branch lengths (the distance among species on the tree)



- Some methods available:
  - BAMM (Bayesian estimation of different models of diversification)
  - Species-specific diversification rates (DR statistics)
  - MEDUSA: detect shifts in diversification in a tree (accepts incomplete trees)

Diversification rates

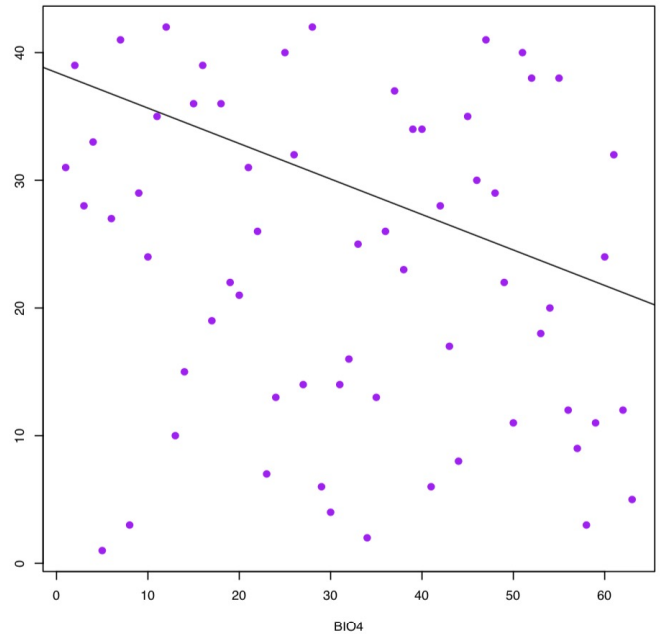
- Species-specific diversification rates (DR statistics)
  - estimate of the present-day rate of speciation or extinction for an individual lineage, conditional on past evolutionary history
  - several ways of calculating it, using information from different parts of the tree, we'll use the method proposed by Jetz et al. 2012



Chresta_filicifolia	0.15269725
Chresta_pacourinoides	0.25424341
Chresta_heteropappa	0.25424341
Chresta_subverticillata	0.20858697
Chresta_hatschbachii	0.20858697
Chresta_harleyi	0.18106719
Chresta_artemisiifolia	0.15423625
Chresta_curumbensis	0.19901707
Chresta_pycnocephala	0.24776732
Chresta_scapigera	0.29697164
Chresta_sphaerocephala	0.33717403
Chresta_exsucca	0.33717403
Chresta_angustifolia	0.19757658
Chresta_souzae	0.23982804
Chresta_plantaginifolia	0.28362043
Chresta_speciosa	0.28362043

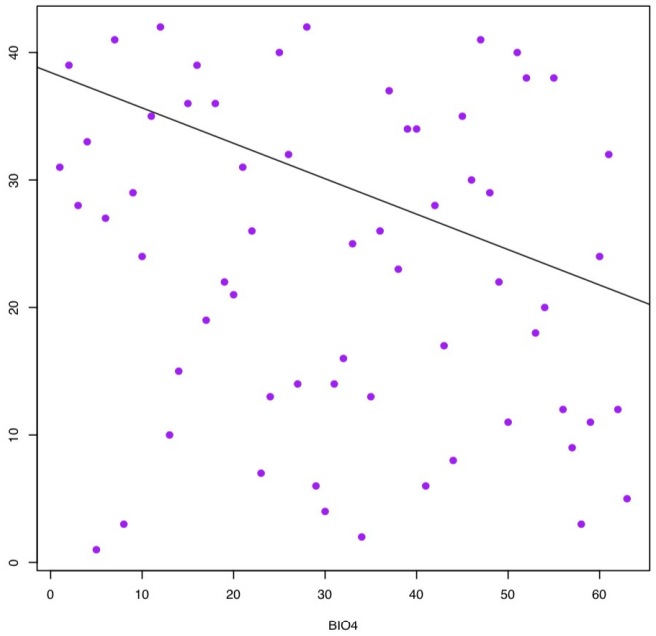
Linear regressions and PGLS

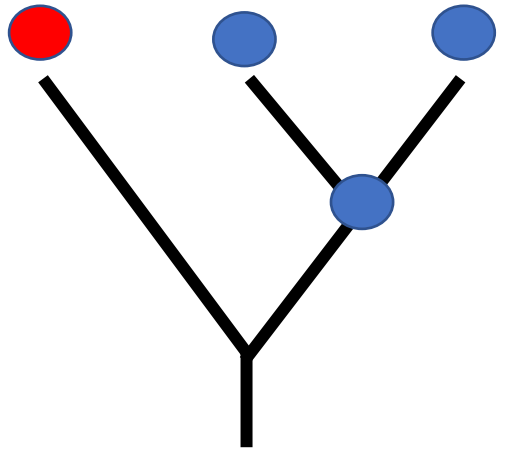
- We can use regressions to investigate if two variables are correlated, e.g.,
  - Is floral tube length related to nectar production?
  - Is aridity related to diversification?
  - Is a polyploidy associated to the occupation of drier environments?
  - Is nitrogen fixation related to poor soils?



Linear models and PGLS

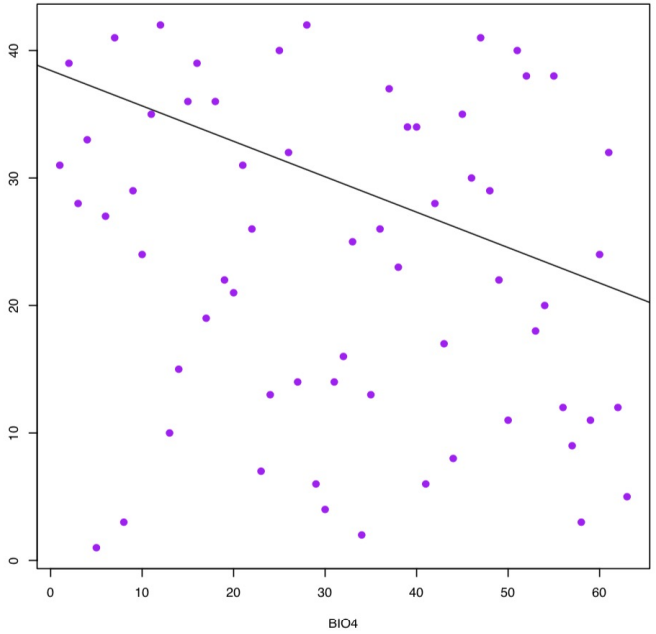
- We can use linear regressions to investigate if two variables are correlated
  - However, because of shared ancestry between species, they do not provide independent datapoints
  - Violates fundamental assumptions of several statistical tests





Linear models and PGLS

- PGLS is a modification of traditional generalized least squares
- It assumes that closely related species will have more similar values
- Accounts for specific autocorrelation due to phylogenetic history
- Several ways to calculate, pgls function implemented on the R package caper



- Response variable: continuous trait
- Predictor variable: continuous or binary discrete data

SSE Models

- Character state-dependent speciation and extinction
- Birth-death process where the diversification rates are dependent on the state of an evolving character

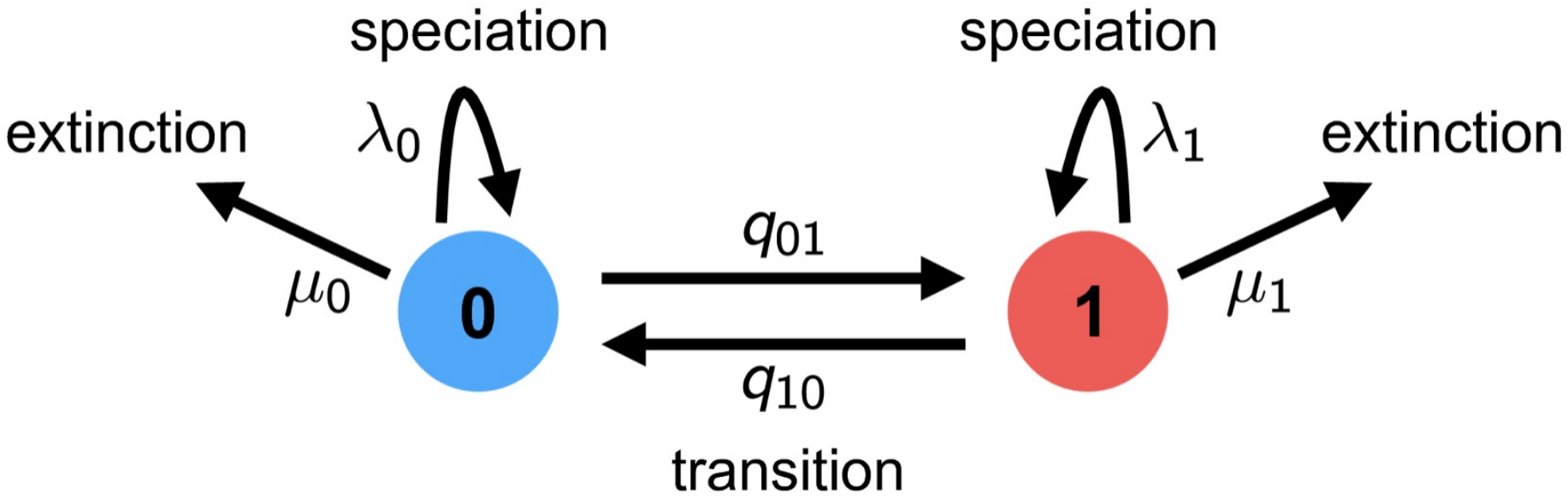


Figure 1. A schematic overview of the BiSSE model. Each lineage has a binary trait associated with it, so it is either in state 0 (blue) or state 1 (red). When a lineage is in state 0, it can either (a) speciate with rate λ_0 which results into two descendant lineage both being in state 0; (b) go extinct with rate μ_0; or (c) transition to state 1 with rate q_{01}. The same types of events are possible when a lineage is in state 1 but with rates λ_1, μ_1, and q_{10}, respectively.

Figure from RevBayes website

SSE Models

- Initial model (BiSSE) uses ordinary differential equations to calculate the probability of observing changes in descendent clades in a branch of a phylogeny
    - Only binary discrete characters
- Other models were subsequently developed:
    - MuSSE: multistate characters
    - HiSSE: hidden-state speciation and extinction
    - ClaSSE: cladogenetic state change (biogeographic range evolution)
    - ChromoSSE: chromosome evolution
    - QuaSSE: quantitative speciation and extinction (continuous traits)
    - FiSSE: non-parametric test for binary characters

SSE Models

- Several models implemented on the R package diversitree or on their own packages (e.g., hisse)
- Several models also implemented on RevBayes (great intro on SSE methods here: https://revbayes.github.io/tutorials/sse/bisse-intro.html)

- There's been some criticism of some of these methods, due to concerns of sensitivity to model inadequacy and phylogenetic pseudoreplication. Some methods were created to control these issues (e.g., FiSSE, HiSSE).