

Towards holistic insect monitoring by accelerating species discovery, description, and identification using robots, nanopore sequencing, and Artificial Intelligence



Biodiversity research: 1990s versus 2020s

1990s: "Ivory Tower" discipline for biologists out of touch with modernity

2020s: Biodiversity loss threatens planetary health and society now cares

Who cares?

Re	e Global Risks port 2023 h Edition
1	Failure to mitigate climate change
2	Failure of climate-change adaptation
3	Natural disasters and extreme weather events
4	Biodiversity loss and ecosystem collapse
5	Large-scale involuntary migration

Who else?

Goldman Sachs



Biodiversity Assessing the Financial Links to Natural Capital

Biodiversity is the cornerstone around which most sustainability objectives sit and an area to which we think corporates and investors will likely pay increased attention in order to meet sustainability and economic *Re/insurance and Biodiversity: Insights on a Swiss Re Report*



Allianz

Trade

Allianz 🕕

The new risk frontier in finance: biodiversity loss

DISCOVER OUR REPORT

But is society interested in "Ivory Tower biodiversity"?



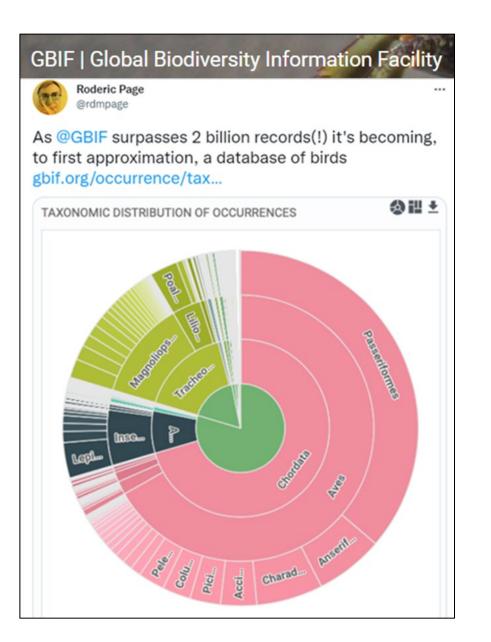
1990s: Ivory Tower taxon biases:

- pretty
- pretty large
- pretty diurnal
- pretty endangered
- pretty impressive on a tree
- pretty manageable taxonomically
- pretty unusual in many other ways...

What is missing is:

• pretty important but ugly...

How bad is the data gap for the uglies?

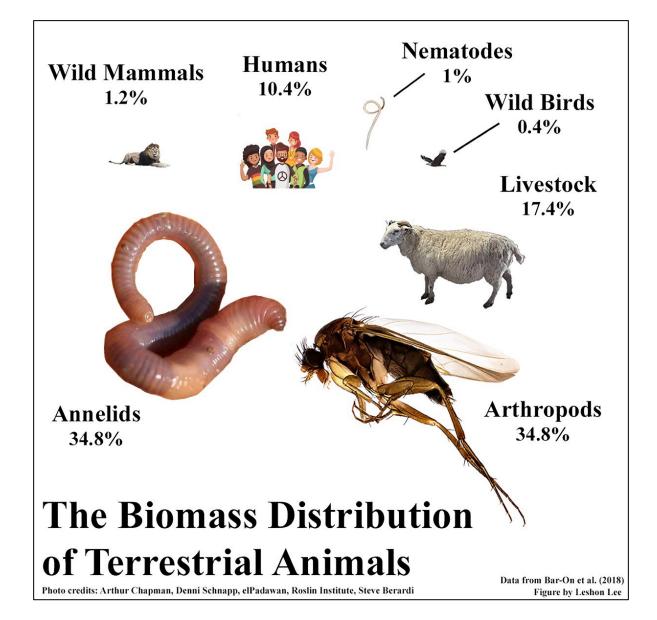


2024 update:

- 2.5 billion occurrence records
- ca. 2 billion for bird species
- Insecta: 238 million
 - 42 million butterfly records
 - top 10 bird species: 230 million

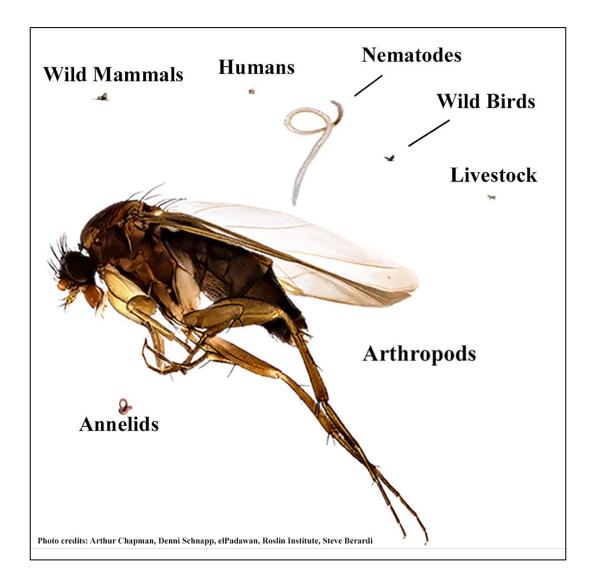
=> we need a new field:

• Ugly Zoology



Terrestrial animal biodiversity: biomass

Terrestrial animal biodiversity: species diversity



Shortcomings of the old ways...



- >80% of all species undescribed
- 95% of all described species only identifiable for taxonomic experts
- very little baseline biodiversity data from standardized samples
- natural history data only available for <5% of species

<u>Triage</u>: which taxa need our attention?



Dr. Amrita Srivathsan



nature ecology & evolution

Article

https://doi.org/10.1038/s41559-023-02066-0

0

Convergence of dominance and neglect in flying insect diversity

Received: 8 August 2022

Accepted: 6 April 2023

Wan F. A. Jusoh @^{2,4}, Sujatha Narayanan Kutty^{5,6}, Jayanthi Puniamoorthy @⁵, Darren Yeo⁵, Tomas Roslin [®]⁷ & Rudolf Meier [®] ¹⁵ Published online: 18 May 2023

Check for updates



Most of arthropod biodiversity is unknown to science. Consequently, it has been unclear whether insect communities around the world are dominated by the same or different taxa. This question can be answered through standardized sampling of biodiversity followed by estimation of species diversity and community composition with DNA barcodes. Here this approach is applied to flying insects sampled by 39 Malaise traps placed in five biogeographic regions, eight countries and numerous habitats (>225,000 specimens belonging to >25,000 species in 458 families). We find that 20 insect families (10 belonging to Diptera) account for >50% of local species diversity regardless of clade age, continent, climatic region and habitat type. Consistent differences in family-level dominance explain two-thirds of variation in community composition despite massive levels of species turnover, with most species (>97%) in the top 20 families encountered at a single site only. Alarmingly, the same families that dominate insect diversity are 'dark taxa' in that they suffer from extreme taxonomic neglect, with little signs of increasing activities in recent years. Taxonomic neglect tends to increase with diversity and decrease with body size. Identifying and tackling the diversity of 'dark taxa' with scalable techniques emerge as urgent priorities in biodiversity science.

Amrita Srivathsan @1, Yuchen Ang2, John M. Heraty3, Wei Song Hwang @2,

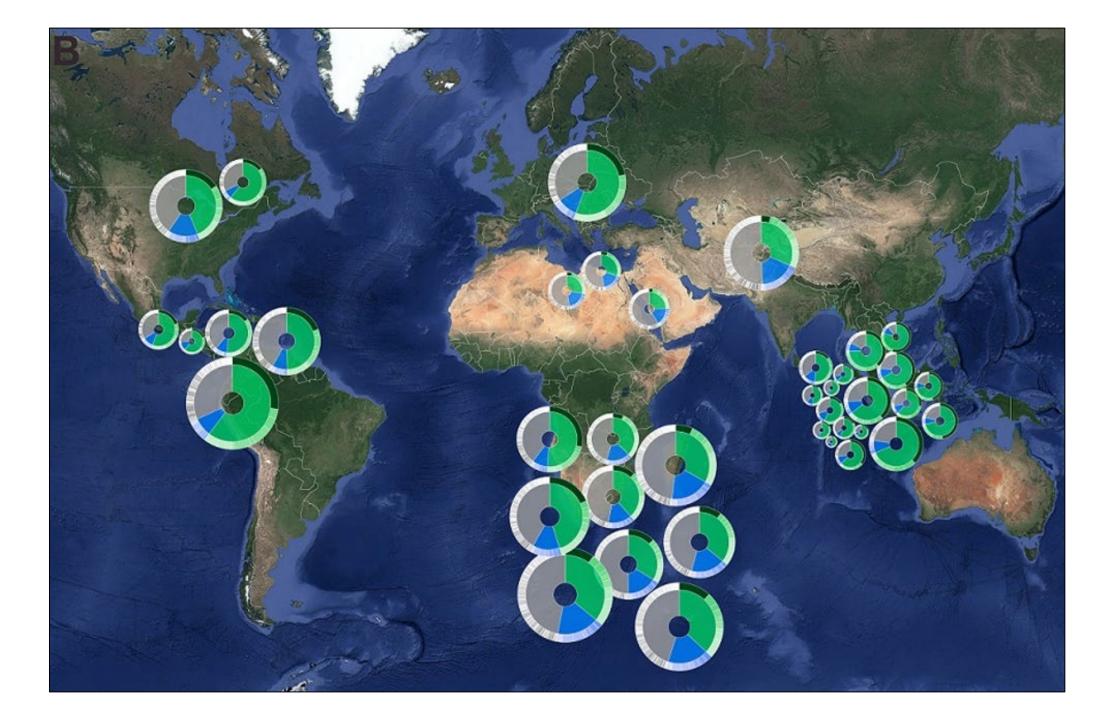
=> same 20 family-ranked insect clades contribute >50% of species diversity

=> same 20 clades also account for >50% of specimen abundance

Habitat, climate, etc. don't matter

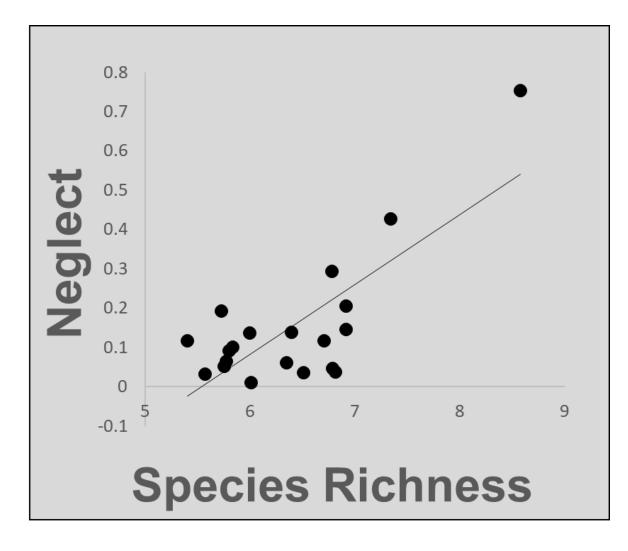


Dr Amrita Srivathsa



The top 10 A Top 4: Plat NI=0.11 NI=0.75 Top 5: Psychodidae NI=0.42 Top 1: Cecidomyiidae Top 2: Ceratopogonidae Top 3: Chironomidae NI=0.09 0.2 mm Top 7: Phoridae Top 8: ants NI=0.29 Top 9: Cicadellidae NI=0.04 NI=0.20 NI=0.04 NI=0.06 Top 10: Ichneumor ic Top 6: Sciaridae

Most of these top taxa suffer from taxonomic neglect



'Neglect Index' (NI): ratio between the number of mOTUs found across Malaise traps and number of species described

- 1. Neglect increases with species richness
- 2. Neglect is not improving over time
- 3. Small body size leads to neglect

Clearly, we need "Ugly Zoology"! But we also need rebranding...



Dr Amrita Srivathsa

"Dark Taxon Zoology": resolving biology's dark matter

Roderik Page in 2011: "dark taxa"

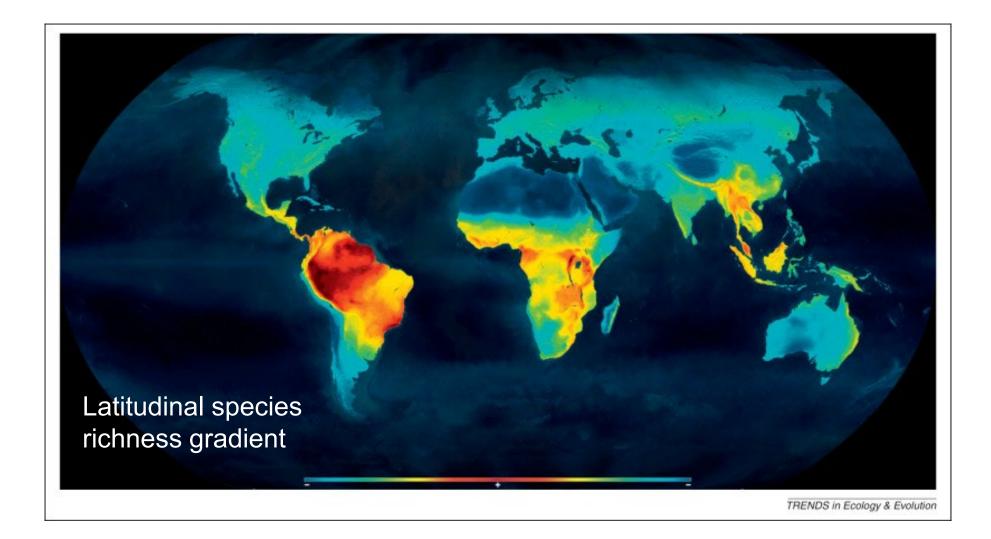
- 1. groups of organisms present in genetic databases
- 2. lacking formal scientific names

Hartop et al. 2022: "dark taxon"

- 1. clade whose diversity exceeds 1,000 species
- 2. of which fewer than 10% of species described

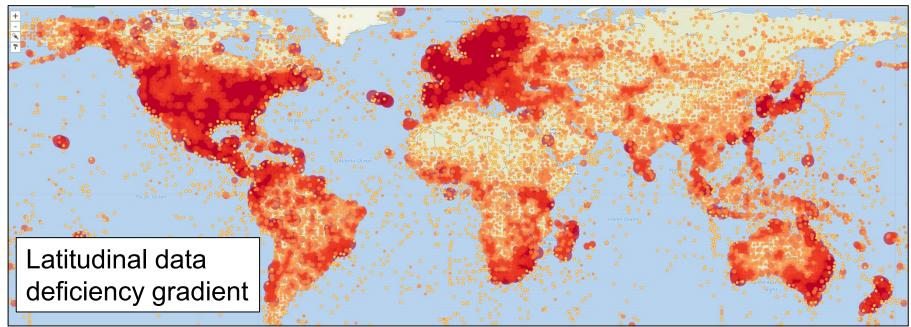
Biodiversity data crisis: more problems...

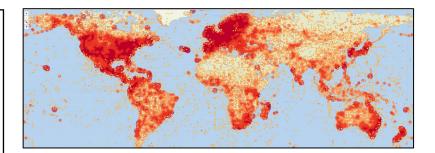
Geographic biases in biodiversity data



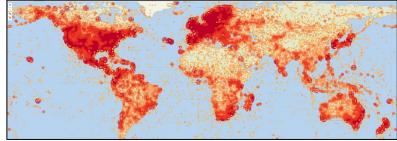
GBIF | Global Biodiversity Information Facility

Insect data in GBIF



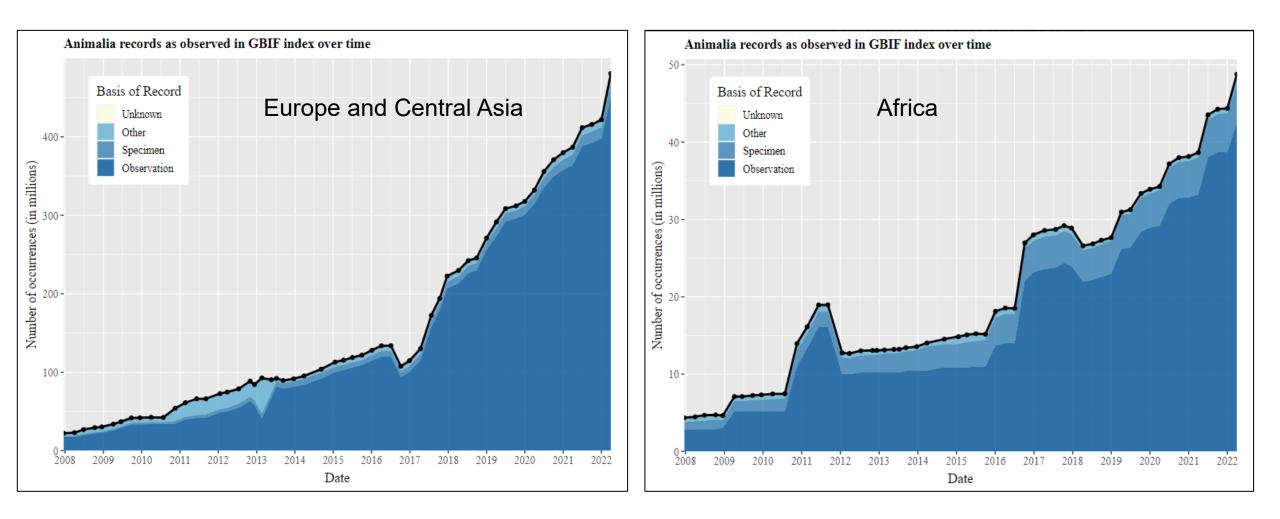




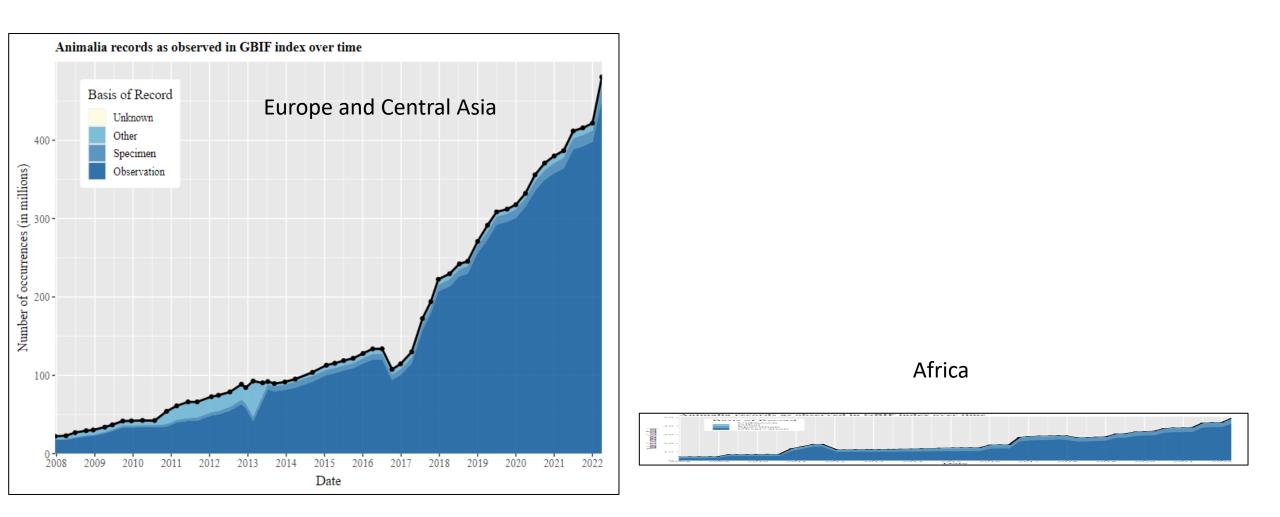


20% of insect data are for the UK

Is the situation improving?



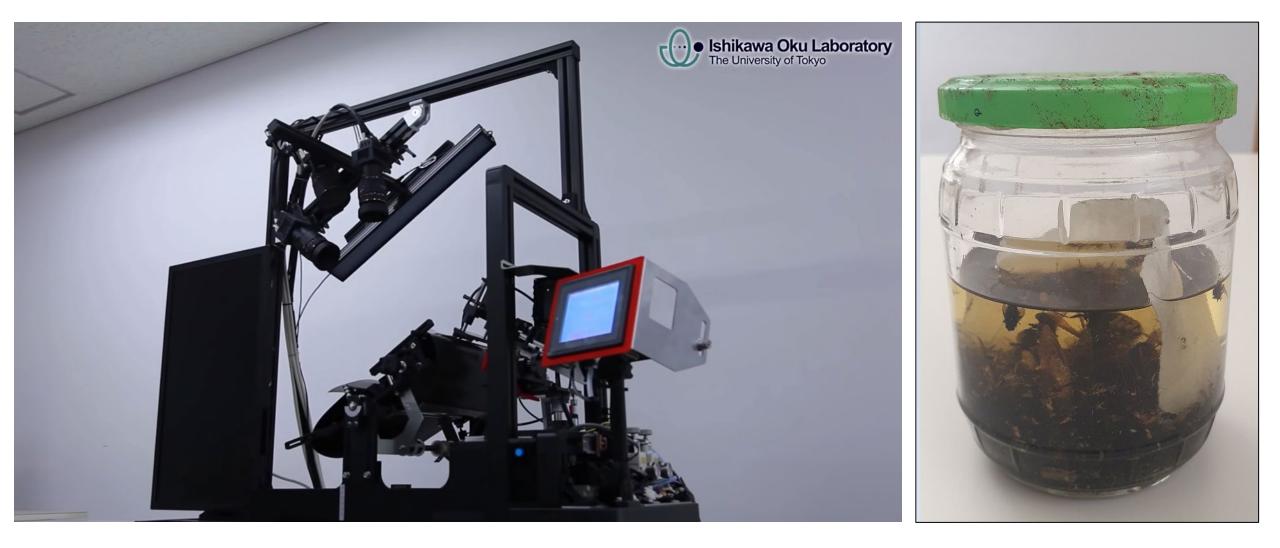
Is the situation improving?



Biodiversity data crisis

- we need to generate more data fast
- data should be for standardized samples

Let's analyze biodiversity samples like libraries



Species Discovery Factory



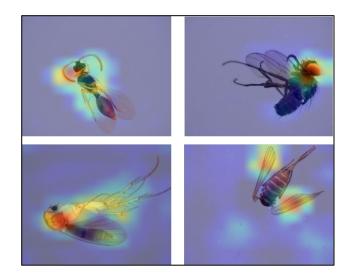
1. DiversityScanner: Digitization with robots



2. Nanopore Sequencing: Sorting with DNA barcodes



3. Identified images: biomonitoring with AI





Lorenz Wührl, KIT





- images each specimen
- classifies the specimen
- measures specimen
- prepares it for DNA sequencing

MOLECULAR ECOLOGY RESOURCES

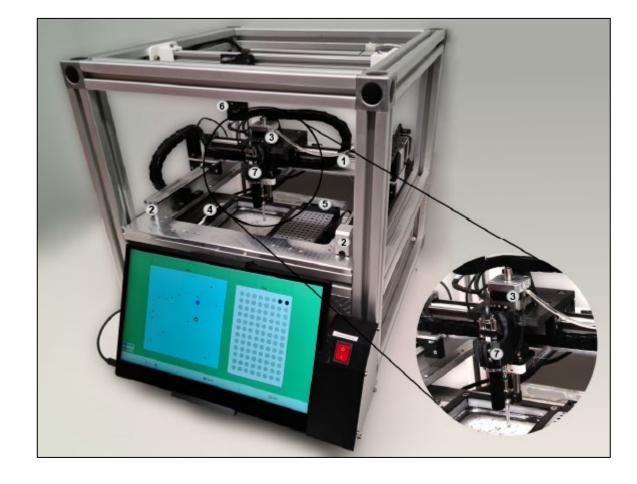
RESOURCE ARTICLE 🛛 🔂 Open Access

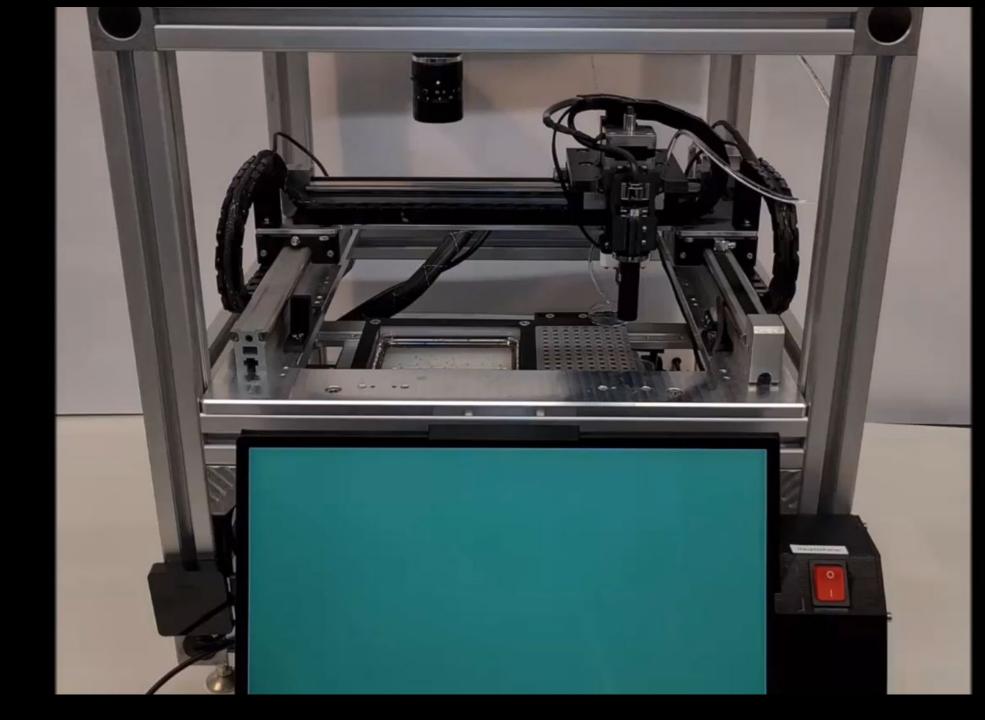
DiversityScanner: Robotic handling of small invertebrates with machine learning methods

Lorenz Wührl, Christian Pylatiuk 🔀, Matthias Giersch, Florian Lapp, Thomas von Rintelen, Michael Balke, Stefan Schmidt, Pierfilippo Cerretti, Rudolf Meier 🗙

First published: 04 December 2021 | https://doi.org/10.1111/1755-0998.13567

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi:10.1111/1755-0998.13567





Karlsruher Institut für Technologie (KIT)



Prof. C. Pylatiuk



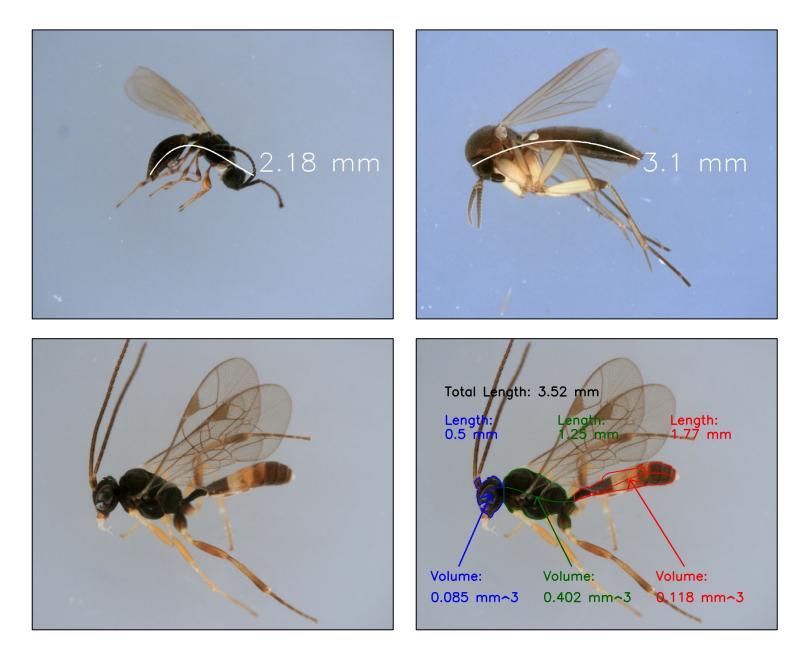
Lorenz Wührl

Detailed images for:

- taxonomic work
- training of AI models
- => image-based biomonitoring

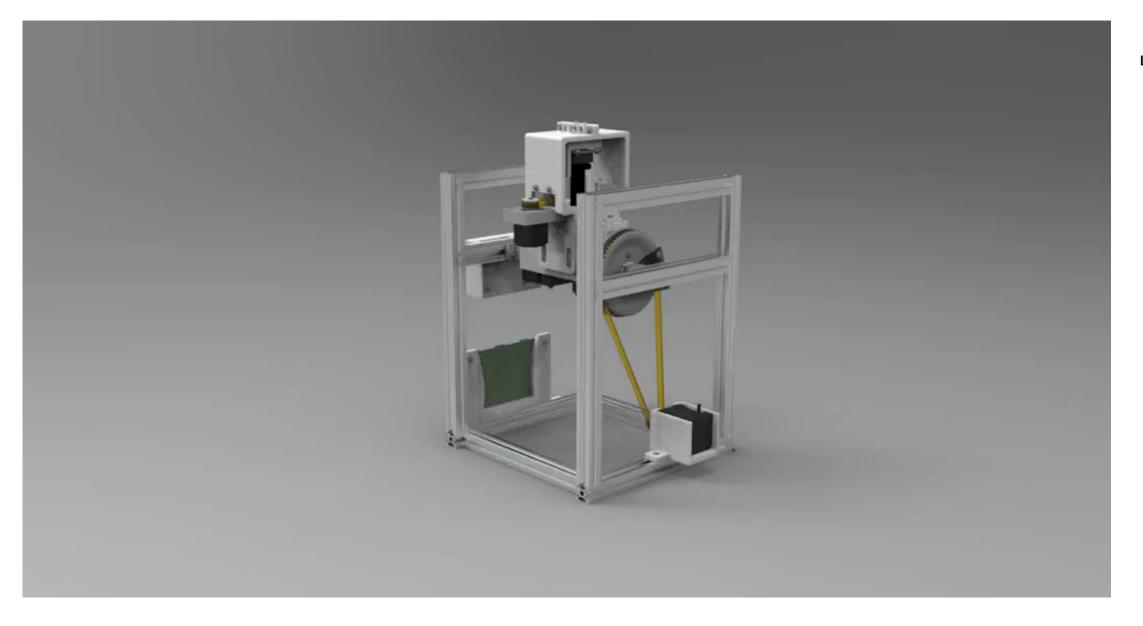


DiversityScanner Biomass estimation





Lorenz Wührl, KIT





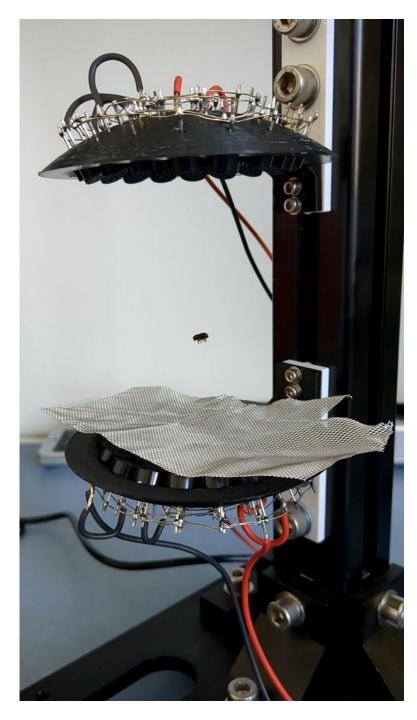
Lorenz Wührl, KIT













Imaging from multiple perspectives also important for training AI identification models



Nathalie Klug, KIT

Different "DiversityScanners" for different purposes



Entomoscope: An Open-Source Photomicroscope for Biodiversity Discovery

LORENZ WÜHRL¹, LUCA RETTENBERGER¹, RUDOLF MEIER², EMILY HARTOP², JULIEN GRAF², and CHRISTIAN PYLATIUK¹,

¹Institute for Automation and Applied Informatics, Karlsruhe Institute of Technology, Karlsruhe, Germany (Iorenz.wuehrl, luca.rettenberger, pylatiuk) @kit.edu ²Center for Integrative Biodiversity Discovery, Leibniz Institute for Evolution and Biodiversity Science, Museum für Naturkunde Berlin, Germany (rudolf.meier, emily.hartop, julien.graf)@mfn.berlin

Corresponding author: Lorenz Wührl (e-mail: lorenz.wuchrl@kit.edu).

This research was supported by the Center for Integrative Biodiversity Discovery at the Museum für Naturkunde Berlin and the program Natural, Artificial and Cognitive Information Processing (NACIP) of the Helmholtz-Association.

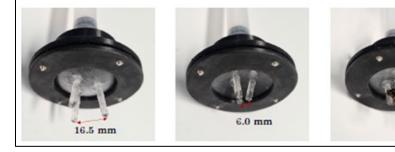




Lorenz Wührl, KIT

Lorenz Wührl*, Leonard Keller, Nathalie Klug, Hossein Shirali, Rudolf Meier, and Christian Pylatiuk

Automated Handling of Biological Objects with a Flexible Gripper for Biodiversity Research

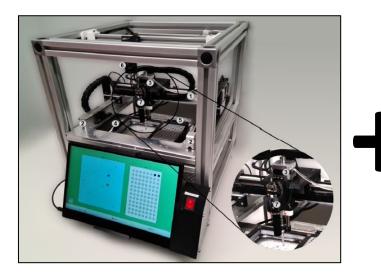




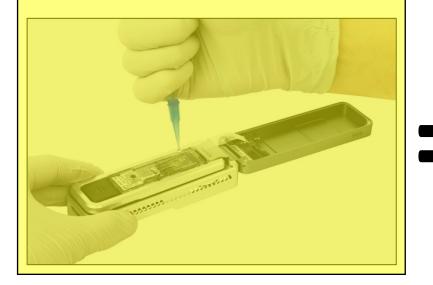
Species Discovery Factory



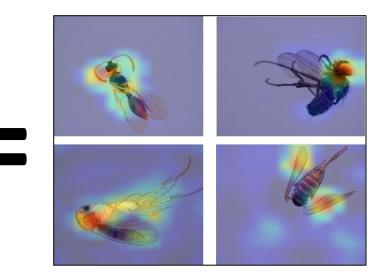
1. DiversityScanner: Digitization with robots



2. Nanopore Sequencing: Sorting with DNA barcodes



3. Identified images: biomonitoring with AI



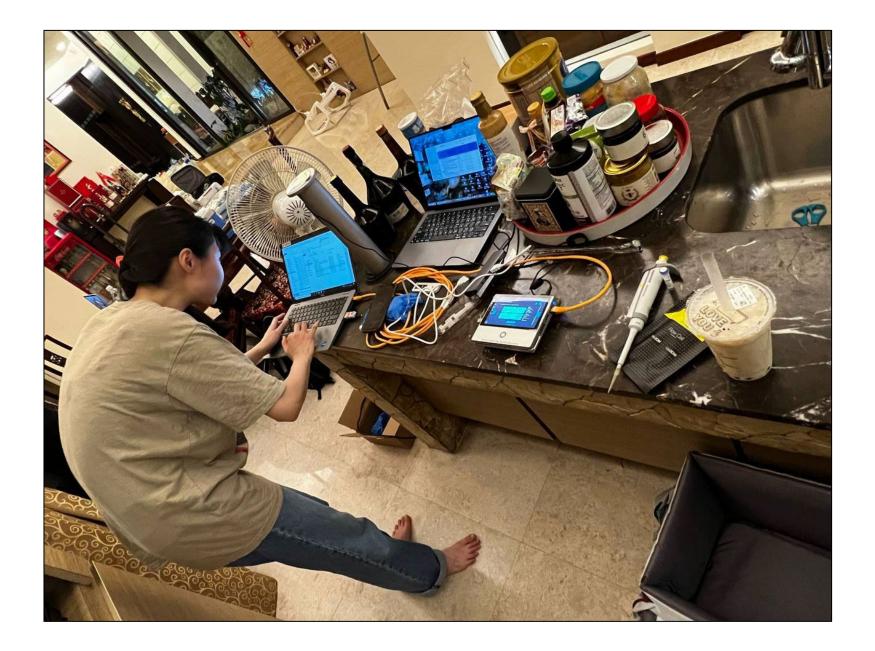
2. Nanopore Sequencing



"Megabarcoding":

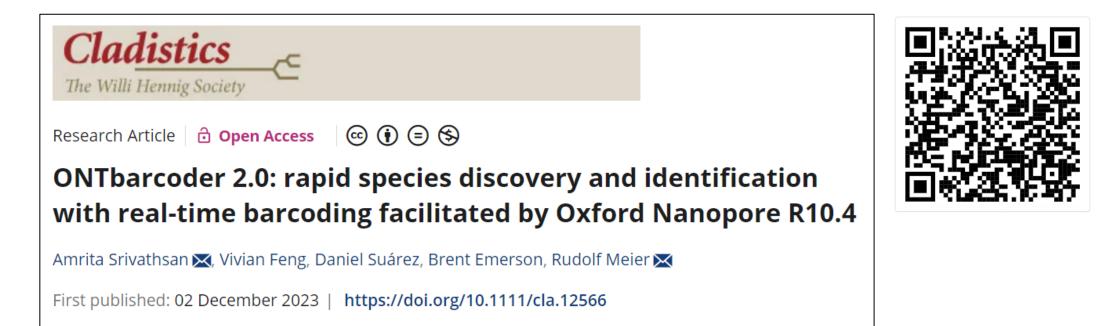
- simplified, fast, and robust techniques for specimen-based barcoding
- 10 cents/specimen







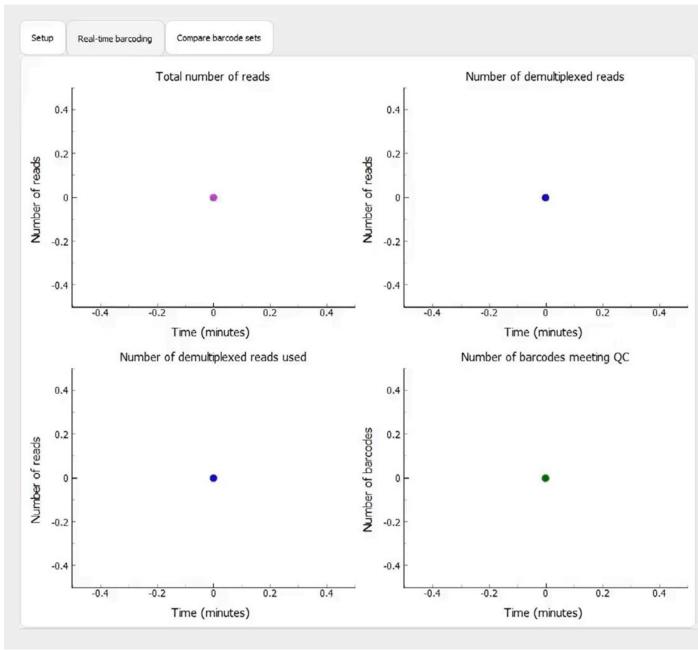
Matching real-time sequencing with real-time DNA barcoding





Dr. Amrita Srivathsan

Flongle R10.4



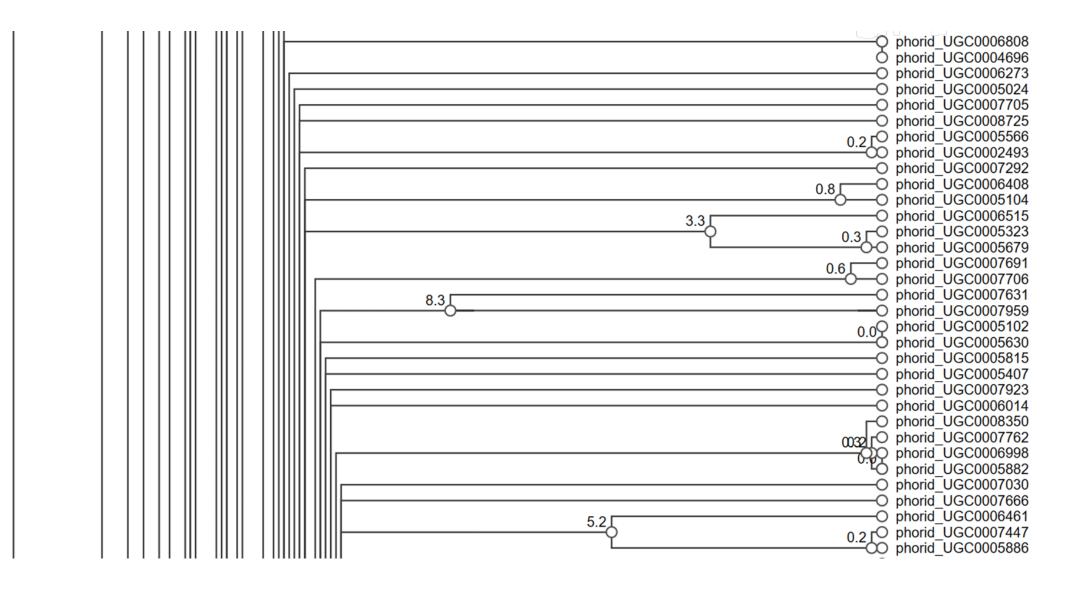


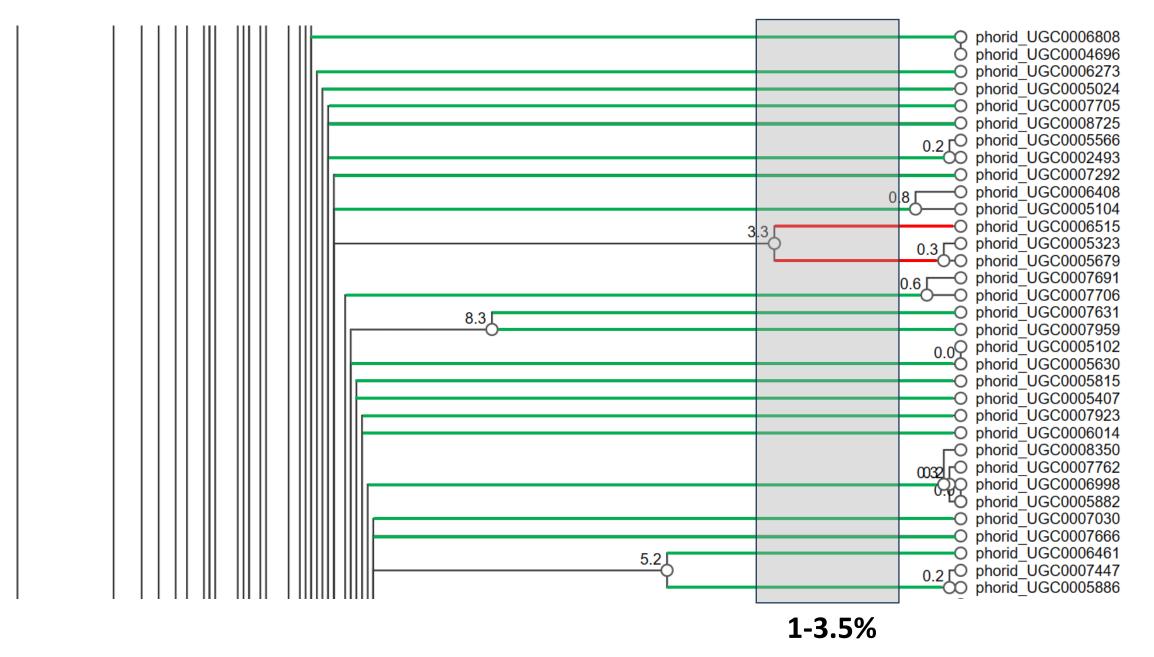
AliView - 658_minion_sep2_Filtered_barcodes_1percamb_upto5err_arthro.fa

File Edit Selection View Align Tools External commands Help

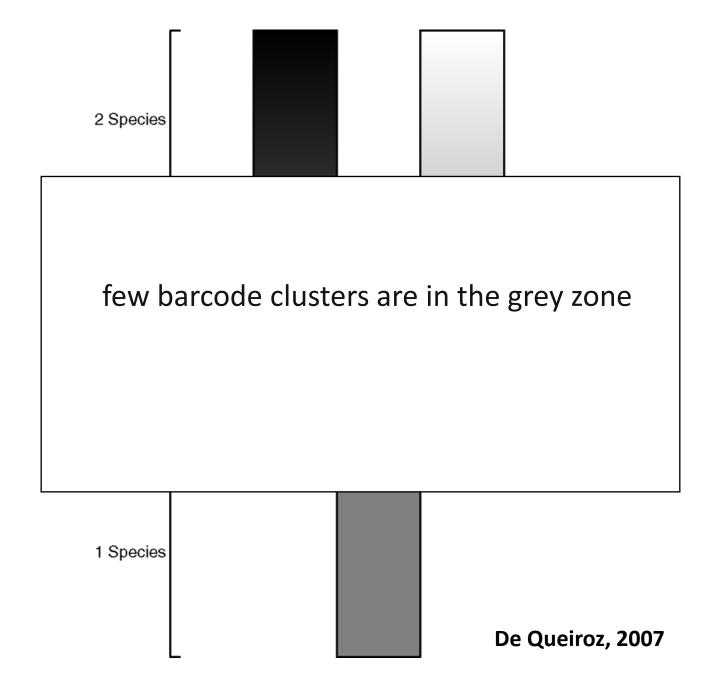
	File Edit Selection View Align Tools External commands Help						
	10 20 30 40 50 60 70 80 90 100 110 120 130 140						
enya_CN2281_CNMyc025_22AUG2 🔥 🔒							
enya CN5195 CNMyc056 26JUN2(A	A A TITT A TATTT CATTITT G GT GTTTG AT CA G GAATA GTTG GAACTT CATTAAGA GTTATTATT CGAACAGAAATTA GGACATTT CTGATGTAATG GTAATG GTA						
nya_CN1952_CNMyc021_15AUG2 T	A T C I T A T A T T T T A T T T T A A C A A T A C A A C A T A C T A A C A A C A T A A C A A C A T A C C A C A						
nya_CN2548_CNMyc028_22AUG2 A nya_CN0300_CNMyc004_01AUG2 A	A A TATTATATTTTTTTT GA G C C T G A A TA G T A G T A G TA C T T CATT A A G TA C T T CATT A A T T C G A G C C G C A T T A G C C C G C A T T A G C C C G C A T T A T T C C A A A T T T A T A A T G T T A T T G T A A T T C A T A A T C A T A A T C A T A A T T C A T A A T T C A T A A T C A A A A						
	A A LII AI A CII CAI III GUI CAI AI AI CII A A A A A A A A A A A A						
nya CN3353 CNMyc0036 29Augi 🔉	AATATTATATTTTATATTTTA GAAGCTTG AGCTTG AGCTTG GAATAGTTG GTACTT CATTAAGTATATTTG GAGCAGAATTAG GACATCCA 6 GAGCATTAATTG GAGAT GAT GAT GAT GAT AATTTATAAT GTTACTG CT CATG CATTATTC GAACTTC AATTTTATAAT GTTACTG CT CATGCATTATTATATTC GAACTTC AATTTTATAATTC GAACTTC AATTTTATAATTC GAACTTC AATTTTATAATTC GAACTTC AATTTTTATAATTC GAACTTC AATTTTTTTT AATTTTTATAATTC GAACTTC AATTTTTTTTTT						
nya CN0785 CNMyc009 08AUG2	A TATT TTAT ET ATTTTATETTE GA & CEEDA & CEEDA & CEETA & CAETTE A TTTATE A ATTTATE A ATTTATE A ATTTE A ATTE A AT						
nya_CN2598_CNMyc028_22AUG2	A TATATATATATATATATATA CONCEPTION CONTACTATATA A CONTACTATA A CONTACTATA A CONTACTATA A CONTACTATA A CONTACTATA A CONTACTATA A CONTACTA A CONTA						
nya_CN3590_CNMyc039_29AUG2 A nya_CN4254_CNMyc046_12SEP2(T	A A TATTATATTTTATATTT GA G G C T G GAATAATTG GA A C T T C C T TAAG C A T C C T G GA C C T C GA C C A A A T T T A TAAT G T A A T T G C A A C A C C T C A A G C T C C A C A C C A C C A A A T T C C A A C A C						
nya_CN5876_CNMyc063_redo1_2							
nya_CN3472_CNMyc037_29AUG2 A	A TACT & TATTTTATTTT & GAAG CTT & ATT & TT & GAATAGT C & GAACTT CTT TAAGAATAATTATT C & T & C & C & C & GAATTA & GACACCCC & G & C TTTTAATT & GAAAC GAT CAAATCTATAAT & TAGTA & G & C & C & C & C & C & C & C & C & C						
nya CN4701 CNMyc051 redo1 0 🛛 🗛	$\mathbf{x} \mathbf{x} \mathbf{x} \mathbf{x} \mathbf{x} \mathbf{x} \mathbf{x} \mathbf{x} $						
nya_CN4560_CNMyc049_26SEP2(A	A TATTATATTTTATTTTC 6 GA CCTT 6 AT CA 6 GAATA 6 T 6 6 GACAT C C CCT A 6 T A ATTATT C 6 T 6 CT 6 AATTA 5 6 C CA CCCT 6 6 GT CA CT A ATTT C 6 AA ATTTATAT AT 6 T A GT CA ATTTATAAT 6 T A GT CA A CT 6 CT CA T 6 CT CA T 6 CT CA T 6 CT CA T 6 CT CA CT 6 CT CT CA CT 6 CT CT 6 CT CT						
nya_CN2511_CNMyc027_22AUG2 A nya_CN2554_CNMyc028_22AUG2 A	ΛΑΤΑΤΙΑΤΑΤΙΤΙΑΙ ΤΙΤΙ Ο ΝΑ Ο COT Ο ΝΑΤΙΑ Ο ΤΑ Ο ΤΑ Ο ΤΑ Ο ΤΑ ΑΤΙΑ ΑΤΙ ΛΑΤΙ CO Α Ο CON Ο ΑΤΙ ΛΑΤΙ ΤΑ ΤΙ ΤΑ ΤΙ Ο ΤΟ ΛΑΤΙΑΙ ΤΟ ΤΑ ΤΙ Ο ΤΑ Ο ΤΟ ΛΑΤΙΑΙ ΤΟ ΤΑ Ο ΤΟ ΛΑΤΙΑΙ Ο ΤΟ ΛΑΤΙΑΙ Ο ΤΟ ΛΑΤΙΑΙ ΤΟ ΤΑ Ο ΤΟ ΛΑΤΙΑ						
nya_CN5958_CNMyc064_12OCT2	A CATTATATTTTATATTTTA GA COTTAA GA ATA TTC TA A GA ATTTTAATTC GA ATTA ATT						
venport_maggots3_maggots_NA_N A	A A C C TT A T A TT TT A T TT T A G A B C A B C A B C A T A G T A G C A A C A T C A C T A A T T C C A B C C D A A T A T T A T A C A B C C C A C A C A C A C A C A C A						
nya CN4754 CNMyc051 redo1 0 T	ACTITATATITIATITIATITIACAACTIAAATAAACTIACAACTIACAACTIACAACTIACAACTIACAACTIACAACTIACAACTIACAACTIACAACCAAC						
nya_CN1256_CNMyc0014_8Augus	CACATTATACTTCATTTTTE GAGCATGATTATAGTGGGAACATCTTTAAGGATATTAATCGGGGTTCCTAATGGCGTGTCCTAATGGCGAGATTATATATTATATTATACTACATGTATATAATCGAGATATTATAATCGGGGACATGATGTATATATA						
nya_CN2017_CNMyc022_redo2_2 A nya_CN5860_CNMyc063_09SEP2(A	a = correct a tattita to tata tattice a general correction and tattice a general correction and tattice a second correction						
nya_CN0720_CNMyc008_01AUG2 A	A A CTITIATATITIA TATITIA TATITA TATITA A TITATITA						
nya_CN0324_CNMyc004_01AUG2	A A CTITIAT A CTITIATITITI 5 5 5 5 CCT5 A T CAA GAATA ST C 5 6 TA CAA CT CT CTIA 5 A ATA CT A ATA CT A ATA CT A ATA C T A ATA C TA CAA C A C						
nya_CN1196_CNMyc013_08AUG2	A TATTTTATATTTTATTTTATATTTACAACTCCACCACCAC						
nya_CN1066_CNMyc012_08AUG2 T nya_CN4818_CNMyc052_10OCT2 A	A CTTTATACTTTATACTTTATACTAGGAACTTGATCAAGTATCCATGCATAAGTATCCATGCATTCGAACTGCATTCGAACTGCATCGCATTCGGAACTGCATGCTTTCT						
nya_CN4616_CNMyc052_100C12 A							
nya_CN3767_CNMyc0041_29Augi A	A A T A T A T A T A T A T A T A T A T A						
iya_CN1457_CNMyc016_15AUG2 A	A A CTITA I A CTITATITI 5 5 C C CATS A T CAT A F A S F A S A A C A T A CTATA A S A A T A CT A A T T C S S C T S A A T A C S S C C A A A T S C A T A T T A T S T A T A T T A T S A T A T						
nya_CN0713_CNMyc008_01AUG2	, AT ATT AT ATT T ATT TT I GO A GO CT GO A AT A GT <u>A</u>G GT A CT T C ATT A AGT ATT A AGT A TATT A ATT C A A GT C A ATT A GO A GO A T A ATT F G A GAT O AT C A A ATT T A T ATT A TAT F T A A CT O C A C ATG C T TT T						
nya_CN3702_CNMyc0040_5Septe A nya_CN3437_CNMyc037_29AUG2 c	A TATT AT ATTTTATTTTC 6 GA GCTTG A 6 CA 5 GA AT A GT G G GTA CAT CATT A A GT ATA CTTATATT CG A GA CACCCCC 6 GA GCTTT A ATTC AT CAA AT CTATATG TA ATTG TA CT GCT CATG CCTTT A						
nya_CN0600_CNMyc007_01AUG2	A A T G TT A TATT TT ATT TT ATT TT A G A A C C T C A TT A A G A A C C TT T A A TT C G A G C T T T A A TT C G A G C T T T A A TT C G A G C T T T A A TT C G A G C T T T A A TT C G A G C T T T A A TT C G A G C T T T A A TT C G A G C T T T A A TT C G A G C T T T A A TT C G A G C T T T A A TT C G A G C T T T A A TT C G A G C T T T A A TT C G A G C T T T A A TT C G A G C T T T A A TT C G A G C T T T A A TT C G A G C T T T A A TT C G A G C T T T A A TT C A G C A A T T T A T T A T T A T T C A G C A A T T T A T T C A G C A A T T T A T T A T T A T T A T T C A G C A A T T A T T A T T A T T A T A T T C A A G C C T T T A A T T C A A G C C T T T A A T T A T T A T T A T A T T A T A T T A T A T T A T A T T A T						
	AATATTATATTTTATATTTTGGAACTTGATCAGGTATAGTAGGAACTTCTCTTAGTATAATTATTCGTGCTGAATTAGGACACCCCCGGATCCTTAATTGGTAACGATCAAATTTATAATGTAGTAGTTGTAACTGCCCATGCTTTA						
enya_CN2307_CNMyc025_22AUG2 T	A CTITATA CTITATTETA GOAACTIGAT CAGOAATATAGOAACAT CCATAAGTATCITAATT COAACTIGAACTAGOCATTACCACGACATTACTATTAGOCACTIGATAATTAGOCAATCAAATTAAATGTAACTGCTCACG						
nya_CN4834_CNMyc052_10OCT2	A TTTTATA CTTTATTTTT 6 GAATTT 6 AT CA 6 GAATA GTA 6 GAACCT CATTAA 6 A ETTATT C GAACAATAA 6 A CACCCA 6 GT CTAATT 6 GAAATTA 6 GAAATTTATATATTTTTT 6 GAAATTTATATATTTTTT 6 GAAATTTATATTATTATTATTATT 6 GAATTT 6 TTATT 7 TTATT 6 TTATT 6 TTATT 6 TTATT 7 8 TTATT 8 TTA 7 8 8 TTA 7 8 8 TTA 7 8 8 TTA 7 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8						
enya_CN2002_CNMyc022_redo2_2 T enya_CN1378_CNMyc015_15AUG2 A	ΓΑΣΤΤΙΑΤΑΣΤΤΑΙΤΕΤΑ ΘΑΑΣΤΤΑΙΤΑ ΘΑΑΤΑΤΙΑΘΕΑΤΕ ΤΑΑΤΤΟ ΘΑΑΤΙΑΘΕΤΑΙΤΑΤΙΑΘΕΤΑΙΤΑΙΤΟ ΜΑΑΤΑΙΤΑΙΤΟ ΜΑΑΤΑΙΤΟ ΤΑΑΤΙΘΙ ΤΙ ΑΣΤΙΔΙΤΙΔΙΤΟΙ ΤΑΙ ΤΑΙ ΤΑΙ ΤΑΙ ΤΑΙ ΤΑΙ ΤΟ ΣΕΙΤΑΙΤΟΙΟ ΤΙ ΑΙ ΤΟ ΣΕΙΤΑΙΔΙΟΙ ΤΙΔΙΟΙΑΤΙΘΙΑΤΙΘΙΑΤΙΘΙΑΤΙΘΙΑΤΙΘΙΑΤΙΘΙΑ						
enya_CN5532_CNMyc060_24JUL20	A TTTT AT ATT TT ATT ATT ATT ATT ATT AT						
nya_CN1221_CNMyc013_08AUG2 A	A A CATTAT ATTATATETTATETTA CATTA A CAACATA CAACAACTT CATTA A CTATA A CTATA A CTATE CAACCT CCAACTT CAACAT CTATE ATTATATTAT						
nya_CN5021_CNMyc054_08MAY2	A T A T A T A T T T T A T T T T A T T T T C 6 6 A F C T T A A T A F T 6 6 6 A X T A F T 6 C T A A T T A F T A T T						
nya_CN4884_CNMyc053_redo1_2 T	A TATTATATTTTATTTTTE GASCTTE AS CAT STATASTCE SAACTTCTTTAAS AATATTAASTCE SACTTCCTE SACATCCTE SACATCATTATATTE SA						
nya_CN4740_CNMyc051_03OCT2 A nya_CN4976_CNMyc054_redo1_0 A	Α ΤΑ ΤΑ ΤΑ ΤΑ ΤΑ ΤΑ ΤΑ ΤΕ ΤΑ ΤΕ ΑΤΑ Ε ΑΤΑ Ε ΑΛΑ ΤΙ ΑΤΕ ΤΕ ΤΑ ΑΤΑ ΤΙ ΑΤΕ Ε ΑΛΑ ΤΙ ΑΤΕ Ε ΑΛΑ ΤΙ ΤΑ ΤΑ ΤΕ ΤΑ ΑΤΕ Ε ΤΑ ΤΑ ΤΙ ΑΤΑ ΤΙ ΤΕ Ε ΑΛΑ Ε ΤΙ ΑΤΑ ΤΕ Ε ΤΑ ΑΤΑ ΤΙ ΑΤΕ Ε ΤΑ ΑΤΑ ΤΙ ΑΤΕ Ε ΑΛΑ ΤΙ ΤΑ Ε ΑΛΑ ΤΙ ΤΑ ΤΑ ΤΕ ΤΑ ΑΤΕ Ε ΤΑ Τ ΤΑ ΤΙ ΑΤΑ ΑΤΑ						
nya CN0234 CNMyc003 18JULY2 T	A AUTTTATAUTTTATAUTTTATAUTTAG GAAUTTGAT CAG GAAUTATTAG GAAUTTGATTGUAAUTTGUAAUTTGUAAUTTGUAAUTTGUAAUTTATAATGUAAUTGUAAUTGUAUTGUTTU						
iya CN1978 CNMyc0022 22Augi T	A RETTTATATTTTATTTTATTTTE S A RETT S AT A A T S S A A C A T A A T T S S A A C T T C C T T A A S T A T A T T A S T A T A T T A S A A C T T S S C A C C C A S S C S C T T A A T T A T A T S A T A T A T T A T A						
nya_CN4697_CNMyc050_03OCT2	. A T A T T T T A T T T T A T T T T A T T C A B G T A T A A T <u>B G C A A C A T C C T T A S T C C T A G T T A T T A C C C C A T C C A A C T A A T T A C C A A A T T A C C C T A A C T A A T T A C C A A A T T A C C C T A A C C A A A C T A C T A A C T A C T A A C T A C T A A C T A C A C</u>						
nya_CN1075_CNMyc012_08AUG2 A nya_CN4581_CNMyc049_redo1_2 A	A CUT T MT ATT T T A TT T T A T GA MAA TA GA A TA GA A TA TT A A TT C G G G C T GA A T A T T G GA MA A TT GA C CA A A TT TA TA A T G T A A TT G T A A T A T						
nya_CN5043_CNMyc049_red01_2							
nya_CN5022_CNMyc054_redo1_0 🔉 🔉	A A TATTATTATTATTATATTATATTATATATAGAA COATCATCATAGTAGAACAACTAAGTACTAGTACTAGTACTAGTACTAGTACTAGTACTAGTACTAGTACTACTACTACTACTACTACTACTACTACTACTACTACT						
nya_CN0321_CNMyc004_01AUG2	A RETTIATA ETTEATTITE FECTERATION TO A TAFA FA FA A TAFA A A TAFA FA A TAFTA FE A A TA TA FE CATA A TAFA TA FA A TATA TA A FA A TATA A FA						
	A A TATTATATTTTATTTTATTTTTA G G A A TA G CA A CATT CTTTA A G A A TTTT BATT C G A G CT G A A TTTATATTATATTAT C G A C C C T G G A A TTTTATAATT G G G A A TTTATAAT G T A A TA T G T A A TA CA C C T C A TG C TTTA						
enport_SK00057_SK1_NA_NASK_ A iya_CN5357_CNMyc058_10JUL20 A	A CULLA TATATTTATATTTATATTATATATTATATTATATT						
ya_CN1585_CNMyc017_15AUG2	A A CTTT A TA CTTC A TTTT E GT CC A GA T A CT A TA GT A G A A A A T CTTT A A GA A TA TT C G G C C GA A T C TTT A A TT C G A C A A T C T TT A A GA A TA TT C T A A GA A TA TT A TT						
ya CN3591 CNMyc039 29AUG2 T	A RETTTATATTTTATTTTATTTTE CAASTTS AT CAS CAATATTS CAACTT CETTA ASTATATTATT CEATTA CATECACCA AS CECTTA AT CACCAAATTTATAT AT CTTATAT CE CACATS CITTA						
a CN3762 CNMvc0041 29Augu A	A A CITIATA CITIATATITI 6 CITCAT 6 A CA F 6 A A TA CIA CIT CAATA A GA A TICITATIC CA TA COATACIC T 6 CA CITIAATI 6 CA A A TICAA A TA A TA A CIA A CITA A TA CAATA COATACIC T 6 CA A A TITIATA TA CIA A TO TA A TO TA A TO CITA CI CAATA A CAATA COATACIC T 6 CA A CITIAATI 6 CA TA CIA A TITIATA TA CITA A TA CITA						
/a_CN4151_CNMyc0045_5Septe A	ATATTATATTTTATATTTTATATTC G GAAGATCA C GTATAGTA C C CTT AGTAT GATTATT C GTG CT GAATTA C GC CAT CCT C GTT CTTT AAT C G GAAAATTTATAT GT AGTAACT G CTCAT G CTCAT G CTTTT /						
ra_CN3639_CNMyc039_05SEP2(T ra_CN3663_CNMyc039_05SEP2(T	Α ΤΤΙΤΑΤΤΙΤΙΤΙ ΑΤΙΤΙΤΙ ΑΤΙ ΤΙΤΙ ΑΤΤΟ Α ΜΑΤΤΑΤΤΟ Ο ΑΧΤΤΑΤΤΟ Ο ΑΧΤΤΑΤΤΟ Ο ΑΧΟΤΤΙΑΤΤΟ Ο ΑΧΟΤΤΙΑΤΤΟ Ο ΑΧΤΤΑΤΑΤΤΟ Ο ΑΧΑΤΤΑΤΑΤΤΟ Ο ΑΧΟΤΤΑΤΑΤΤΟ Ο ΑΧΟΤΤΑΤΑΤΟ Ο ΑΧΑΤΤΑΤΑΤΟ Ο ΑΧΟΤΤΑΤΑΤΟ Ο ΑΧΟΤΑΤΟ Ο ΑΧΟΤΟ Ο Α						
ya_CN3865_CNMyc039_055EP20 T ya_CN3216_CNMyc035_29AUG2 T							
/a_CN0273_CNMyc003_25JULY2 A	A TACTATACTATACTATATTTT & GA C CAT & G C C G G A TA & TT & C C C TA A G A A TATT A A TT C G A & C C C TT TA TT C G A & G A & C TT TA TT & C A & C						
/a CN3757 CNMyc0040 5Septe A	. AT A T A T A T T T A T A T T T A T T T C F A F A T A F A F A F A A A T T A C T A C T A C T A A A T T A T C F A C A C A C C C C C C C C C C C C C						
ya_CN4266_CNMyc046_12SEP2(T	ACTITATATTTTATTTTTATTTTTGAAATTATATTGATAAATTGAACTICCTTAAGTATAATTATTCGAACCCCACCCC						
iya_CN0605_CNMyc007_01AUG2 T iya_CN0778_CNMyc009_08AUG2 A	A CATTATATTTTTATTTTTGATTGATGAAGAATAGTCGGAACATCATTAAGTCTATAATTGGAGGATAATTGGAGGATACCCAAATTCTTTAATTGGAGCCAAATCTATAACGTATAGGAATAGCCTCATGCTTTTA						
nya_CN3323_CNMyc0036_29Augi							
nya_CN5942_CNMyc064_12OCT2	A ATTTTATATTTTATATTTTATATTTATATTATATTA						
nya CN1265 CNMyc0014 8Augus A	ATATATATATTTATTTTATTTTTATATTTTAACACTEGTATAGTAGGAACCTCATTAACACTTTTAATTCGAGCTGAATTAGCAEGACTTTAATTGAAATTTTAATGAATTTTAATGAACACCTCAAGCTCTAATT						
nya_CN0943_CNMyc011_08AUG2	A A CTTTATA CTTCATTTTTG GTGCATGAT CAGAATAGTA G GA A CATCTTTA A GA A TA CTA A TTCG G GCTGA A TTAG G GTGCCCA A A TGCTTTA A TTGGA GA A CA A TTGTATA A TGTA A TGTA A TGTA A TTGTA A TGTA A TGTA A TTGTA A TGTA A TTGTA A TTG						

"Cluster Fusion Diagram"





=> most haplotype distinct and rare => almost certainly species



Singapore dataset: 350,000 specimens

=> ca. 9,000 species with distribution and abundance information

=> 45 minute-movie

Few species are in the "grey zone" where species limits are unclear



1-3.5%

DNA barcoding or Integrative Taxonomy?

<u>Step 1</u>. DNA barcoding for presorting specimens to putative species

<u>Step 2</u>. Morphological checks using a subset of specimens representing

- main haplotypes for distinct barcode clusters
- main and divergent haplotypes for "grey-zone clusters"
- <u>Step 3</u>: Resolving incongruence

Step 4: Identification or Description

JOURNAL ARTICLE

Towards Large-Scale Integrative Taxonomy (LIT): Resolving the Data Conundrum for Dark Taxa 3

Emily Hartop, Amrita Srivathsan, Fredrik Ronquist, Rudolf Meier 🐱

Systematic Biology, Volume 71, Issue 6, November 2022, Pages 1404–1422, https://doi.org/10.1093/sysbio/syac033

Published: 13 May 2022 Article history •





Dr. Emily Hartop

Managing specimens: voucher handling



Vending machine for specimens

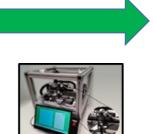
Carousel system for high density, low footprint storage

Pneumatic technology for vial storage and retrieval

function	comPOUND		
temperature	-20°C, +4°C, ambient		
dimension (w x d x h)	1.2 x 1.65 x 2.4 m (48 x 65 x 93 inches)		
capacity	100,000 x 1.4 mL vials, 200,000 x 0.5 mL vials		
throughput	600 vials/hour (1 sample/6 secs/unit)		













Managing data: from Entomoscope to Zenodo and GBIF

Automatic data upload to CERN's Zenodo

- Entomoscope and DiversityScanner software has Zenodo upload feature
- each specimen gets Zenodo entry with DOI (images, barcodes)



18 September 2024, 6 pm at the Embassy of Switzerland in Germany



Dr. Amrita Srivathsan



Dr. Donat Agosti, Plazi

Automatic data upload to Zenodo/GBIF

& ENIMAS - Entomoscope Imaging Software



STOP

-

\$

+

0,80 mm

Step size





Dr. Amrita Srivathsan



Dr. Donat Agosti, Plazi

 \square

SpeciesID: Prediction: Vespula_vulgaris Probability: 0.9302 Open in Wikipedia How quickly can we generate good biodiversity data?

How quickly can we generate good biodiversity data?

5000 new data points for Berlin Kreuzberg

Bachelor's thesis

- 1 Malaise trap: 2-week July sample
- 5951 individuals
- 750 species





Let's imagine: 1000 BSc theses per year

- => 5 million new data points
- distribution of thousands of species known

imagine doing this for several years and in many places...



Cem Kir

My challenge to you: be the first country that knows its arthropods

Step 1: barcode 1,000 specimens for 500 sites/sample (500,000 specimens) => species discovery rate can be 5-10% (i.e., 25,000 species) Step 2: image numerically important species to train AI models => AI tools for common species

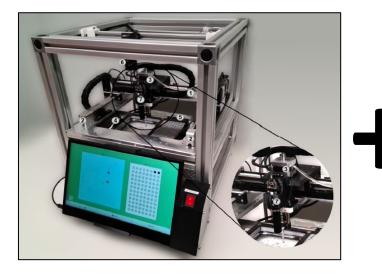
This would yield:

- unbiased abundance and diversity data for 500 sites
- barcodes for 25,000 species to interpret metabarcoding data
- vouchers, barcodes, and AI models for the most abundant species
- critical information on distributions and ecology of these species

Species Discovery Factory

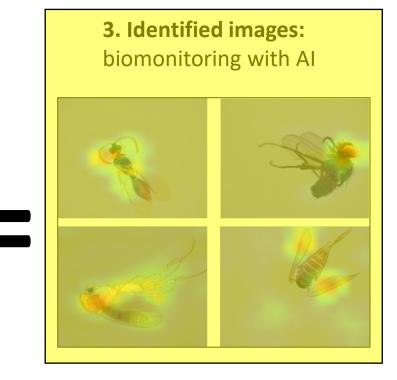


1. DiversityScanner: Digitization with robots



2. Nanopore Sequencing: Sorting with DNA barcodes





3. Identification with AI

Class (Taxon)	Result	Class (Taxon)	Result
Diptera Acalyptratae	91%	Diptera Psychodidae	89%
Diptera Calyptratae	83%	Diptera Sciaridae	92%
Diptera Cecidomyiidae	91%	Hemiptera Cicadellidae	100%
Diptera Chironomidae	97%	Hymenoptera Braconidae	82%
Diptera Dolichopodidae	86%	Hymenoptera Diapriidae	100%
Diptera Empididae & Hybotidae	87%	Hymenoptera Ichneumonidae	75%
Diptera Keroplatidae & Mycetophilidae	99%	Other	81%
Diptera Phoridae	97%		

AI identification tools evolve

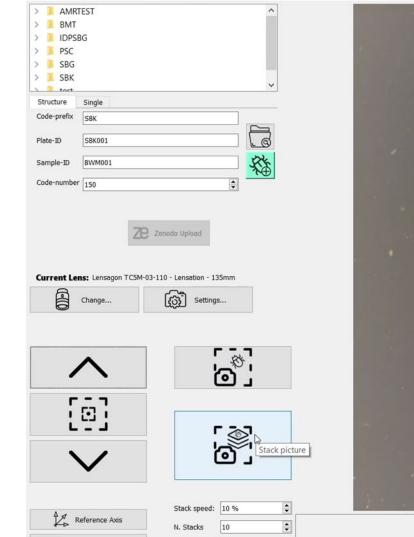
- common taxa benefit first
- tools evolve from family => genus => species
- latest version of DiversityScanner/Entomoscope has an AI model for common species in Berlin



Hossein Shirali, KIT



STOP



0,80 mm

Step size

\$





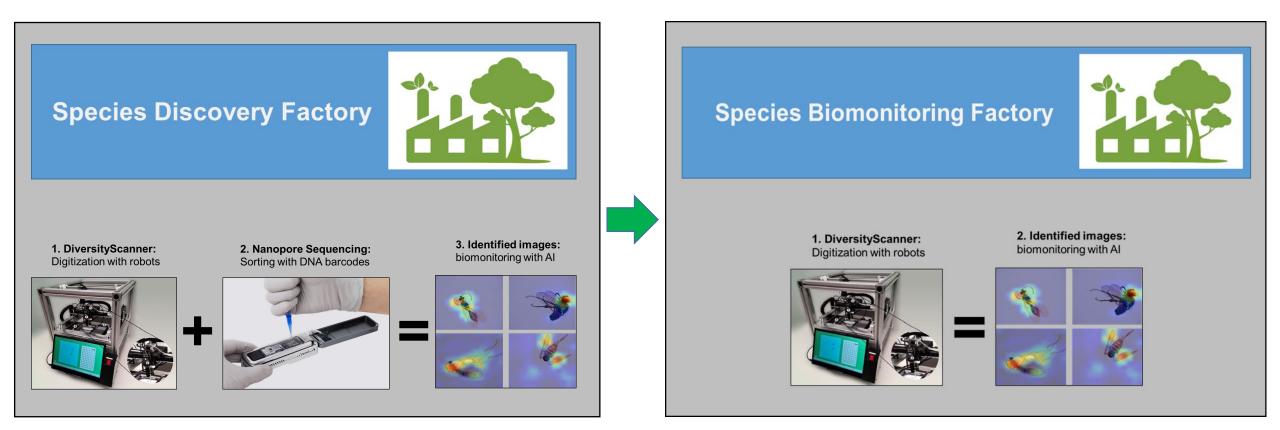
O

_

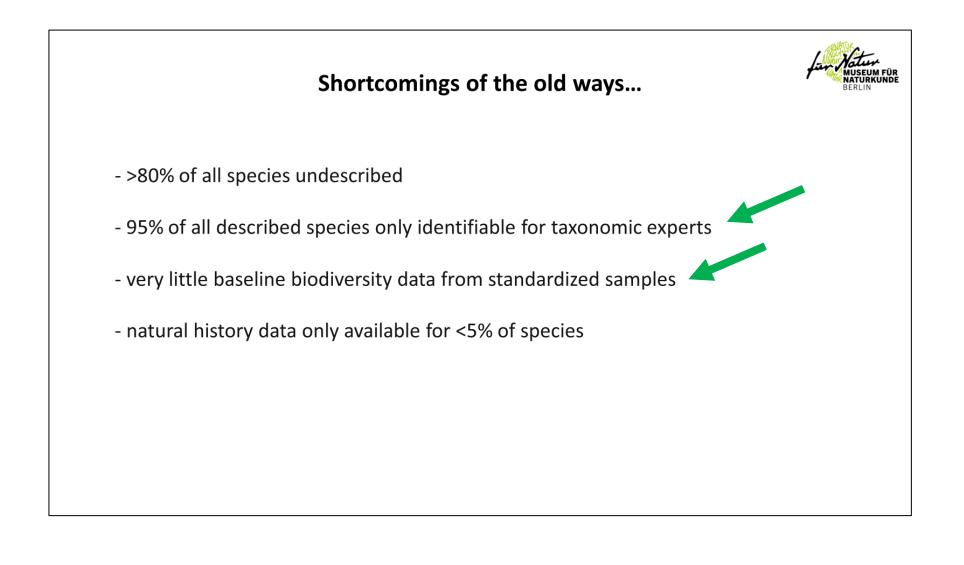
×

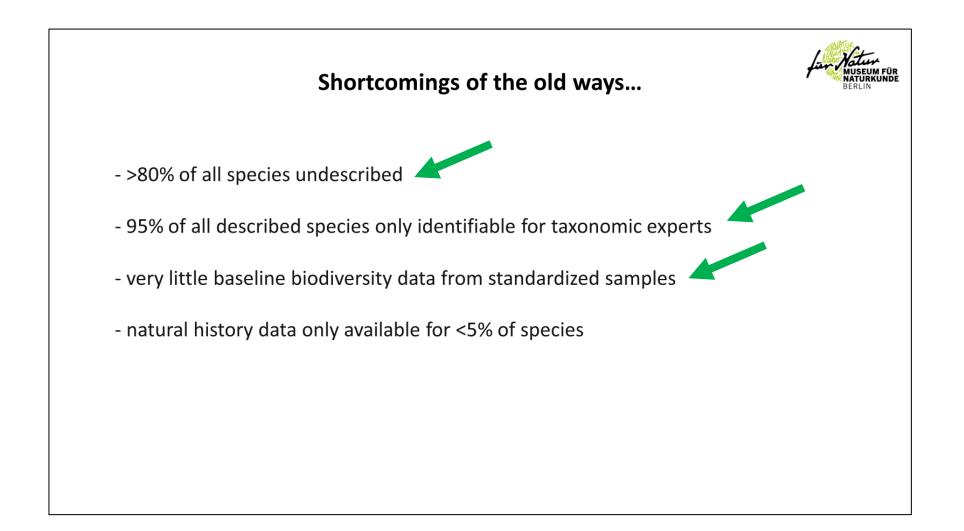
Hossein Shirali, KIT

Does the model really work? some species are easy some are hard



Could generate large amounts of data quickly





Too many species... isn't it hopeless?



"Dark Taxonomy": focus taxonomic work on species important for biomonitoring



"Dark Taxonomy" with biomonitoring focus

<u>Step 1</u>: Get fresh material for a species-rich taxon using traps also used for biomonitoring



Taxonomic sample

"Dark Taxonomy"

<u>Step 1</u>: Get fresh material from many traps used in biomonitoring

<u>Step 2</u>: Pre-sort specimens with DNA barcodes



"Dark Taxonomy": phase 1 of taxonomic exploration

<u>Step 1</u>: Get fresh material from many traps used in biomonitoring

<u>Step 2</u>: Pre-sort specimens with DNA barcodes

<u>Step 3</u>: Integrative taxonomy with two data sources

"Dark Taxonomy"

<u>Step 1</u>: Get fresh material from many traps used in biomonitoring

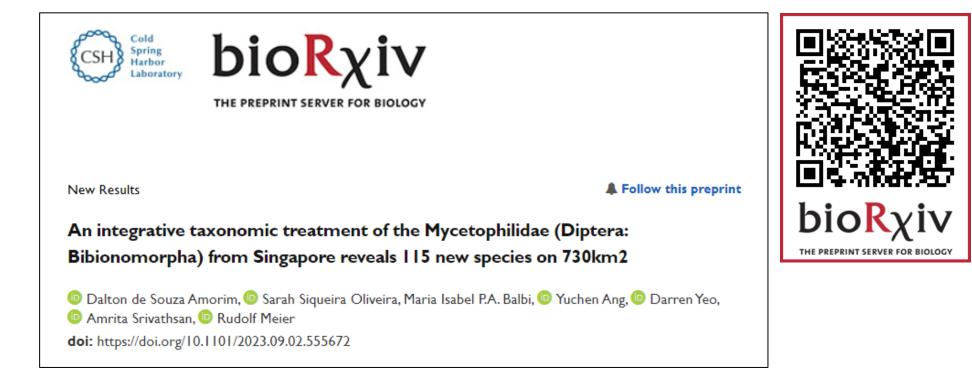
<u>Step 2</u>: Pre-sort specimens with DNA barcodes

<u>Step 3</u>: Integrative taxonomy with two data sources

<u>Step 4</u>: Identify/describe species

120 species of fungus gnats (Diptera: Mycetophilidae) discovered in Singapore:

- 5 described species
- 115 new species



"Dark Taxonomy"

<u>Step 1</u>: Get fresh material from many traps used in biomonitoring

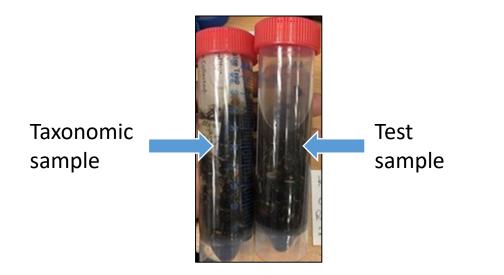
<u>Step 2</u>: Pre-sort specimens with DNA barcodes

<u>Step 3</u>: Integrative taxonomy with two data sources

<u>Step 4</u>: Identify or describe species

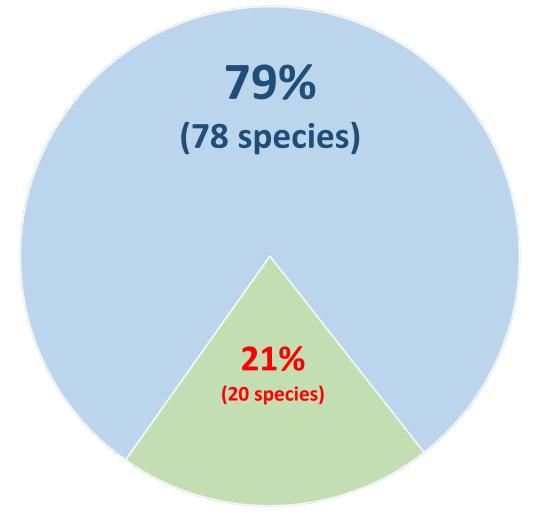
<u>Step 5</u>: Test whether all species important for biomonitoring are covered



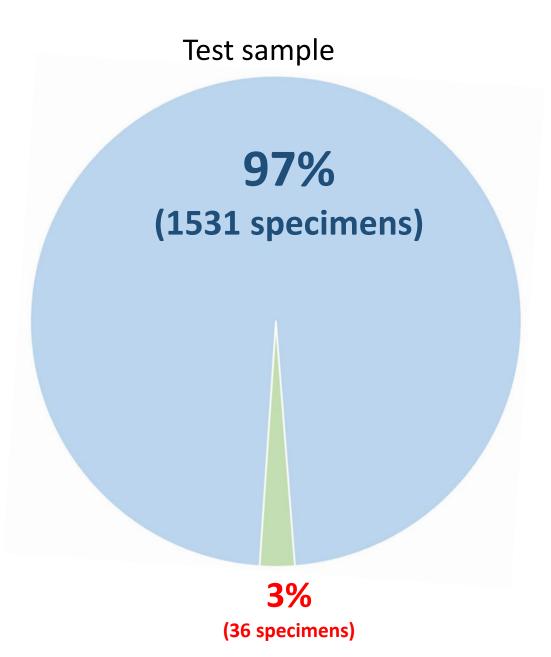


Mycetophilidae would be "biomonitoring-enabled" if test sample consists mostly of described species also in taxonomic sample

Test sample



What about specimens?



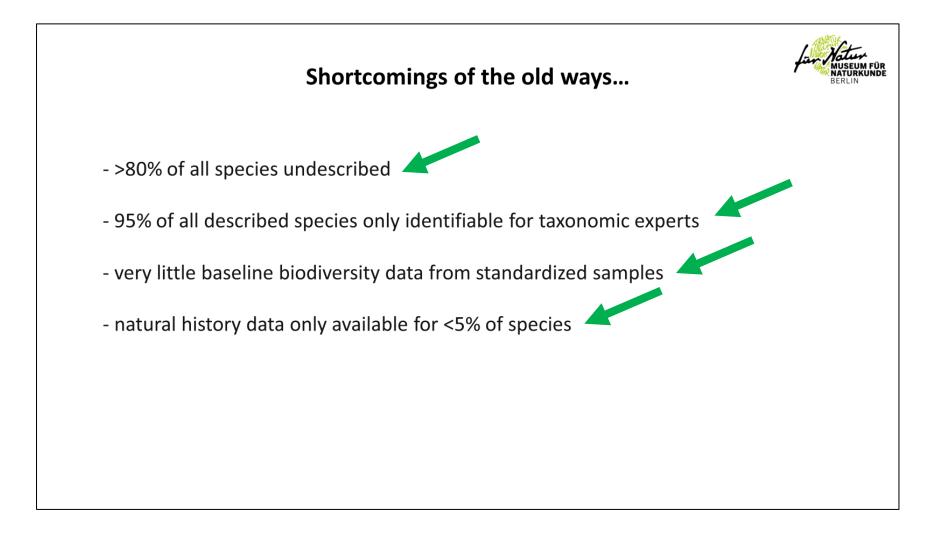
Dark Taxonomy:

from unknown to ready for biomonitoring in 5 steps



=> insect samples and faunas aren't hopeless

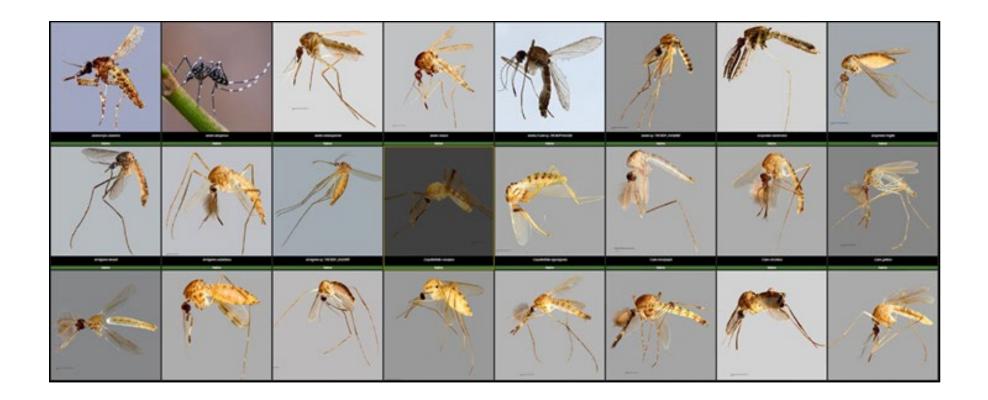




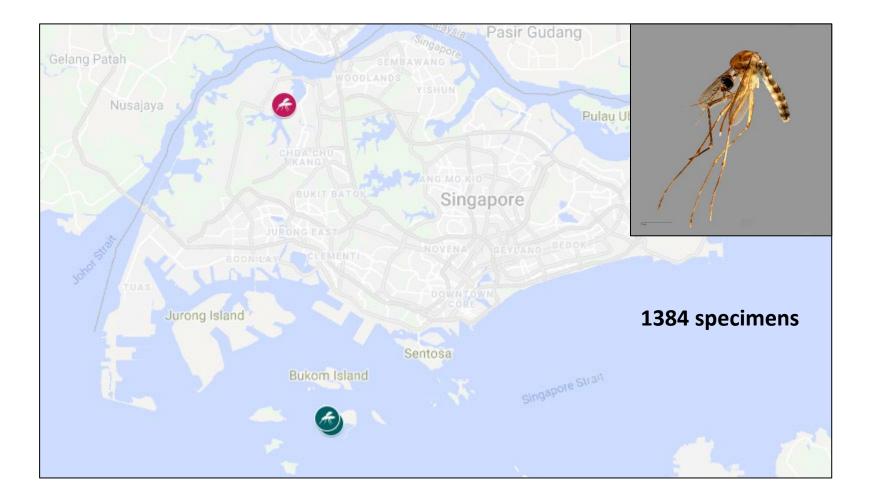
Our recommendation: resolve the natural history of common species first

Example 1: Mosquito diet

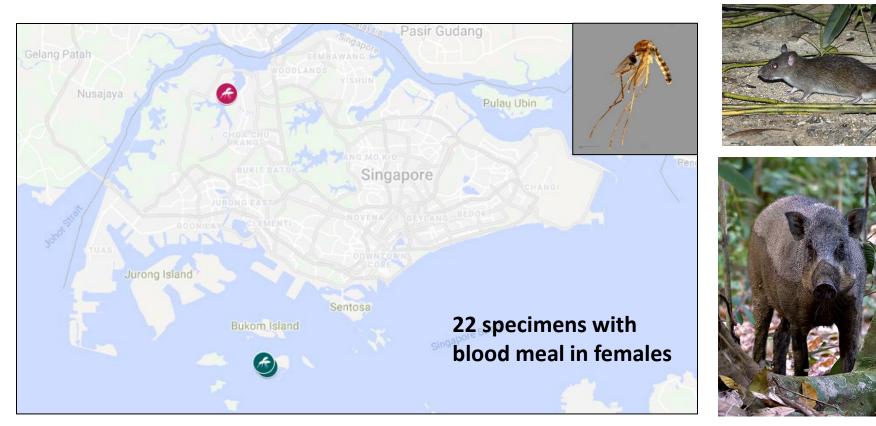
- 4000 mosquitoes barcoded
- => ca. 120 species detected (1/3rd identifiable via BLAST)
- => females for ca. 10 species had visible blood meals
- => metabarcoding of gut content



Aedes sp. ZRCBDP00009685



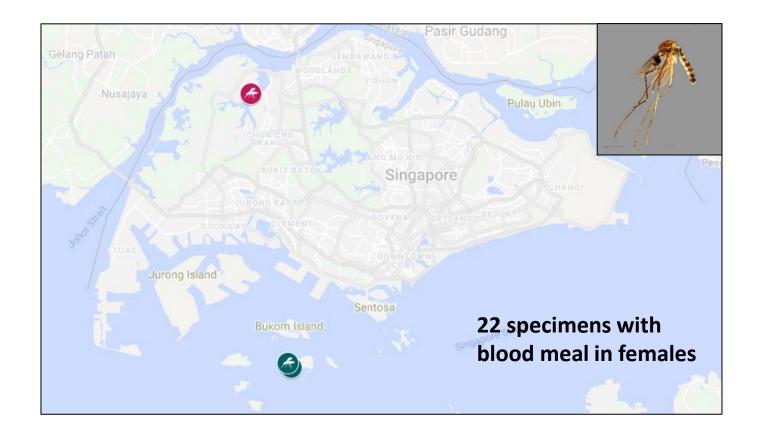
Aedes sp. ZRCBDP00009685







Aedes sp. ZRCBDP00009685



=> high potential as an agent transmitting viruses between distantly related hosts, but fortunately with restricted distribution

Uranotaenia sp. ZRCBDP0009257



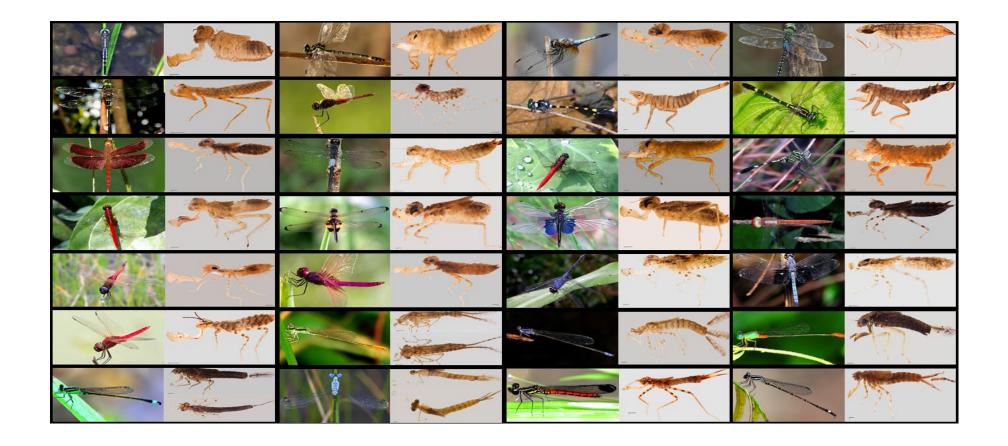


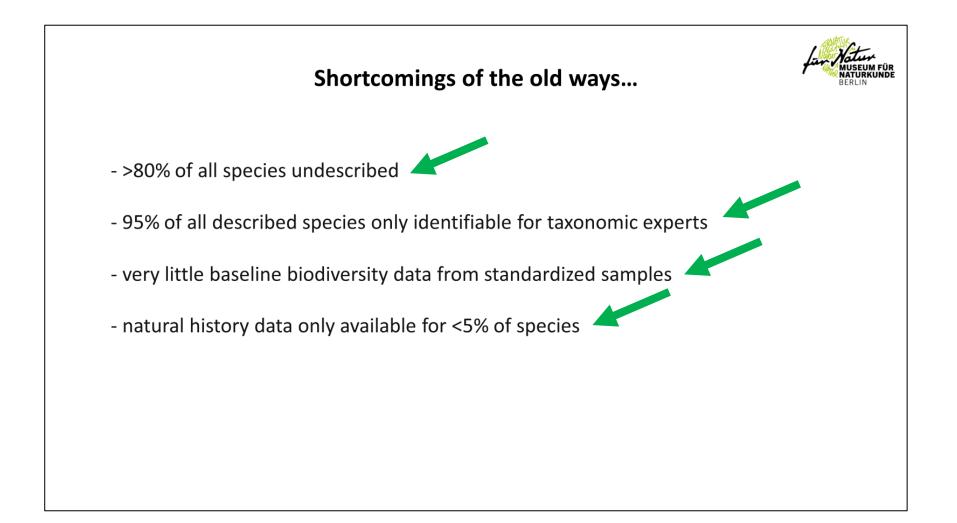
Example 2: Dragonfly nymph ecology

Singapore's odonates (122 extant spp)

- 1500 specimens sequenced
- => larval and adult matches for 59 species

=> molecular work: 1-week





Conclusions

- Society expects us to monitor biodiversity holistically: we are failing
- We can catch up quickly as long as we use robotics, new DNA sequencers, and AI
- But we need "Dark Taxon Biology" to complement existing fields in biology
- Goal should be that biodiversity samples can be scanned like libraries
 - already realized for some samples and taxa
 - needs to be expanded to dark taxa









