Is Bray-Curtis differentiation meaningful in Molecular Ecology?

William B. Sherwin[1*]

[1]Evolution and Ecology Research Centre, School of BEES, UNSW-Sydney, NSW, Australia.
*e-mail: W.Sherwin@unsw.edu.au (corresponding author)

Running Title: Bray-Curtis in Molecular Ecology

ABSTRACT

A popular measure of differentiation in biodiversity is the Bray Curtis index of dissimilarity. It has recently also been proposed for use in molecular ecology. However, this measure currently cannot be predicted under specified conditions of population size, dispersal and speciation or mutation. Here I show forecasts for Bray-Curtis for two-variant systems such as single-nucleotide polymorphisms (SNPs) (or two species ecosystems). These are derived from well-known equations in population genetics, for forecasting measures such as G_{ST}, and shown to be appropriate by simulation. Thus, Bray-Curtis can now be used for assessment of differentiation, in order to understand natural or artificial processes, thus complementing other measures with different sensitivities, such as Morisita-Horn/D_{EST}, G_{ST} and Shannon Mutual Information/Shannon Differentiation.

Keywords: Bray-Curtis; genetic differentiation; community differentiation; biodiversity; Single nucleotide polymorphism (SNP); allele frequency difference (AFD)

1. INTRODUCTION

Comparisons of biodiversity between regions are important aspects of understanding both ecological and genetic systems. As in all science, it is important to test for departure from predicted values, because any departure reveals either incorrect assumptions about a wild population, or failure to achieve expected results in a managed population. It is therefore surprising that until recently, even some very popular measures of biodiversity have had very poor ability to either assess biodiversity, or to be forecast from the underlying biological processes (Nei 1973, Jost 2008, Jost et al. 2010, Chao et al. 2014, Sherwin et al. 2017). Recently, there have been attempts to rectify problems of measurement (Jost, DeVries et al. 2010, Leinster & Cobbold 2012, Sherwin et al. 2017, Chao et al. 2019), and new methods have been developed to derive expectations for various biodiversity measures, from an understanding of the underlying biological processes such as population dynamics, dispersal and mutation (or the parallel process in macroecology, speciation) (Hubbell 2001, Rosindell et al. 2010, Sherwin et al. 2017, Sherwin 2018). Much of this work has focused on the family of biodiversity measures derived from, or closely related to, the 'Hill Numbers', such as Gini-Simpson, Heterozygosity, nucleotide diversity, Shannon Entropy, Mutual Information, Shannon differentiation, F_{ST}, G_{ST}, Morisita-Horn, and Jost's D_{EST} (Jost 2008, Jost et al. 2010, Chao et al. 2014, Sherwin et al. 2017, Gaggiotti et al. 2018).

This article will concentrate upon an extremely popular method of assessing differentiation which is not part of the Hill-number family, but has recently been proposed for use in molecular ecology (Shriver et al. 1997, Berner 2019a,b, Price et al. 2020), following a trend for unification of ecological and genetic work (Rosindell et al. 2015, Sherwin 2018). This is

the Bray-Curtis index (Bray & Curtis 1957), which was originally used to compare diversity between forests (11), but is now used very widely, including for metagenomics (Peng et al. 2020). During 2020 alone, this index was cited throughout biology, medicine, and other sciences, being mentioned over 800 times in Google Scholar. Bray-Curtis (*B*) can be expressed in a way that facilitates comparison with differentiation measures derived from Hill numbers (Chao & Chiu 2016, Ricotta & Podani 2017, Ricotta et al. 2021):

$$B = \frac{\sum_{j=1}^{S}|a_{1j}-a_{2j}|}{\sum_{j=1}^{S}(a_{1j}+a_{2j})} \qquad \text{Equation 1}$$

where a_{1j} and a_{2j} are the abundances in each of two locations (1,2), for variant *j* ($1 \leq j \leq S$) and *S* is the total number of species or allelic types. This measure satisfies many of the requirements of a good measurement of differentiation between assemblages (Chao & Chiu 2016, Ricotta & Podani 2017). Its connection to other biodiversity measures has been explored (Ricotta et al. 2021).

Recently, two authors have also proposed that Bray-Curtis should be used for differentiation in molecular ecology and evolution, particularly for studies based on SNPs (two-allele single-nucleotide polymorphisms) (Shriver, Smith et al. 1997, Berner 2019a,b, Price et al. 2020). In these papers, Bray-Curtis was referred to as *AFD*, allele frequency difference (although it was admitted that *AFD* is really differentiation of proportion between zero and unity, rather than frequency between zero and infinity). In this two-variant case, Bray-Curtis simplifies to the unsigned difference of proportions of either of the two allelic variants between locations 1 and 2 (Berner 2019a,b)

$$B = |p_1 - p_2| \qquad \text{Equations 2 and A1.1}$$

where $p_1 = a_1/(a_1 + a_2)$ and $q_1 = 1 - p_1$, and similarly for p_2 and q_2.

This paper deals with genetic use of Bray-Curtis, responding to the suggested use in molecular ecology (Shriver et al. 1997, Berner 2019a,b, Price et al. 2020). Therefore the focus of this paper is on making forecasts for Bray-Curtis for SNPs, under various scenarios of population size, mutation, and dispersal, so that measures of Bray-Curtis can be used to evaluate competing models of population history, or make projections for the future. I then test these predictions by simulation, in comparison to Bray-Curtis' closest competitor measure, G_{ST}. Additionally, although Bray-Curtis is known to conform to many of the basic desirable properties of differentiation measures (Magurran 2004, Ricotta & Podani 2017), this paper also assesses Bray-Curtis' ability to satisfy another important property of differentiation measures – independence of alpha (within location) and beta (between location) variation (Jost 2008, Jost et al. 2010, Sherwin et al. 2017). A between location (beta) differentiation measure can be confounded by two aspects of within-location (alpha) diversity: proportions of variants, and number of variant types. With the restriction to two-variant SNPs, the latter is not a problem, but the effect of proportions of variants will be examined in this paper.

2. MATERIALS AND METHODS

This article constructs the forecasting apparatus for the simplest possible case of a single neutral biallelic SNP locus, with two locations (1,2); the measure can be averaged over multiple loci, and can be applied to haploids, or to diploids with linkage equilibrium.

When there are only two variants, the Bray-Curtis equation is

$$B = |p_1 - p_2|$$ (Berner 2019a,b) Equation 2, above

where p_1 p_2 are proportions of one of the two alleles at each location ($q_1 = 1 - p_1$; $q_2 = 1 - p_2$).

The quantity in equation 2 is a transform of two well-known differentiation measures

$G_{ST} = [H_T - \overline{H_1, H_2}]/H_T \quad \approx \quad F_{ST} = \sigma_p^2/pq$ ((Halliburton 2004) Box 9.5)

Equations 3 and A1.2

where σ_p^2 is the variance of p between locations, H is the Hardy-Weinberg (Binomial) expected heterozygosity eg $H_T = 1 - p^2 - q^2$; $H_1 = 1 - p_1^2 - q_1^2$; and p is the average p over the two locations (1,2); $q = 1 - p$. The measures G_{ST} and F_{ST} in equation 3 are identical in the two-allele, two location case ((Halliburton 2004) Box 9.5). Appendix A1 shows that $B^2 = 4pqG_{ST} = 2H_TG_{ST}$ Equations 4, A1.4

Because Bray-Curtis is closely related to G_{ST} or F_{ST}, Bray-Curtis forecasts can be based on well-known forecasts for these measures (Appendix A1). The expectation for diploid Bray-Curtis is:

$$B = \sqrt{\frac{2\,{}^2D - 2}{{}^2D(1+8N(2m+\mu))}}$$ Equation 5, A1.7

Where m is symmetrical dispersal between the two locations ($0 \le m \le 1$); μ is the rate of mutation (or speciation; $0 \le \mu \le 1$); N is the effective population size at each location (identical); and ${}^2D$ is the second order diversity, or effective number of alleles ${}^2D = 1/(1 - H_T)$.

The equivalent equation for the haploid SNPs simulated in this article is:

$$B = \sqrt{\frac{2\,{}^2D - 2}{{}^2D(1+4N(2m+\mu))}}$$ Equation 6, A1.8

This haploid equation is also appropriate for a pair of species variants in two local communities, if the mutation rate is replaced by the speciation rate, or considered to be negligible relative to the dispersal rate.

Using equation 6, forecasts of equilibrium Bray-Curtis (*B*) were devised for biallelic neutral single-nucleotide polymorphisms (SNPs) in two haploid subpopulations, for scenarios covering all possible combinations of symmetric dispersal (rate m = 0.01, 0.03, 0.1, 0.3), mutation rate ($\mu = 10^{-9}, 10^{-6}$) subpopulation effective sizes (N = 1000, 10000, 100000) and starting allele proportion in each subpopulation ($p = 0.1, 0.5$; $q = 1-p$). The latter allows examination of the effects of alpha (within locality) variation on Bray-Curtis.

For each scenario, the predictions of Bray-Curtis (*B*, equation 6) were tested by comparison with the output of the haploid simulation programs (MATLAB, Appendix A2, (Dewar et al. 2011)), which also assessed ability to predict G_{ST} (equation A1.5). There were 100 iterations of each scenario. Each iteration was run for 200 generations, and each generation included stochastic binomial sampling of the parents to establish the allele proportions for the offspring, followed by symmetrical dispersal to create the parent populations for the next generation. At the final generation, Bray-Curtis index *B* and G_{ST} were calculated, and regression was used to compare the simulation output to the predictions of equations 6 and A1.5 respectively.

3. RESULTS

Figure 1a shows the result of Bray-Curtis at the final generation of a MATLAB simulation of two equal-sized populations with two neutral (non-adaptive) variants such as SNP alleles, in 48 different scenarios with various: starting allele proportions $p = 0.1, 0.5$; effective population sizes *N*=1000, 10000,100000; mutation (or speciation) rates $\mu = 10^{-6}, 10^{-9}$; and symmetrical dispersal rates *m*= 0.01, 0.03, 0.1, 0.3 (further details are in Methods, or Appendix A1 for forecasts, A2 for simulations). Figure 1a shows simulated Bray-Curtis, (Equation 2), regressed against algebraic predictions of Bray-Curtis (*B*) (Equation 6). Three things are apparent in Figure 1a:

- there is an extremely good regression of simulated Bray-Curtis on predicted (*P*=3.5*10^{-23} see caption of Figure 1a)
- however, the slope is slightly below the expected 45 degree line for perfect prediction (slope = 0.80 see caption of Figure 1a; the 95% confidence limits for the slope were 0.766 to 0.840).
- Therefore the empirically best forecasting equation for haploids would be, combining equation 6 and the correction for regression slope:

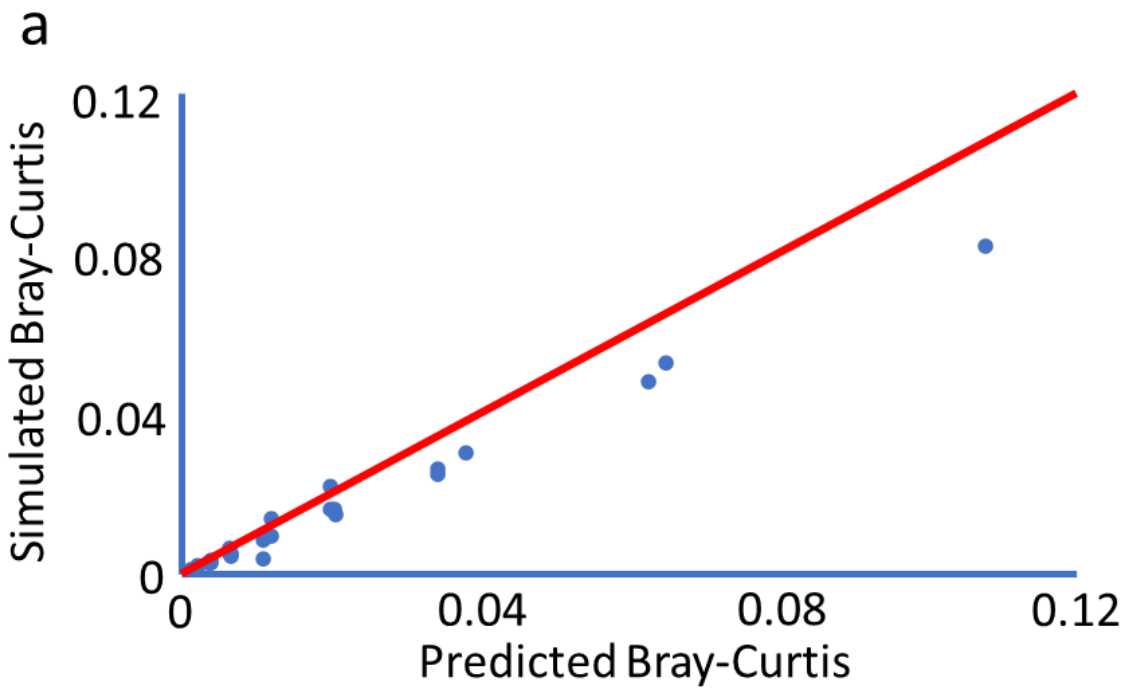
$$B = 0.8\sqrt{\frac{2\,{}^2D-2}{{}^2D(1+4N(2m+\mu))}} \qquad \text{Equation 7}$$

or the same for unlinked diploid loci, replacing $4N$ with $8N$:

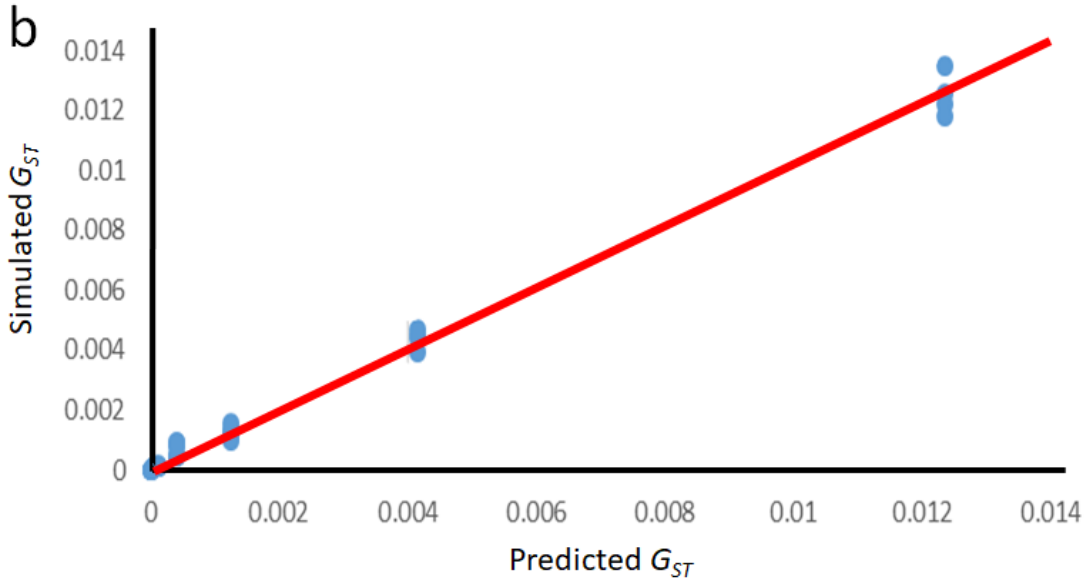
$$B = 0.8\sqrt{\frac{2\,{}^2D-2}{{}^2D(1+8N(2m+\mu))}} \qquad \text{Equation 8}$$

Figure 1b shows the result of regressions of simulated G_{ST} on the algebraic predictions of G_{ST} (Takahata 1983) (equation A1.5). Several things are apparent:

- there is also a very good regression of simulated G_{ST} on predicted ($P = 1.4*10^{-54}$)
- unlike Bray-Curtis, the slope is almost exactly the expected 45 degrees (1.01116 see caption of Figure 1b).



176



177

178 **FIGURE 1 | Comparison of simulation results with algebraic predictions.** (a) Bray-Curtis,
179 with regression equation Simulated-Bray-Curtis= 0.80*Predicted-Bray-Curtis + 0.0007;
180 Significance p = 3.5*10^{-23}; predicted Bray-Curtis from equation 6 in methods, A1.8 in
181 appendix; simulation result calculated by equation 2. (b) G_{ST}, with regression equation
182 Simulated-G_{ST} = 1.01116*Predicted-G_{ST} + 0.000006; p = 1.4*10^{-54}; predicted G_{ST} from
183 equation A1.5 (Takahata 1983). The red lines are the expected 1:1 relationships.

184

4. DISCUSSION

It is obvious from equations 7 and 8, and Figure 1a, that Bray-Curtis can now be used either for biological-inventories, or for studying underlying biological processes such as population size, speciation/mutation, reproduction, and dispersal (Vellend 2016, Sherwin 2018). These are the processes which some conservation initiatives aim to conserve (Anonymous 1988), and of course underly all biology. This paper shows that we can now have some ability to use these processes to predict Bray-Curtis, in a simplified two-location two-variant system, based upon equation 8 for diploid genes, or equation 7 for haploids (or for species).

However, two caveats apply here. Firstly, the forecasts in equations 7 and 8 are based upon selectively neutral assumptions, which sounds far-fetched, yet these forecasts have proved very useful in genetics despite the strong likelihood of intermittent selection. Secondly, Equations 5,6,7,and 8 show that Bray-Curtis has strong dependence on average within-location heterozygosity $H_T = 1 - 1/\ ^2D$, and thus on variant proportion *p*, which are aspects of within-location (alpha) variation, and therefore should not influence a differentiation (beta) measure such as Bray-Curtis (Jost et al. 2010, Chao et al. 2014, Sherwin et al. 2017). How the influence of within-location allele proportion occurs can be demonstrated with a simplified example: it is apparent from equation 2 that if either p_1 or p_2 is zero, then the value of Bray-Curtis will be equal to the other, more abundant, proportion. It should be noted that G_{ST} cannot be used to remedy this failing of Bray-Curtis, because it also unfortunately has dependence on alpha within-locality diversity (Nei 1973, Nei 1977, Jost 2008, Meirmans & Hedrick 2010). The latter paper offers a correction for the unwanted dependency of G_{ST}, but using this correction in the theory for Bray-Curtis would

have two drawbacks: it considerably complicates the correspondence to theoretical expectations; and it does not remove the effect of alpha variation on Bray-Curtis, but simply makes the effect more explicit (Appendix equation A1.14).

Biological scientists are now able to use the Bray-Curtis measure to either catalogue differentiation between-locations (or times) or even to investigate possible mechanisms of population dynamics, mutation, and dispersal in natural or managed systems. Thus Bray-Curtis can now complement other measures with different sensitivities, becoming part of a spectrum to represent biodiversity fully, as advocated by a number of authors (eg, Sherwin et al. 2017). These complementary measures derived from Hill-numbers for alpha and beta diversity have been well investigated, with many having good predictions from underlying factors such as population size, speciation/mutation, and dispersal, as well as showing independence of alpha and beta diversity (Sherwin et al. 2017). Shannon Mutual Information/Shannon Differentiation and Morisita-Horn/D_{EST} are differentiation measures that have available forecasts, and avoid errors such as dependency on within-location variation; the Shannon measures also avoid the heavy emphasis of effects of common variants, such as is seen with Morisita-Horn/D_{EST} (Magurran 2004, Jost 2008, Sherwin et al. 2017). It should also be noted that unlike the Hill-family of diversity measures, which can be corrected for incomplete sampling by the Good-Turing method (Chao & Jost 2015), the general Bray Curtis measure cannot currently use this optimum correction (A. Chao pers. comm.). However, this correction method is also inapplicable to any two-variant system such as SNPs.

231 This paper is the first introduction of predictive modelling for Bray Curtis in molecular
232 ecology. It can be extended in many ways. The equations for G_{ST} are based upon a number
233 of assumptions (Whitlock & McCauley 1999, Semenov et al. 2019, Ochoa & Storey 2021) and
234 each of these needs to be investigated if it is proposed to apply the Bray-Curtis equation 7
235 or 8 to any particular case. Firstly, it was assumed that there are only two locations, of
236 approximately equal effective size, which may be the case especially in some conservation
237 applications, but other possibilities would require further theory. Secondly, it was assumed
238 that there is symmetric dispersal m, the same for both locations, so that addressing a
239 source-sink situation would require further theory based on the continent-island model.
240 Thirdly, it was also assumed that there are only two alleles, as is often the case for SNPs, but
241 not for haplotypes. In future, all the theory in this paper might be extended to cases with
242 multiple alleles, broadening its use. With greater than two variants, there may be a need
243 for correction for *S*, number of variant alleles or species, as well as correction for variant
244 distribution. Fourthly, it was assumed that during filtering of data, these SNPs are chosen to
245 be neutral and unaffected by strong selection at nearby locations in the genome. Additional
246 theory would be required for loci under selection, which of course are very important in
247 evolution and conservation (Teixeira & Huber 2021). Fifthly, mutation rates are probably
248 negligible compared with dispersal rates; for example, typical SNP mutation rates are 10^{-9}
249 to 10^{-6}. However if the mutation (or speciation) rate is not negligible, then it needs to be
250 estimated. Finally, the equilibrium calculations presented above are appropriate in many
251 cases, with Tables A2.1 and A2.2. showing that there is a wide window of generation times
252 for which equilibrium is a reasonable assumption. However, in both natural and modified
253 habitats, often there is a non-equilibrium situation such as a sudden reduction in
254 connectivity, eg due to new human infrastructure. Therefore, dynamic (non-equilibrium)

255 equations are also needed, and one such equation is shown in equation A1.11, for time t
256 generations after a complete cessation of dispersal between two locations.
257
258 The other major direction for future development of this theory is to species-assemblages –
259 the original use of Bray-Curtis (Bray & Curtis 1957). The haploid case above is also
260 equivalent to species in an ecosystem, using the same underlying concepts: population
261 dynamics, dispersal, selection and speciation (in place of mutation). However, this might
262 require various refinements. Firstly, the extension to multiple species/alleles discussed
263 above will be very important. Secondly, the simulations use effective population size, not
264 actual size. Those using Bray-Curtis in evolution would be very familiar with effective
265 population size and its calculation, but this measure may not be so familiar to ecologists
266 dealing with arrays of species rather than alleles. Effective size is the reciprocal of the rate
267 at which variation is lost by random processes (eg loss of allele- or species-diversity through
268 stochastic drift (Vellend 2016)). It is best calculated from demographic data such as
269 reproduction and mortality (Engen et al. 2005), but can also be back-calculated from its
270 effect on genetic variation, and is typically much smaller than the actual number of
271 individuals of all types in the assemblage (Frankham 1995). There is a precedent for
272 calculating an equivalent of effective size for assemblages of species J_M (Hubbell 2001).
273 Thirdly, the simulations used a binomial mechanism because of the initial focus on 2-allele
274 SNPs. However other mechanisms such as Poisson or negative binomial might give different
275 dependency (Warton & Hui 2017), and this might be appropriate in other cases, including
276 where the underlying biological process for generating variants (speciation) is not fully
277 understood at present. Finally, as mentioned before, the forecasts in equations 7 and 8 are
278 based upon selectively neutral assumptions, and although some neutral genetic theory has

279 been applied to species assemblages (Hubbell 2001, Rosindell et al. 2010), it is best to add
280 selection to these models (Rosindell et al. 2010, 2015).
281

282 ACKNOWLEDGEMENTS

283 Advice and comments on the manuscript were kindly provided by: Juliet Byrnes, Anne Chao,
284 Lou Jost, Luis Mijangos, Greg Parry, Lee Ann Rollins, Alex Sentinella, John Sved, David
285 Warton and Jia Zhou.
286

287 ORCID

288 https://orcid.org/0000-0002-1578-8473
289

290 REFERENCES FOR MAIN AND APPENDIX

291
292 Anonymous. (1988). *The Flora and Fauna Guarantee Act 1988 (Victoria, Australia),* from
293 https://www.environment.vic.gov.au/conserving-threatened-species/flora-and-fauna-guarantee-
294 act-1988.
295
296 Berner, D. (2019a). Allele Frequency Difference AFD–an intuitive alternative to Fst for quantifying
297 genetic population differentiation. *Genes 10*, 308.
298
299 Berner, D. (2019b). Correction: Berner, D. Allele Frequency Difference AFD—an intuitive alternative
300 to Fst for quantifying genetic population differentiation. Genes 2019, 10, 308. *Genes 10*, 810.
301
302 Bray, J. R. & Curtis, J. T. (1957). An Ordination of the Upland Forest Communities of Southern
303 Wisconsin. *Ecological Monographs 27*, 325-349.
304
305 Chao, A. & Chiu, C.-H. (2016). Bridging the variance and diversity decomposition approaches to beta
306 diversity via similarity and differentiation measures. *Methods in Ecology and Evolution 7*: ,19-928.
307
308 Chao, A., Chiu, C.-H., & Jost, L. (2014). Unifying species diversity, phylogenetic diversity, functional
309 diversity, and related similarity and differentiation measures through Hill numbers. *Annual Review*
310 *of Ecology, Evolution, and Systematics 45*, 297-324.
311
312 Chao, A., Chiu, C.-H., Villéger, S., Sun,. I.-F., Thorn, S., Lin, Y.-C., … & Sherwin, W. B. (2019). An
313 attribute-diversity approach to functional diversity, functional beta diversity, and related
314 (dis)similarity measures. *Ecological Monographs 89*, e01343.
315

Chao, A. & Jost, L. (2015). Estimating diversity and entropy profiles via discovery rates of new species. *Methods in Ecology and Evolution 6*, 873–882.

Crow, J. F. & Kimura, M. (1970). *Introduction to population genetics*. New York: Harper and Row.

Dewar, R. C., Sherwin, W. B., Thomas, E., Holleley, C. E. and Nichols, R. A. (2011). Predictions of single-nucleotide polymorphism differentiation between two populations in terms of mutual information. *Molecular Ecology 20*, 3156–3166.

Engen, S., Lande, R., & Saether, B.-E. (2005). Effective size of a flucuating age-structured population. *Genetics 170*, 941-954.

Falconer, D. S., & Mackay, T. F. C. (1996). *Introduction to quantitative genetics*. Harlow: Addison Wesley Longman.

Frankham, R. (1995). Effective Population Size Adult Population Size Ratios in Wildlife - a Review. *Genetical Research 66*, 95-107.

Gaggiotti, O., Chao, A., Peres-Neto, P., Chiu, C.-H., Edwards, C., Fortin, M.-J., Jost, L.,…. & K. Selkoe (2018). Diversity from genes to ecosystems: A unifying framework to study variation across biological metrics and scales. *Evolutionary Applications 2018*, 1-18.

Halliburton, R. (2004). *Introduction to population genetics*. Upper Saddle River, NJ: Pearson.

Hubbell, S. P. (2001). *The unified neutral theory of biodiversity and biogeography*. Princeton, NJ: Princeton University Press.

Jost, L. (2008). Gst and its relatives do not measure differentiation. *Molecular Ecology 17*, 4015-4026.

Jost, L., DeVries, P., Walla, T., Greeney, H., & Ricotta, C. (2010). Partitioning diversity for conservation analyses. *Diversity and Distributions 16*, 65–76.

Leinster, T. & Cobbold, C. (2012). Measuring diversity: the importance of species similarity. *Ecology 93*, 477-489.

Magurran, A. E. (2004). *Measuring biological diversity*. Oxford: Blackwell.

Maruyama, T. (1970). On the fixation probability of mutant genes in a subdivided popualtion. *Genetical Research Cambridge 15*, 221-225.

Meirmans, P. G. & Hedrick, P.W. (2010). Assessing population structure: *Fst* and related measures. . *Molecular Ecology Resources 11*, 5–18.

Nei, M. (1973). Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Science USA 70*: 3321-3323.

Nei, M. (1977). F statistics and analysis of gene diversity in subdivided populations. *Annals of Human Genetics 41*, 225-234.

366 Ochoa, A., & Storey, J. D. (2021). Estimating Fst and kinship for arbitrary population structures.
367 *PLoS Genetics 17*, e1009241.
368
369 Peng, W., Huang, J., Yang, J., Zhang, Z., Yu, R., Fayyaz, S., … & Q, Y.-H. (2020). Integrated 16S rRNA
370 Sequencing, Metagenomics, and Metabolomics to Characterize Gut Microbial Composition,
371 Function, and Fecal Metabolic Phenotype in Non-obese Type 2 Diabetic Goto-Kakizaki Rats.
372 *Frontiers of Microbiology 10*, 3141.
373
374 Price, N., Lopez, L., Platts, A.E., & Lasky, J.R. (2020). In the presence of population structure: From
375 genomics to candidate genes underlying local adaptation. *Ecology and Evolution 10*, 1889–1904.
376
377 Ricotta, C. & Podani, J. (2017). On some properties of the Bray-Curtis dissimilarity and their
378 ecological meaning. *Ecological Complexity 31*, 201–205
379
380 Ricotta, C., Szeidl, L., & Pavoine, S. (2021). Towards a unifying framework for diversity and
381 dissimilarity coefficients. *bioRxiv* doi: https://doi.org/10.1101/2021.01.23.427893
382
383 Rosindell, J., Cornell, S.J., Hubbell, S. P., & Etienne, R. S. (2010). Protracted speciation revitalizes the
384 neutral theory of biodiversity. *Ecology Letters 13*, 716–727.
385
386 Rosindell, J., Harmon, L.J., & Etienne, R. S. (2015). Unifying ecology and macroevolution with
387 individual-based theory. *Ecology Letters 18*, 472–482.
388
389 Semenov, G. A., Safran, R. J., Smith, C. C. R., Turbek, S. P., Mullen, S. P., & Flaxman, S. M. (2019).
390 Unifying Theoretical and Empirical Perspectives on Genomic Differentiation. *Trends in Ecology &*
391 *Evolution 34*, 987-995.
392
393 Sherwin, W. B. (2018). Entropy, or Information, Unifies Ecology and Evolution. *Entropy 20*, 727.
394
395 Sherwin, W. B., Chao, A., Jost, L., & Smouse, P. E. (2017). Information Theory Broadens the
396 Spectrum of Molecular Ecology and Evolution. *Trends in Ecology and Evolution 32*, 948 - 963.
397
398 Shriver, M. D., Smith, M. W., Jin, L., Marcini, A., Akey, J. M., Deka, R. E., & Ferrell, R. (1997). Ethnic-
399 affliation estimation by use of population-specific DNA markers. *American Journal of Human*
400 *Genetics 60*, 957–964.
401
402 Takahata, N. (1983). Gene identity and genetic differentiation of populations in the finite island
403 model. *Genetics 104*, 497.
404
405 Teixeira, J. C. & Huber, C. D. (2021). The inflated significance of neutral genetic diversity in
406 conservation genetics. *Proceedings of the National Academy of Sciences USA 118*, e2015096118.
407
408 Vellend, M. (2016). *The Theory of Ecological Communities (MPB-57).* Princeton: Princeton University
409 Press.
410
411 Warton, D. I. & Hui, F. K. C. (2017). The central role of mean-variance relationships in the analysis of
412 multivariate abundance data: a response to Roberts (2017). *Methods in Ecology and Evolution* 8,
413 1408–1414.
414
415 Whitlock, M. C. (1992). Temporal fluctuations in demographic parameters and the genetic variance
416 among populations. *Evolution 46*, 608-615.

Whitlock, M. C. & McCauley, D. E. (1999). Indirect measures of gene flow and migration: Fst ≠ 1/(4Nm+1). *Heredity 82*, 117-125.

DATA ACCESSIBILITY

MATLAB program, and data for Figure 1, will be on DRYAD.

APPENDICES: Is Bray-Curtis differentiation meaningful in Molecular Ecology? Sherwin

A1 Forecasting equilibrium Bray-Curtis with mutation, dispersal and drift due to small population size, for two locations, with a single neutral biallelic SNP locus.

i = 1,2, – indices for locations. Where there is no index, or the index is T, it is the value calculated for the pooled locations (metapopulation), eg pooled allele proportion, overall heterozygosity.

B - Bray-Curtis between locations "1" and "2", the unsigned difference of proportions, ie $B = |p_1 - p_2|$ (Berner 2019a,b) (equation 2 in main article). (This is also called *AFD* – Difference of Allele "Frequency" ie proportion). The algebra below deals with a single locus, but Bray-Curtis can be averaged over loci.

^2D – Second order diversity, or effective number of alleles $^2D = 1/(1-H)$ or $H = 1 - 1/\ ^2D$

F_{ST} – Wright's measure of differentiation for biallelic SNPs $G_{ST} = F_{ST} = \sigma_p^2/pq = [H_T - \overline{H_1, H_2}]/H_T$ ((Halliburton 2004) Box 9.5)

G_{ST} – See F_{ST}; these are equivalent in the 2-allele, 2-location case.

H – Binomial (Hardy-Weinberg) expected heterozygosity eg $H_T = 1 - p^2 - q^2$; $H_1 = 1 - p_1^2 - q_1^2$

m – dispersal per generation between the two populations, symmetrical ($0 \le m \le 1$)

μ – mutation (or speciation) rate per generation ($0 \le \mu \le 1$)

N – effective population size at each location (identical)

p_1 p_2 – proportions of the chosen allele at each location $0 \le p_i \le 1$ (for the other allele, q_1=1-p_1 etc) at generation *t*

p – average *p* over the two locations at beginning of generation *t* : $p = \overline{p_i} = (p_1 + p_2)/2$; $q = 1 - p$

p' – proportions partway through generation *t*.

p" etc – proportions one generation after time *t* (at time *t"*).

s – number of localities (always two unless stated otherwise)

t - generation index (*t"* after one full generation).

T – is the index for the pooled locations (metapopulation), eg overall heterozygosity.

I restricted analysis to cases where there are two locations:

- with identical effective population size,
- reproduction with stochastic drift in each population is followed by dispersal
- deterministic symmetric dispersal between the two locations
- locations were followed for a single generation *t* to *t"*, during which the expected change of proportions is zero when the system is at equilibrium
- two alleles per locus (eg conventionally filtered SNP data)

	Location 1	Location 2
Generation *t*, initially	p_1, q_1	p_2, q_2
After Drift	p'_1, q'_1	p'_2, q'_2
After Dispersal	$p"_1 = p'_1 - mp'_1 + m'p_2.$ $q"_1 = 1 - p"_1$	$p"_2 = p'_2 - mp'_2 + m'p_1.$ $q"_2 = 1 - p"_2$

BRAY-CURTIS/AFD AT DRIFT-DISPERSAL EQUILIBRIUM

Bray-Curtis between locations "1" and "2", is

$$B = |p_1 - p_2| \text{(Berner 2019a,b)} \qquad \text{Equations 2, A1.1}$$

At any time, for 2 localities with 2 alleles per locus,

$G_{ST} = F_{ST} = [H_T - \overline{H_1, H_2}]/H_T = \sigma_p^2/pq$ Equation A1.2

where $\sigma_p^2 = \overline{p_i^2} - (\overline{p_i})^2$ ((Halliburton 2004) Box9.5 (Falconer & Mackay 1996) p56, Eq3.4)

IE $\sigma_p^2 = [(p_1^2 + p_2^2)/2] - \left[\left(\frac{(p_1+p_2)}{2}\right)^2\right] = (p_1 - p_2)^2/4 = B^2/4$ Equation A1.3

So $G_{ST} = F_{ST} = B^2/4pq$

Or $B^2 = 4pqG_{ST} = 2H_T G_{ST}$ Equation A1.4

Now at dispersal-drift-mutation equilibrium for *s* localities,

$G_{ST} = 1/\left(1 + \frac{4s}{s-1}\left(N\mu + \frac{sNm}{s-1}\right)\right)$ (equations 8 and 20 in (Takahata 1983)) Equation A1.5a

So with one pair of localities, $s = 2$

$$G_{ST} = 1/(1 + 8N(2m + \mu)) \qquad \text{Equation A1.5b}$$

So inserting eqn A1.5b into eqn A1.4, at equilibrium,

$$B^2 = 2H_T/(1 + 8N(2m + \mu)) = \frac{2\,{}^2D - 2}{{}^2D(1 + 8N(2m + \mu))} \qquad \text{Equation A1.6}$$

we get for diploid: $B = \sqrt{\frac{2\,{}^2D - 2}{{}^2D(1 + 8N(2m + \mu))}}$ Equation A1.7

and for haploid $B = \sqrt{\frac{2\,{}^2D - 2}{{}^2D(1 + 4N(2m + \mu))}}$ Equation A1.8

DYNAMIC (NON-EQUILIBRIUM) BRAY-CURTIS/AFD OVER TIME AFTER DISPERSAL IS REDUCED TO ZERO

At time t after dispersal is reduced to zero

$\sigma_p^2(at\ time\ t) = pq[1-(1-1/2N)^t]$ ((Falconer & Mackay 1996) eqn 3.2) Equation A1.9

From equation A1.3 above, $\sigma_p^2 = B^2/4$ or $B = \sqrt{4\,\sigma_p^2}$ Equation A1.10

If we are averaging over many loci, it is reasonable to assume that average allele proportions for the metapopulation (*p,q*) do not change over time. Then at time *t* after dispersal is reduced to zero, combine equations A1.9 and A 1.10:

$$B(at\ time\ t) = \sqrt{4p_{init}q_{init}[1-(1-1/2N)^t]} \quad \text{Equation A1.11a}$$

where $p_{init}$and q_{init} are the starting allele proportions for the metapopulation, ie:

$$B(at\ time\ t) = \sqrt{2H_{T\ (init)}[1-(1-1/2N)^t]} \quad \text{Equation A1.11b}$$

In equation A1.11, for haploids, 2*N* is replaced by *N.*

CAN WE CORRECT FOR DEPENDENCE ON ALPHA?

Note that equations A 1.4, A1.7 and A1.8 explicitly show the dependence of Bray-Curtis on within-locality (alpha) variation, H_T or ${}^2D$, and such dependence is not a desirable property for a measure of between-locality (beta) differentiation. This is additional to the dependence of G_{ST} on (alpha) heterozygosity. There is a correction for this unwanted dependency of G_{ST} (Meirmans & Hedrick 2010), so it is interesting to ask whether using this correction would remove the effect of alpha variation on Bray-Curtis. For a pair of locations, the corrected G_{ST} is:

$$G"_{ST} = \frac{2(H_T - \overline{H_1,H_2})}{(2H_T - \overline{H_1,H_2})(1-\overline{H_1,H_2})} \quad \text{Equation A1.12}$$

Combining equations A1.2 and A1.12,

$$G_{ST} = \frac{(2H_T - \overline{H_1,H_2})(1-\overline{H_1,H_2})}{2H_T}\, G"_{ST} \quad \text{Equation A1.13}$$

Combining equations A1.4 and A 1.13

$$B = \sqrt{(2H_T - \overline{H_1,H_2})(1-\overline{H_1,H_2})\, G"_{ST}} \quad \text{Equation A1.14}$$

Thus although this new formulation of Bray-Curtis uses $G"_{ST}$, which is free of influence of heterozygosity, Bray-Curtis is still heavily dependent upon heterozygosity *H*. Additionally, using this formulation in equation A1.14 for Bray-Curtis would considerably complicate the derivation of theoretical expectations, equations A1.5 to A1.8.

A2 The MATLAB simulation program

This MATLAB program was modified from the one previously described (Dewar et al. 2011), to include calculation of the Bray-Curtis Index (equations 2, A1.1), as well as the previously calculated G_{ST}(equation A1.2).

The simulation dealt with two biallelic haploid subpopulations, for scenarios with every possible combination of levels of symmetric dispersal (rate m = 0.01, 0.03, 0.1, 0.3), mutation rate (μ = 10^{-9}, 10^{-6}), effective subpopulation sizes (N = 1000, 10000, 100000) and starting allele proportion in each subpopulation (p = 0.1, 0.5). There were 100 iterations of each scenario. Each generation included stochastic binomial sampling of the parent alleles to establish the allele proportion for the offspring, followed by deterministic symmetrical dispersal to create the parent populations for the next generation. For each scenario (combination of m, μ, N, p), there were 100 independent iterations, whose results were averaged.

Each iteration was run for 200 generations, which was expected to be sufficient time to allow drift-dispersal equilibration without fixation of loci (see Tables A2.1, A2.2 below). Because the calculations in appendix A1 are for equilibrium m, μ, N without fixation (ie, loss of all alleles except one), it was important to run the simulations for times that are consistent with these two conditions. This is also important because most researchers, or the companies that do their genotyping, will filter out invariant (fixed) SNPs from the data. The two subsections below show that it is possible to choose simulation generation numbers that are sufficiently large to give approximate equilibrium, but short enough to give minimum fixation (see below). All simulations were run for the same time, 200 generations. The program included a trap for fixation, and it was designed to then restart (ie replace) any iterations where fixation occurred, in line with the filtering normally applied to such data. Because of the relatively short number of generations (200), there were no restarts for fixation.

<u>Expected Time to half equilibrium (for F_{ST}) (Whitlock 1992)</u>

Time to half equilibrium in generations for diploid is

$$t_{1/2\,eq} = \frac{\ln 0.5}{ln[(1-m)^2(1-{}^1/_{2N})]} \qquad \text{Equation A2.1a}$$

and for haploid is

$$t_{1/2\,eq} = \frac{\ln 0.5}{ln[(1-m)^2(1-{}^1/_{N})]} \qquad \text{Equation A2.1b}$$

where symbols are as in appendix A1. Maximum time to half-equilibrium is 69 generations for the scenarios trialled in the main paper (Table A2.1). Given that Bray Curtis is a function of F_{ST}, is seems reasonable to assume that this will also approximate the time to half-equilibrium for Bray-Curtis. The simulations should be run for several times this $t_{1/2\,eq}$. A time of 200 generations was chosen, and applied to all simulated scenarios. Iterations were each also inspected to ensure that each scenario had asymptoted to a stable value for Bray-Curtis, well before the final generation, and had a variance between-generations that was much lower than variance between replicate iterations (typically one tenth).

N	*m*	$t_{1/2\ eq}$
100000	0.01	68.95041
100000	0.03	22.75471
100000	0.1	6.578657
100000	0.3	1.943345
1000	0.01	67.29324
1000	0.03	22.57127
1000	0.1	6.563236
1000	0.3	1.941997
100000	0.01	68.95041
100000	0.03	22.75471
100000	0.1	6.578657
100000	0.3	1.943345

TABLE A2.1 Time to half-equilibrium $t_{1/2\ eq}$ generations for the scenario conditions simulated; see A1 for definitions of other symbols.

Expected Time to fixation

In this case with two equal-sized subpopulations making up a metapopulation with dispersal, *N* for metapopulation ≈ 2**N*-subpopulation; for haploid we use 4*N(metapop)* instead 8*N* in Maruyama's equation of expected time to fixation t_{fix}. (Maruyama 1970); (Crow & Kimura 1970 eqn 8.9.4 p 431).

Thus $$t_{fix} = -\frac{4Npln(p)}{1-p}$$ Equation A2.2

where symbols are as in A1. Minimum time to fixation is 1023 generations, for the scenarios trialled in main paper (Table A2.2). In an extreme case where *N* for the metapopulations was equal to the *N* for either subpopulation, the fixation times would be halved, so these times would all still be more than double the 200 generations simulated. Note that no fixations occurred in any iterations of the simulations.

Initial *p*	*N*	Fixation time
0.5	100000	277258.9
0.1	100000	102337.1
0.5	10000	27725.89
0.1	10000	10233.71
0.5	1000	2772.589
0.1	1000	1023.371

TABLE A2.2 Expected time to fixation the scenario conditions simulated; see A1 for definitions of symbols.