FST and its relatives GST and Jost's D

Summary

GST is the multi-allelic equivalent of Wright's FST as defined by Nei (1973). Estimates of GST and of Wright's FST as defined by Nei, hereafter referred to as Nei's estimator, are therefore identical when based on biallelic data, such as the barcode and WGS data used in the current study (Fig 2 and 3), and do not correlate with distance, neither naively (Fig 4 and 5) nor when using the transformation predicted to give a linear relationship under a two-dimensional isolation by distance model (Rousset 1997) (Fig 6 and 7).

Estimates generated by Nei's estimator (Nei 1973) differ to estimates generated using Hudson's FST estimator (Fig 8 and 9). This is because Hudson's estimator is based on Weir and Hill's definition of Wright's FST (B. S. Weir and Hill 2002), where the denominator is defined relative to an ancestral population, whereas Nei defined FST relative to present populations (as reviewed by Bhatia et al. (2013)). The latter is also true of the more commonly used estimator of B. Weir and Cockerham (1984). We use Hudson's estimator to estimate FST because it is robust for unbalanced sample sizes (Willing, Dreyer, and Oosterhout 2012; Bhatia et al. 2013), where as the more commonly used estimator of B. Weir and Cockerham (1984) is not. Moreover, Hudson's estimator was designed for multi-loci data, being asymptotically consistent in the number of loci (Reich et al. 2009; Bhatia et al. 2013).

In contrast to bi-allelic FST, which is bounded by 0 and 1, within population diversity (such as that caused by high mutation rates) limits the upper bound of multi-allelic GST, hampering its interpretation, as discussed in (Whitlock 2011; Meirmans and Hedrick 2011). Jost's D is an alternative measure of differentiation, intended to overcome problems associated with GST and its relatives (Jost 2008). According to Meirmans and Hedrick (2011), Jost's perceived problems included the aforementioned restricted range, as well as additive variance and the heterozygosity basis of diversity. Jost thus proposed a normalized measure of allelic differentiation, assuming multiplicative variance, where diversity is based on the effective number of alleles, essentially resulting in a logistic relationship between diversity and heterozygosity (Meirmans and Hedrick 2011).

There are numerous issues with Jost's D (Whitlock 2011), especially when assuming convergence under the island model of migration (Meirmans and Hedrick 2011). However, the most relevant issue here is its application to SNP data, which can lead to counterintuitive results, since SNPs have at most four allelic states (Meirmans and Hedrick 2011). Another complication is that Jost's D was designed as a single-locus measure (Josts'D is highly locus-specific, being dependent on the mutation rate, which varies over multiallelic loci, e.g. microsatellites, for which Jost's D was formulated). Nevertheless, following suggestions in (Meirmans and Hedrick 2011), we calculated Jost's D using average heterozygosities (as described at length below). Fig 11 to 13 show that there is no relationship between estimates of Jost's D and distance based on the barcode nor WGS data, neither when considered naively, nor when using the transformation predicted to give a linear relationship under the two-dimensional model of isolation by distance (Rousset 1997).

In summary, we have shown here that there is little evidence of a relationship between distance and estimates of GST and Jost's D based on data from the Thai-Myanmar border. We have not included these results in the revised manuscript, however, since we are working with biallelic data sets (both barcode and WGS) over multiple loci, and thus do not feel GST nor Jost's D are appropriate measures.

Methods and plots

F_{ST} definition

There are many different definitions and estimators of FST. In our manuscript, we used Hudson's estimator (Hudson, Slatkin, and Maddison 1992), whose explicit formulation was outlined first in the supplementary information of Reich et al. (2009), and later in (Bhatia et al. 2013), where it was fist coined "the Hudson

estimator". In the limit of large sample sizes, the Hudson estimator does not depend on the ratio of sample sizes (Bhatia et al. 2013), whereas the most commonly used estimator of B. Weir and Cockerham (1984) does (assuming population-wise FST estimates as defined B. S. Weir and Hill (2002) are unequal), and so is liable to inflate FST estimates (Bhatia et al. 2013). The Hudson estimator as also been shown numerically to perform well when the ratio is high and the minimum sample size is small (Willing, Dreyer, and Oosterhout 2012). This is important because Hudson's estimator is typically used by the Reich lab to study ancient DNA samples, which are rare (personal communication). Its relationship to other commonly used estimators is explored both theoretically and numerically in Bhatia et al. (2013). Here we show its relationship to estimates of G_{ST} (Nei 1973) and Jost's D (Jost 2008). We focus on multi-locus estimates, because estimates from single SNPS are highly liable to inflation given the small and unequal sample sizes.

Parameter definition

Parameters G_{ST} (Nei 1973), G'_{ST} (Hedrick 2005) and Jost's D (Jost 2008) (reviewed in Whitlock (2011)) can all be defined using total and within population heterozygosity (H_T and H_S, respectively) as outlined in Meirmans and Hedrick (2011) (but here taking k of Meirmans and Hedrick (2011) to be equal to 2, since we are only interested in pairwise comparisons),

$$G_{ST} = {}^{H_T - H_S}\!/_{H_T}, \tag{1}$$
$$G'_{ST} = {}^{G_{ST}(1 + H_S)}\!/_{1 - H_s}, \tag{2}$$
$$D = 2{}^{H_T - H_S}\!/_{1 - H_S}. \tag{3}$$

Based on Nei and Chesser (1983), taking $w_i = {}^1\!/_2$ (assuming equal relative population size), $r = 2$ and $s = 2$ (since we are working with pairwise comparisons of biallelic data), and using p_i and q_i to denote $p_{i\ k=1}$ and $p_{i\ k=2}$ of Nei and Chesser (1983) (that is to say, p denotes one allele frequency and $q = 1 - p$ the other, while numerical subscripts denote populations),

$$H_S = {}^1\!/_2\big\{(1 - p_1^2 - q_1^2) + (1 - p_2^2 - q_2^2)\big\}, \tag{4}$$
$$H_T = 1 - \Big({}^{p_1 + p_2}\!/_2\Big)^2 - \Big({}^{q_1 + q_2}\!/_2\Big)^2. \tag{5}$$

We note that

$$\begin{aligned} H_S &= {}^1\!/_2\big\{(1 - p_1^2 - q_1^2) + (1 - p_2^2 - q_2^2)\big\}, \\ &= {}^1\!/_2\big\{2p_1(1 - p_1) + 2p_2(1 - p_2)\big\} \\ &= {}^1\!/_2\big\{2h_1 + 2h_2\big\}, \\ &= h_1 + h_2. \end{aligned} \tag{6}$$

Where $2h_1$ and $2h_2$ are the heterozygosities for a given marker in populations 1 and 2, respectively, following notation in Reich et al. (2009). And that

$$\begin{aligned} H_T &= 1 - \Big({}^{p_1 + p_2}\!/_2\Big)^2 - \Big({}^{q_1 + q_2}\!/_2\Big)^2, \\ &= 1 - {}^1\!/_4\big((p_1 + p_2)^2 - (q_1 + q_2)^2\big). \end{aligned} \tag{7}$$

Parameter estimation

Following Meirmans and Hedrick (2011), we estimate G_{ST}, G_{ST} and Jost's D using unbiased estimates of H_S and H_T,

$$\hat{G}_{ST} = {}^{\hat{H}_T - \hat{H}_S}/_{\hat{H}_T}, \tag{8}$$

$$\hat{G}'_{ST} = {}^{\hat{G}_{ST}(1+\hat{H}_S)}/_{1-\hat{H}_s}, \tag{9}$$

$$\hat{D} = {}^{2\hat{H}_T - \hat{H}_S}/_{1-\hat{H}_S}, \tag{10}$$

where $\hat{H}_T = \sum_{k=1}^{M} \hat{H}_T^{[k]}$ and $\hat{H}_S = \sum_{k=1}^{M} \hat{H}_S^{[k]}$ and $\hat{H}_T^{[k]}$ and $\hat{H}_S^{[k]}$ are unbiased estimates at the kth SNP of the within-population and total heterozygosity, respectively. For consistency with our estimates of F_{ST}, we construct $\hat{H}_T^{[k]}$ and $\hat{H}_S^{[k]}$ following Reich et al. (2009). Specifically, fixing k, we estimate H_S (as defined in equation (6)) using,

$$\hat{H}_S = \hat{h}_1 + \hat{h}_2, \tag{11}$$

where $\hat{h}_1$ and $\hat{h}_2$ are unbiased estimators of h_1 and h_2 introduced via equation 9 of Reich et al. (2009). Note that $2\hat{h}_i = \hat{H}_{S_i}$ of equation 7 in Nei and Chesser (1983) if H_{0i} of Nei and Chesser (1983) is zero. However, equation (11) above does not equal equation 9 of Nei and Chesser (1983), where the latter is only approximately unbiased as remarked upon in Reich et al. (2009). Similarly, our estimate of H_T outlined below differs from equation 11 of Nei and Chesser (1983).

We unbiasedly estimate H_T for a given marker (as defined in equation (7)), using unbiased estimators of $(p_1+p_2)^2$ and $(q_1+q_2)^2$, which we denote $\hat{U}$ and $\hat{V}$, respectively,

$$\hat{H}_T = 1 - \frac{\hat{U}}{4} - \frac{\hat{V}}{4}. \tag{12}$$

The unbiased estimates $\hat{U}$ and $\hat{V}$,

$$\hat{U} = \left({}^{a_1}/_{n_1} + {}^{a_2}/_{n_2}\right)^2 - {}^{\hat{h}_1}/_{n_1} - {}^{\hat{h}_2}/_{n_2}, \tag{13}$$

$$\hat{V} = \left({}^{b_1}/_{n_2} + {}^{b_2}/_{n_1}\right)^2 - {}^{\hat{h}_1}/_{n_1} - {}^{\hat{h}_2}/_{n_2}, \tag{14}$$

are constructed following the same reasoning used to construct $\hat{N}$ of equation 10 in Reich et al. (2009)).

$\hat{U}$ and $\hat{V}$ are unbiased

Analytical results

Let $Y = \left({}^{a_1}/_{n_1} + {}^{a_2}/_{n_2}\right)^2$ be a naive estimate of $(p_1+p_2)^2$ and $X = \left({}^{b_1}/_{n_1} + {}^{b_2}/_{n_2}\right)^2$ be a naive estimate of $(q_1+q_2)^2$, then

$$
\begin{aligned}
E[Y] &= E\Big[\Big({}^{a_1}\!/_{n_1} + {}^{a_2}\!/_{n_2}\Big)^2\Big], \\
&= E\Big[\Big\{\Big({}^{a_1}\!/_{n_1} - p_1\Big) + \Big({}^{a_2}\!/_{n_2} - p_2\Big) + (p_1+p_2)\Big\}^2\Big], \\
&= E\Big[\Big({}^{a_1}\!/_{n_1} - p_1\Big)^2 + \Big({}^{a_2}\!/_{n_2} - p_2\Big)^2 + (p_1+p_2)^2 \\
&+ 2\Big({}^{a_1}\!/_{n_1} - p_1\Big)\Big({}^{a_2}\!/_{n_2} - p_2\Big) + 2\Big({}^{a_1}\!/_{n_1} - p_1\Big)(p_1+p_2) + 2\Big({}^{a_2}\!/_{n_2} - p_2\Big)(p_1+p_2)\Big], \qquad (15) \\
&= E\Big[\Big({}^{a_1}\!/_{n_1} - p_1\Big)^2\Big] + E\Big[\Big({}^{a_2}\!/_{n_2} - p_2\Big)^2\Big] + (p_1+p_2)^2, \qquad (16) \\
&= Var\Big[{}^{a_1}\!/_{n_1}\Big] + Var\Big[{}^{a_2}\!/_{n_2}\Big] + (p_1+p_2)^2, \\
&= {}^{1}\!/_{n_1^2} Var[a_1] + {}^{1}\!/_{n_2^2} Var[a_2] + (p_1+p_2)^2, \qquad (17) \\
&= {}^{p_1(1-p_1)}\!/_{n_1} + {}^{p_2(1-p_2)}\!/_{n_2} + (p_1+p_2)^2, \qquad (18) \\
&= {}^{h_1}\!/_{n_1} + {}^{h_2}\!/_{n_2} + (p_1+p_2)^2,
\end{aligned}
$$

assuming from equation (15) to (16), that the covariance $E\big[\big({}^{a_1}\!/_{n_1} - p_1\big)\big({}^{a_2}\!/_{n_2} - p_2\big)\big] = 0$ and noting that $E\big[2\big({}^{a_1}\!/_{n_1} - p_1\big)(p_1+p_2)\big] = 2(p_1+p_2)\big\{E\big[{}^{a_1}\!/_{n_1}\big] - p_1\big\} = 0$ and likewise for $E\Big[2\big({}^{a_1}\!/_{n_1} - p_1\Big)(p_1+p_2)\big]$, and assuming from (17) to (18), that a_1 and a_2 are binomially distributed, where h_1 and h_2 are defined above. And similarly for X, such that,

$$
\begin{aligned}
E[X] &= E\Big[\Big\{\Big({}^{b_1}\!/_{n_1} - q_1\Big) + \Big({}^{b_2}\!/_{n_2} - q_2\Big) + (q_1+q_2)\Big\}^2\Big], \\
&= E\Big[\Big({}^{b_1}\!/_{n_1} - q_1\Big)^2 + \Big({}^{b_2}\!/_{n_2} - q_2\Big)^2 + (q_1+q_2)^2 \\
&+ 2\Big({}^{b_1}\!/_{n_1} - q_1\Big)\Big({}^{b_2}\!/_{n_2} - q_2\Big) + 2\Big({}^{b_1}\!/_{n_1} - q_1\Big)(q_1+q_2) - 2\Big({}^{b_2}\!/_{n_2} - q_2\Big)(q_1+q_2)\Big], \\
&= Var\Big[{}^{b_1}\!/_{n_1}\Big] + Var\Big[{}^{b_2}\!/_{n_2}\Big] + (q_1+q_2)^2, \\
&= {}^{h_2}\!/_{n_2} + {}^{h_1}\!/_{n_1} + (q_1+q_2)^2,
\end{aligned}
$$

noting that $Var[b_1] = Var[a_1]$ and likewise for $b2$. Given $E[Y]$ and $E[X]$, and the fact that $\hat{h}_1$ and $\hat{h}_2$ are unbiased estimators of h_1 and h_2, respectively, it is easy to show that $\hat{U}$ and $\hat{V}$ are unbiased estimates of $(p_1+p_2)^2$ and $(q_1+q_2)^2$, respectively,

$$
\begin{aligned}
E\big[\hat{U}\big] &= E\Big[\Big({}^{a_1}\!/_{n_1} + {}^{a_2}\!/_{n_2}\Big)^2 - {}^{\hat{h}_1}\!/_{n_1} - {}^{\hat{h}_2}\!/_{n_2}\Big], \\
&= E\Big[Y - {}^{\hat{h}_1}\!/_{n_1} - {}^{\hat{h}_2}\!/_{n_2}\Big], \\
&= E[Y] - E[{}^{\hat{h}_1}\!/_{n_1}] - E[{}^{\hat{h}_2}\!/_{n_2}], \\
&= {}^{h_1}\!/_{n_1} + {}^{h_2}\!/_{n_2} + (p_1+p_2)^2 - {}^{h_1}\!/_{n_1} - {}^{h_2}\!/_{n_1}, \\
&= (p_1+p_2)^2.
\end{aligned}
$$

$$
\begin{aligned}
E\big[\hat{V}\big] &= E\Big[\Big({}^{b_1}\!/_{n_1} + {}^{b_2}\!/_{n_2}\Big)^2 - {}^{\hat{h}_1}\!/_{n_1} - {}^{\hat{h}_2}\!/_{n_2}\Big], \\
&= E\Big[X - {}^{\hat{h}_1}\!/_{n_1} - {}^{\hat{h}_2}\!/_{n_2}\Big], \\
&= E[X] - E[{}^{\hat{h}_1}\!/_{n_1}] - E[{}^{\hat{h}_2}\!/_{n_2}], \\
&= {}^{h_1}\!/_{n_1} + {}^{h_2}\!/_{n_2} + (q_1+q_2)^2 - {}^{h_1}\!/_{n_1} - {}^{h_2}\!/_{n_1}, \\
&= (q_1+q_2)^2.
\end{aligned}
$$

Numerical results

To illustrate the unbiased nature of $\hat{H}_S^{[k]}$ and $\hat{H}_T^{[k]}$, in this section we plot average results from increasingly many independent experiments, each with independent draws of a_1 and a_2 from binomial distributions with fixed p_1 and n_1, and p_2 and n_2.

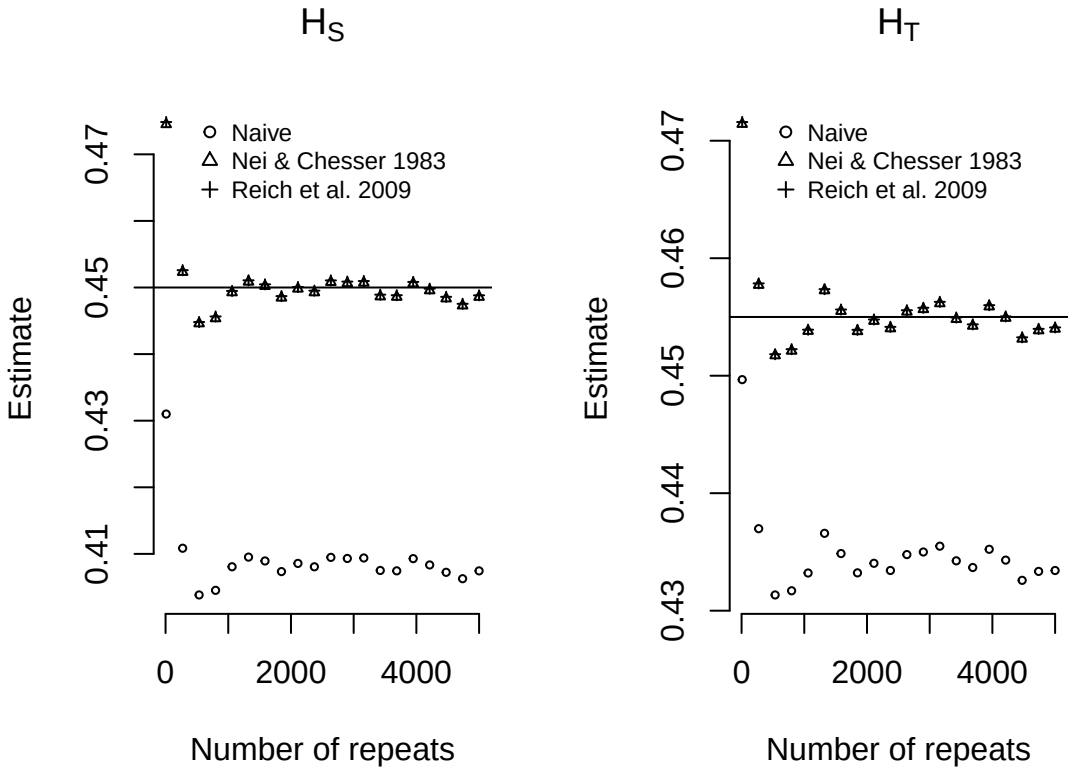

Figure 1: Numerical verification of unbiased heteroygosity estimates

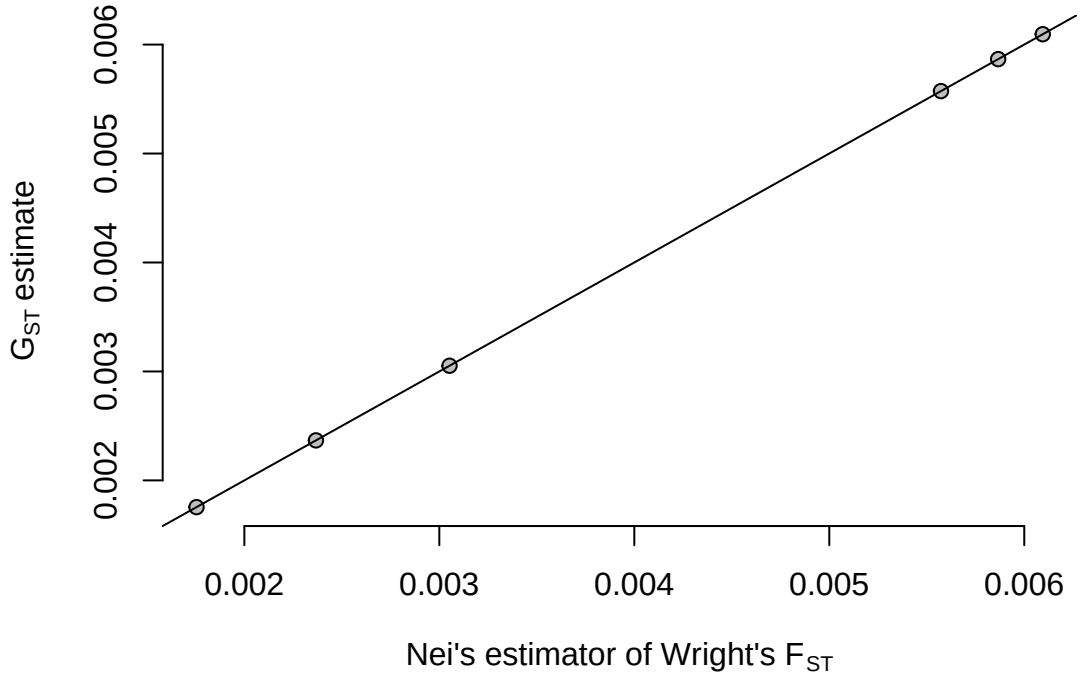

Figure 2: Equality of estimates of Nei's definition of Wright's FST and GST based on barcode data.

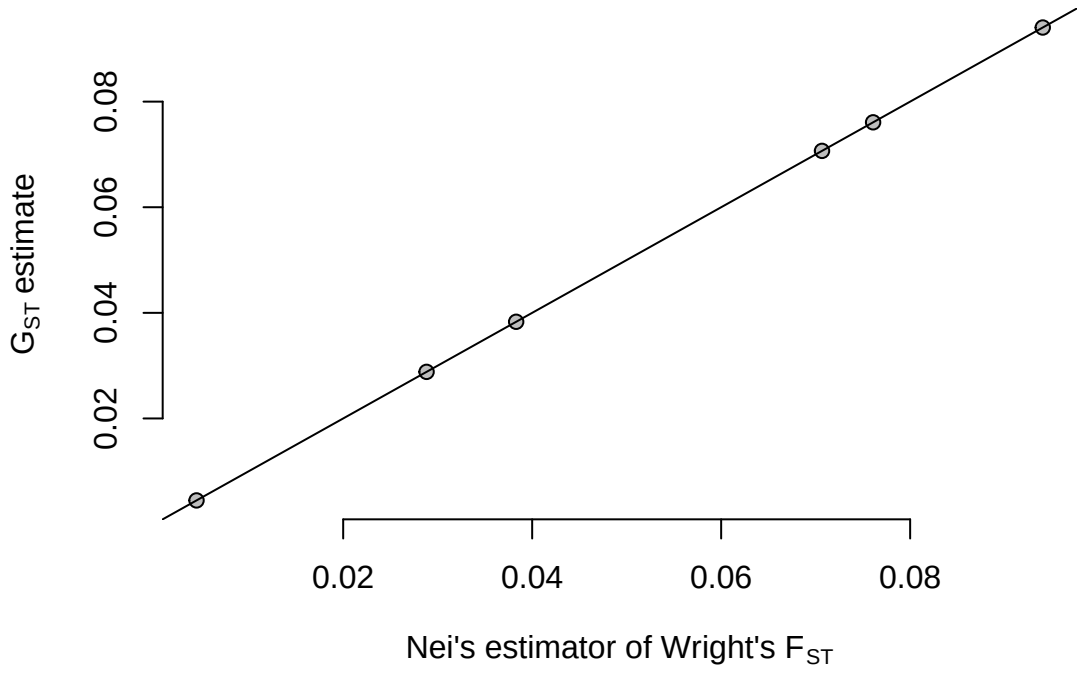

Figure 3: Equality of estimates of Nei's definition of Wright's FST and GST based on WGS data.

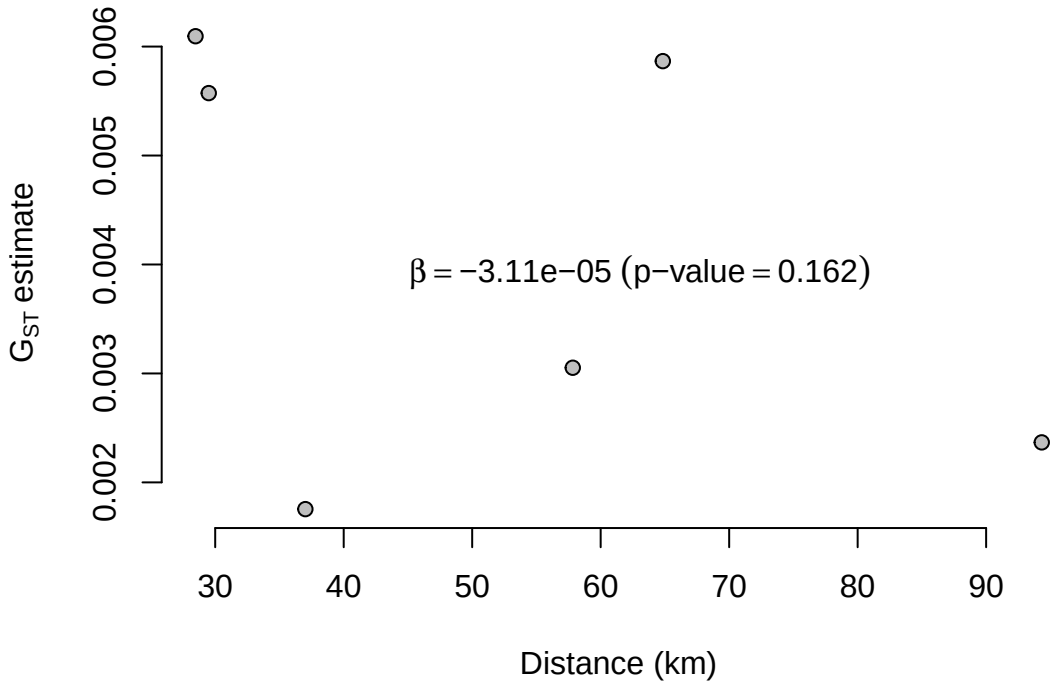

Figure 4: GST estimates and distance: barcode data, no transformation.

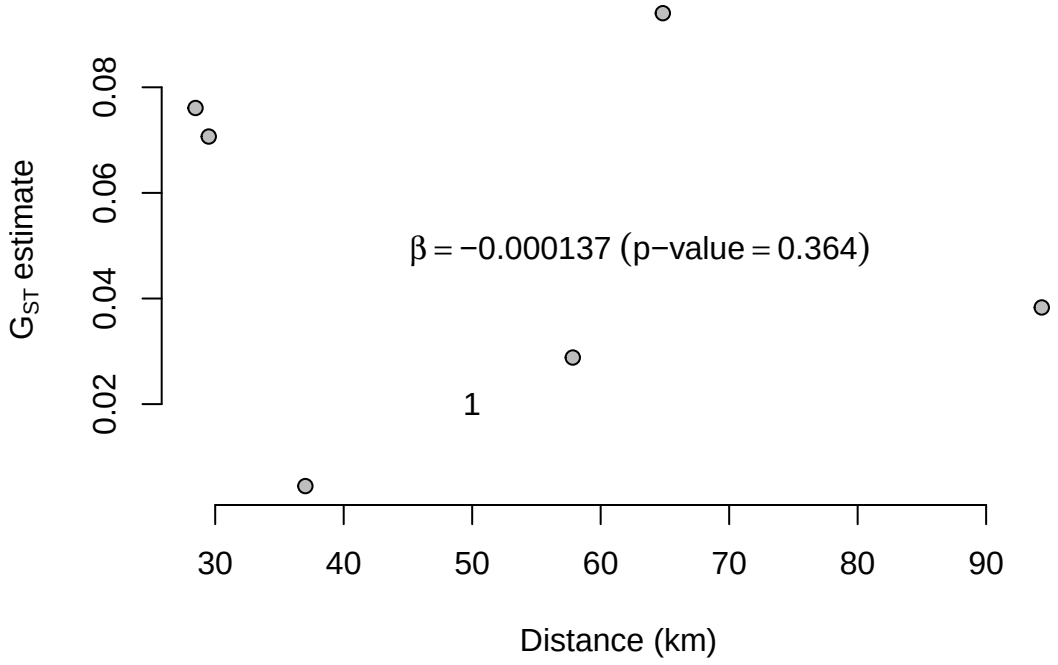

Figure 5: GST estimates and distance: WGS data, no transformation.

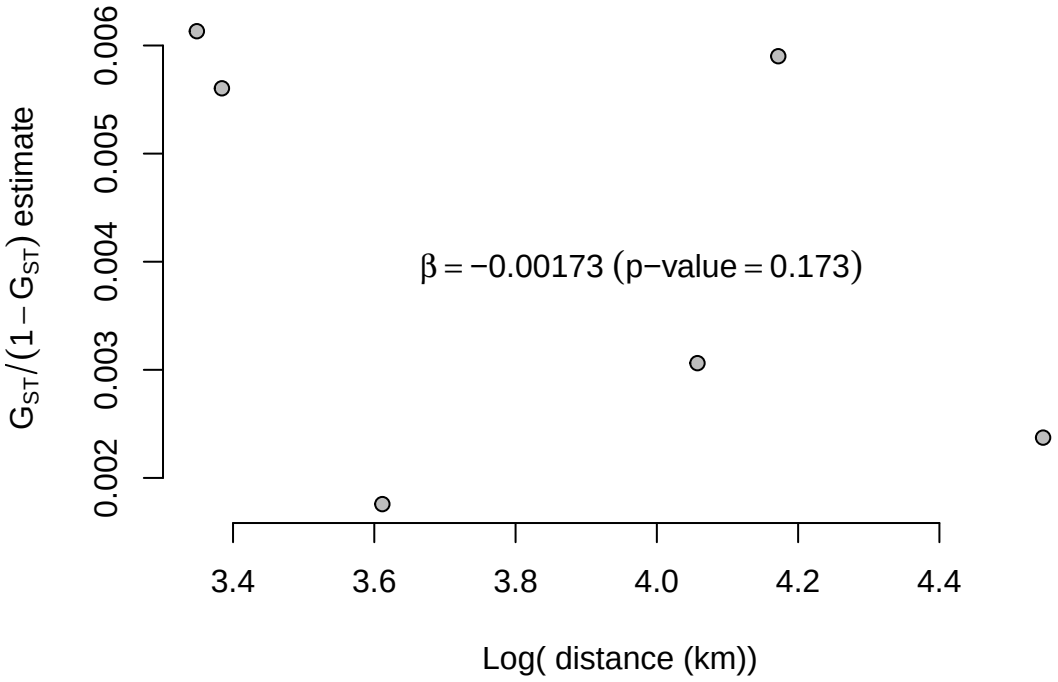

Figure 6: GST estimates and distance: barcode data, using the transformation predicted to give a linear relationship under a two-dimensional isolation by distance model (Rousset 1997).

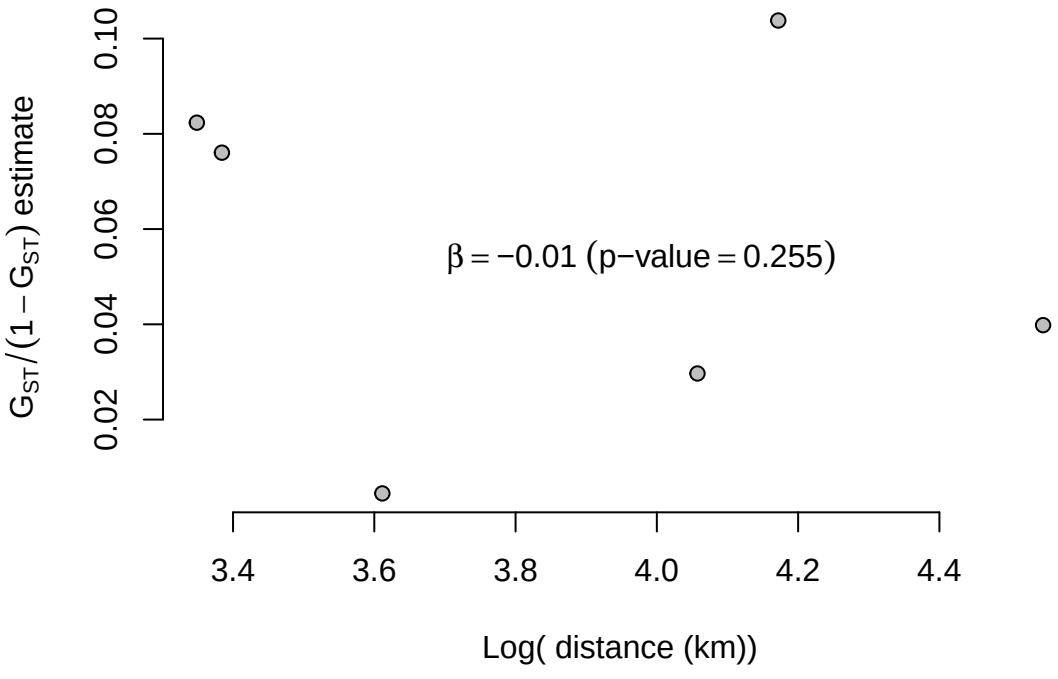

Figure 7: GST estimates and distance: WGS data, using the transformation predicted to give a linear relationship under a two-dimensional isolation by distance model (Rousset 1997).

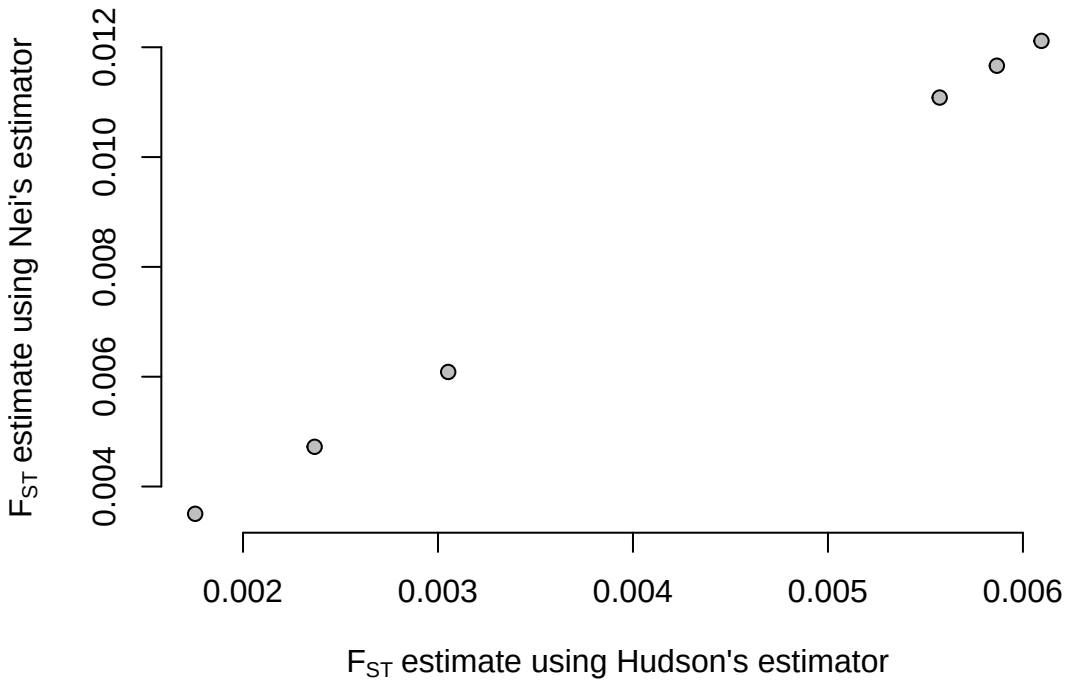

Figure 8: FST estimates: barcode data.

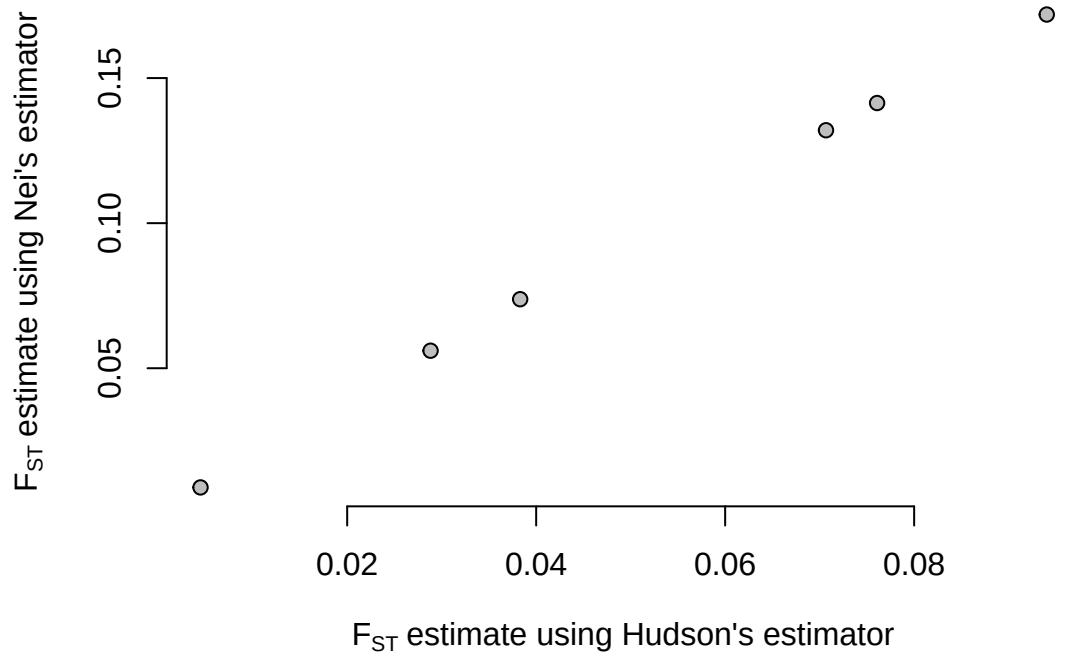

Figure 9: FST estimates: WGS data.

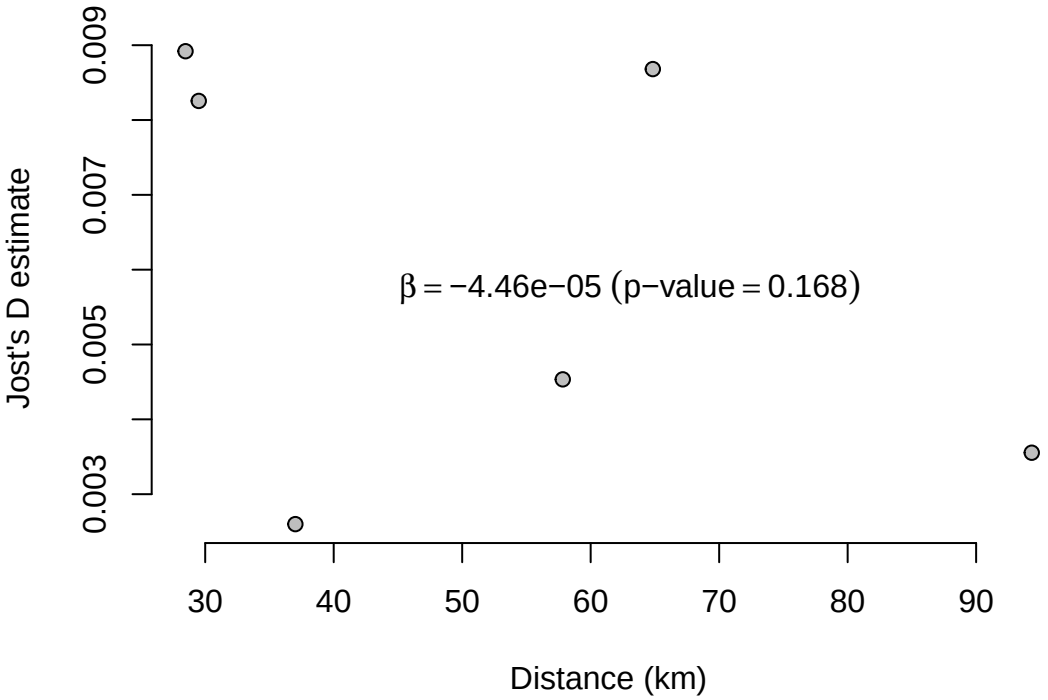

Figure 10: Jost's D estimates and distance: barcode data, no transformation.

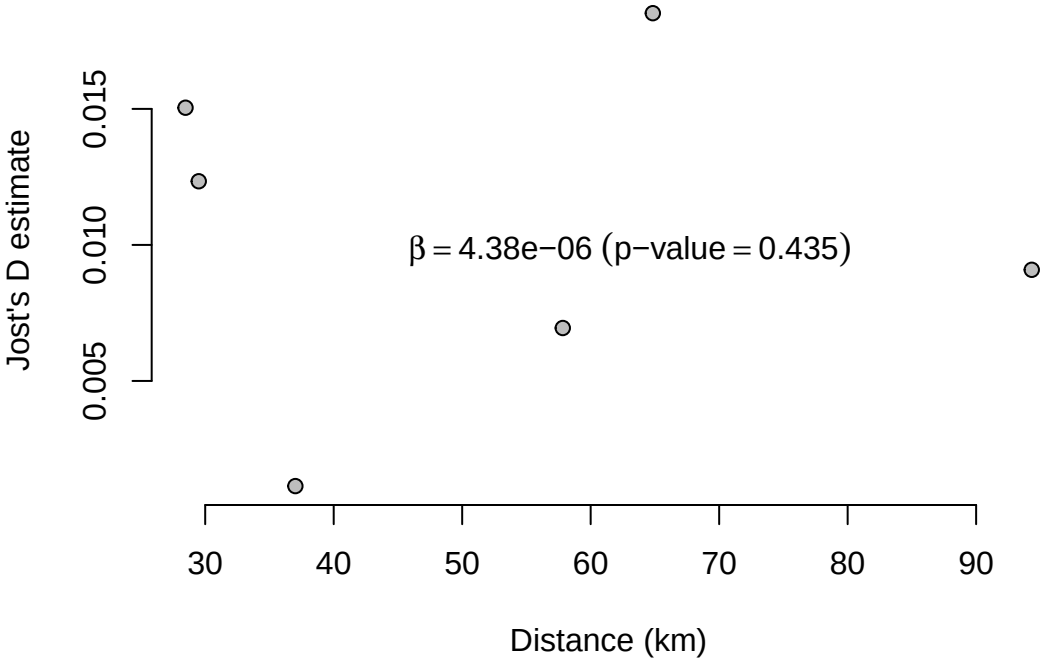

Figure 11: Jost's D estimates and distance: WGS data, no transformation.

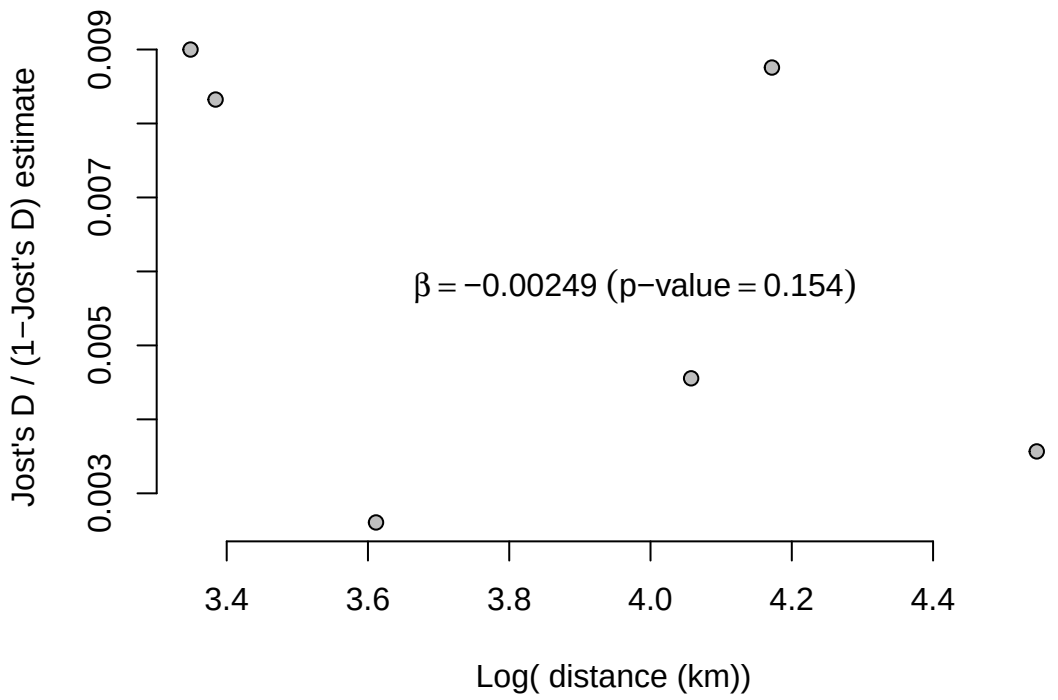

Figure 12: Jost's D estimates and distance: barcode datausing the transformation predicted to give a linear relationship under a two-dimensional isolation by distance model (Rousset 1997).

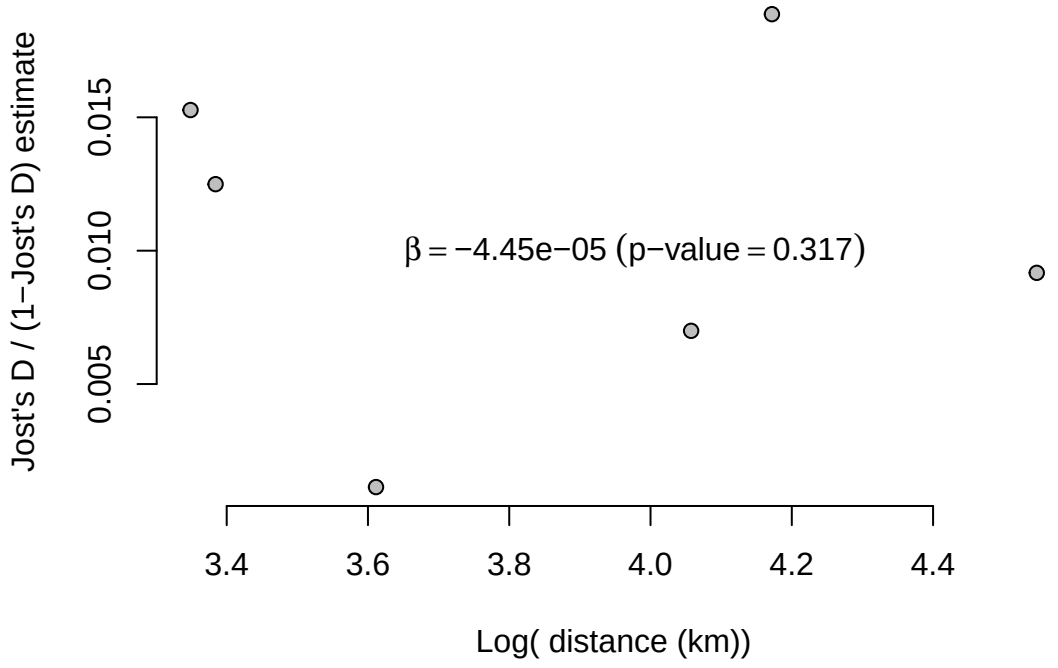

Figure 13: Jost's D estimates and distance: WGS data using the transformation predicted to give a linear relationship under a two-dimensional isolation by distance model (Rousset 1997).

References

Bhatia, Gaurav, Nick Patterson, Sriram Sankararaman, and Alkes L Price. 2013. "Estimating and interpreting F ST: The impact of rare variants." *Genome Research* 23: 1514–21. doi:10.1101/gr.154831.113.1514.

Hedrick, Philip W. 2005. "A standardized genetic differentiation measure." *Evolution* 59 (8): 1633–8.

Hudson, R. R., M. Slatkin, and W. P. Maddison. 1992. "Estimation of levels of gene flow from DNA sequence data." *Genetics* 132 (2): 583–89. doi:PMC1205159.

Jost, Lou. 2008. "GST and its relatives do not measure differentiation." *Molecular Ecology* 17: 4015–26. doi:10.1111/j.1365-294X.2008.03887.x.

Meirmans, Patrick G., and Philip W. Hedrick. 2011. "Assessing population structure : FST and related measures." *Molecular Ecology* 11: 5–18. doi:10.1111/j.1755-0998.2010.02927.x.

Nei, Masatoshi. 1973. "Analysis of Gene Diversity in Subdivided Populations." *Proceedings of the National Academy of Sciences* 70 (12): 3321–3.

Nei, Masatoshi, and R. K Chesser. 1983. "Estimation of fixation indices and gene diversities." *Annals of Human Genetics* 47 (3): 253–59. doi:10.1111/j.1469-1809.1983.tb00993.x.

Reich, David, Kumarasamy Thangaraj, Nick Patterson, Alkes L Price, and Lalji Singh. 2009. "Reconstructing Indian Population History." *Nature* 461 (7263): 489–94. doi:10.1038/nature08365.Reconstructing.

Rousset, François. 1997. "Genetic Differentiation and Estimation of Gene Flow from F-Statistics Under Isolation by Distance." *Genetics* 145 (4): 1219–28. doi:10.1002/ajmg.c.30221.

Weir, B. S., and W. G. Hill. 2002. "Estimating F-Statistics." *Annual Review of Genetics* 36 (1): 721–50. doi:10.1146/annurev.genet.36.050802.093940.

Weir, B.S., and C. Clark Cockerham. 1984. "Estimating F-Statistics for the Analysis of Population Structure." *Evolution* 38 (6): 1358–70.

Whitlock, Michael C. 2011. "G'ST and D do not replace FST." *Molecular Ecology* 20: 1083–91. doi:10.1111/j.1365-294X.2010.04996.x.

Willing, Eva Maria, Christine Dreyer, and Cock van Oosterhout. 2012. "Estimates of genetic differentiation measured by fst do not necessarily require large sample sizes when using many snp markers." *PLoS ONE* 7 (8): e42649. doi:10.1371/journal.pone.0042649.