Additional file 1: Comparison between JSD and some classical genetic distances

Additional material for the paper *Genetic flow directionality and geographical segregation in a Cymodocea nodosa genetic diversity network*, by Paolo Masucci[1], Sophie Arnaud-Haond[2], Víctor M. Eguíluz[3], Emilio Hernández-García*[3]and Ester A. Serrão[4]

[1] Centre for Advanced Spatial Analysis, University College of London, London, UK
[2] Institut Français de Recherche pour l'Exploitation de la MER, Département Étude des Ecosystèmes Profonds-DEEP, Laboratoire Environnement Profond-LEP, Centre de Brest, France
[3] IFISC (CSIC-UIB), Instituto de Física Interdisciplinar y Sistemas Complejos, Consejo Superior de Investigaciones Científicas - Universitat de les Illes Balears, Palma de Mallorca, Spain
[4] CCMAR, CIMAR-Laboratório Associado, Universidade do Algarve, Gambelas, 8005-139, Faro, Portugal

Email: Paolo Masucci - a.masucci@ucl.ac.uk; Sophie Arnaud-Haond - sophie.arnaud@ifremer.fr; Víctor M. Eguíluz - victor@ifisc.uib-csic.es; Emilio Hernández-García*- emilio@ifisc.uib-csic.es; Ester A. Serrão - eserrao@ualg.pt;

*Corresponding author

Here we compare JSD as defined in Eq.1 of the main text with some of the most used genetic distances. In particular we compare it with the Nei distance NEI [1], the Cavalli-Sforza distance CS [2], the Goldstein distance GD [3] and the average square distance ASD [4]. D, NEI and CS are distances defined in a symbolic space, while GD and ASD are defined in metrical space where the metric is defined by the allele repetitions.

The **Cavalli-Sforza distance** CS is defined as

$$CS = \sqrt{4\frac{\sum_l (1 - \sum_i \sqrt{x_i y_i})}{\sum_l (a_m - 1)}}. \tag{1}$$

Here x_i is the fraction of allele i in the first population, y_i is the fraction of allele i in the second population. A second sum is over the loci l and a_m is the total number of alleles, but if we work in the gamete space, then x_i and y_i refer to gamete frequencies and $l = 1$.

NEI is defined as:

$$NEI \equiv -\log\left(\frac{\sum_l \sum_i x_i y_i}{\sqrt{\sum_l \sum_i x_i^2 \sum_l \sum_i y_i^2}}\right). \tag{2}$$

The main statistical difference between NEI and CS with JSD, as we are going to show, is that JSD incorporates a weighting system for the different population sizes, while NEI and CS don't.

In Fig.1 we show the correlations between those measures, as measured in the CN dataset. For a reason that will be clear below we define the parameter $\Delta\pi = |\pi_1 - \pi_2|$ as the absolute value of the difference of the statistical weight between population 1 and 2. We represent with black triangles the distances between populations whose weight difference $\Delta\pi$ is larger or equal than 0.75 and with white circles the distances between those populations whose weight difference $\Delta\pi$ is less than 0.75.

In the top-left panel we show the plot for NEI versus CS. The main difference between NEI and CS is that NEI considers the variance of the gamete distributions, while CS doesn't. We see that the measures are well correlated with the differences given by the variances of the populations.

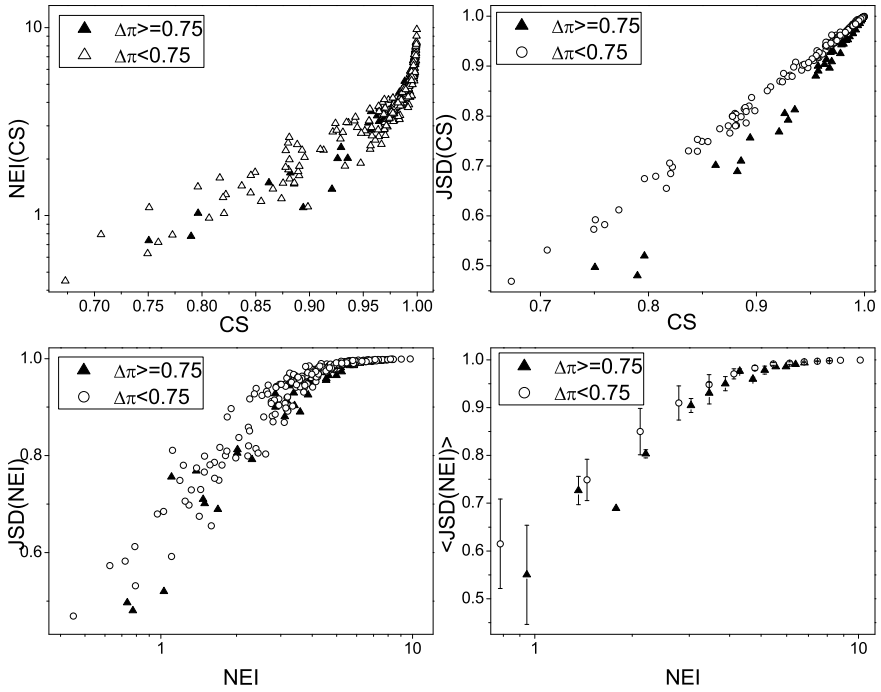

Figure 1: Measures performed for the CN dataset. Top-left: correlations between NEI and CS. Top-right: JSD versus CS. Bottom-left: JSD versus NEI. Bottom-right: $< JSD(NEI) >$ versus NEI.

In the top-right panel we show the plot of JSD versus CS. Also in this case the measures are positively correlated, even if two different branches appear in the plot. Those two branches appear to be well represented by the two different categories for $\Delta\pi$. In fact when the population sizes are very different, $\Delta\pi \geq 0.75$, we see that CS overestimates the distance between them in respect to JSD.

In the bottom-left panel we show the plot of JSD versus NEI. Again we can see that the measures are well correlated, but also in this case NEI overestimates the distance between two populations when $\Delta\pi \geq 0.75$, even if this is less evident that in the previous case. Nevertheless if we look at the plot of the average value of the JSD corresponding to the same NEI, $< JSD(NEI) >$, on the bottom-right panel, we can see that it is true in average, since each black triangle is below the correspondent white circle.

The **average square distance** ASD between population A and B is defined as [4]:

$$ASD_k \equiv \sum_{i,j}(i-j)^2 f_i f_j = (\mu_A^k - \mu_B^k)^2 + V_A^k + V_B^k = (\delta\mu^k)^2 + V_A^k + V_B^k, \tag{3}$$

where i, j are the repetition numbers of allele $i\epsilon A$ and $j\epsilon B$ and $f_{i,j}$ its frequency at locus k. Then $\mu_X^k = \sum i \cdot f_i$ is the average repetition number at locus k for population X and $V_X^k = \sum f_i(i - \mu_X^k)$ is the variance in the repetition number of population X at locus k. Eq.3 has to be averaged on the loci then.

$$ASD = \frac{\sum_l^n ASD_k}{l}. \tag{4}$$

The **Goldstein distance** GD between two populations A and B defined by a set of alleles is defined at the locus k as [3]:

$$(\delta\mu^k)^2 \equiv (\mu_A^k - \mu_B^k)^2, \tag{5}$$

where μ_X^k is the average number of repetitions for population X at locus k.

In the case of n different loci Eq.5 is averaged on the different loci:

$$GD \equiv \frac{\sum_{k=1}^{n}(\delta\mu^k)^2}{n}. \tag{6}$$

GD was introduced as an improvement of ASD *"because distances based on the infinite-alleles model are nearly linear with time immediately following isolation, it is not worthwhile to use ASD with very closely related groups"* [3].

The difference between GD and ASD, as it is clear from Eq.3 and Eq.6, resides on the presence in ASD of the variance of the attributes.

To understand the difference between JSD and GD in the gamete space, we have to keep in mind the representation of the genet in the gamete space as explained in the main paper. The gamete space is a n-dimensional space, each dimension representing a locus, where a diploid genet is represented by a set of 2^n points. For a population X such a distribution of points has a centre of mass, whose coordinates μ_X^k $(k = 1, .., n)$ are given by the average repetition number for each locus $\mu_X^k = \sum i \cdot f_i$. Hence, given two populations A and B, the distance between their average point in the attribute space is given by $\sqrt{\sum_{i=k}^{n}(\mu_A^k - \mu_B^k)^2} = \sqrt{n \cdot GD}$ (see Eq.6).

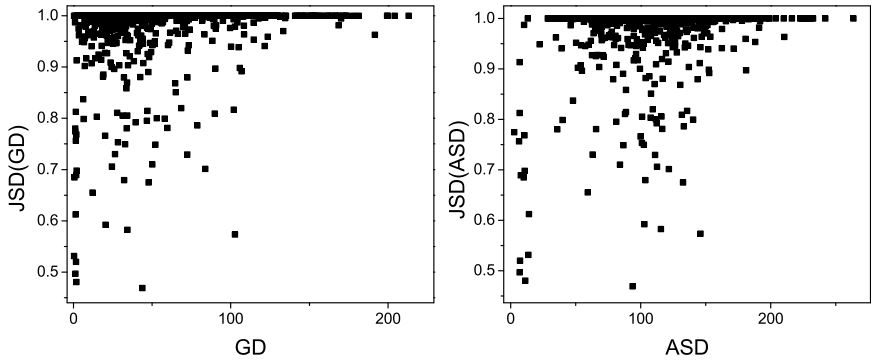

Figure 2: Measures performed for the CN dataset. Left panel: JSD versus GD. Right panel: JSD versus ASD

Hence GD is the square root of the distance between the centre of mass of the populations as represented in the gamete space. JSD instead is a punctual information measure that considers all the point correlations between the two populations. Then JSD and GD could be correlated, this does not occur always. For instance there is an extreme case where population A and B have the same centre of mass in the gamete space, so that $GD = 0$, but not a single common gamete so that $JSD = 1$. We can observe this in Fig.2, where we show evidence for not very strong correlations between JSD and GD in the left panel and between JSD and ASD in the right panel, as measured for the CN dataset.

References

1. Nei M: **Genetic distance between populations**. *Am Nat* 1972, **106**(949):283–292.
2. Cavallis L, Edwards A: **Phylogenetic analysis models and estimation procedures**. *Am J Hum Genet* 1967, **19**:233–257.
3. Goldstein DB, Linares AR, Cavallisforza L, Feldman M: **Genetic absolute dating based on microsatellites and the origin of modern humans**. *Proc Natl Acad Sci USA* 1995, **92**(15):6723–6727.
4. Goldstein DB, Linares AR, Cavallisforza L, Feldman M: **An evaluation of genetic distances for use with microsatellite loci**. *Genetics* 1995, **139**(1):463–471.