Best practices for genetic indicator estimation from DNA-based data

Marie-Gabrielle Harribey, Joachim Mergeay, Gernot Segelbacher, Anja Westram, Peter Galbusera, Alexander Kopatz, Cristiano Vernesi, Leslie Noble, Frédéric Raspail, Sean Hoban, et al.

HAL Id: hal-04627693
https://hal.inrae.fr/hal-04627693v1

Submitted on 27 Jun 2024

Best practices for genetic indicator estimation from DNA-based data

Marie-Gabrielle Harribey[1], Joachim Mergeay[2], Gernot Segelbacher[3], Anja Westram[4], Peter Galbusera[5], Alexander Kopatz[6], Cristiano Vernesi[7], Leslie Noble[4], Frédéric Raspail[1], Sean Hoban[8], Joost Raeymaekers[4], Myriam Heuertz[1], Pauline Garnier-Géré[1]

[1] Biogeco, INRAE, Univ. Bordeaux, France; [2] Evolution and Biodiversity Conservation, KU Leuven, Belgium
[3] Chair of Wildlife Ecology and Management, Univ. Freiburg, Germany
[4] Faculty of Biosciences and Aquaculture, Nord Univ., Norway
[5] Centre for Research and Conservation, Royal Zoological Society of Antwerp, Belgium
[6] Norwegian Institute for Nature Research, Trondheim, Norway
[7] Forest Ecology Unit, Research and Innovation Centre - Fondazione Edmund Mach, Italy
[8] The Morton Arboretum, Center for Tree Science, Lisle, United States
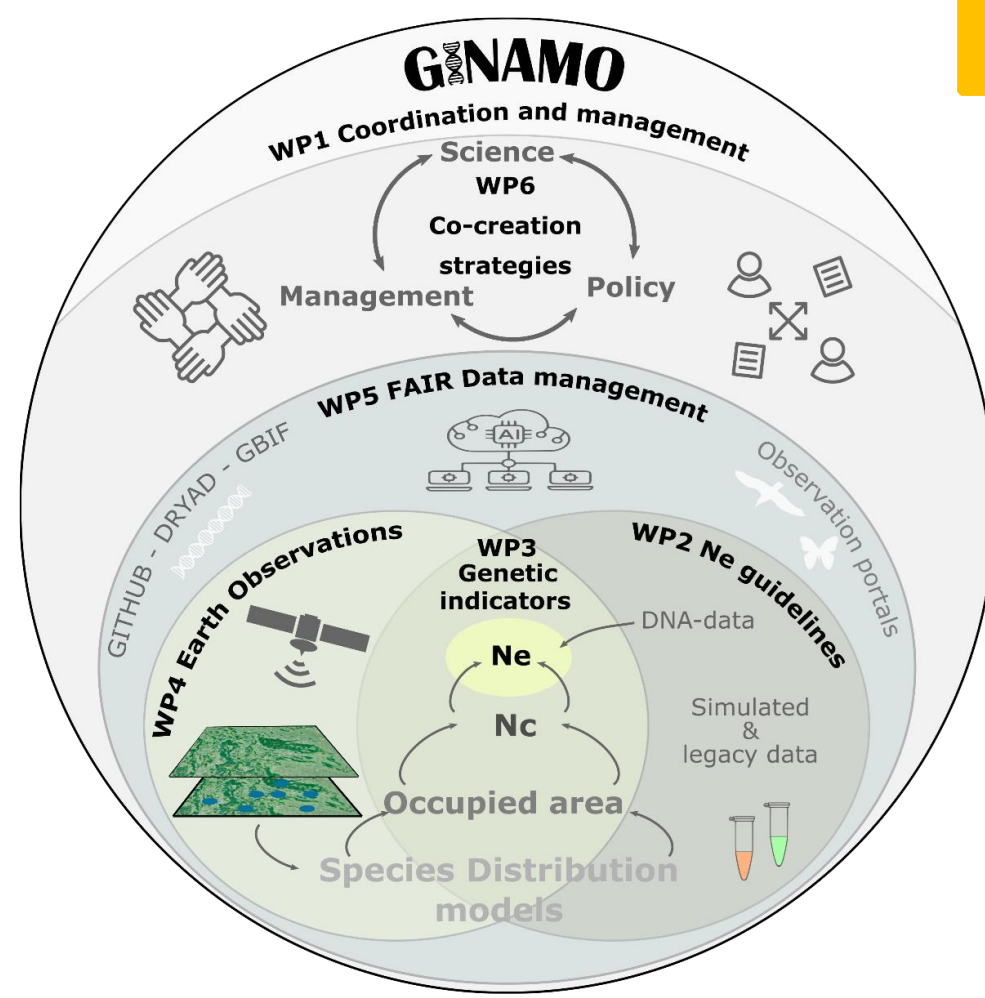
Recent progress in biodiversity conservation policies

- In 2022, a timely **new headline genetic diversity indicator** for application to natural populations (Kunming-Montreal Global Biodiversity Framework - KMGBF): **the proportion of populations with an effective population size (Ne) > 500** [1]
- A Ne > 500 has long been recognized as a minimum threshold to ensure species' survival in the long-term, and potential for adaptation and resilience [2-3]

→ How to estimate recent Ne? Either with demographic data or with genetic data: recent methods from one sample only, using **Linkage Disequilibrium statistics** (*i.e.* statistical association between alleles at different loci, *e.g.* LDNe)

GINAMO Genetic **I**ndicators for **NA**ture **MO**nitoring
Coord. Christina Hvilsom, Copenhagen Zoo - 11 Partners in Europe and the USA - 2024 - 2027

- General aim: facilitate the implementation of the KMGBF by improving the estimation, application and integration into biodiversity observation workflows of genetic diversity indicators
- This work focuses on developing best practices for conservation managers to obtain accurate and robust Ne estimates from DNA-based data (**WP2**)

Data for Q1: Choosing species across life-history trait diversity and population demographic history

Choice strategy for legacy datasets (*i.e.* large population genomic datasets with metadata - *e.g.* census sizes, pedigrees)

→ Categorization framework proposed that includes **3 organizing principles** based on meta analyses/reviews of multidimensional representations [9-10-11-12], combined across plant and animal phyla

The GINAMO "cube"

more isolated populations

Population dispersal capacity & isolation

More continuous distribution, less Isolation-by-distance (IBD)

Ursus arctos — *Pinus pinaster*
Canis lupus
Passer domesticus
Fagus sylvatica
Pinus sylvestris

Fast-slow continuum
30 – 45% of var. explained
Fast / Short lived
Slow / Long lived

Reproductive strategies
15 – 30% of var. explained
Semelparous — Iteroparous

Generation time, Survivorship curve, Age at maturity, Development, Mean sexual reproduction

Interbirth interval, Survival, Age at first/last reproduction, Annual fecundity ...

Mean sexual reproduction, degree of iteropary, Net reproductive rate

Fecundity, Offspring mass, Juvenile survival, Interbirth interval ...

Correlated LHT traits calculated from Matrix Population Models from COMPADRE/COMADRE databases

Organisms: plants and animals

Correlated LHT traits

Organisms: animal taxa (mammals, fishes, birds)

Coordinates on axis "Fast-slow continuum" and "reproductive strategies" obtained from Paniw et al. 2018
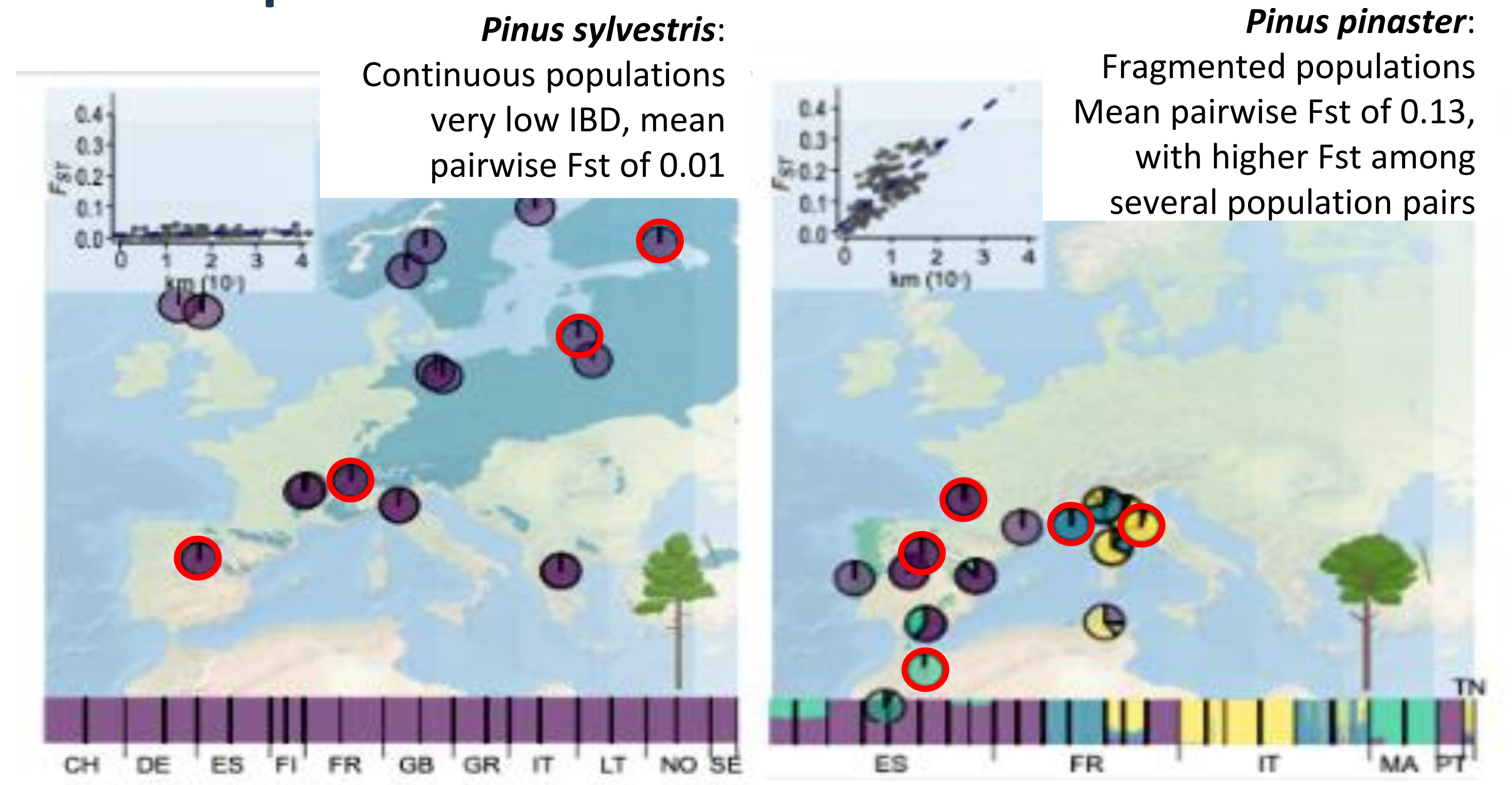
Scientific Issues

- Data from natural populations depart from LD-based method assumptions
- Several factors can bias estimation of Ne:
  - **Population structure,** demography history (*e.g.* migration rates in a metapopulation)[4-5-6]
  - **Complex life cycles** (*e.g.* overlapping generations) [7-8]
  - **Technical data features** (*e.g.* number of sampled individuals and SNPs, quality of SNPs, missing data…) [7-8]

Questions

Q1 - How do these factors affect Ne estimates in species with contrasted life-history traits and evolutionary scenarios?

Q2 - Knowing the true Ne, what are the impacts of these factors on genetic Ne estimates, as well as method assumptions and practical sampling schemes?

Preliminary results for Q1: Ne estimates in 2 contrasted forest tree species

Pinus sylvestris: Continuous populations very low IBD, mean pairwise Fst of 0.01

Pinus pinaster: Fragmented populations Mean pairwise Fst of 0.13, with higher Fst among several population pairs

Figures after Milesi et al. 2023 (BioRxiv[13]). Gene pools defined with STRUCTURE analyses.

DATA
- Putatively neutral SNPs with <50% missing data across samples at each SNP, filtered for quality & depth (> 8)
- Ne estimated with: 1000 randomly chosen SNPs per location (≃20 individuals), and for 20 individuals across the range (harmonic mean, 3 replicates)

Species	*Pinus sylvestris*					*Pinus pinaster*					
Location (country)	Spain	Swiss	Lithuania	Finland	Random inds.	Spain	France (West)	France (East)	Italy	Moroc-co	Random inds.
LDNe	7247.9 (32-Inf)	Inf (32-Inf)	Inf (32.3-Inf)	Inf (3.3-Inf)	Inf (18.3-Inf)	Inf (30.1-Inf)	Inf (28.9-Inf)	851.1 (36.2-Inf)	102.6 (25.8-Inf)	2.5 (2.3-5.2)	14.4 (5.7-265.5)
CurrentNe	13.7 (10.2-18.2)	15,5 (11.4-21.1)	15.2 (11.1-20.6)	5.3 (5-7.6)	9.7 (7.2-13)	12.2 (9.2-16.2)	9.8 (7.3-13.2)	7.4 (5.9-9.3)	15.3 (11.5-20.3)	4 (5-5)	5.9 (5.1-7.6)

LDNE (in NeEstimator[14]): singletons pruned, jackknife confidence intervals (CIs).
currentNe[15]: polymorphic SNPs, assuming random pairing. CIs determined with Artificial Neural Network

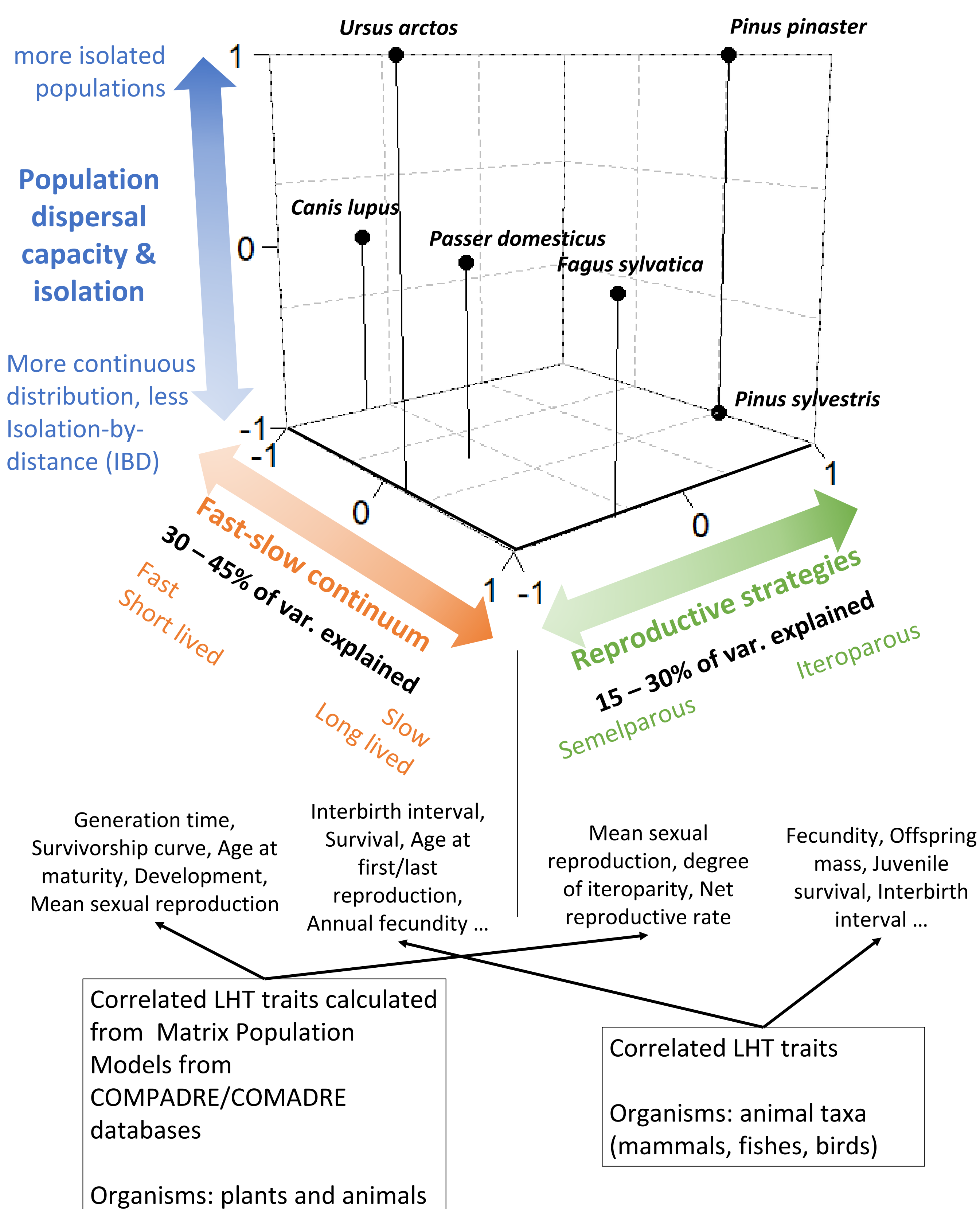First observations and **possible** interpretations:

- **Downward bias in currentNe** vs LDNe → technical issues, *e.g.,* SNP quality (testing higher depth & lower % missing data filters), dealing with pseudoreplication? [7-8]
- **Inf. values with LDNe** : genetic drift signal might be very low or not captured in small sample sizes in a continuous distribution? [7]
- **Downward bias due to localized sampling** in populations with IBD → may reflect the **neighbourhood size** [5-6]

Ongoing work and Perspectives

- **For Q1**: Choose additional legacy datasets (aim at ~25-50) that would be both representative of a large range of taxa (i.e. in the "Cube" space) & of IUCN conservation status → compare Ne estimates and how they are impacted by various factors across contrasted species
- **For Q2**: Design simulation scenarios matching case studies from legacy datasets → study interactions between factors affecting Ne estimates
- Provide workflows to biodiversity observation networks (*e.g.* GEO BON, French BON)

References [1] Hoban et al. 2020, *Biological Conservation* 248(8):108654; [2] Franklin 1980, *Conservation Biology: An Evolutionary-Ecological Perspective* 135-49; [3] Frankham et al. 2013, *Trends in Ecology & Evolution* 28 (4): 187-88; [4] Ryman et al. 2019, *Molecular Ecology* 28 (8): 1904-18; [5] Santos-del-Blanco et al. 2022, *Biological Conservation* 273 (9):109704 ; [6] Neel et al. 2013, *Heredity* 111 (3): 189-99 ; [7] Waples 2024, *Molecular Ecology Resources* 24 (1): e13879; [8] Gargiulo et al. 2024, *Evolutionary Applications* 17 (5): e13691 ; [9] Paniw et al. 2018, *Ecology Letters* 21 (2): 275-86; [10] Capdevila et al. 2020, *Functional Ecology* 34 (8): 1613-25 ; [11] Gaillard et al. 2016, *Encyclopedia of Evolutionary Biology*, 312-23; [12] De Kort et al. 2021, *Nature Communications* 12 (1): 516 ; [13] Milesi et al. 2013, BioRxiv; [14] Do et al. 2014, *Molecular Ecology Resources* 14 (1): 209-14 ; [15] Santiago et al. 2024, *Molecular Ecology Resources* 24 (1): e13890