Collaborators: Tom Artois, Marlies Monnens, Laura Vanstraelen

Talk structure

I. How we assembled the data

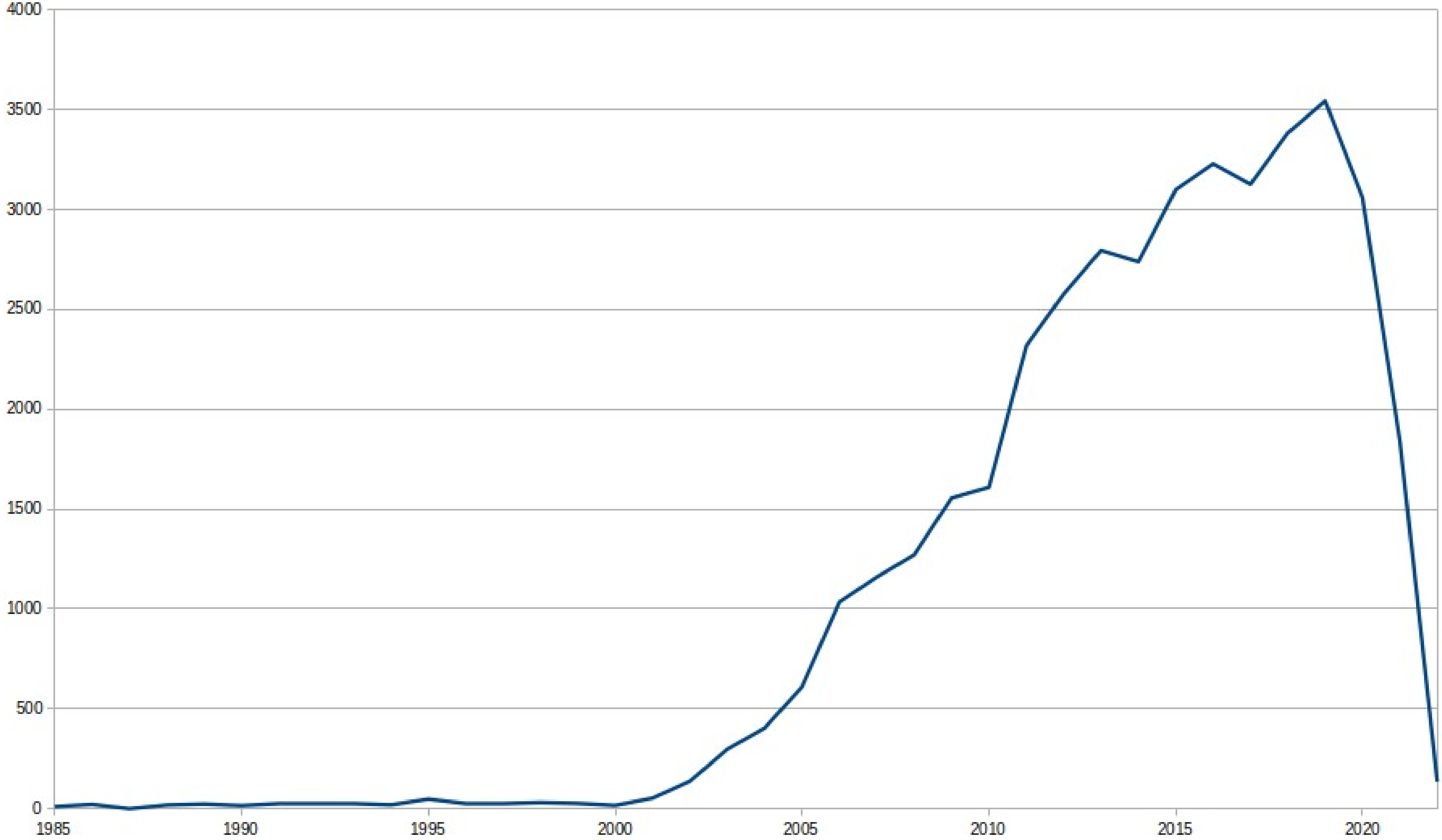
- Corpus assembly
- Extracting taxa and locations
- Topic modelling
- Disagreement
- Taxonomic methods

II. What we might do with the data

- Molecular vs Morphological taxonomy
- North/South divide
- Bias in taxonomy: forests and terrestrial species
- Disagreement in taxonomy

Corpus Content

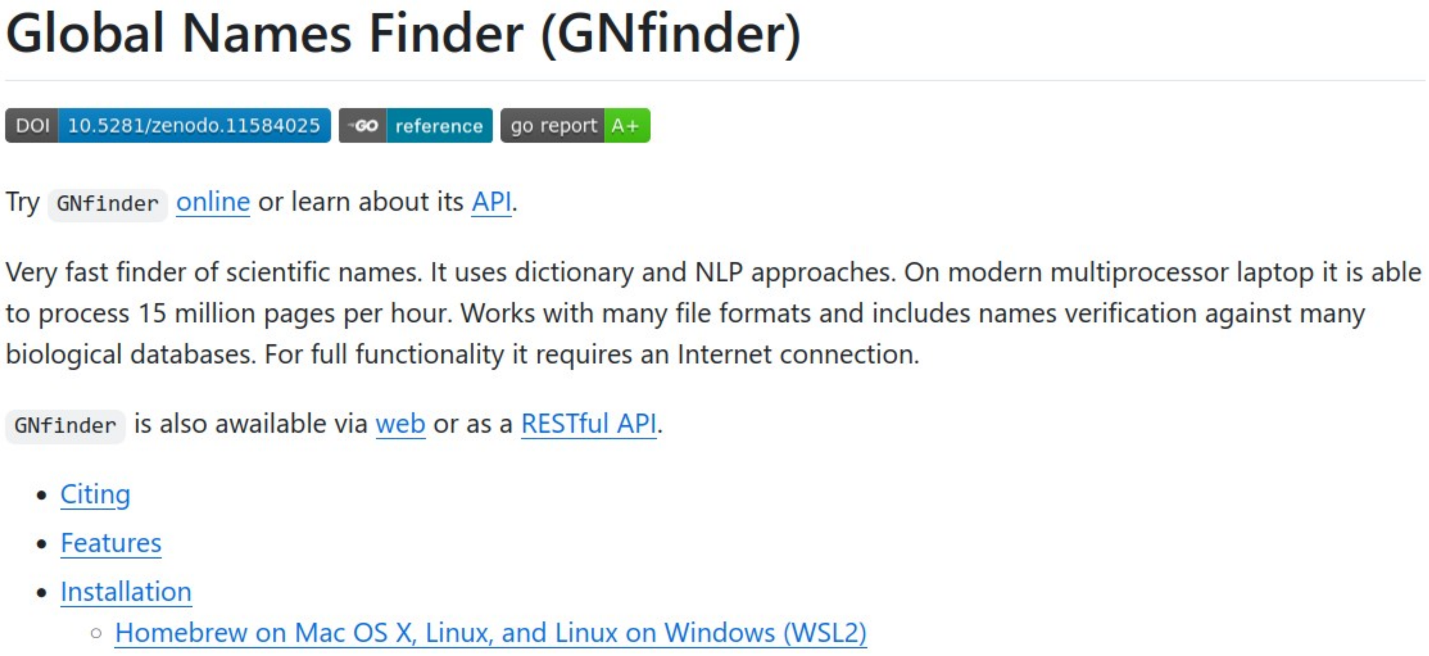
Journal	Publisher	Size
Zootaxa	Magnolia Press	31,348
ZooKeys	Pensoft	4,940
PhytoKeys	Pensoft	820
Journal of Hymenoptera Research	Pensoft	382
MycoKeys	Pensoft	315
Zoosystematics and Evolution	Pensoft	153
Insecta Mundi	Center for Systematic Entomology	1,367
European Journal of Taxonomy	Muséum National d’histoire naturelle	1,105



4000
3500
3000
2500
2000
1500
1000
500
0
1985
1990
1995
2000
2005
2010
2015
2020

Extracting Taxa

Global Names Finder (gnfinder): match species names in text, both against dictionaries and by looking for “likely” species names



Global Names Finder (GNfinder)

Try `GNfinder` online or learn about its API.

Very fast finder of scientific names. It uses dictionary and NLP approaches. On modern multiprocessor laptop it is able to process 15 million pages per hour. Works with many file formats and includes names verification against many biological databases. For full functionality it requires an Internet connection.

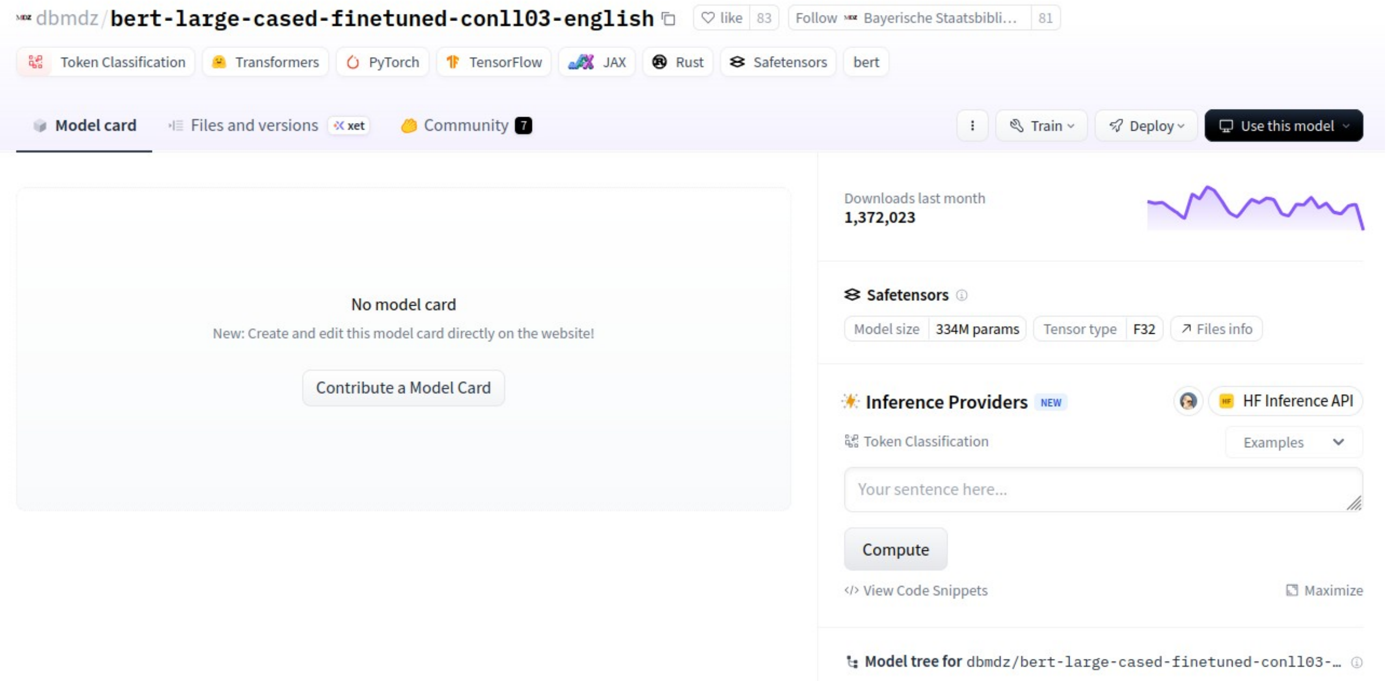
`GNfinder` is also available via web or as a RESTful API.

- Citing
- Features
- Installation
  - Homebrew on Mac OS X, Linux, and Linux on Windows (WSL2)

https://github.com/gnames/gnfinder

Extracting Locations

A case-sensitive language model trained on the "CoNLL-2003" dataset for recognizing locations, organizations, and persons in English text by the Bayerische Staatsbibliothek:



dbmdz / bert-large-cased-finetuned-conll03-english

No model card

New: Create and edit this model card directly on the website!

Downloads last month
1,372,023

Safetensors

Model size 334M params
Tensor type F32

Inference Providers NEW

Your sentence here...

Model tree for dbmdz/bert-large-cased-finetuned-conll03-...

A Problem

We can’t do any automated analysis of pieces of text that describe place names! We have to convert them to latitude and longitude coordinates.

And proper reverse geocoding is too expensive (€€).

So let’s use a gazette! Which works great, but is *very* slow.

Gazette Location Matching

1) Download a list of place names and their corresponding latitudes and longitudes

2) Load the whole thing into a database

3) Try matching unambiguous hits for location names in the database

4) If that doesn’t work, try approximate matches

5) If there’s more than one, try to compute the “geographic center” for the things that already matched, and return the hit closest to that

Topic Modeling

Embed documents into a 400-dimensional vector space using the doc2vec algorithm, and then examine the pattern of clusters within that high-dimensional space.

Less useful in this dataset: Very often seems to pick out topics that describe how scientists talk about different groups of organisms (“fin, rays, gill, pectoral…”) but occasionally some topics might have other meaning (“taxonomists, barcoding, biodiversity, dna…”).

Measuring Disagreement

Three lists of terms: *disagreement*, *epistemic value*, and *pejorative* language terms, extracted from journal articles

e.g., **disagreement:** critique, doubt, opinion, disagree, redundant, reject, rebuttal, debate, object, invalid, misunderstanding, misconception, allegation, allegedly, mistake, obsolete, error, misclassify, erroneously, contentious

Measuring Disagreement

e.g., **disagreement:** critique, doubt, opinion, disagree, redundant, reject, rebuttal, debate, object, invalid, misunderstanding, misconception, allegation, allegedly, mistake, obsolete, error, misclassify, erroneously, contentious

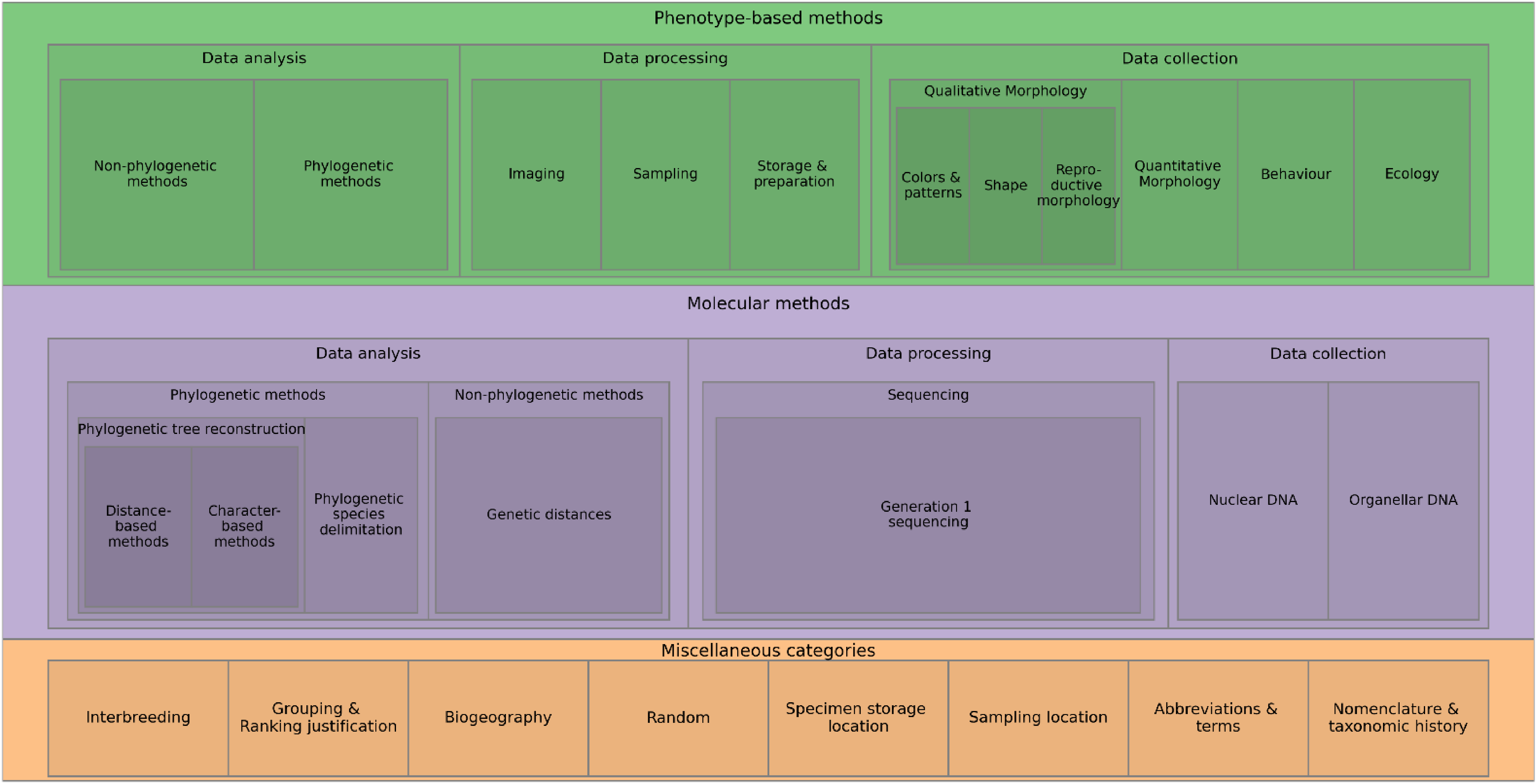
Measure the relative frequency of those terms within each journal article to give each paper a “disagreement index.”

Extracting taxonomic methods

- Methods as an interesting proxy to the kinds of science that are done and the kinds of knowledge that are created
- Tricky (in taxonomy):
  - No ‘standard’ references for methods
  - Different research traditions (taxa) ➔ different terminologies
  - No tradition of extensive reporting: exploratory science
  - Many amateurs and researchers from the south
- Interesting:
  - 1995 – 2020: rapid change towards ‘a new taxonomy’

Extracting methods: approach

1. Choosing the general structure of a hierarchical classification
2. Isolating methods sections
3. Exploratory analysis through topic modelling
4. Annotating random methods paragraphs
5. Finalizing and reviewing the classification
6. Training classifiers
7. Annotating targeted methods paragraphs
8. Training classifiers
9. Comparing with LLMs

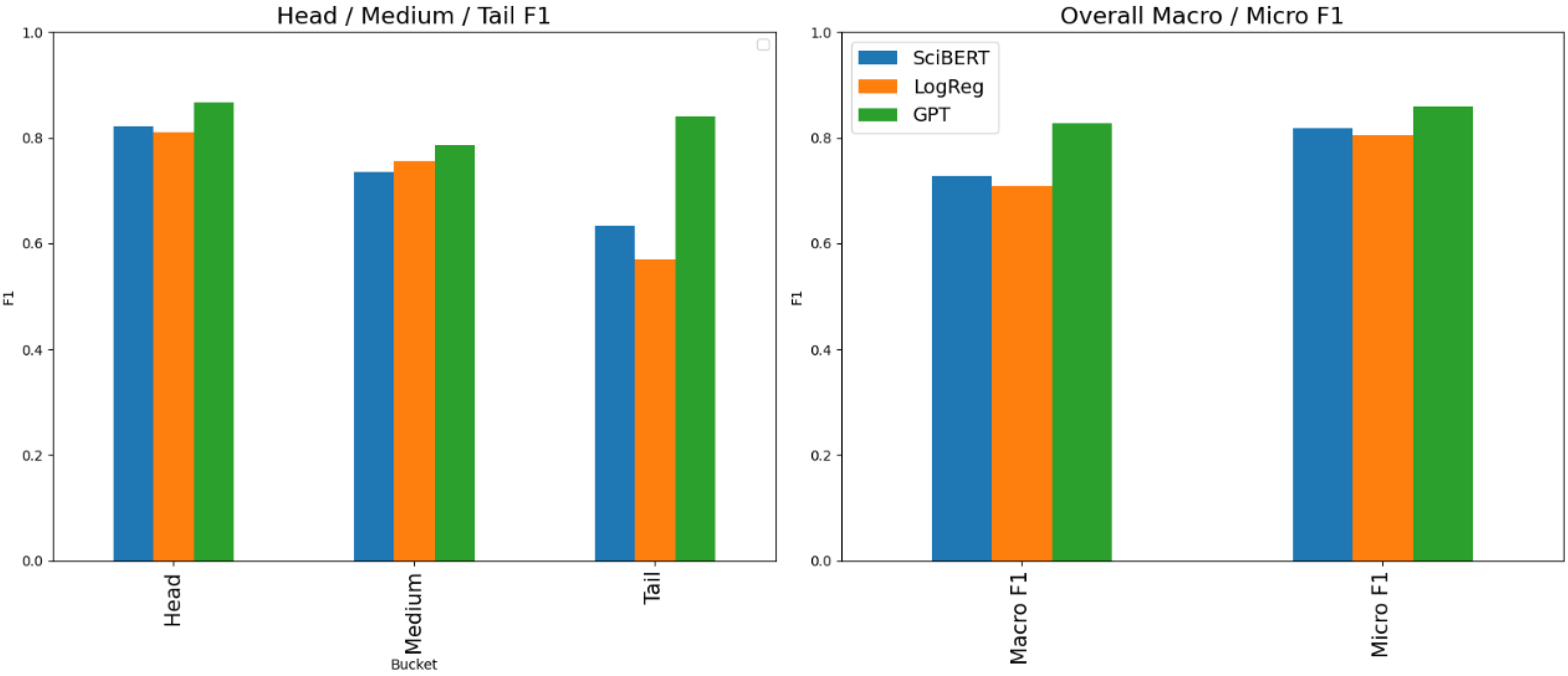

Phenotype-based methods
Data analysis
Non-phylogenetic methods
Phylogenetic methods
Data processing
Imaging
Sampling
Storage & preparation
Data collection
Qualitative Morphology
Colors & patterns
Shape
Repro-ductive morphology
Quantitative Morphology
Behaviour
Ecology
Molecular methods
Data analysis
Phylogenetic methods
Phylogenetic tree reconstruction
Distance-based methods
Character-based methods
Phylogenetic species delimitation
Non-phylogenetic methods
Genetic distances
Data processing
Sequencing
Generation 1 sequencing
Data collection
Nuclear DNA
Organellar DNA
Miscellaneous categories
Interbreeding
Grouping & Ranking justification
Biogeography
Random
Specimen storage location
Sampling location
Abbreviations & terms
Nomenclature & taxonomic history

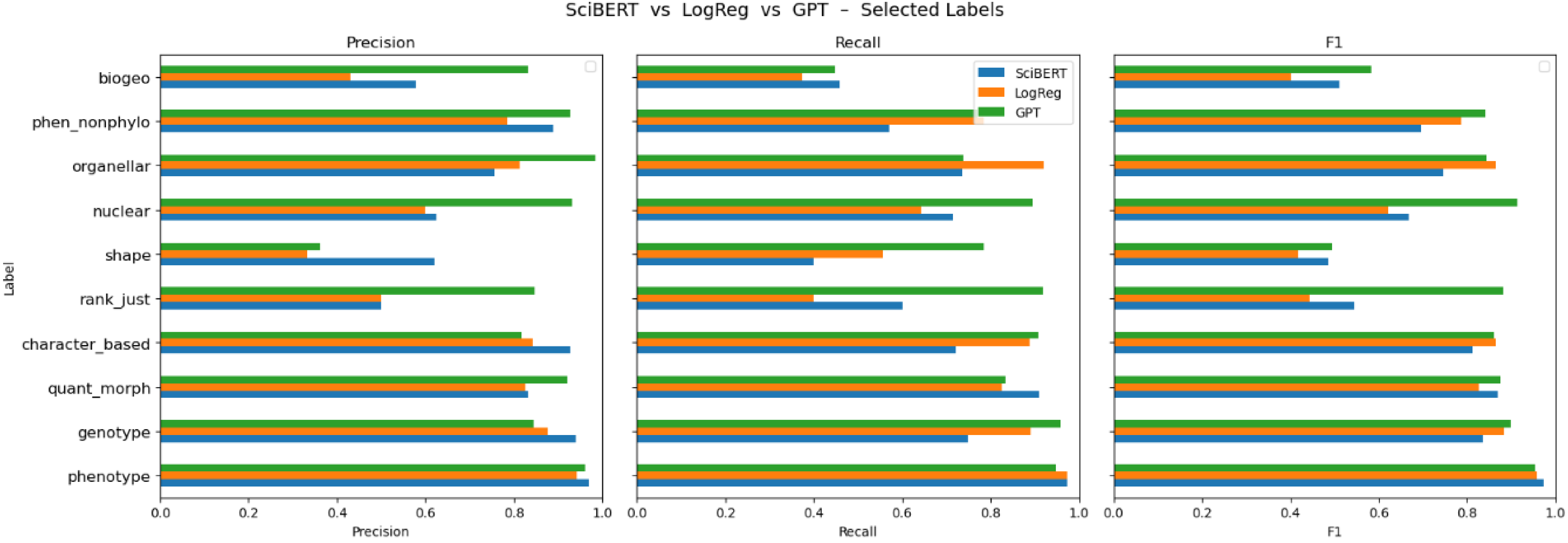
Extracting methods: classifiers

1) Regex baseline
2) Classic ML binary relevance classifiers (SVM, LR)
3) Classic ML classifier chain
4) Transformer model: DistilBERT and SciBERT
5) LLMs: GPT 4o-mini and gpt4

Difficulties:

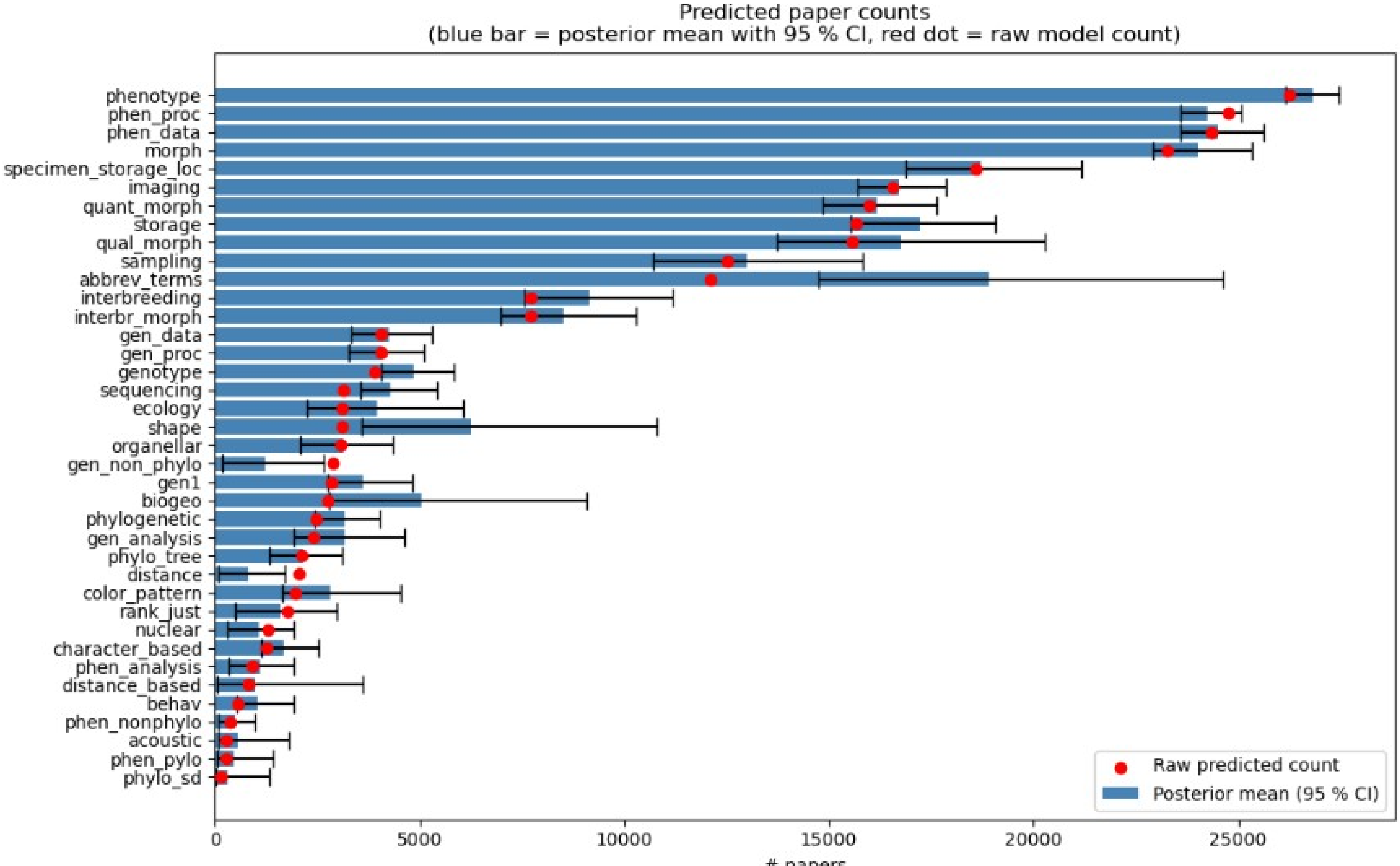
- Very few annotated samples
- Sparse categories combined with common categories
- Obscure language
- Not much compute


Head / Medium / Tail F1
F1
1.0
0.8
0.6
0.4
0.2
0.0
Head
Medium
Tail
Bucket
Overall Macro / Micro F1
SciBERT
LogReg
GPT
Macro F1
Micro F1


SciBERT vs LogReg vs GPT – Selected Labels
Precision
Recall
F1
Label
biogeo
phen_nonphylo
organellar
nuclear
shape
rank_just
character_based
quant_morph
genotype
phenotype
SciBERT
LogReg
GPT
0.0
0.2
0.4
0.6
0.8
1.0
Precision
Recall
F1

Estimating corpus-level proportions

1. Get TP, FP, FN, TN from the test zet
2. Get the predicted counts per label by using SciBERT model on the entire corpus
3. Generative Bayesian model of these observed predictions:
   - Turn step 1 into posteriors of TPR and FPR (keep uncertainty!)
   - p_true as the true rate (flat prior)
   - p_pred = p_true*TPR + (1-p_true)*FPR
   - Observed counts ~ Binomial(N = n_papers, p=p_pred)
4. Use the posterior of p_true to estimate the true count

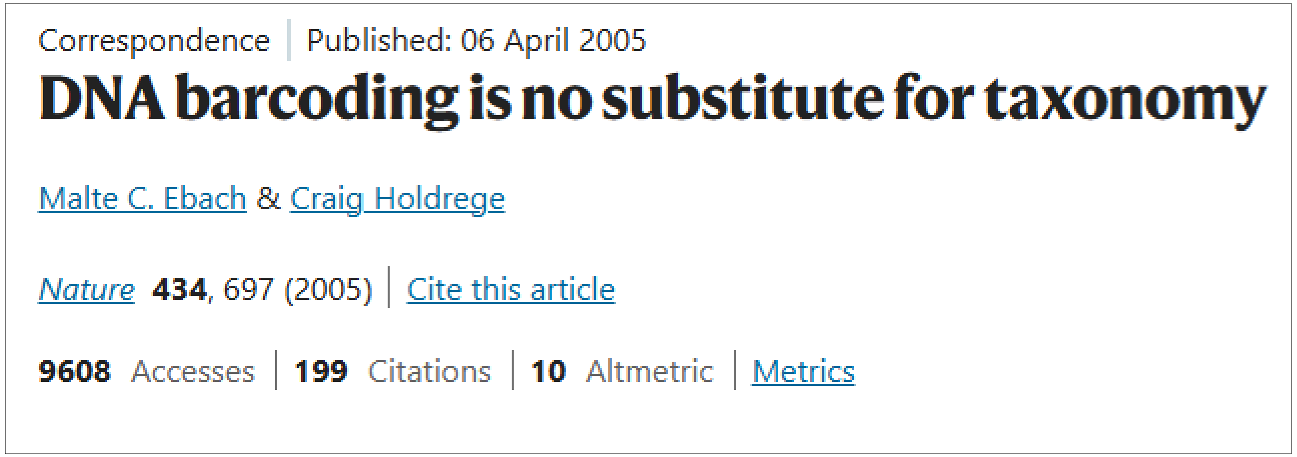

Predicted paper counts
(blue bar = posterior mean with 95 % CI, red dot = raw model count)
phenotype
phen_proc
phen_data
morph
specimen_storage_loc
imaging
quant_morph
storage
qual_morph
sampling
abbrev_terms
interbreeding
interbr_morph
gen_data
gen_proc
genotype
sequencing
ecology
shape
organellar
gen_non_phylo
gen1
biogeo
phylogenetic
gen_analysis
phylo_tree
distance
color_pattern
rank_just
nuclear
character_based
phen_analysis
distance_based
behav
phen_nonphylo
acoustic
phen_pylo
phylo_sd
● Raw predicted count
Posterior mean (95 % CI)
0
5000
10000
15000
20000
25000
# papers

Molecular vs Morphological work

Molecular revolution in the early 2000s:

- Barcoding debate
- In principle academic consensus: ‘Integrative taxonomy’
- But: taxonomy is broader than academy…

➔ Integration? Traditional taxonomy disolved?



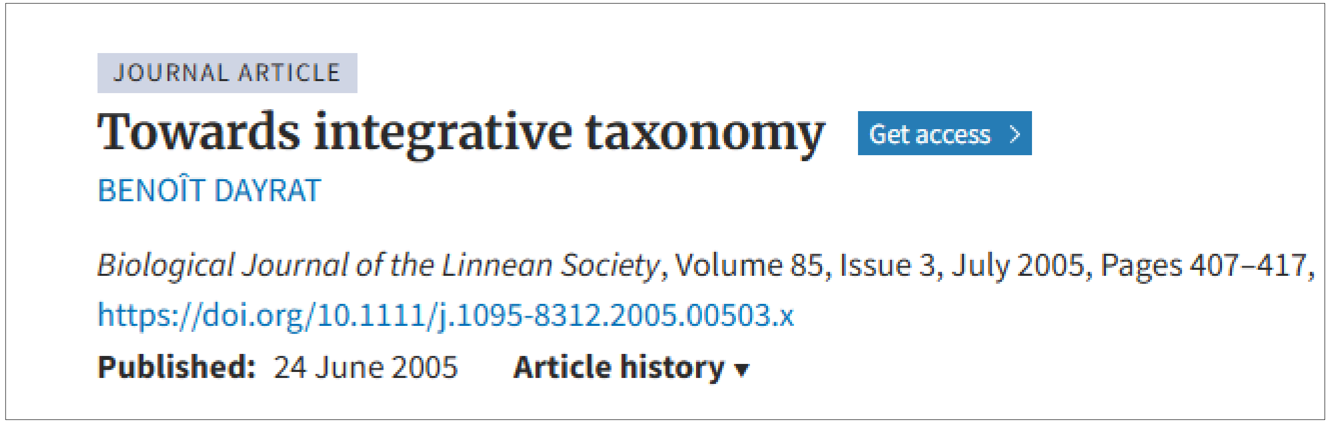
Correspondence | Published: 06 April 2005

DNA barcoding is no substitute for taxonomy

Malte C. Ebach & Craig Holdrege

Nature **434**, 697 (2005) | Cite this article

9608 Accesses | **199** Citations | **10** Altmetric | Metrics



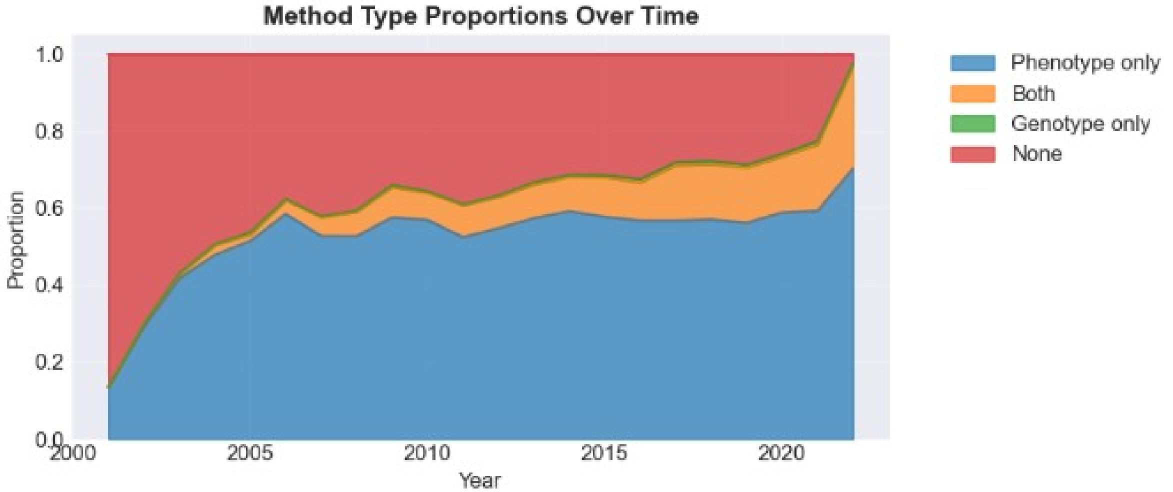
JOURNAL ARTICLE

Towards integrative taxonomy

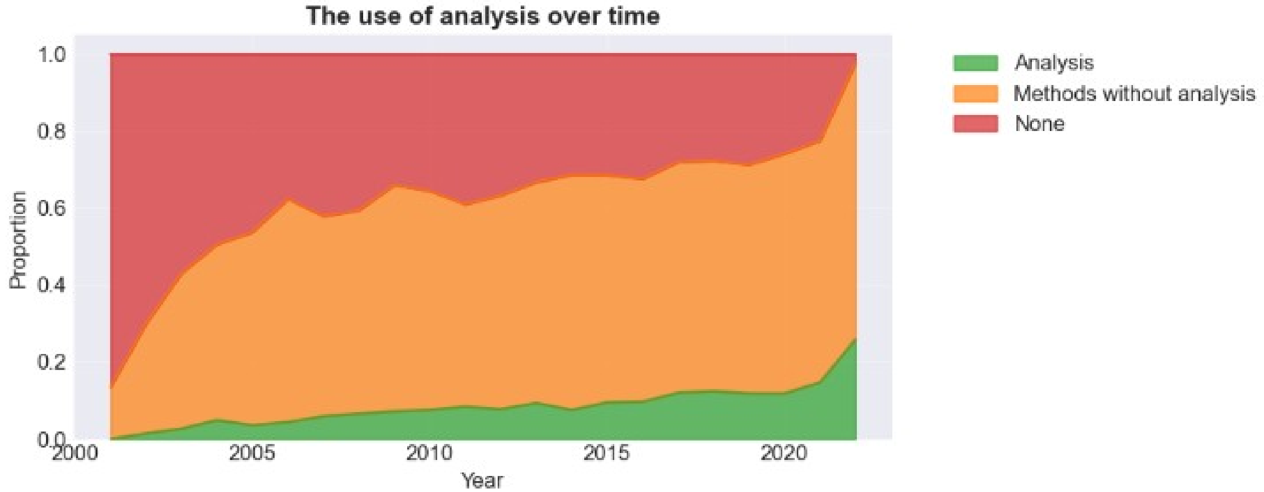
BENOÎT DAYRAT

Biological Journal of the Linnean Society, Volume 85, Issue 3, July 2005, Pages 407–417, https://doi.org/10.1111/j.1095-8312.2005.00503.x

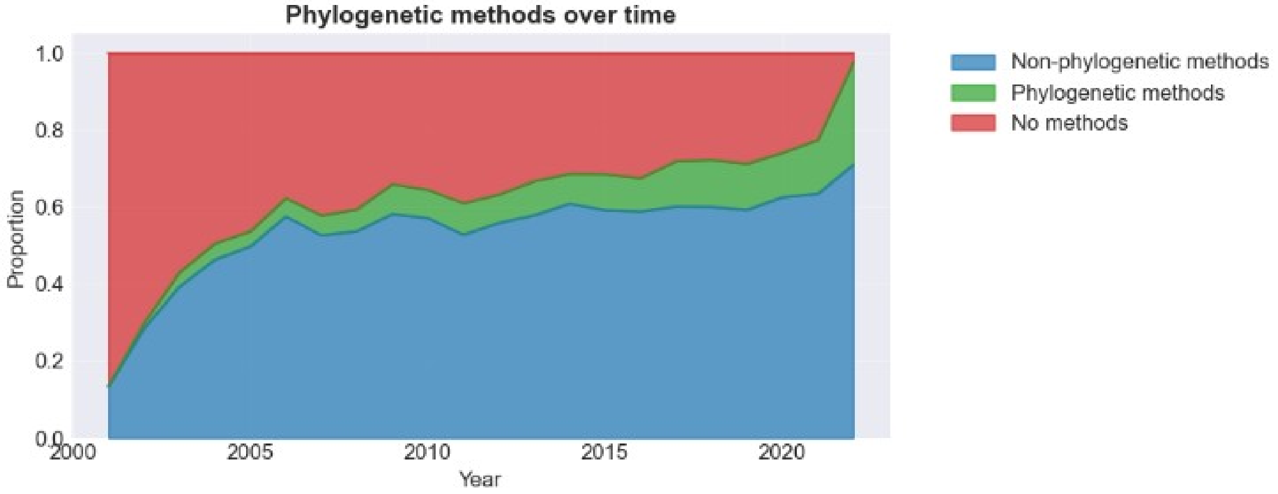
Published: 24 June 2005 **Article history ▾**



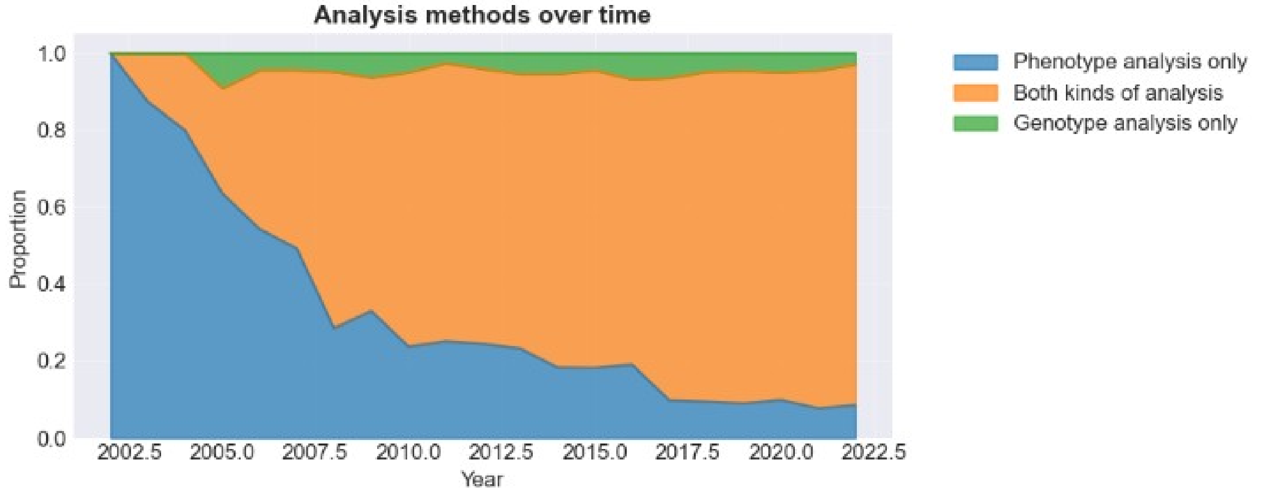
Method Type Proportions Over Time



The use of analysis over time

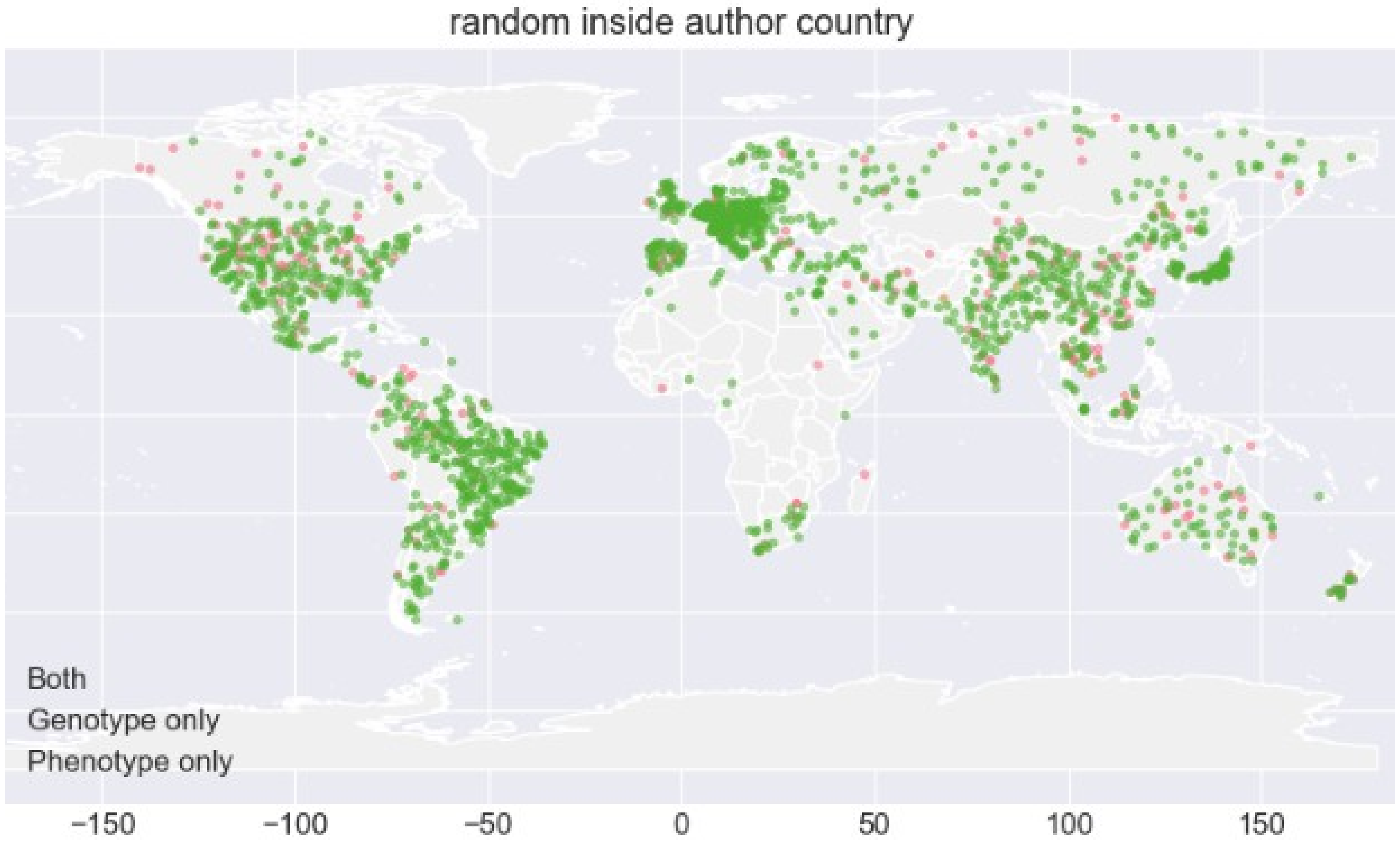


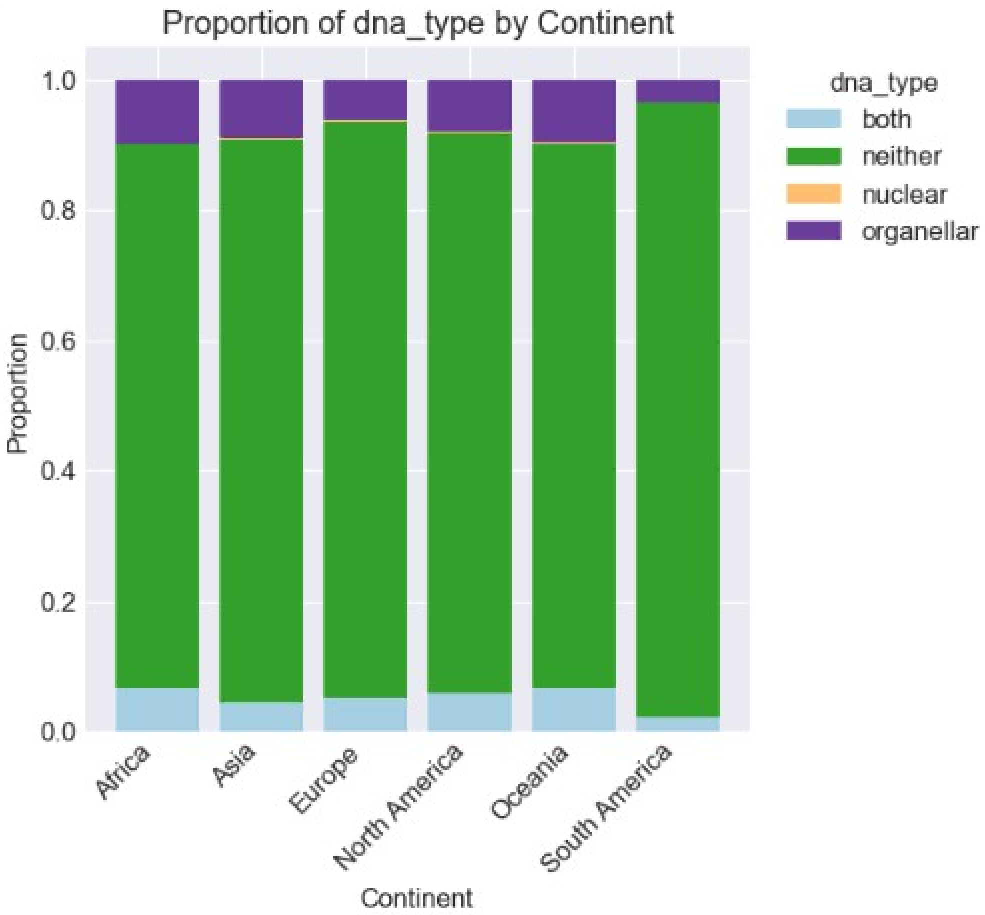
Phylogenetic methods over time



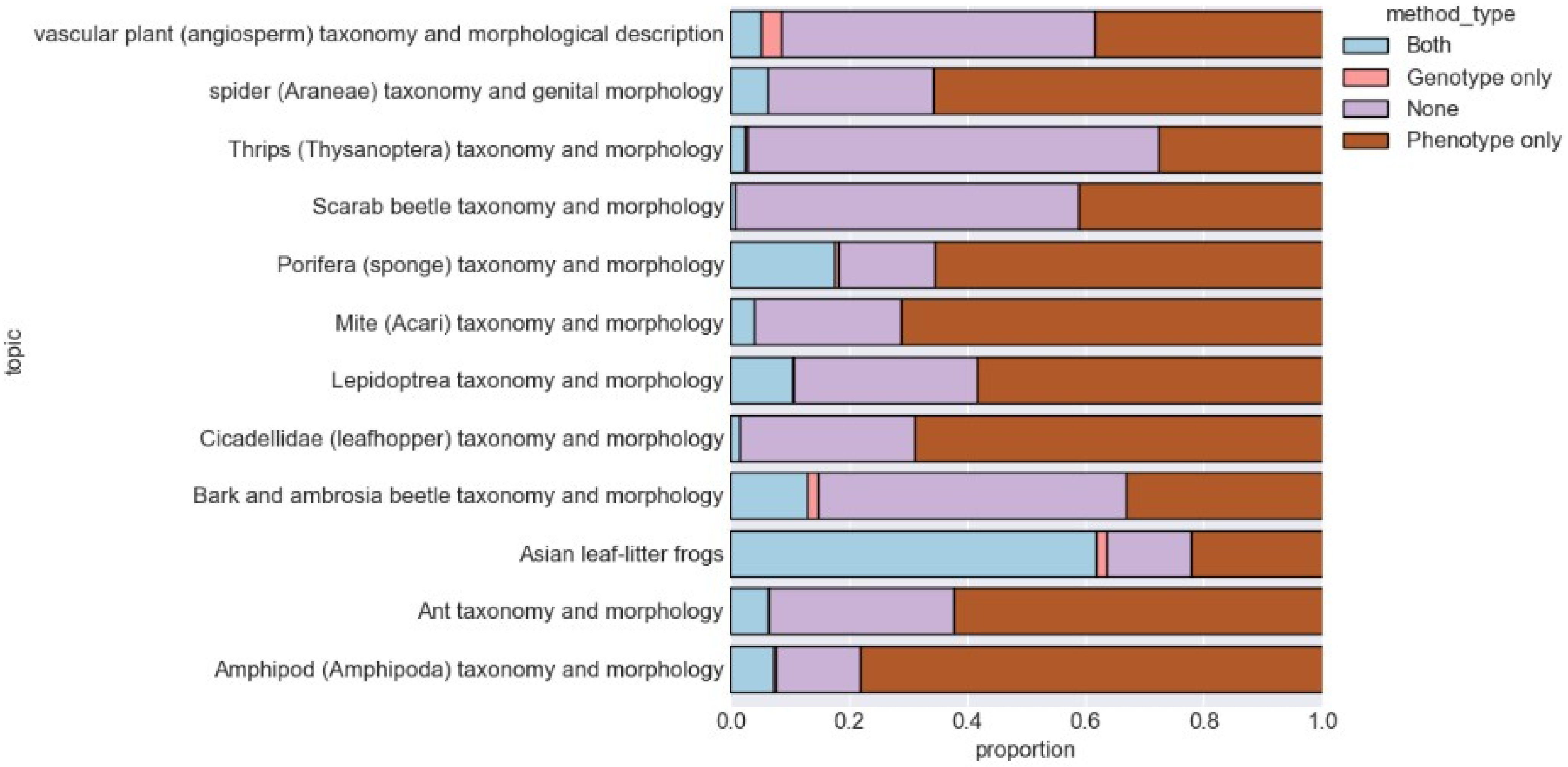
Analysis methods over time

Are molecular methods limited to richer countries?



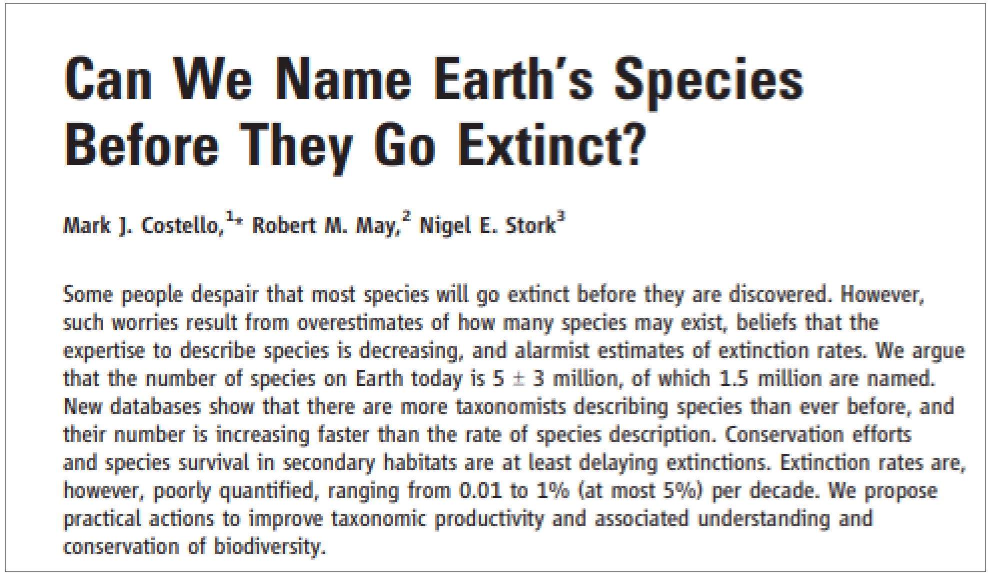


Methods by communities



North/South imbalance

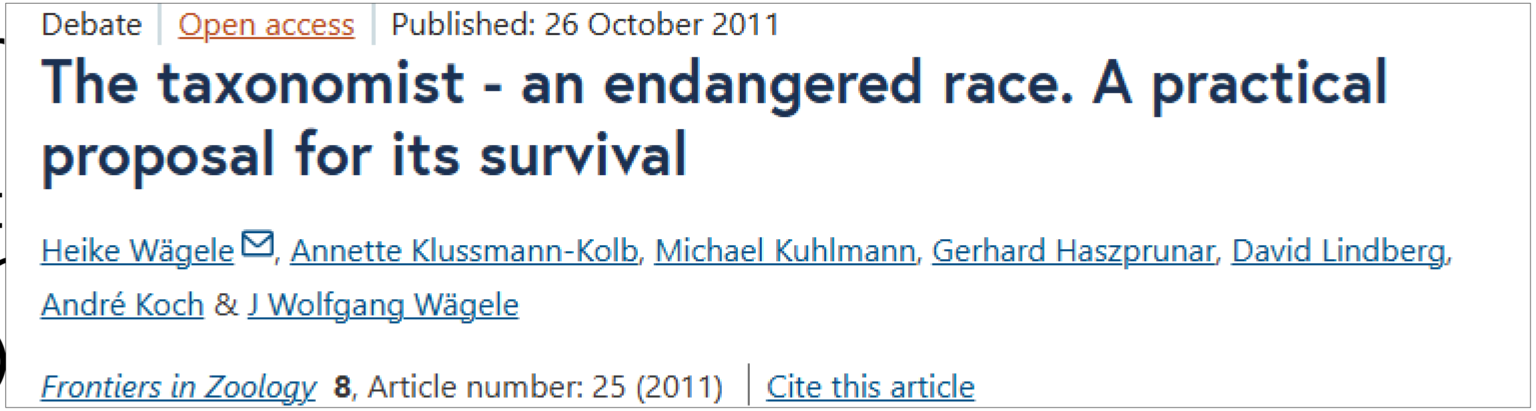
- ‘Global taxonomy initiative’ (CBD in 1998)
  - Need more taxonomy
  - in megadiverse parts of the world
- 1980s – 2000s:
  - Increasing proportion of authors from Asia & latin-america (costello et al. 2012)
  - Continued pleas for mor in diverse regions
- Note: tricky operationalizat
  - Author countries (which
  - Study locations (which?)
  - How to aggregate?



Can We Name Earth’s Species Before They Go Extinct?

Mark J. Costello,[1*] Robert M. May,[2] Nigel E. Stork[3]

Some people despair that most species will go extinct before they are discovered. However, such worries result from overestimates of how many species may exist, beliefs that the expertise to describe species is decreasing, and alarmist estimates of extinction rates. We argue that the number of species on Earth today is 5 ± 3 million, of which 1.5 million are named. New databases show that there are more taxonomists describing species than ever before, and their number is increasing faster than the rate of species description. Conservation efforts and species survival in secondary habitats are at least delaying extinctions. Extinction rates are, however, poorly quantified, ranging from 0.01 to 1% (at most 5%) per decade. We propose practical actions to improve taxonomic productivity and associated understanding and conservation of biodiversity.



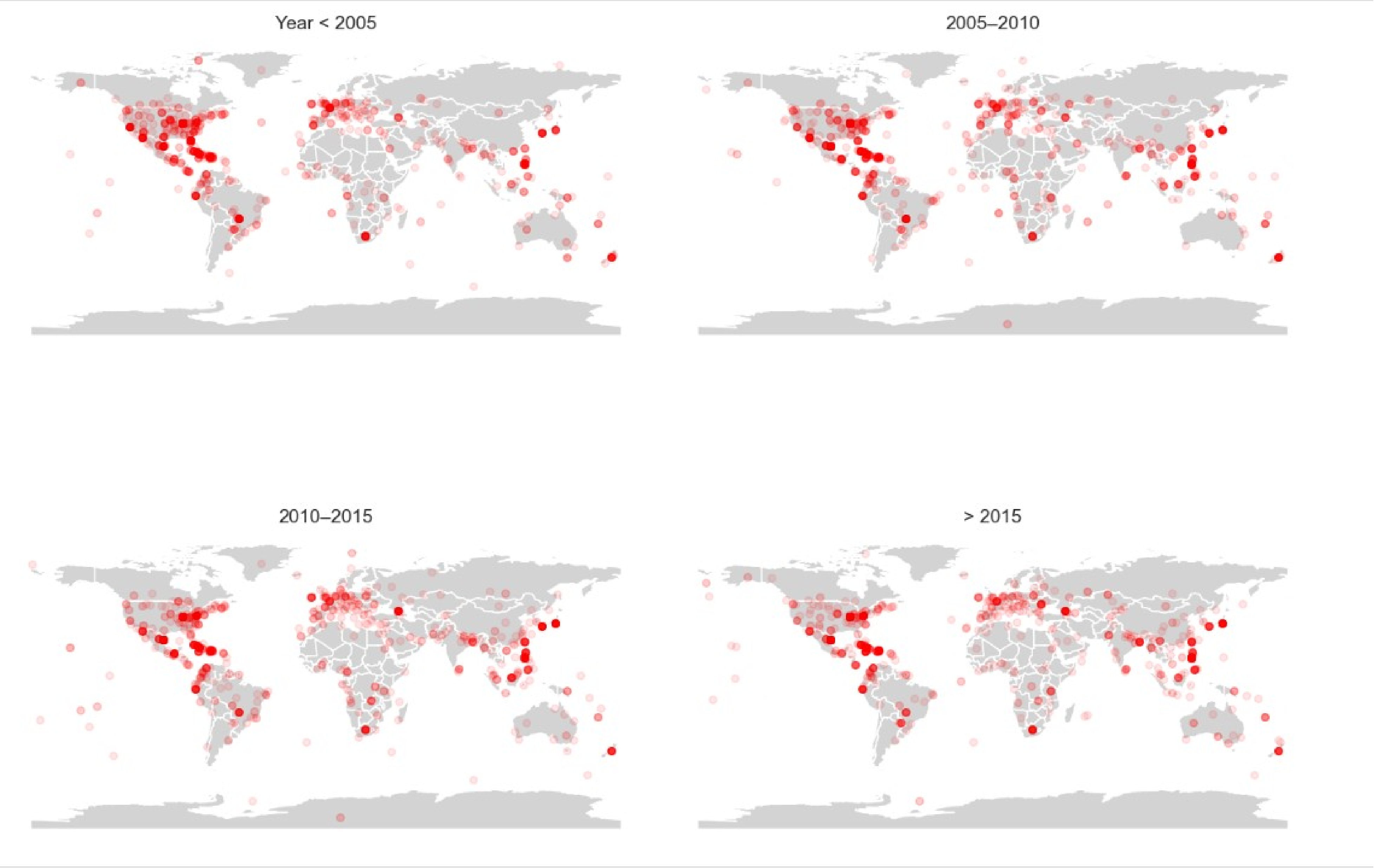
Debate | Open access | Published: 26 October 2011

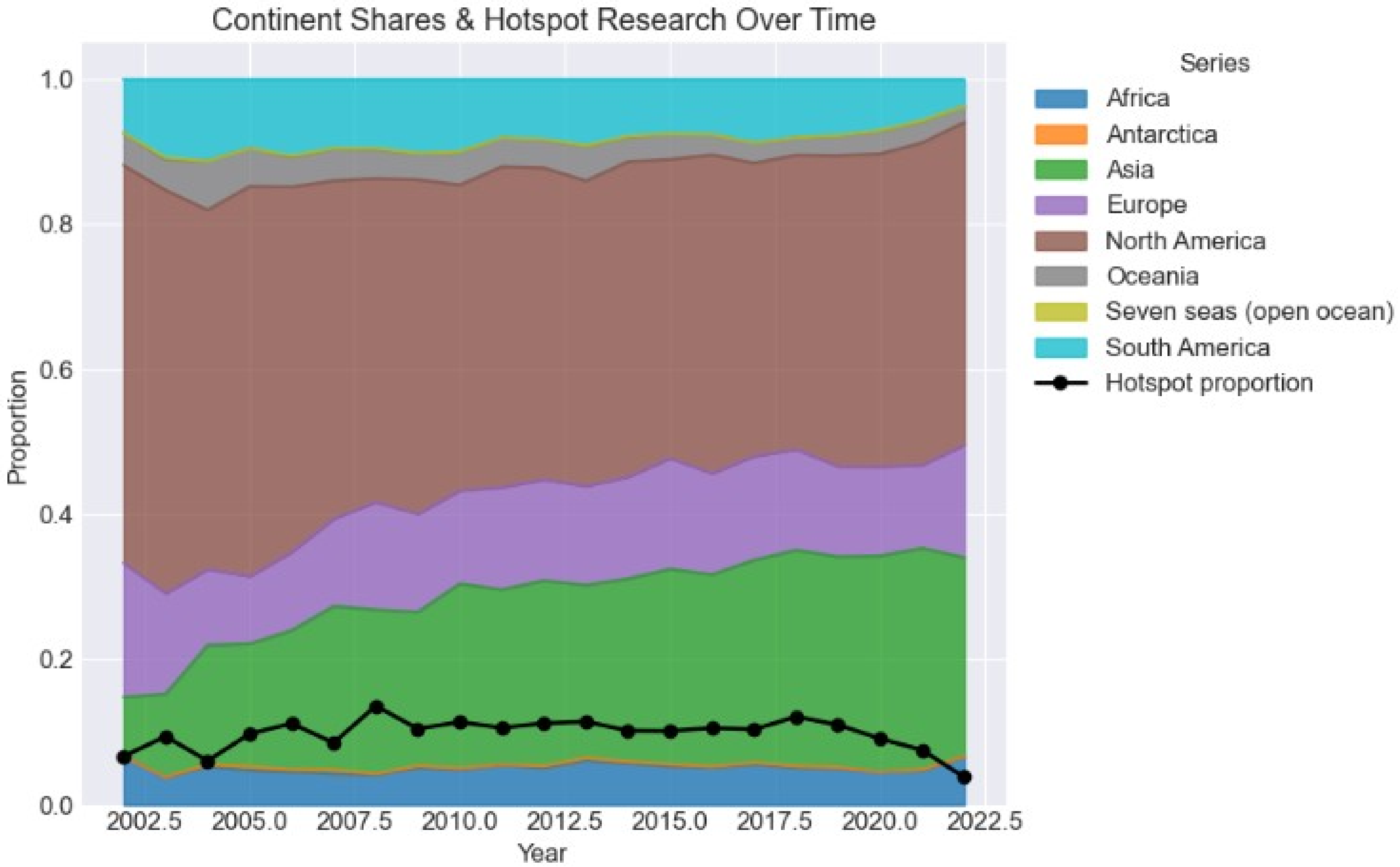
The taxonomist - an endangered race. A practical proposal for its survival

Heike Wägele, Annette Klussmann-Kolb, Michael Kuhlmann, Gerhard Haszprunar, David Lindberg, André Koch & J Wolfgang Wägele

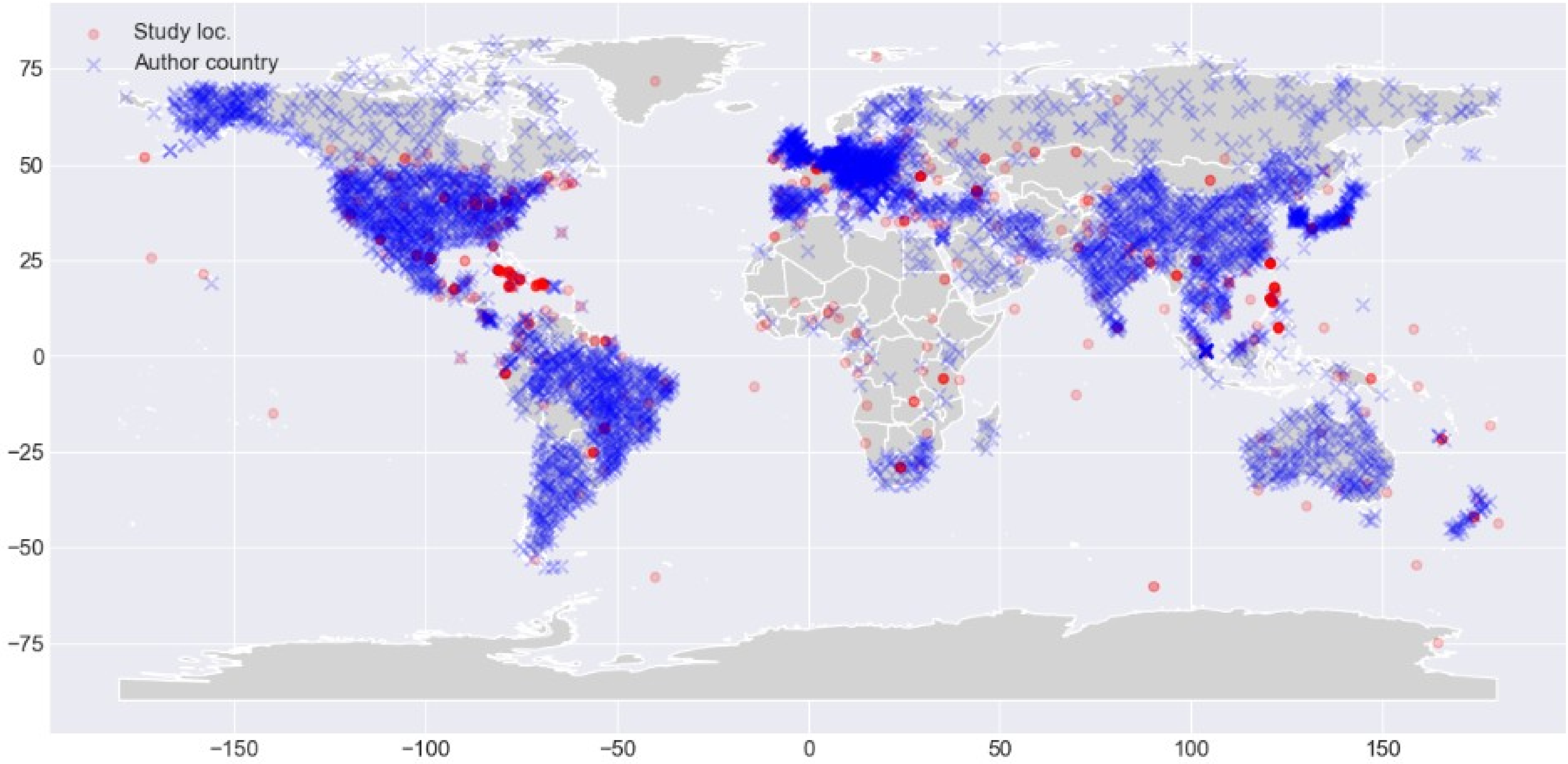
Frontiers in Zoology **8**, Article number: 25 (2011) | Cite this article

Has taxonomy in megadiverse regions increased?



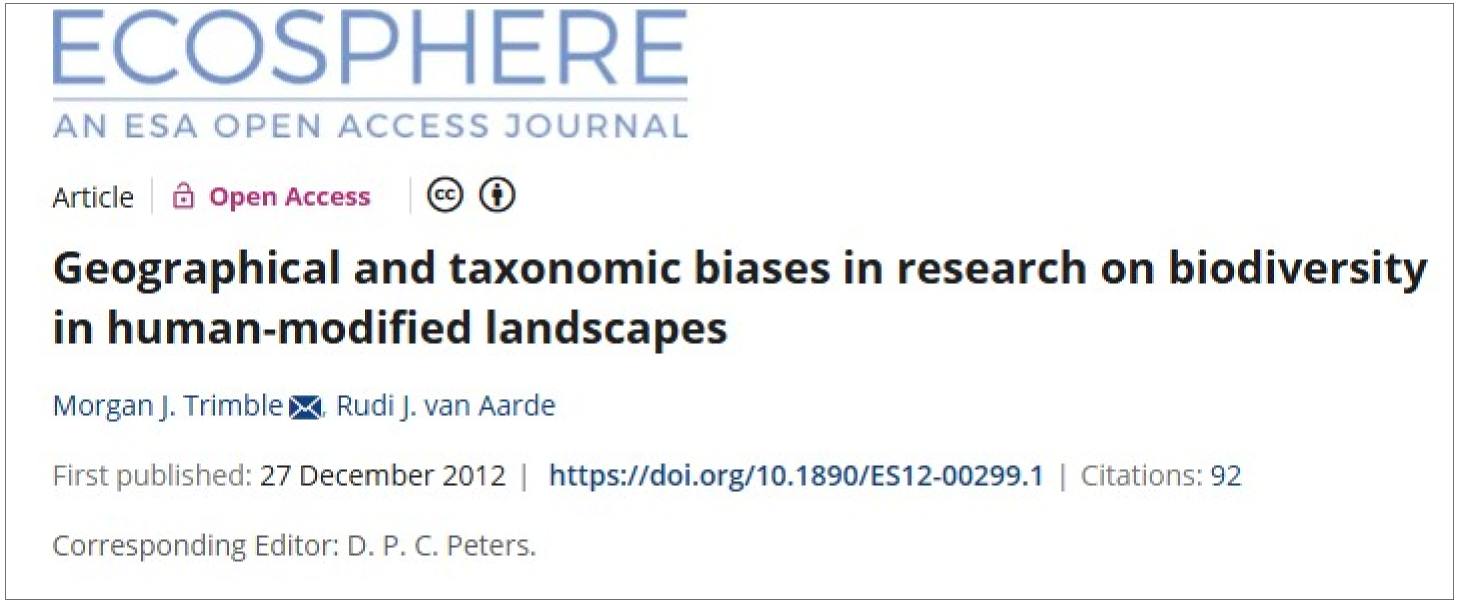

Continent Shares & Hotspot Research Over Time
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
2002.5
2005.0
2007.5
2010.0
2012.5
2015.0
2017.5
2020.0
2022.5
Year
Series
Africa
Antarctica
Asia
Europe
North America
Oceania
Seven seas (open ocean)
South America
Hotspot proportion

Local expertise?



Biomes: (temperate) forest vs the rest

- Well-documented forest-bias in biodiversity research
- Mix of preferences and geographical bias?
- Need to control:
  - Species richness?
  - Area?



ECOSPHERE

AN ESA OPEN ACCESS JOURNAL

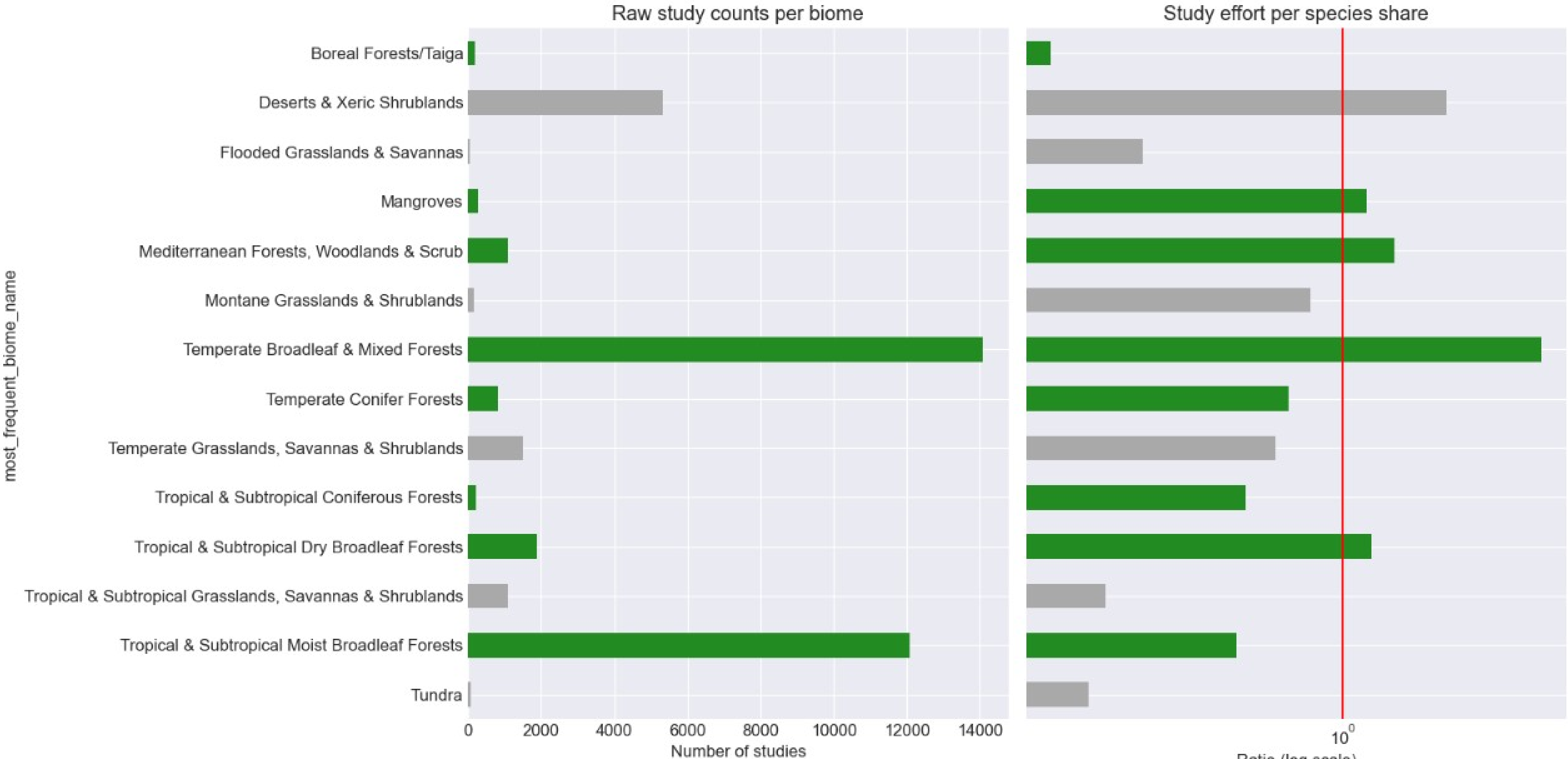
Article | Open Access

Geographical and taxonomic biases in research on biodiversity in human-modified landscapes

Morgan J. Trimble, Rudi J. van Aarde

First published: 27 December 2012 | https://doi.org/10.1890/ES12-00299.1 | Citations: 92

Corresponding Editor: D. P. C. Peters.


Raw study counts per biome
Study effort per species share
most_frequent_biome_name
Boreal Forests/Taiga
Deserts & Xeric Shrublands
Flooded Grasslands & Savannas
Mangroves
Mediterranean Forests, Woodlands & Scrub
Montane Grasslands & Shrublands
Temperate Broadleaf & Mixed Forests
Temperate Conifer Forests
Temperate Grasslands, Savannas & Shrublands
Tropical & Subtropical Coniferous Forests
Tropical & Subtropical Dry Broadleaf Forests
Tropical & Subtropical Grasslands, Savannas & Shrublands
Tropical & Subtropical Moist Broadleaf Forests
Tundra
0
2000
4000
6000
8000
10000
12000
14000
Number of studies
10^0
Ratio (log scale)

Disagreement

First hypothesis: does disagreement vary in function of group studied?

- Much more disagreement (>2x): birds (*n* = 333); mollusuks (*n* = 1064)
- Slightly more (>1.25x): mammals (*n* = 396)
- Slightly less (<0.75x): fish (*n* = 2132); non-insect arthropods (*n* = 7285)
- Much less (<0.5x): prokaryotes (but *n* = only 13!)

Disagreement

Second hypothesis: What about the **age** of the group studied? Test for a correlation between disagreement index and the **year** in which the paper’s main genus was described.

Hypothesis: should be a **negative** correlation: the older the group is, the more likely you are to fight about it.

Disagreement

Confirmed: a significant negative correlation

A paper discussing a genus described in 1750 (the oldest description date in our corpus) should have around 0.003 more disagreement index compared to a new genus (and 0.003 is approximately the *mean* disagreement value).