Why biologists should care about the mathematics of diversity

- Biologists are all passionate about conserving biological diversity. We are seldom as passionate about the math we use to guide us in this task.

- Yet what if the mathematical tools we have always used to measure diversity, set conservation priorities, and monitor impacts of climate change or pollution, are systematically flawed? *What if much of biology inadvertently promotes the extinction of species and the destruction of unique ecosystems?*

-

- If so, should we not be equally passionate about reforming the way we measure diversity?

Probing our diversity concept

Mathematicians should be detectives, not dictators

Species richness $D_{rich} = S = \sum_{i=1}^{S} p_i^0$

Shannon-Wiener index $H_{Sh} = -\sum_{i=1}^{S} p_i \ln p_i$

Gini-Simpson index $H_{GS} = 1 - \sum_{i=1}^{S} p_i^2$ = Heterozygosity

Exponential Shannon-Wiener index $D_{expSh} = \exp[H_{Sh}]$

Inverse Simpson concentration $D_{invS} = 1/(\sum_{i=1}^{S} p_i^2)$ = Effective number of alleles

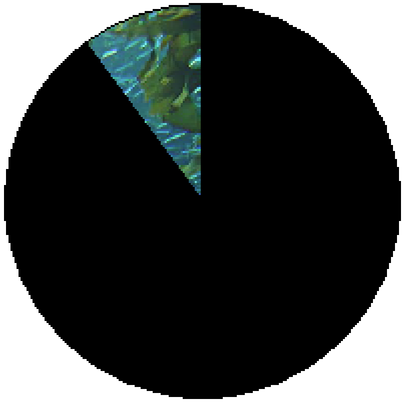
where p_i is the true population frequency of the i-th species, and i ranges from 1 to S, the total number of species present.

1000 equally common species or alleles before impact.

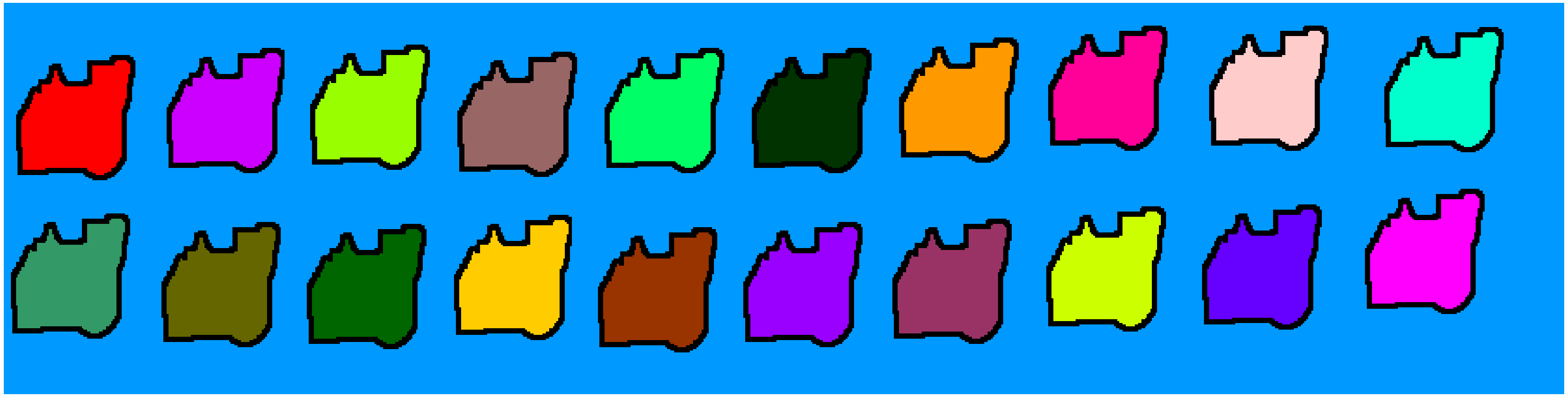
Something kills 90% of the species or removes 90% of the alleles.

Pre-impact

Post-impact



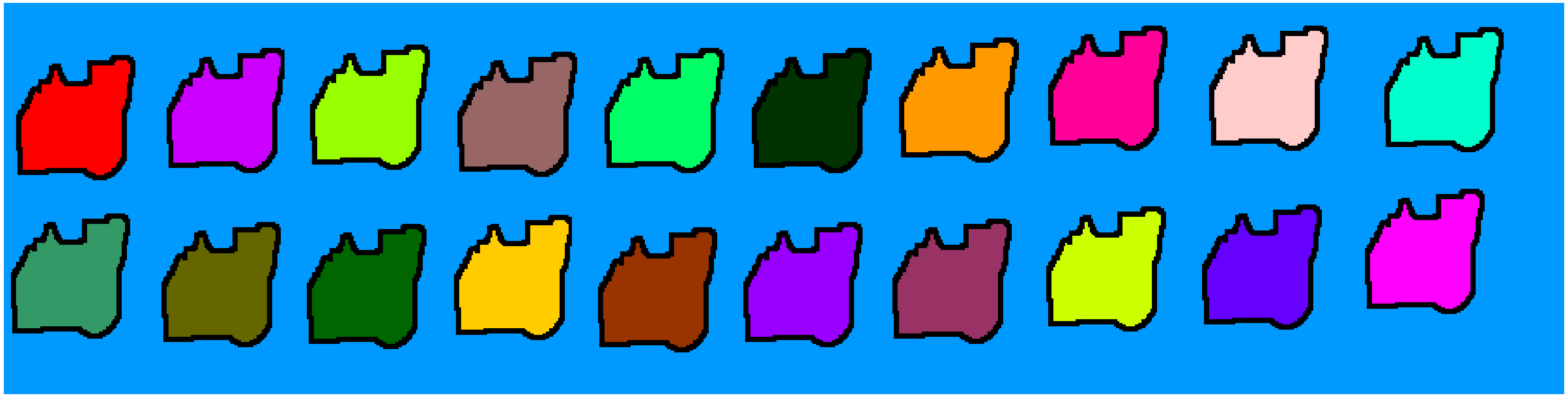
	1000 species or alleles	100 species or alleles
Species richness Allele number	1000	100 (10%)
Shannon-Wiener index Shannon entropy	6.91	4.61 (67%)
Gini-Simpson index Heterozygosity	0.999	0.99 (99.1%)
Exponential of entropy Exponential of entropy	1000	100 (10%)
Inverse Simpson concentration Effective number of alleles	1000	100 (10%)



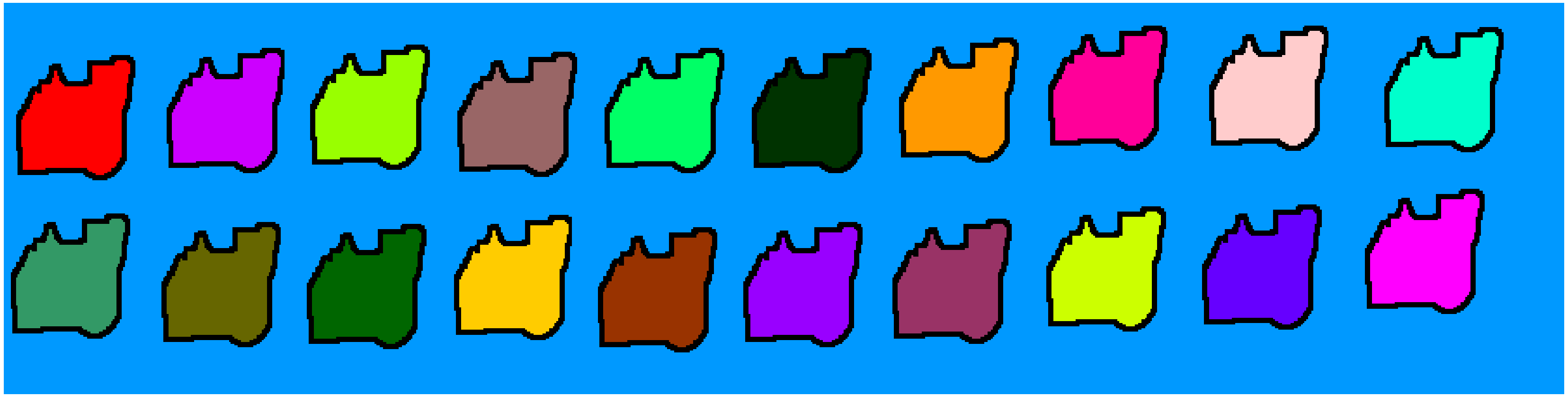
Triage

- Conservation biologists must often evaluate conservation plans or make judgments about how many sites to preserve in a given region, given limited conservation resources.

- Geneticists must often decide how many subpopulations of an endangered species need protection to preserve its genetic diversity.



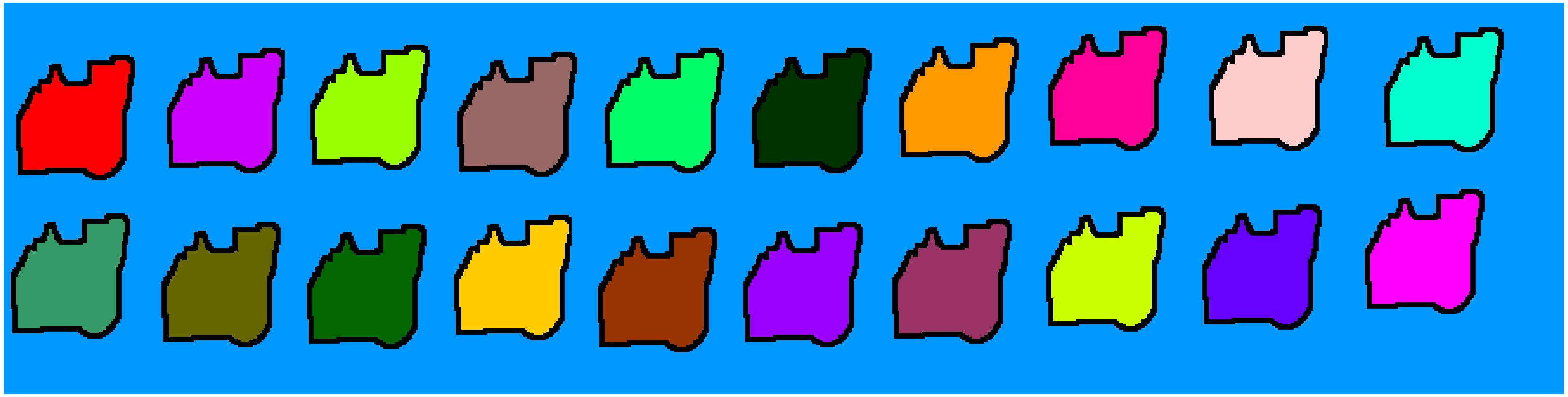
- 20 islands: Each island has the same number of individuals, the same number of species, and the same set of species frequencies. Assume that there are no shared species between islands; each island has a completely distinctive set of species.



- 20 islands: Each island has the same number of individuals, the same number of species, and the same set of species frequencies. Assume that there are no shared species between islands; each island has a completely distinctive set of species.

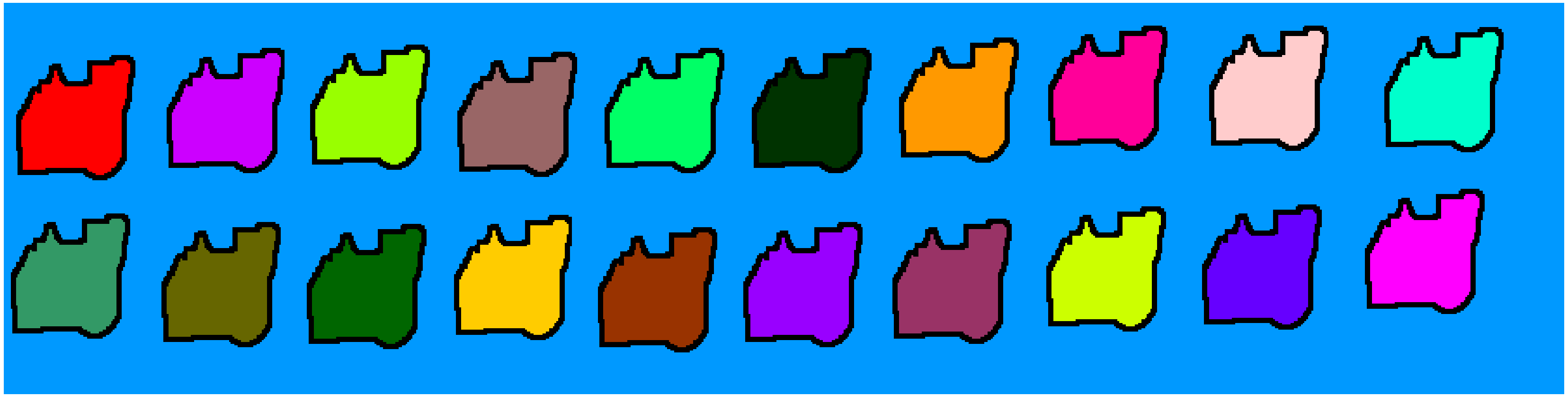
- Their diversities are all equal regardless of one’s definition of diversity.

- **How many islands must we protect in order to conserve half of the diversity of the archipelago?**



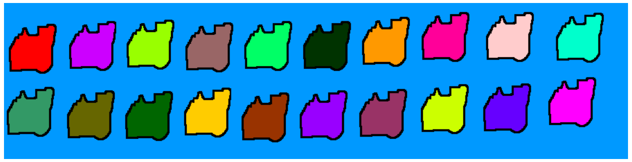
- Since each island has a completely different set of species and each is equally diverse, we should save at half of the islands (10 islands) in order to preserve at half of the diversity of the archipelago.

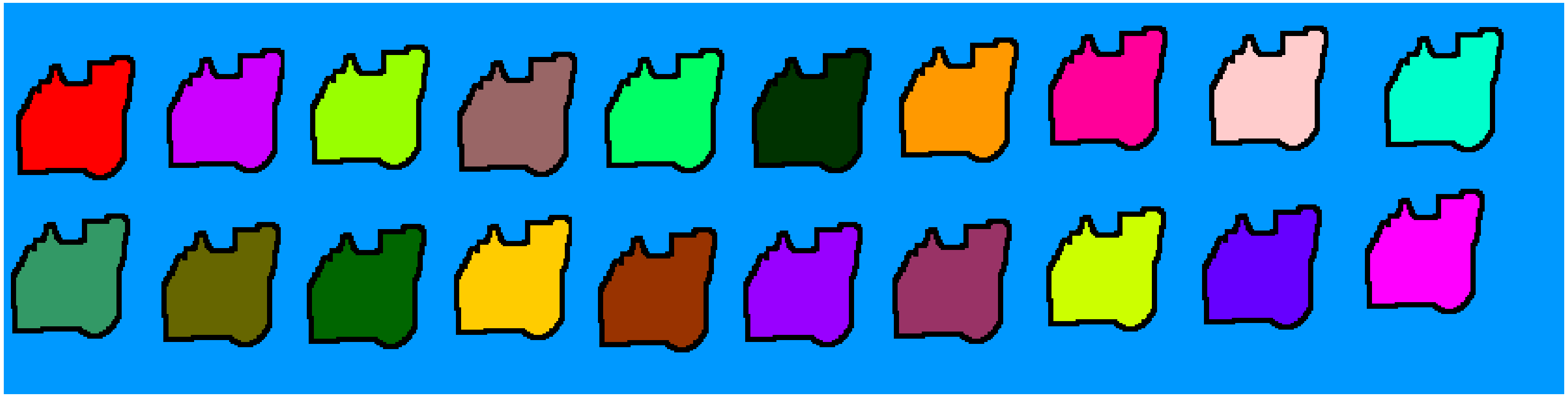
- (We assume all species are equally important)



- For definiteness suppose the number of species and the species frequencies on each island are similar to those actually observed for the trees of Barro Colorado Island, Panama (Hubbell et al. 2005).

	One island or deme	Pooled islands or demes	Half of pooled diversity	Number of islands or demes needed to preserve half of pooled diversity
Species richness Allele number	300	6000	3000	10
Shannon-Wiener index Shannon entropy	3.96	6.96	3.48	1 (saves 57% of total “diversity”!) Also sacrifices 99% of the “diversity”
Gini-Simpson index Heterozygosity	0.951	0.998	0.499	1 (saves 95% of total “diversity”!) Also sacrifices 99% of the “diversity”!
Exponential of entropy Exponential of entropy	52.3	1046	523	10
Inverse Simpson concentration Effective number of alleles	20.3	406	203	10

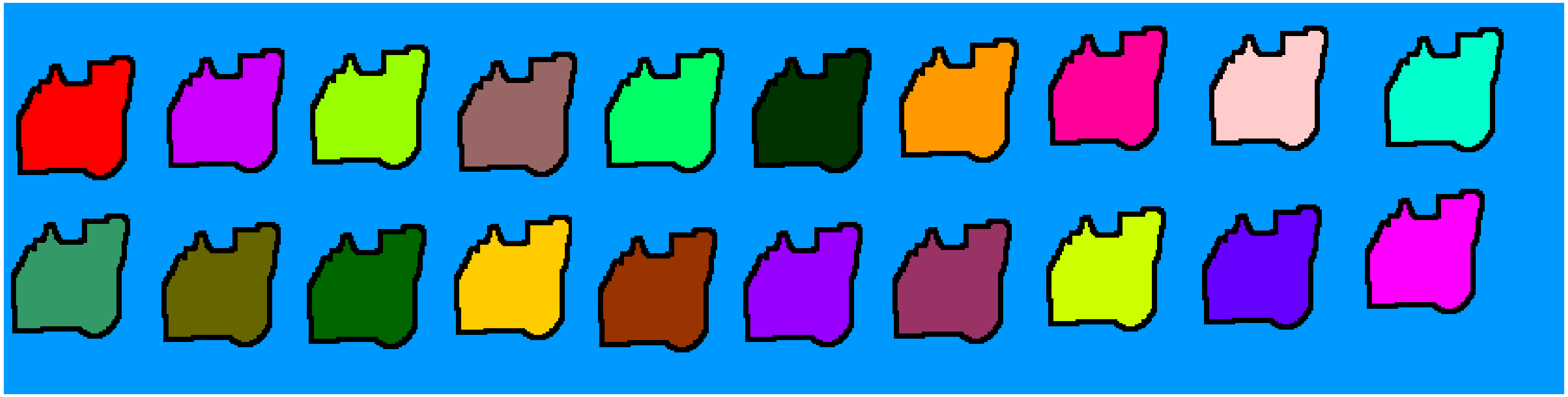




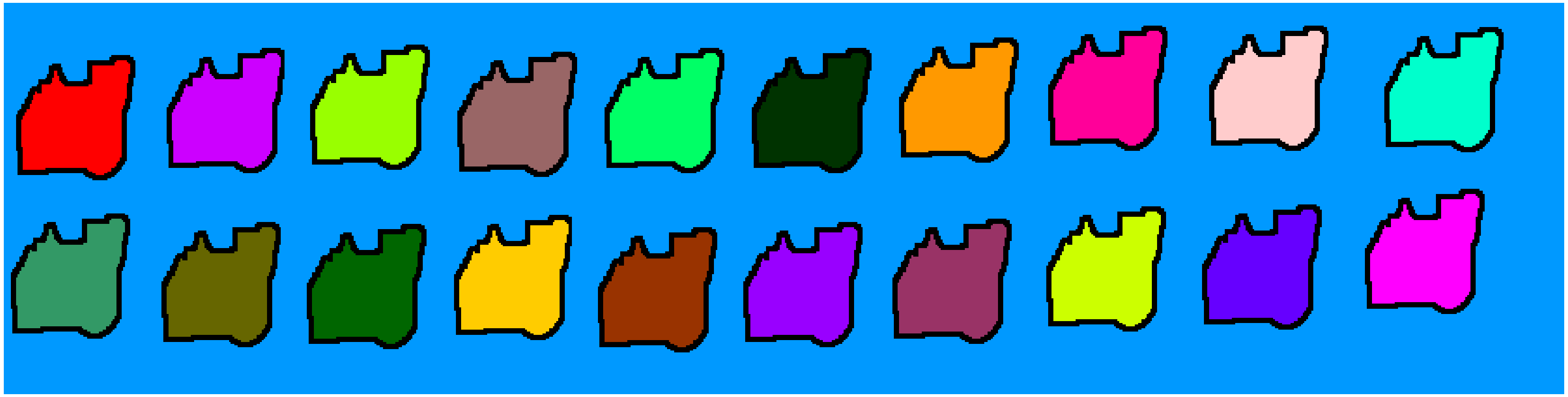
- Shannon entropy and the Gini-Simpson index give bad, **self-contradictory** conservation advice.

	“Diversity” saved	“Diversity” lost
Shannon entropy	57% (1 island)	99% (19 islands)
Gini-Simpson	95% (1 island)	99% (19 islands)
Effective number of species	5% (1 island)	95% (19 islands)
Effective number of species	50% (10 islands)	50% (10 islands)

- Why do some measures give self-consistent reasonable answers in this symmetric case?



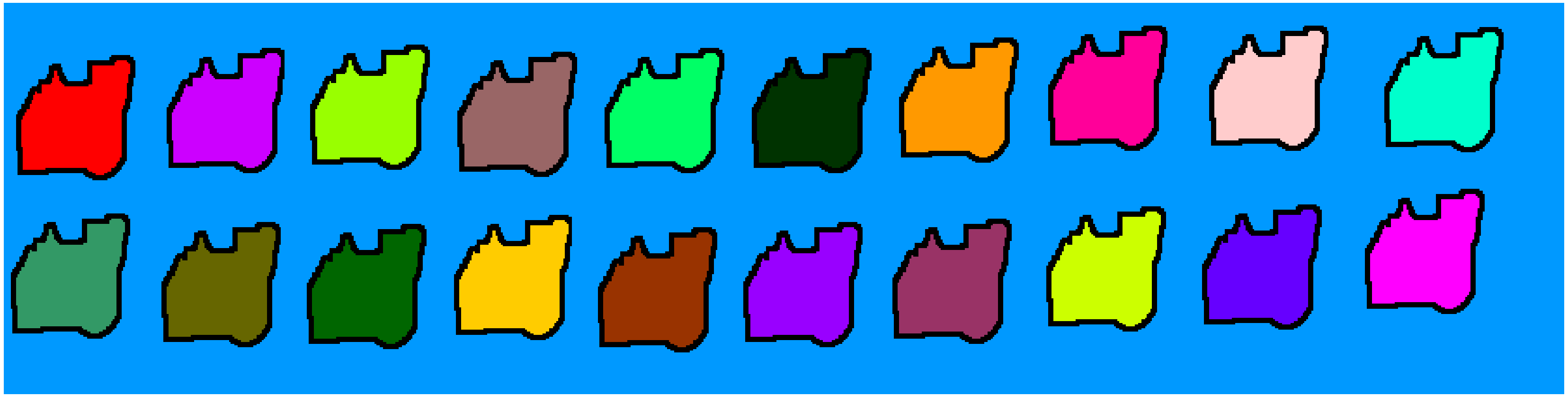
- Why do some measures give the reasonable answer?
- Each island must contribute equally to total diversity.
- For N completely distinct, equally large islands of equal diversity, pooled diversity must equal N* individual diversity.
- **This strengthened "replication principle" from economics is the requirement for diversity to be self-consistent in these kinds of inferences.**



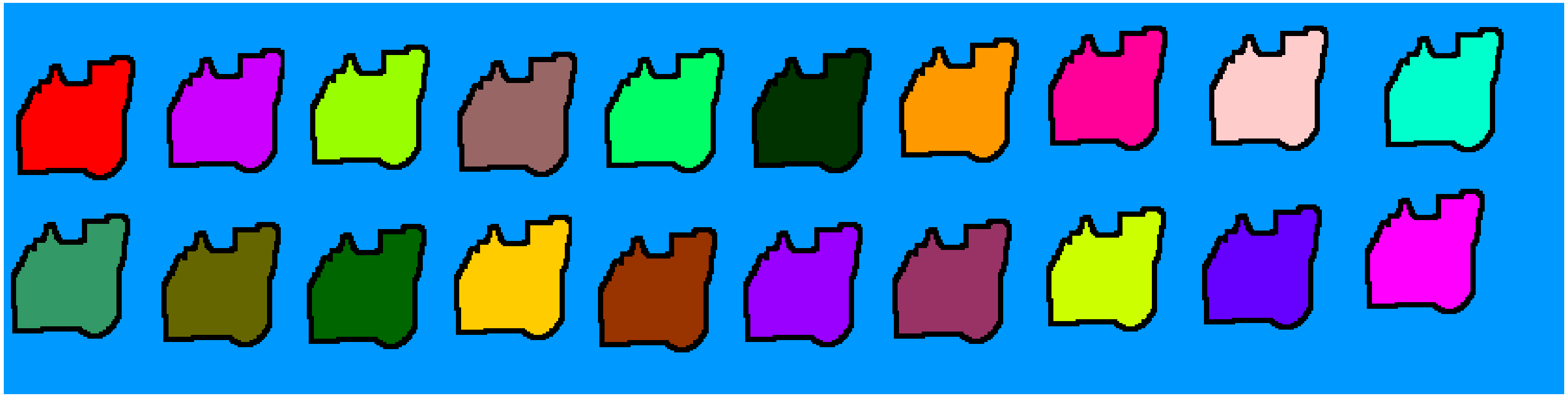
- **This property is implicit in our intuitive concept of diversity, and our rules of inference presuppose this property.**

- Shannon entropy and the Gini-Simpson index (heterozygosity) do not have this property.

- Hill numbers, like species richness (allele number), exponential of Shannon entropy, and inverse Simpson concentration (effective number of alleles) do have it.



- **Fundamental question: How is diversity related to compositional similarity and differentiation?**

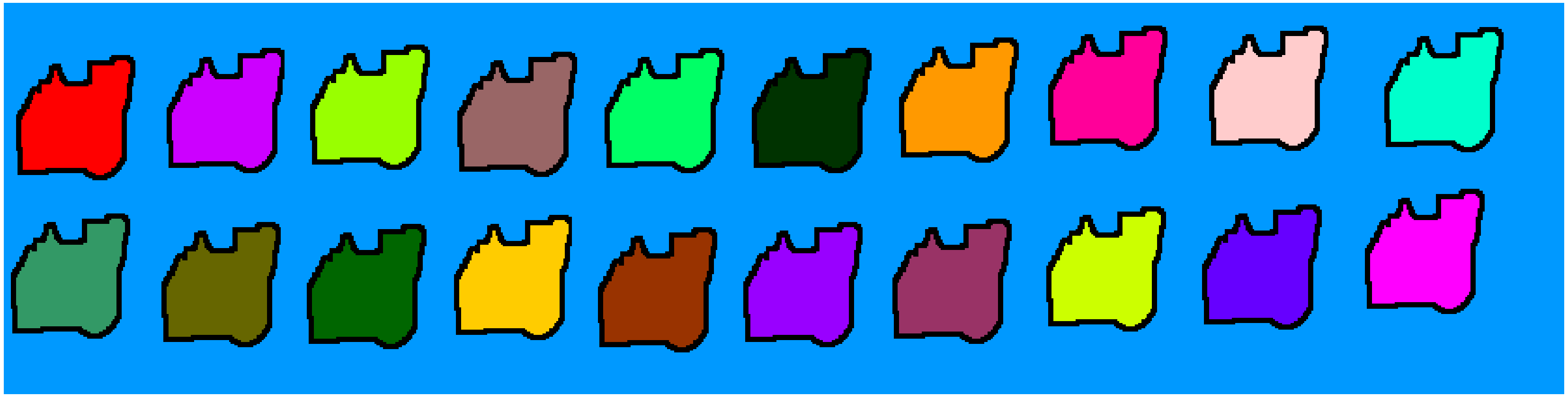


- **Compositional similarity between communities or demes**

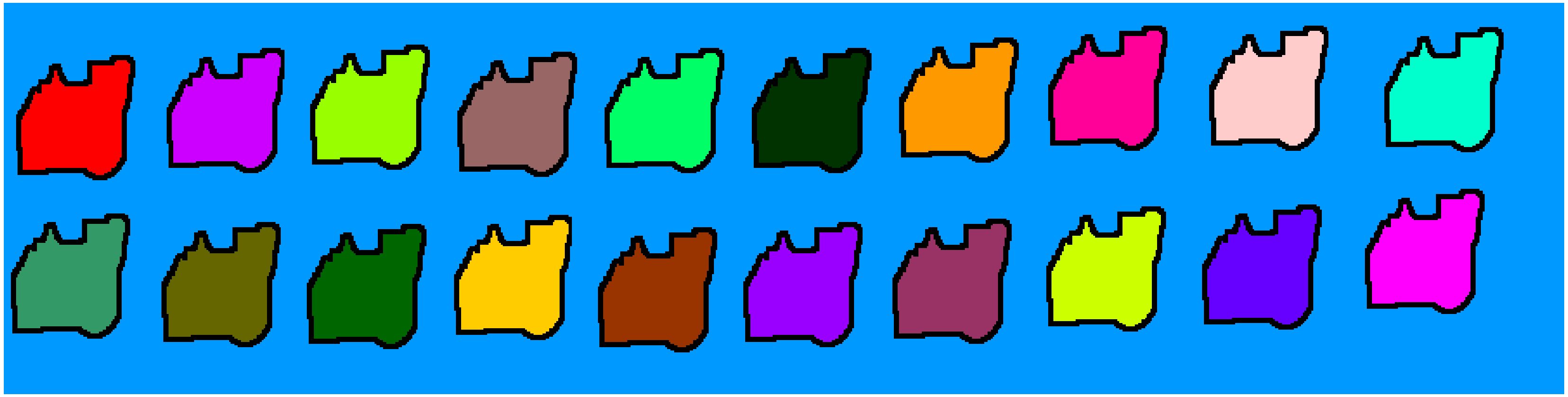
- Ecologists (Lande 1996) and geneticists (e g Lewontin 1972) often use the ratio of within-group "diversity" Hs to total pooled "diversity" Ht to measure the compositional similarity of the groups.

- Hs/Ht or its complement, (Ht-Hs)/Ht.

- The latter may be interpreted as the proportion of total "diversity" contained in the average single community or deme. The complement of this similarity is G_{ST}, the principle measure of differentiation in population genetics.



- Common phrase: “95% of the diversity is within groups”
- Usual conclusion: differences between groups are small; groups are similar in composition.
- This conclusion is often false when standard diversity measures are used.

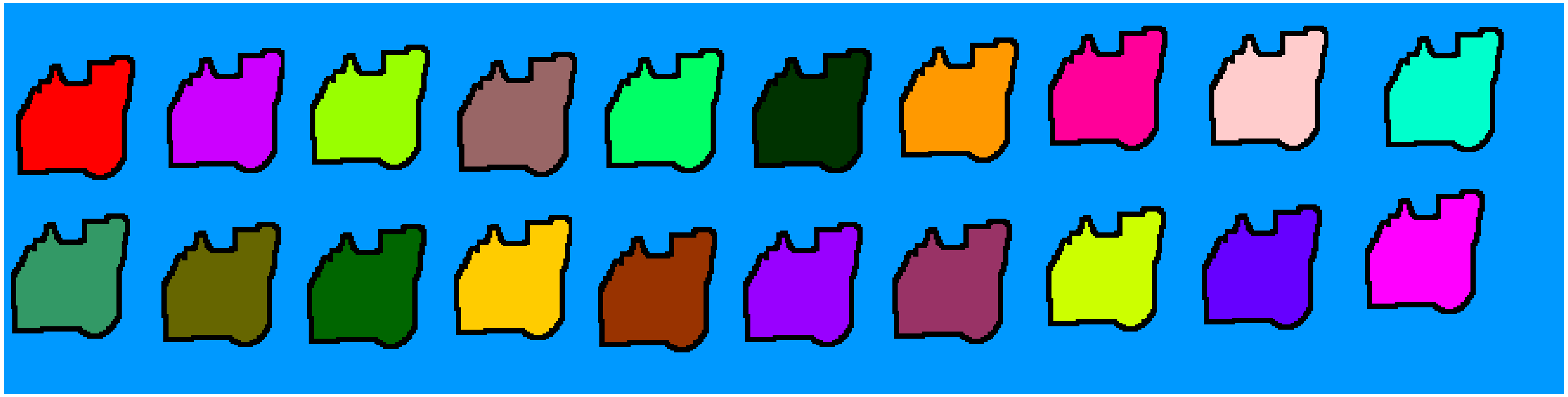


Compositional similarity between communities or demes Hs/Ht

- Species richness (allele number) gives the reasonable answer, 0.05 = 1/20, the smallest possible value for 20 equally large communities. This shows correctly that the islands are completely dissimilar.

- Shannon entropy gives a similarity ratio is 0.57, wrongly suggesting considerable similarity.

- The Gini-Simpson index or heterozygosity gives a within-group/total diversity of 0.95, close to its maximum of 1.00, suggesting that these islands or demes -- which have nothing in common – do not differ much in composition!!

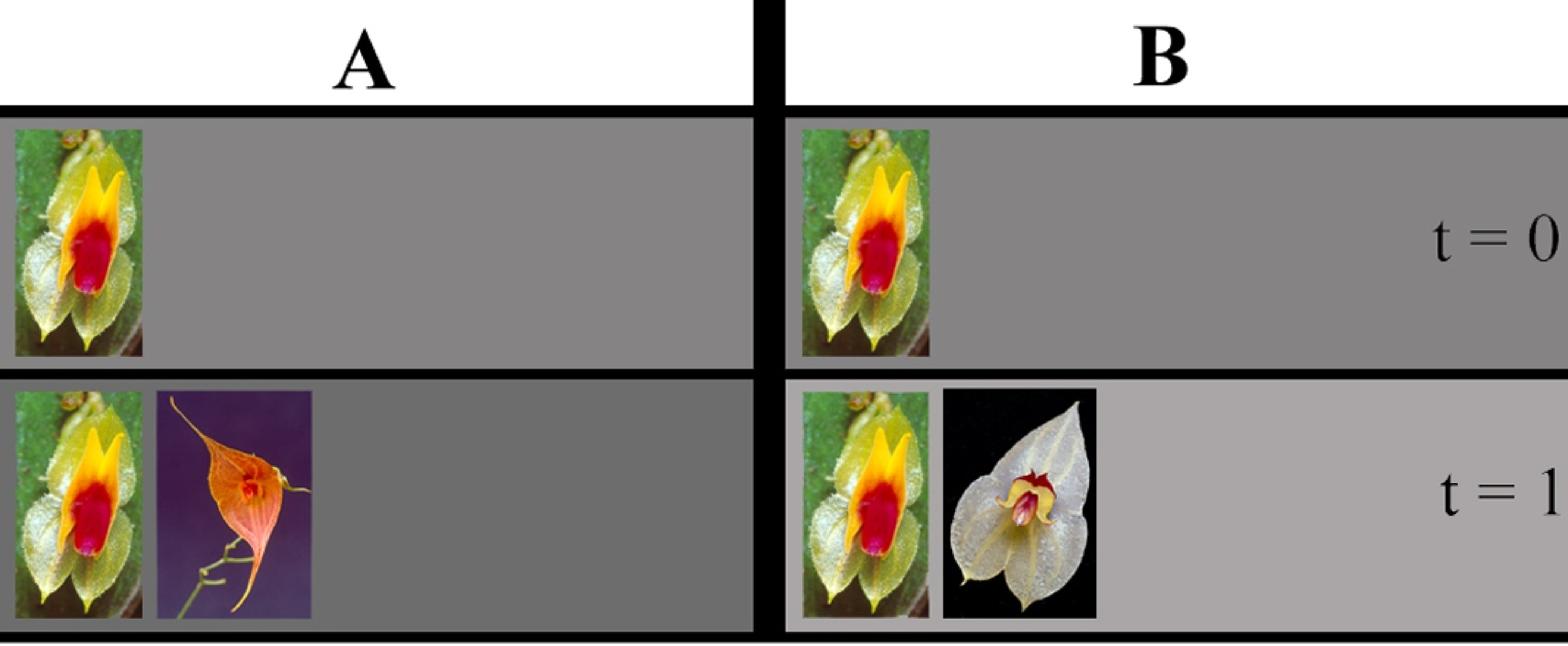


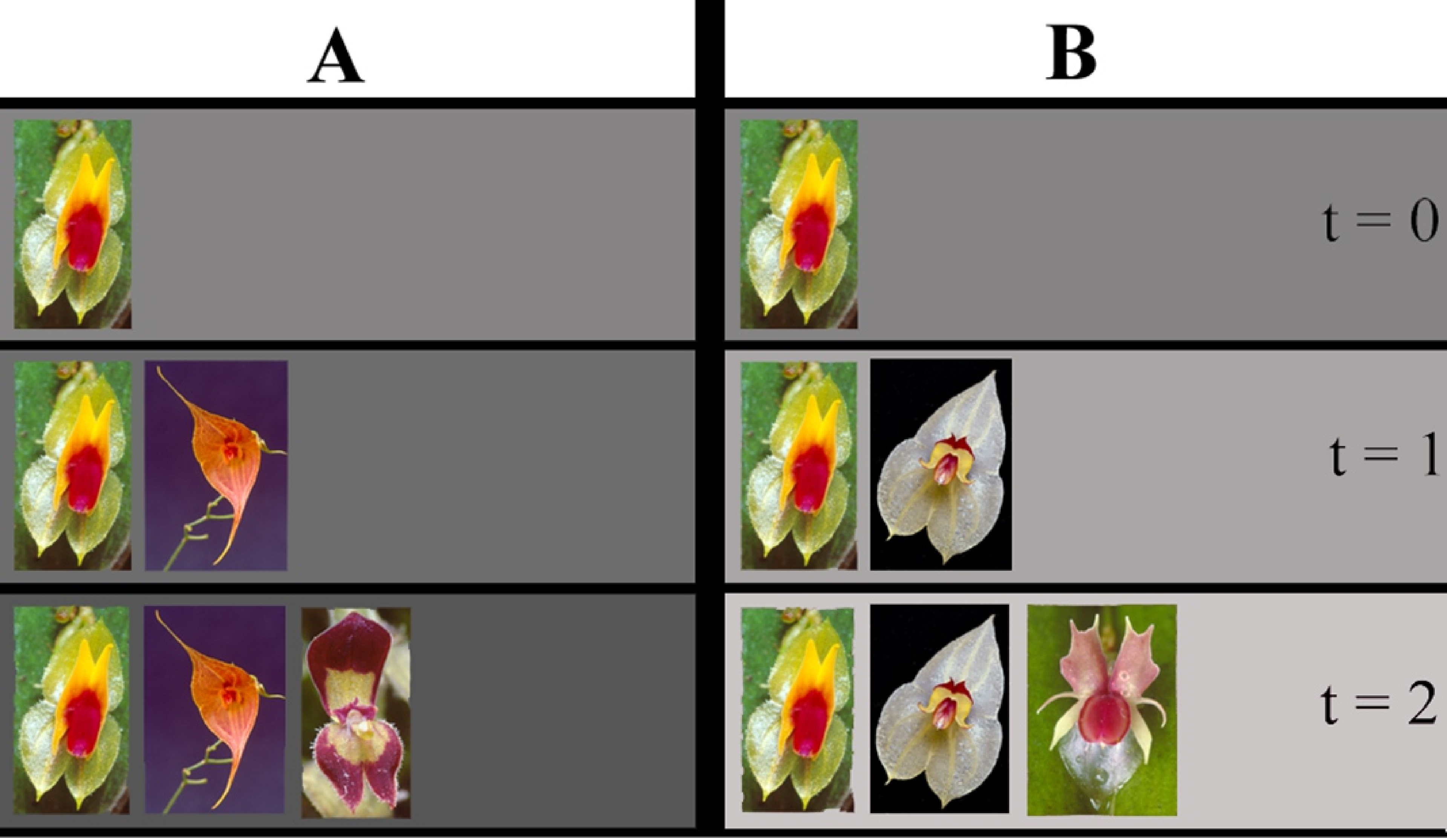
- **Compositional similarity between communities or demes**

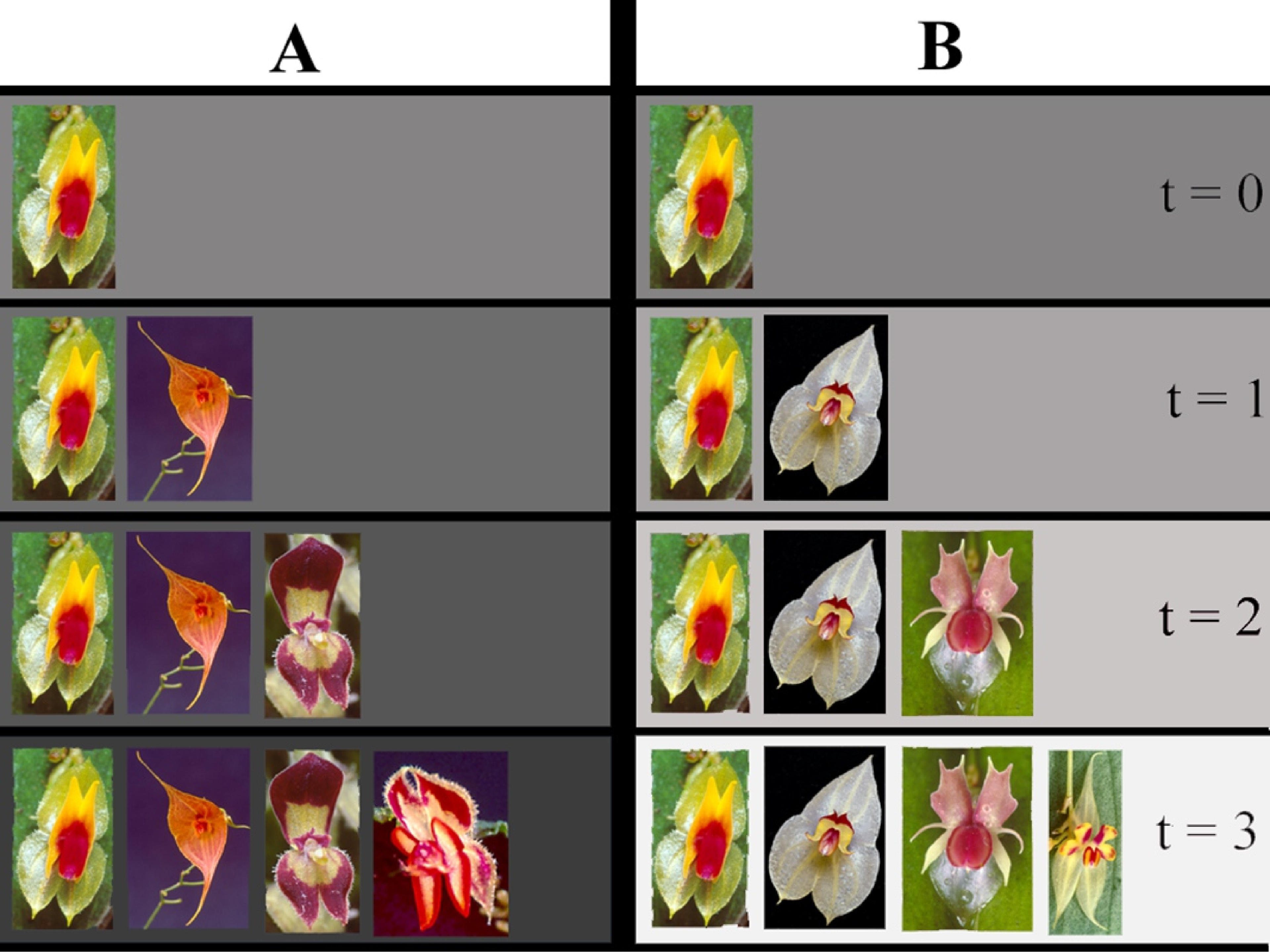
- Exponential of Shannon entropy and the inverse Simpson concentration (effective number of alleles) gives the same reasonable answer as species richness (allele number): 0.05.

- They give reasonable answers because these measures obey the replication principle. Gamma = 20*alpha. Their ratio alpha/gamma (within-group diversity / total diversity) is therefore replication-invariant and equals 1/20. *We could make the alpha diversities anything we wanted, and this number would not change*.

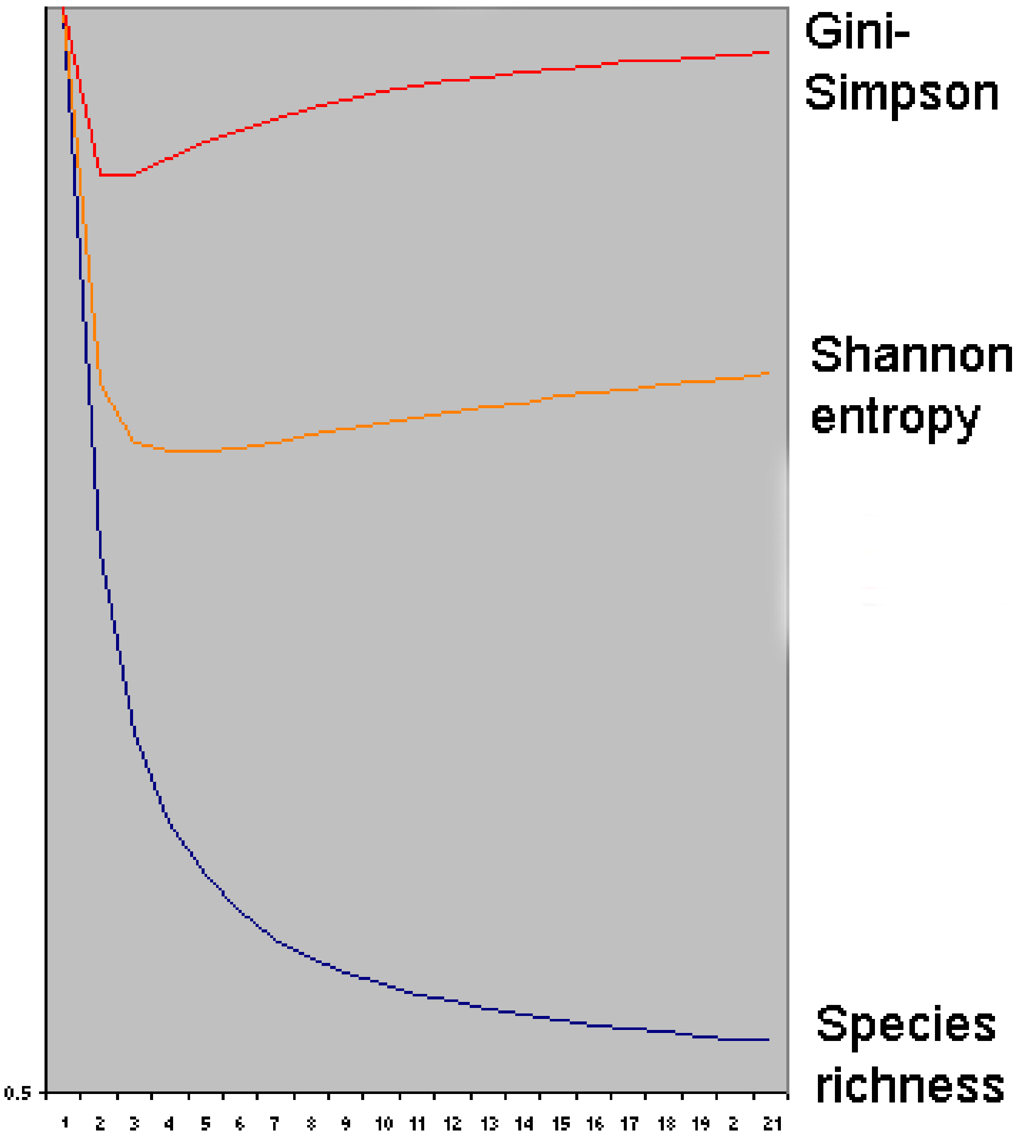
A
B
t = 0


A
B
t = 0
t = 1

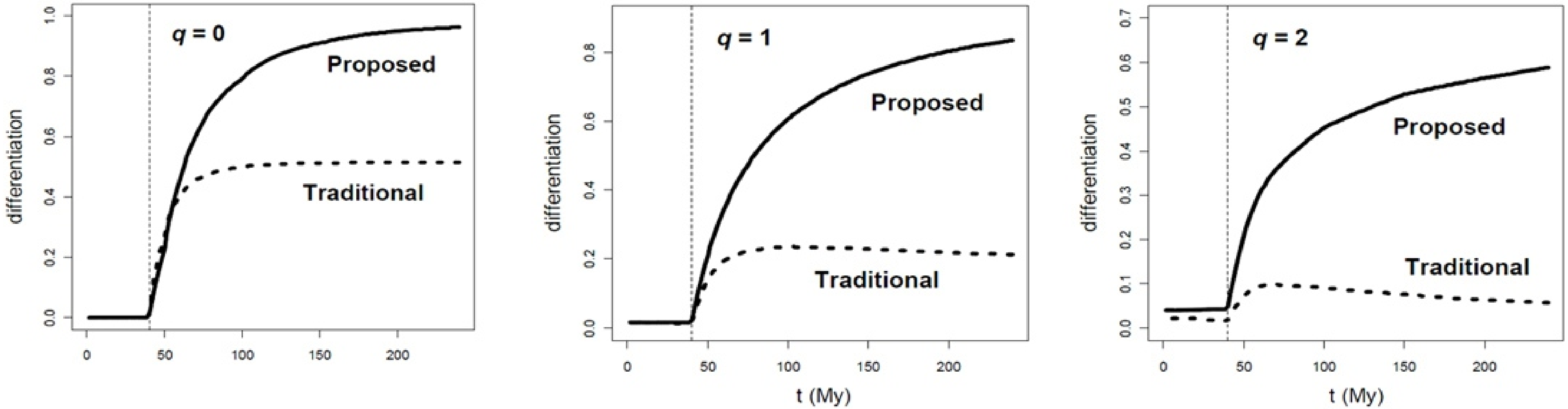

A
B
t = 0
t = 1
t = 2


A
B
t = 0
t = 1
t = 2
t = 3

“Similarity” alpha/gamma



Relative differentiation for phylogenetic entropy and diversity



"...phylogenetic additive "beta" (i.e., Qg - Qa based on Rao's quadratic entropy Q) is confounded with alpha through the constraint Qg-Qa<=(1-1/N)(T-Qa). A high alpha quadratic entropy means that Qa → T. In this case the right side of the inequality tends to 0, so the additive "beta" Qg-Qa is thus bound to be small, even if the assemblages involved share no lineages whatsoever. This implies that the "differentiation" measure $J_2(T)$ is necessarily close to 0, and the "similarity" measure $1-J_2(T)$ is close to its maximum value of unity, when Qa is high (close to T), for any set of assemblages, even assemblages that share no species or lineages."

----Chiu, Chao, Jost, Phylogenetic beta diversity, similarity, and differentiation measures based on Hill numbers (under review)

See also Ricotta and Szeidl (2009), de Bello et al (2010), Hardy and Jost (2008)

Example: Tropical butterfly canopy and understory communities

- Similarity (alpha/gamma) of canopy and understory communities in a series of long-term trapping studies in various tropical countries: 97%, 95%, 97% .

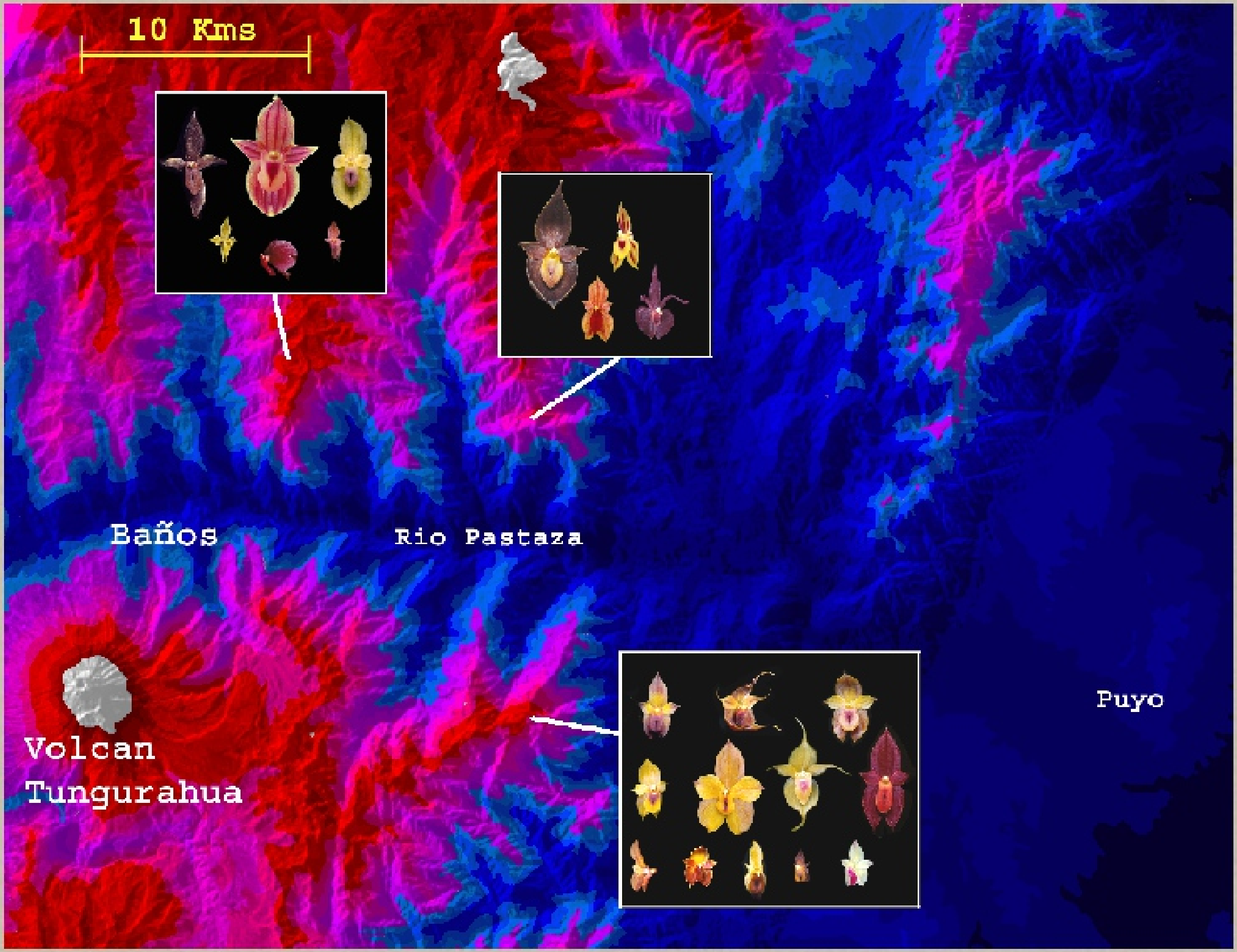
Example: Tropical butterfly canopy and understory communities

- Similarity (alpha/gamma) of canopy and understory communities in a series of long-term trapping studies in various tropical countries: 97%, 95%, 97% .

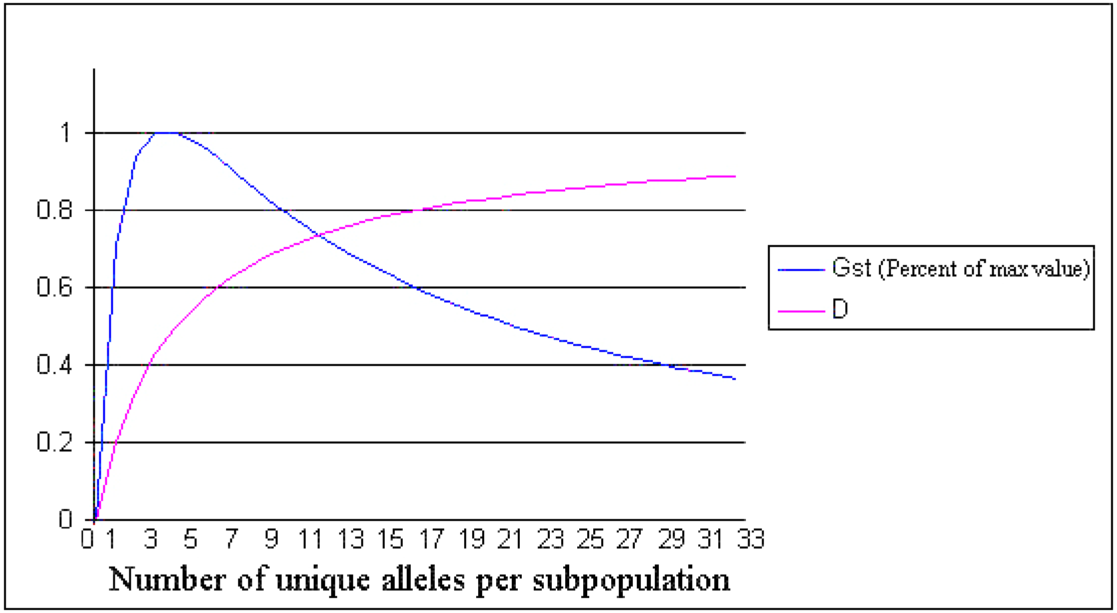
Example: Tropical butterfly canopy and understory communities

- Similarity (alpha/gamma) of canopy and understory communities in a series of long-term trapping studies in various tropical countries: 97%, 95%, 97% .

	Canopy	Understory
Historis acheronta	100	1
Bia actorion	0	98
Morpho achilles	1	89
Catonephele orites	45	31
Historis odius	58	3
Taygetis sp-1	1	54
Nessea obrina	1	47
Cf Euptychia sp.	0	36
Panacea duvalis	30	4
Nessea hewitsonii	1	32


10 Kms
Baños
Rio Pastaza
Volcan
Tungurahua
Puyo

Behavior of G_{ST} as differentiation increases.



Behavior of G_{ST} as differentiation increases. We start with two identical subpopulations (four equally common alleles, 10000 individuals per allele per subpopulation). We then successively add unique alleles to each subpopulation (10000 individuals per allele) and graph G_{ST} and D (the measure of differentiation defined later). Even though real differentiation increases steadily from left to right, G_{ST} reaches a maximum (0.0345) and then falls back to zero. G_{ST} is calculated from exact population allele frequencies, so this is not a sampling issue.

	Species A		**Species B**		**Species C**	
Allele:	**Subpop. 1**	**Subpop. 2**	**Subpop. 1**	**Subpop. 2**	**Subpop. 1**	**Subpop. 2**
1	0.5	0.5	0.15	0.8	0.095	0
2	0.5	0.5	0.85	0.2	0.08	0
3	0	0	0	0	0.11	0
4	0	0	0	0	0.08	0
5	0	0	0	0	0.095	0
6	0	0	0	0	0.06	0
7	0	0	0	0	0.07	0
8	0	0	0	0	0.096	0
9	0	0	0	0	0.094	0
10	0	0	0	0	0.08	0
11	0	0	0	0	0.03	0
12	0	0	0	0	0.06	0
13	0	0	0	0	0.05	0
14	0	0	0	0	0	0.15
15	0	0	0	0	0	0.16
16	0	0	0	0	0	0.12
17	0	0	0	0	0	0.13
18	0	0	0	0	0	0.17
19	0	0	0	0	0	0.14
20	0	0	0	0	0	0.13

Measures of differentiation; should increase with increasing differentiation:

	Species A:	Species B:	Species C:
D_{ST}	0	0.24	0.06(!)
G_{ST}	0	0.49	0.06(!)
D	0	0.66	1.00

Measures of similarity; should decrease with increasing differentiation:

	Species A:	Species B:	Species C:
H_S/H_T	1.00	0.51	0.94(!)

- Genetic rules of thumb:

- G_{ST} = 0-0.05: little differentiation
- G_{ST} = 0.05-0.15: moderate differentiation

- These rules of thumb are not true, since G_{ST} can be arbitrarily close to zero even for completely differentiated subpopulations.

- F_{ST} and G_{ST} values are constantly interpreted and compared without reference to the value of H_S (though see Hedrick 2005)

- Low values of F_{ST} and G_{ST} are still often interpreted as indicating low differentiation, even though this is easily shown to be false.

- **A value of G_{ST} = 0.001 could mean subpopulations are almost identical, completely differentiated, or anything in between.**

Empirical evidence

- Microsatellites in two subspecies of the fish *Tripterygion delaisi* that were distinguishable behaviourally, geographically, and via mtDNA (Carreras-Carbonell et al. 2006)
- The authors were puzzled that F_{ST} was very low between subspecies, even at loci with no shared alleles between subspecies.
- Like many other authors, they observed that “estimates of F_{ST} seem to decline with increasing polymorphism”.
- The correct treatment shows that the between-subspecies differentiation at their most diverse locus (Td06) is not 0.05 (the value of F_{ST}) but 1.00.

Surprise:

- We are not really free to choose our diversity measures.

- If we are going to use these rules of inference, we have to use measures that obey the replication principle!

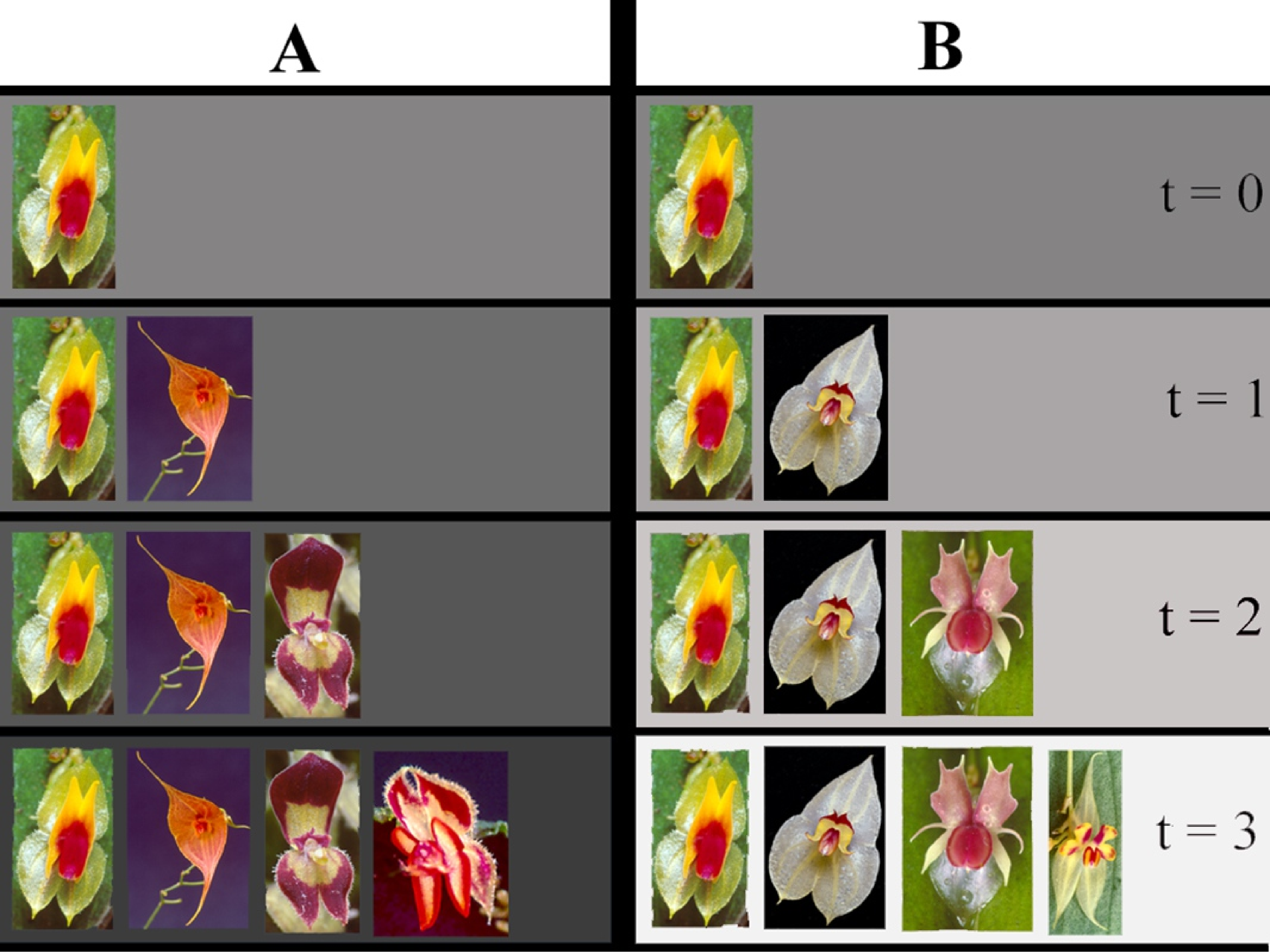
How to partition diversity into within- and between-group components?

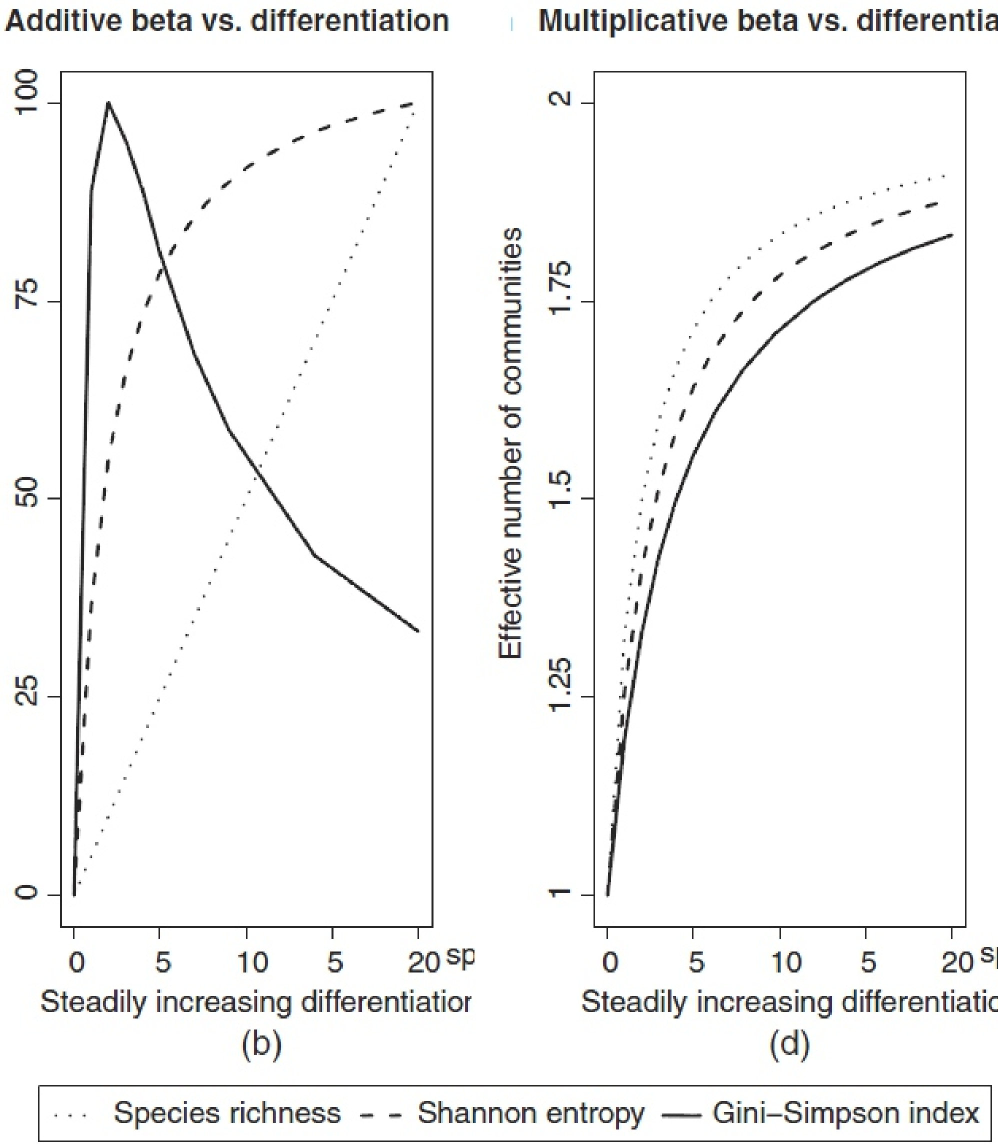
- A method commonly used in ecology and universally used in genetics is “additive partitioning” (Nei 1973, Lande 1996).

- Total (gamma) diversity = mean within-group diversity (alpha) + between-group (beta) diversity
- gamma = alpha + beta

- These measures are used to set conservation priorities and understand evolutionary processes.

- Additive partitioning of the Gini-Simpson index (heterozygosity) produces a between-group component that is confounded with the within-group component. This is fatal.


A
B
t = 0
t = 1
t = 2
t = 3



Jost, Chao, DeVries, Walla, Greeney, Ricotta (2010) in Diversity and Distributions

Modified from Summerville et al (2003) in Conservation Biology

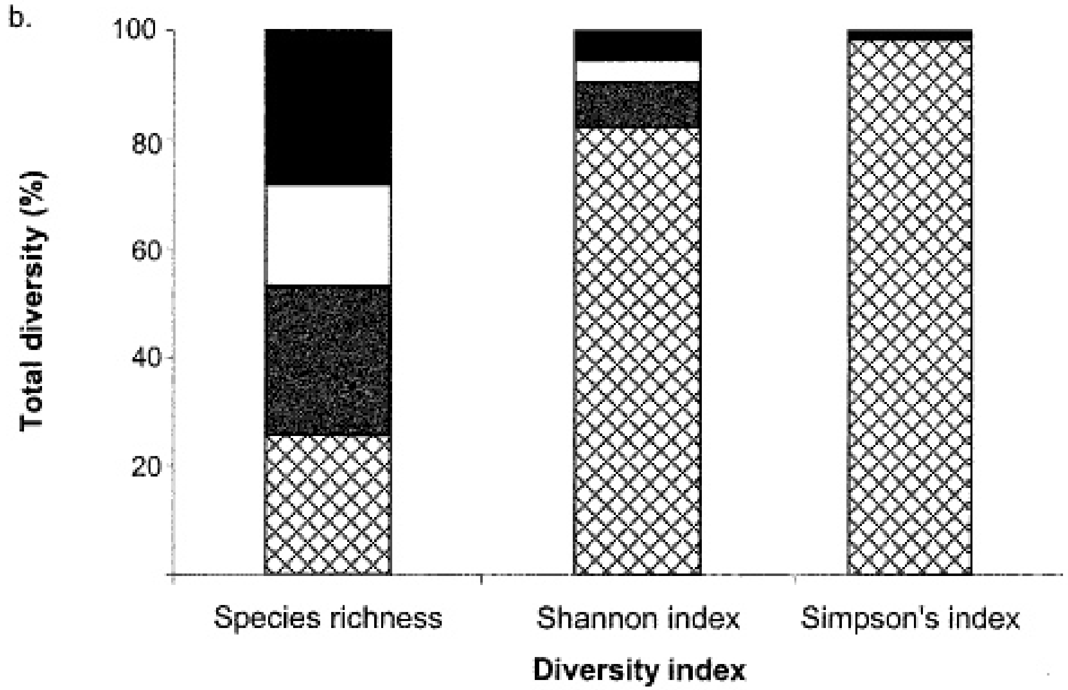


Figure 3. Percentage of total moth species richness, Simpson diversity, and Shannon diversity explained by α and β components of regional diversity: within and among forest stands (α_1 and β_1), among sites (β_2), and between ecoregions (β_3).

Partitioning diversity

- Partitioning into within- and between-group components should be complete: the within-group component should contain no information about the between-group component, and vice versa.

- The two components will then be independent: knowledge of one component (and only that component—I exclude other empirical knowledge) should give no information, nor impose any mathematical constraint, on the value of the other component.

- This is “mathematical independence”, similar to orthogonality, like the x and y components of vectors. A high value of one component should not force the other component to be either high or low.

- In particular situations in the real world, of course, within- and between-group diversity can be correlated or anticorrelated. We want these correlations to reflect only the behavior of the real world. We don’t want them to be artefacts of our mathematics.

Partitioning diversity

Shannon entropy: $H_p(p_1, p_2, ..) + H_q(q_1, q_2, …) = H_{pq}(p_1, p_2,…, q_1, q_2,…)$ for two unrelated probability distributions **p**, **q**.

The numbers equivalent (effective number of species) of Shannon entropy is obtained by taking the exponential. Doing that to both sides of the above, we get $\exp(H_p)*\exp(H_q) = \exp(H_{pq})$

Now suppose we didn’t know anything about q or H_q, but we could measure H_p and H_{pq}. We could still figure out the magnitude of H_q and $\exp(H_q)$.

Now suppose we have H_a and H_g, with $H_g>=H_a$. H_g includes spatial structure, H_a does not. By including spatial structure the entropy increases by H_g-H_a. It is as if we had added a new, unrelated probability distribution to H_a, with entropy H_g-H_a. We call this entropy the beta component of entropy. Convert to diversity by taking its exponential. So $D_aD_b = D_g$.

Partitioning diversity

Tsallis entropy for q=2: $H_p + H_q - H_p H_q = H_{pq}(p_1, p_2,\ldots, q_1, q_2,\ldots)$ for two unrelated probability distributions **p**, **q**. Suppose $H_{pq} >= H_p$.

The numbers equivalent (effective number of species) is obtained by taking 1/(1-H). Doing that to both sides of the above, we get $1/(H_p + H_q - H_p H_q) = 1/(1-H_{pq})$ which factors into $[1/(1-H_p)][1/(1- H_q)] = 1/(1-H_{pq})$ or $D_p D_q = D_{pq}$ exactly as in the Shannon case.

We can figure out the magnitude of H_q and D_q from either of these equations.

Now suppose we have H_a and H_g. H_g includes spatial structure, H_a does not. It is as if we had added a new, unrelated probability distribution with entropy H_b. Convert to diversity by taking $1/(1-H_b)$. So if gamma diversity consists of two independent components using this metric, then $D_a D_b = D_g$.

Renyi entropies give same result.

Partitioning diversity

$${}^{q}D_{\alpha}\,{}^{q}D_{\beta} = {}^{q}D_{\gamma}$$

Species richness: $H_\beta = H_\gamma / H_\alpha$

Shannon entropy: $H_\beta = H_\gamma - H_\alpha$

Exponential of Shannon entropy: $H_\beta = H_\gamma / H_\alpha$

Gini-Simpson index: $H_\beta = (H_\gamma - H_\alpha)/(1 - H_\alpha)$.

Simpson concentration: $H_\beta = H_\gamma / H_\alpha$

HCDT entropies: $H_\beta = (H_\gamma - H_\alpha)/(1 - (q-1)(H_\alpha))$

Renyi entropies: $H_\beta = H_\gamma - H_\alpha$

Partitioning diversity

- For N equally-weighted communities, the between-group component ranges from 1 to N and is the effective number of completely distinct communities. This is the normal case in population genetics.

- When communities are not equally weighted, only Shannon measures (q=1) can be decomposed into independent components that are monotonic with respect to increasing differentiation.

- The alpha diversity is not the average diversity of the individual communities or populations. It is the numbers equivalent of the average frequency moments of the individual communities:

$$
\begin{aligned}
{}^qD_\alpha &\equiv D({}^q\lambda_\alpha) \\
&= \left\{\left(\frac{1}{N}\right)\left[\left(\sum_{i=1}^{S} p_{i1}^q\right) + \left(\sum_{i=1}^{S} p_{i2}^q\right) + \cdots + \left(\sum_{i=1}^{S} p_{iN}^q\right)\right]\right\}^{1/(1-q)}
\end{aligned}
$$

Compositional similarity and differentiation

- Between-group diversity (beta diversity) ranges from 1 to N. It can be transformed onto the unit interval to give measures of relative similarity or differentiation.

- There are several families of such measures, each parameterized by q. We can make overlap measures, measures of shared diversity, measures of community turnover, etc.

- These generate many of ecology’s best-known similarity measures: Jaccard, Sorensen, Horn, Morisita-Horn indices.

- Something special happens when q=2. For this value of q, the measure of community overlap and the measure of shared diversity are the same. This is the Morisita-Horn index.

Measure of differentiation to replace G_{ST} or F_{ST}

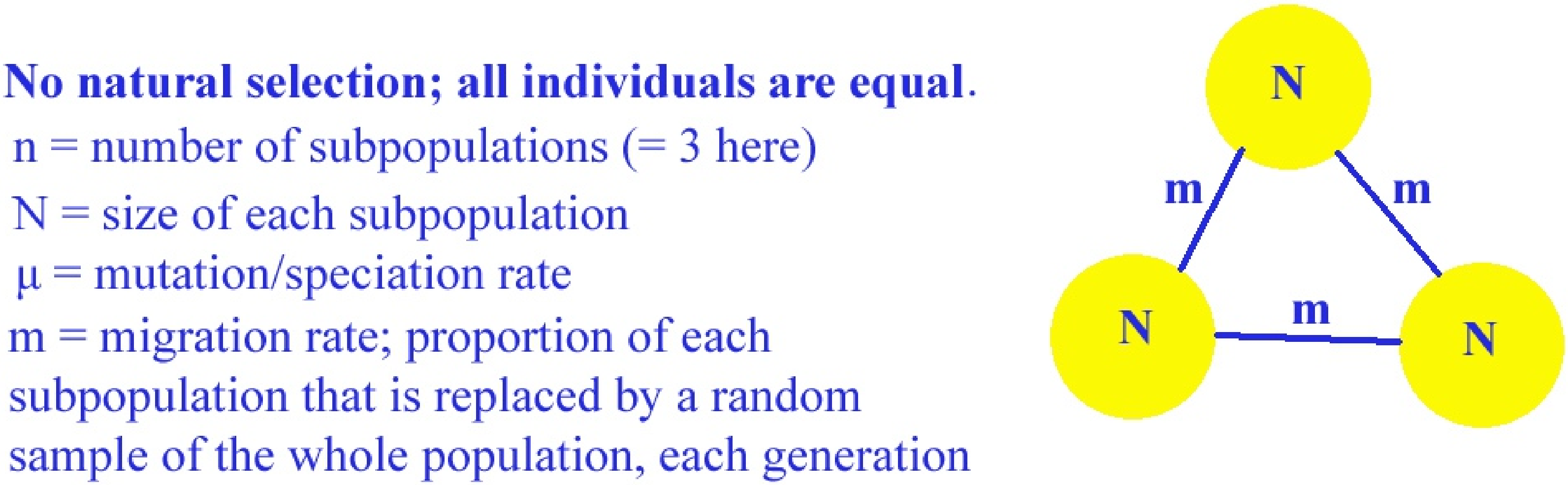
- The complement of the Morisita-Horn index (q=2) is a measure of dissimilarity for genetics:

$$[(H_T - H_S)/(1 - H_S)]\ [n/(n-1)]$$

- If all n subpopulations consist of k equally common alleles, this measure gives the proportion of each subpopulation's alleles that are unique to that subpopulation.
- This is a measure of pure differentiation, independent of average within-subpopulation heterozygosity.
- Linear in shared diversity.
- It should replace G_{ST} when differentiation is the quantity of interest.

Linking diversity to ecological and genetic models

- Hubbell's neutral model of biodiversity
- Finite island model in population genetics
- Kind of like the ideal gas or the two-body problem of physics; they are simple enough that we can solve them analytically.



No natural selection; all individuals are equal.

n = number of subpopulations (= 3 here)

N = size of each subpopulation

μ = mutation/speciation rate

m = migration rate; proportion of each subpopulation that is replaced by a random sample of the whole population, each generation

Speciation

- Real differentiation D at equilibrium is:

$$D \approx (n+1)\ \mu/m$$

These factors control neutral speciation in subdivided populations.

Very different from the standard view!

- Traditional G_{ST} ("the extent of differentiation") at equilibrium under the same model is:

$$G_{ST} \approx 1/(1+ 4Nm)$$

This equation is irrelevant for describing differentiation and neutral speciation.

n = # of subpopulations, N= size of subpopulations, m = migration rate, mu = mutation rate.

Speciation

- Population genetics rule of thumb: more than 1 migrant per generation = little or no differentiation. Wet blanket on speciation.

- This is wrong since it only tells us that G_{ST} will be low, not that real differentiation will be low. G_{ST} can be low even for completely differentiated subpopulations, or can be high even when subpopulations show little differentiation.

Speciation: Conclusion

- Speciation can easily happen even if there is some migration between subpopulations.

- May help explain how diversity arises

Why have people ignored these problems with their measures for so long?

- Ecologists and geneticists often treat measures as mere tools for the extraction of p-values (statistical significance).

- Statistical significance depends at least as much on sample size as on the magnitude of the effect being measured.

- Ecological problems should usually be cast in terms of estimating a meaningful parameter, with confidence intervals, rather than testing an always-false null hypothesis (which will always be rejected if sample size is large enough).

- Biologists need measures whose absolute magnitudes are interpretable.

- Biologists are all passionate about conserving biological diversity. We are seldom as passionate about the math we use to guide us in this task.
- Yet what if the mathematical tools we have always used to measure diversity, set conservation priorities, and monitor impacts of climate change or pollution, are systematically flawed? *What if much of conservation biology is pseudoscience which sometimes inadvertently promotes the extinction of species and the destruction of unique ecosystems?*
-
- If so, should we not be equally passionate about reforming the way we measure diversity?
- loujost@yahoo.com