**Supplemental Information**

**Implementation of a Stirling number estimator enables direct calculation of population genetics tests for large sequence data sets**

Swaine L. Chen

Department of Medicine, Division of Infectious Diseases, Yong Loo Lin School of Medicine, National University of Singapore, Singapore; and Infectious Diseases Group, Genome Institute of Singapore, Singapore

**Supplemental Methods**

To review from the main text, Fu’s Fs, can be calculated from a multiple sequence alignment. One requires the number of alleles (denoted ) and the mean number of pairwise nucleotide differences (denoted ). The statistic is then defined as:

(A1)

where and is the coefficient of in (Fu, 1997). The coefficients are also denoted in other literature as , where they are referred to as Stirling numbers of the first kind (Temme, 1993); hereafter I refer to these simply as Stirling numbers. Fu’s Fs is then defined as:

(A2)

Fu’s Fs is related to another statistic, Strobeck’s S (Fu, 1996; Strobeck, 1987), which is also frequently defined in terms of Stirling numbers. The from Fu’s Fs is the probability of having or more alleles in the sample; Strobeck’s S is the probability of having or fewer alleles (Fu, 1997), so that one useful property is:

(A3)

I use the Stirling number estimator developed by (Temme, 1993) (Equation 3.5 therein; see equations A6-A10 below), which has the advantage of being amenable to calculation as a logarithm and which is uniformly applicable for all values of . I then calculate Fu’s Fs directly using Equation (A1). Note that while the numerator and denominator can both become large, the ratio is typically within range for normal floating point mathematics; therefore, the numerator and denominator are estimated as logarithms and the sum may be calculated normally after taking the antilog of the difference.

The following software programs were used to calculate Fu’s Fs and Strobeck’s S or to deduce algorithmic approaches: DnaSP (versions 5.10.1 and 6.12.01) (Librado and Rozas, 2009; Rozas, et al., 2017), Arlequin (version 3.5.2.2) (Excoffier and Lischer, 2010), PopGenome (version 2.6.1) (Pfeifer, et al., 2014), and PGEToolbox (Cai, 2008). DnaSP and Arlequin were both run in Windows 7. PopGenome and all R code I developed were run on Ubuntu 16.04.5 with R (version 3.4.4) (<http://www.R-project.org/>). PGEToolbox was downloaded from the git repository (<https://github.com/jamesjcai/PGEToolbox>) and run in Matlab R2016a. Calculation methods were inferred from examining source code, where available. DnaSP uses a recursive algorithm for calculating Fu’s Fs that avoids the need for calculating Stirling numbers; this algorithm directly calculates Strobeck’s S and uses the identity in Equation (A3) to calculate Fu’s Fs as:

(A4)

Where

(A5)

The Stirling number estimator I used is Equation 3.5 from (Temme, 1993), modified here to match the notation used in (Fu, 1997):

(A6)

Where

(A7)

(A8)

(A9)

(A10)

And is the unique positive solution to the equation . Here the prime (‘) notation indicates the derivative. As noted in (Temme, 1993), these equations are readily converted to use logarithms for calculation using the gamma function and its derivatives.

I ported the methods used in PGEToolbox and DnaSP to R for comparison and benchmarking. The memoise package (Wickham, et al., 2017) was used to cache results from procedures for the ported DnaSP code. An arbitrary precision implementation of the PGEToolbox algorithm was coded in Perl (v5.22.1) using the bignum package (version 0.39).

Sample sequence alignments were extracted from assembled *E. coli* genomes with a completion level of “Complete”, “Chromosome”, or “Scaffold” available from Genbank as of March 6, 2016 (listed at https://www.ncbi.nlm.nih.gov/genome/genomes/167?). The annotated *fimH* gene from UTI89 (NC\_007946.1) (Chen, et al., 2006) was used to search each genome assembly using TBLASTN (version 2.2.28+) (Camacho, et al., 2009); only the top hit for each genome was kept. All TBLASTN hits that passed a cutoff of ≥ 90% identity over ≥ 90% of the length of the UTI89 gene allele were considered full length genes. The full length genes were translated *in silico*, and the resulting predicted protein sequences were aligned with ClustalW (version 2.1) (Larkin, et al., 2007); the protein alignment was then imposed on the DNA sequences. Gaps and polyallelic sites in the alignment were removed, then alleles were successively added in random order to create the alignments tested in Table 1.

Simulations for underflow were done for all combinations of , , and , where all three parameters were restricted to integers. Simulations for benchmarking and testing code (Figure 1A) were done by creating 100 parameter sets with , , and , where all values were randomly drawn from a uniform distribution, and was not restricted to integers. The benchmark for 100 parameter sets was run 50 times, each in a separate R session.

**Supplemental References**

Cai, J.J. (2008) PGEToolbox: A Matlab toolbox for population genetics and evolution, *J Hered*, **99**, 438-440.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L. (2009) BLAST+: architecture and applications, *BMC Bioinformatics*, **10**, 421.

Chen, S.L., Hung, C.S., Xu, J., Reigstad, C.S., Magrini, V., Sabo, A., Blasiar, D., Bieri, T., Meyer, R.R., Ozersky, P., Armstrong, J.R., Fulton, R.S., Latreille, J.P., Spieth, J., Hooton, T.M., Mardis, E.R., Hultgren, S.J. and Gordon, J.I. (2006) Identification of genes subject to positive selection in uropathogenic strains of Escherichia coli: a comparative genomics approach, *Proc Natl Acad Sci U S A*, **103**, 5977-5982.

Excoffier, L. and Lischer, H.E. (2010) Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows, *Mol Ecol Resour*, **10**, 564-567.

Fu, Y.X. (1996) New statistical tests of neutrality for DNA samples from a population, *Genetics*, **143**, 557-570.

Fu, Y.X. (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection, *Genetics*, **147**, 915-925.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J. and Higgins, D.G. (2007) Clustal W and Clustal X version 2.0, *Bioinformatics*, **23**, 2947-2948.

Librado, P. and Rozas, J. (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data, *Bioinformatics*, **25**, 1451-1452.

Pfeifer, B., Wittelsburger, U., Ramos-Onsins, S.E. and Lercher, M.J. (2014) PopGenome: an efficient Swiss army knife for population genomic analyses in R, *Mol Biol Evol*, **31**, 1929-1936.

Rozas, J., Ferrer-Mata, A., Sanchez-DelBarrio, J.C., Guirao-Rico, S., Librado, P., Ramos-Onsins, S.E. and Sanchez-Gracia, A. (2017) DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets, *Mol Biol Evol*, **34**, 3299-3302.

Strobeck, C. (1987) Average number of nucleotide differences in a sample from a single subpopulation: a test for population subdivision, *Genetics*, **117**, 149-153.

Temme, N.M. (1993) *Asymptotic estimates of Stirling numbers*.

Wickham, H., Hester, J., Mueller, K. and Cook, D. (2017) memoise: Memoisation of Functions.