HHS Public Access

Author manuscript

Nat Ecol Evol. Author manuscript; available in PMC 2019 November 14.

Published in final edited form as:

Nat Ecol Evol. 2018 February ; 2(2): 237–240. doi:10.1038/s41559-017-0425-y.

Evolutionary determinants of genome-wide nucleotide composition

Hongan Long[1,2], **Way Sung**[1,3], **Sibel Kucukyildirim**[1,4], **Emily Williams**[5], **Samuel F. Miller**[6], **Wanfeng Guo**[5], **Caitlyn Patterson**[6], **Colin Gregory**[6], **Chloe Strauss**[6], **Casey Stone**[6], **Cécile Berne**[6], **David Kysela**[6], **William R. Shoemaker**[6], **Mario E. Muscarella**[7], **Haiwei Luo**[8], **Jay T. Lennon**[6], **Yves V. Brun**[6], **Michael Lynch**[5,*]

[2]Institute of Evolution & Marine Biodiversity, Ocean University of China, Qingdao 266003, China

[3]Department of Bioinformatics and Genomics, University of North Carolina, Charlotte, North Carolina, USA 28223

[4]Department of Biology, Hacettepe University, Ankara, Turkey

[5]Center for Mechanisms of Evolution, PO Box 877701, Arizona State University, Tempe, Arizona, USA 85287

[6]Department of Biology, Indiana University, Bloomington, Indiana, USA 47405

[7]Department of Plant Biology, University of Illinois, Urbana-Champaign, Illinois, USA 61801

[8]School of Life Sciences and Partner State Key Laboratory of Agrobiotechnology, The Chinese University of Hong Kong, Shatin, Hong Kong

Abstract

One of the long-standing mysteries of evolutionary genomics is the source of the wide phylogenetic diversity in genome nucleotide composition (G+C vs. A+T), which must be a consequence of interspecific differences in mutation bias, the efficiency of selection for different nucleotides, or a combination of the two. We demonstrate that although genomic G+C composition is strongly driven by mutation bias, it is also substantially modified by direct selection and/or as a by-product of biased gene conversion. Moreover, G+C composition at four-fold redundant sites is consistently elevated above the neutral expectation, more so than for any other classes of sites.

For some classes of genomic sites, G+C nucleotide composition covers nearly the full range of possible variation (frequencies of ~0.0 to ~1.0) across species[1–5]. It is commonly thought that the contribution of mutation to such variation can be determined from the nucleotide

*Corresponding author: mlynch11@asu.edu.

[1]These co-first authors contributed equally to this work.

Author Contributions H.L., W.S., and M.L. conceived and designed the study, performed data analyses, and wrote the manuscript. All authors contributed in data collection and provided input to the manuscript.

Competing interests

The authors declare no competing interests.

content of four-fold redundant (synonymous) sites within codons or from the composition of rare variants, and analyses of this type have led to the idea that mutation is universally biased in the direction of A+T[6–8]. However, selection on such sites can bias such interpretations. To eliminate such issues, we use direct estimates of the mutation spectra derived from mutation-accumulation (MA) experiments and/or parent-offspring trios for 37 diverse species.

Of the data sets analyzed herein, 25 involve published data (summarized in Ref. [9] with respect to mutation rates), and 12 involve long-term MA experiments in diverse microbial species reported here for the first time (Supplementary Dataset 1: Tables 1–3). Each new MA experiment involves the complete genome sequencing of ~50 lines serially transferred through single-cell bottlenecks for thousands of cell divisions, which effectively eliminates the ability of natural selection to significantly modify the accumulation of all but the small fraction of extremely deleterious mutations (which in any case are irrelevant to the following analyses, as they do not accumulate evolutionarily; Ref. [9]). From the resultant spectra for base-substitution mutations (typically based on dozens to hundreds of *de novo* mutations), letting m be the ratio of the per-nucleotide mutation rate in the G+C → A+T direction to the reciprocal rate, the expected equilibrium G+C composition under neutrality (where mutation is the only directional evolutionary force) is

$$\widetilde{P}_n = \frac{1}{1+m} \quad (1)$$

Comparison of the observed genome-wide nucleotide compositions of the study species to these neutral expectations reveals several general patterns (Figure 1). First, mutation biases in unicellular species may be in either the A+T or G+C directions (leading to $\widetilde{P}_n$ less than or greater than 0.5, respectively), although the former is most common, and no characterized multicellular eukaryote has mutation bias in the G+C direction. Second, regardless of the class of DNA or the phylogenetic grouping, with few exceptions, genome-wide G+C composition is close to or substantially above the neutral expectation, implying that the existence of a near universal direction force(s) favoring G+C content. Third, the primary exception to this pattern involves 0-fold redundant sites (where all nucleotide substitutions lead to amino-acid changes) in bacteria with endogenous mutation pressure towards G+C ($\widetilde{P}_n > 0.5$), where selection for amino acids containing A+T in such codon positions apparently takes precedence over other G+C enhancing forces. This tendency is reflected in the diminished slope in the regression involving such sites (Supplementary Dataset 1: Table 4). Fourth, for 2- and 4-fold redundant sites (where 2 and 4 nucleotides encode for the same amino acid), G+C composition is particularly strongly elevated, by an average amount that is essentially independent of the neutral expectation, but with considerable variation. The strong elevation for 4-fold redundant sites implies the existence of general forces favoring G +C independent of the implications for the proteome.

The magnitude of the strength of selection required to account for the deviation of G+C composition at 4-fold redundant sites relative to the neutral expectation can be estimated by noting that in the presence of selection, Equation (1) generalizes to

$$P_s = \frac{1}{1 + me^{-S}} \quad (2)$$

where $S = \phi N_e s$, with N_e being the effective population size, $\phi = 2$ or 4 for haploids and diploids respectively, and s being the average selective advantage of G/C nucleotides over A/T[10–11]. S for each genomic category is shown in Supplementary Dataset 1: Table 5. In Figure 1, lines of expectation for P_s for various values of S (equivalent to the ratio of the power of selection s to the power of drift $1/(\phi N_e)$) show that S (in favor of G+C) is generally in the range of 0.5 to 4. Thus, some selective force in favor of G+C composition is pervasive and relatively strong, although not strong enough to entirely overcome the mutational expectations.

The results for four-fold redundant sites are of relevance to the common usage of measures of standing variation at such positions to estimate N_e under the assumption of neutrality (drift-mutation equilibrium), which leads to an expected average heterozygosity of $\tilde{\pi} \approx \phi N_e u$, where u is the mean mutation rate per nucleotide site. From a rearrangement of Equation (15) in Ref. [12], the ratio of heterozygosity under drift-mutation-selection equilibrium and the neutral expectation is

$$\frac{\pi_s}{\tilde{\pi}_n} = \frac{(1+m)\left(e^S - 1\right)}{S\left(m + e^S\right)} \quad (3)$$

solution of which shows that when mutation is strongly biased towards A+T but selection strongly favors G+C, the expected nucleotide diversity can be several-fold greater than the neutral expectation ($\tilde{\pi}$), which would lead to the same proportional overestimation of N_e when the mutation rate is factored out (Figure 2). When mutation bias and selection operate in the same direction, π can be downwardly biased up to a few-fold with respect to the neutral condition. Thus, relative estimates of N_e derived from silent-site variation can be off by several fold (when compared with each other) if selection is moderately strong and there are strong differences in mutation bias among contrasted species, which based on the wide range of estimated $\tilde{P}_n$ is clearly the case.

Our results imply a near universal pervasive mechanism operating to increase G+C content, as previously inferred indirectly from polymorphism data for G+C-rich genomes[7]. However, the sources of such selection remain unclear. Given that the substantial number of species in this study inhabit a wide range of environments and are derived from a diversity of bacterial and eukaryotic lineages, consistent directional selection in favor of G+C is not readily reconciled by ecological and/or genetic-background arguments. Moreover, given that such selection is experienced by both silent and replacement codon sites, arguments based on protein-sequence constraints and transcription fidelity are not compelling. Likewise, because the pattern extends to intergenic (largely noncoding DNA), arguments based on gene expression and translation speed/accuracy[13–14] do not seem to apply. Although gene expression levels within species are correlated with local gene G+C composition, all but one r^2 values involving these variables are ≤ 0.02, and the signs of the relationships are

inconsistent (Supplementary Dataset 1: Table 6). One general force that may be of relevance is DNA stability, in that G:C pairs involve three hydrogen bonds, whereas A:T pairs involve only two.

An alternative explanation for near universal pressure towards G+C content involves gene conversion, which results from the repair of heteroduplex DNA arising from recombination between two nonidentical sequences and if biased can operate like selection at the population-genetic level. In every organism that has been closely scrutinized, eukaryotes and bacteria, gene conversion has been found to be biased in the direction of G+C (Refs. 15-21), although the molecular mechanisms encouraging such universal behavior are unknown. Most attempts to estimate S associated with codon bias (which may be driven by biased gene conversion) have yielded estimates on the order of 0.1 to 4.0 in diverse phylogenetic groups[4] (although not always in the G+C direction), and our results (Figure 1) are fully compatible with this magnitude of selection.

Because effective population sizes vary among organisms by several orders of magnitude, this small range in S suggests that there must be a roughly inverse relationship between N_e and s, whatever the force encouraging G+C content. Under a scenario of natural selection, such a condition is expected under any concave fitness function for increasing G+C content, as the selective advantage of incremental changes would then diminish with increased G+C (further out on the fitness plateau), and larger population sizes would enhance the efficiency of selection for higher G+C content. However, a scenario of biased gene conversion requires a rather different set of conditions – the magnitude of the biasing force (towards G+C) would have to increase with decreasing N_e. In principle, this might occur if a large fraction of GC conversions were deleterious, as natural selection opposing conversion-driven GC would be reduced in the face of increased random genetic drift[9]. This would, however, also require a very strong increase in the biasing force in small populations because biased gene conversion depends on both the asymmetric force and the recombination rate per nucleotide site, with the latter actually scaling negatively with N_e (Ref. 4).

In summary, our results conclusively support the idea that genome-wide nucleotide composition is strongly influenced by mutation bias at all classes of sites, but that phylogenetically general directional forces beyond mutation (natural selection and/or biased gene conversion) play a role as well. The positive association between neutral G+C-composition expectations and actual utilization at 0-fold redundant sites demonstrates that even amino-acid usage is dictated at least in part by mutation pressure, with the G+C content of such sites differing more than two-fold between genomes with strong mutation bias towards A+T vs. those with bias towards G+C (Figure 1; Supplementary Dataset 1: Table 1). However, despite this gradient, G+C utilization at 0-fold redundant sites is generally substantially greater than the neutral expectation when the latter is < 0.5, so the possibility that such content is influenced by the same selection pressures favoring G+C content at silent sites cannot be ruled out.

Finally, although the ultimate sources of variation in the mutation spectrum (which drives the wide range of variation in nucleotide composition among species) are unknown, they may involve effectively neutral processes. Owing to the predominance of deleterious

mutations, selection is expected to generally drive the genome-wide mutation rate down to some level beyond which further advantages are offset by the power of random genetic drift[9]. However, any particular mutation rate can be compatible with a wide-range of mutational spectra, which may be free to wander over evolutionary time, conditional on the maintenance of a constant genome-wide deleterious rate[22]. Notably, when the prevailing mutation pressure towards A+T is in conflict with the forces favoring G+C content (which is true for most taxa), the average genome-wide mutation rate per nucleotide site will be indirectly inflated, owing to the elevated abundance of more mutable (G and C) nucleotides.

Methods

G/C composition calculation

Mutation spectra, strain culturing, and reference-genome information for the 37 species in this study are presented in Supplementary Dataset 1: Table 1. We enumerated all sites of the genomes to calculate genome-wide G+C nucleotide compositions. For the G/C nucleotide composition at different functional sites of the genomes, we parsed out: 1) the second nucleotides of all codons except stop codons to delineate 0-fold redundant sites; 2) the third nucleotides of codons for Asn, Asp, Cys, Gln, Glu, His, Lys, and Tyr amino acids for 2-fold redundant sites; 3) the third nucleotides of codons for Ala, Arg, Gly, Leu, Pro, Ser, Thr, and Val amino acids for 4-fold redundant sites; and 4) the nucleotides between the start and/or stop codons (or between UTRs when annotated) of two adjacent genes for intergenic DNA. Expression data of each gene were downloaded from the NCBI GEO database and the gene-specific G+C contents for 4-fold redundant sites were parsed as above. The statistical details of the relationship between gene expression and 4-fold redundant site G/C composition are in Supplementary Dataset 1: Table 6.

MA line transfers

For the 12 new microbial mutation-accumulation (MA) projects reported in this study, all lines were cultured under ideal conditions on solid agar plates, using procedures relied on in numerous previous studies summarized in Lynch et al.[9] Within each study, all MA lines initiated from a single-cell progenitor, and were then single-cell transferred daily to weekly (depending on the growth rate, necessary for visual localization of colonies). Each month, numbers of cell divisions during each culturing cycle were estimated using colony-forming units from serial-dilution procedures.

Genome sequencing and raw data

Genomic DNA of the MA lines was extracted using the Wizard Genomic Purification Kit (Promega, Inc.). Illumina libraries for genome sequencing were then constructed using an optimized Nextera DNA library prep kit (Illumina Inc.), and 150 or 250 bp paired-end Illumina sequencing was done on a Hiseq2500 platform (Hubbard Center for Genome Studies, University of New Hampshire). Read trimming, mapping, and mutation rate calculations followed Long et al.[23] Duplicate reads were removed using picard-tools-2.5.0 in GATK 3.6. Unique SNP and indel variants were analyzed with HaplotypeCaller, and standard hard-filtering parameters described by GATK Best Practices recommendations[24–26]. Candidate variants were identified visually with the Integrated

Genome Viewer (IGV v. 2.3.5)[27]. All base-substitution and insertions/deletions identified are in Supplementary Dataset 1: Tables 2, 3.

Data availability

Raw reads of genome sequencing generated in this study are available in the National Center for Biotechnology Information Sequence Read Archive with BioProject no. PRJNA376572.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Support was provided by the Multidisciplinary University Research Initiative awards W911NF-09-1-0444 and W911NF-14-1-0411 from the US Army Research Office to ML; National Institutes of Health awards R01-GM036827 and R35-GM122566 to ML; "Zhufeng Talent Program" startup grant 861701013155 from Ocean University of China to H.L.; R01-GM51986 and R35-GM122556 to YVB, and F32-GM083581 to DTK; and National Science Foundation grant DOB 1442246 to JTL. We thank T. G. Doak, P. Keightley, K. Morris, R. Ness, I. Ruiz-Trillo, S. Simpson, and W. K. Thomas, A. Uchimura, and Z. Ye for providing strains and/or technical help in data acquisition. We thank L. Duret for helpful comments.

Literature Cited

1. Sueoka N. Proc Natl Acad Sci USA. 48:582–592.1962; [PubMed: 13918161]
2. Gu X, Hewett-Emmett D, Li WH. Genetica. 102–3:383–391.1998;
3. Chen SL, Lee W, Hottes AK, Shapiro L, McAdams HH. Proc Natl Acad Sci USA. 101:3480–3485.2004; [PubMed: 14990797]
4. Lynch, M. The origin of genome architecture. Sinauer Associates, Inc; 2007.
5. Rocha EPC, Feil EJ. PLoS Genet. 6:e1001104.2010; [PubMed: 20838590]
6. Hershberg R, Petrov DA. PLoS Genet. 6:e1001115.2010; [PubMed: 20838599]
7. Hildebrand F, Meyer A, Eyre-Walker A. PLoS Genet. 6:e1001107.2010; [PubMed: 20838593]
8. Lynch M. Proc Natl Acad Sci USA. 107:961–968.2010; [PubMed: 20080596]
9. Lynch M, et al. Nat Rev Genet. 17:704–714.2016; [PubMed: 27739533]
10. Li WH. J Mol Evol. 24:337–345.1987; [PubMed: 3110426]
11. Bulmer M. Genetics. 129:897–907.1991; [PubMed: 1752426]
12. McVean GAT, Charlesworth B. Genet Res. 74:145–158.1999;
13. Raghavan R, Kelkar YD, Ochman H. Proc Natl Acad Sci USA. 109:14504–14507.2012; [PubMed: 22908296]
14. Kelkar YD, Phillips DS, Ochman H. G3. 5:1247–1252.2015; [PubMed: 25897009]
15. Marais G, Mouchiroud D, Duret L. Genet Res. 81:79–87.2003; [PubMed: 12872909]
16. Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM. Nature. 454:479–485.2008; [PubMed: 18615017]
17. Galtier N, Duret L, Glemin S, Ranwez V. Trends Genet. 25:1–5.2009; [PubMed: 19027980]
18. Pessia E, et al. Genome Biol Evol. 4:675–682.2012; [PubMed: 22628461]
19. Williams AL, et al. Elife. 4:e04637.2015;
20. Mugal CF, Weber CC, Ellegren H. Bioessays. 37:1317–1326.2015; [PubMed: 26445215]
21. Lassalle F, et al. PLoS Genet. 11:e1004941.2015; [PubMed: 25659072]
22. Lynch M. Proc Natl Acad Sci USA. 109:18851–18856.2012; [PubMed: 23115338]
23. Long H, et al. Genome Biol Evol. 8:3815–3821.2016; [PubMed: 28173099]
24. McKenna A, et al. Genome Res. 20:1297–1303.2010; [PubMed: 20644199]
25. DePristo MA, et al. Nat Genet. 43:491–498.2011; [PubMed: 21478889]

26. Van der Auwera GA, et al. Curr Protoc Bioinformatics. 43(11):10–33. 1–33.2013; [PubMed: 25431634]

27. Thorvaldsdottir H, Robinson JT, Mesirov JP. Brief Bioinform. 14:178–192.2013; [PubMed: 22517427]

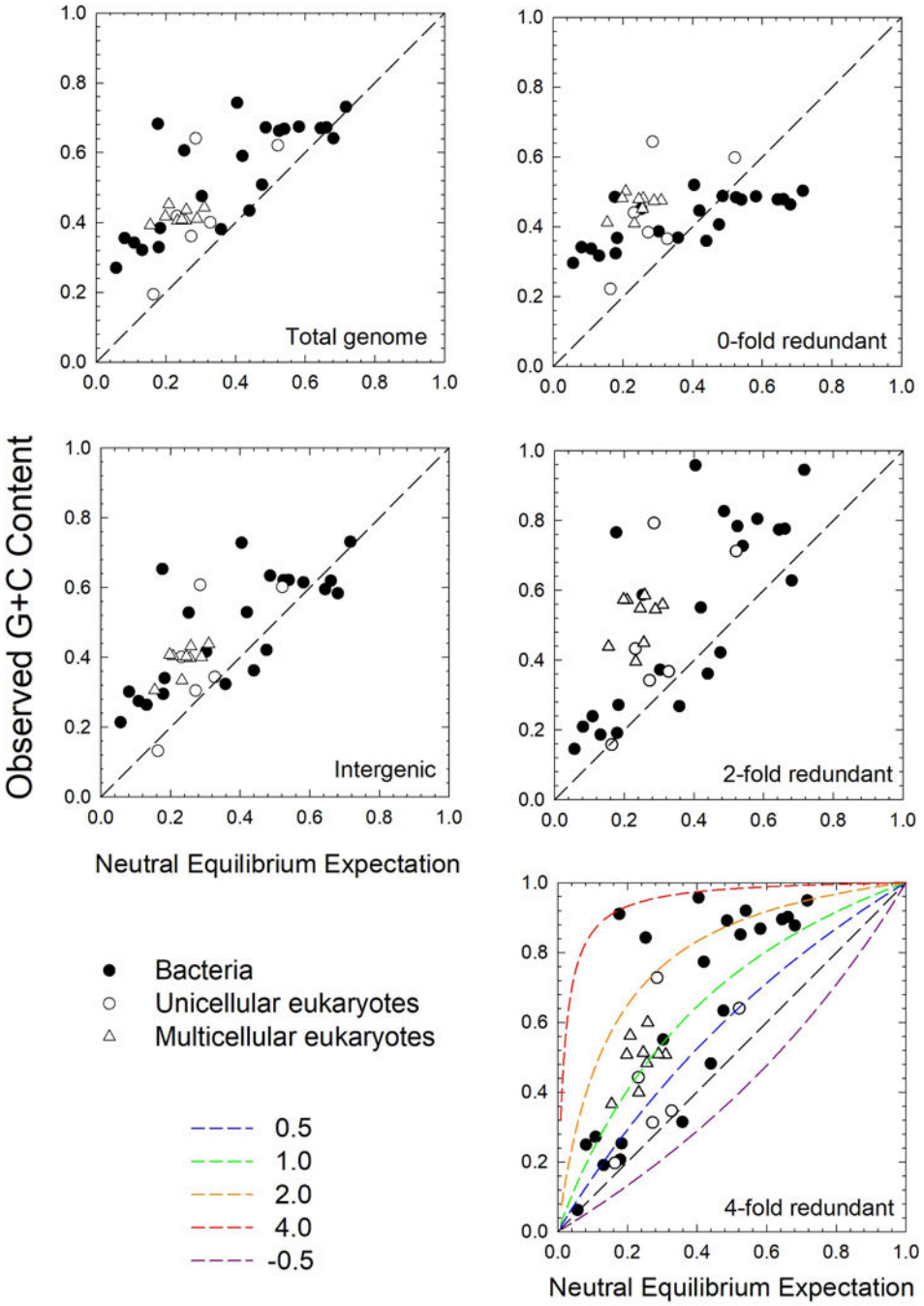

Figure 1. Relationship between genome-wide nucleotide composition and the neutral expectation

The data are subdivided into three major groups of organisms. The diagonal dashed lines denote agreement with the neutral expectation, with points above the diagonal reflecting conditions in which there is selection for elevated G+C content. For reference, the lower panel provides isoclines of expected genome compositions under selection, with values of the composite parameter $S = \phi N_e s$ being equivalent to the ratio of the power of selection in favor of G+C content relative to the power of genetic drift. The neutral equilibrium expectation is calculated from Equation (1), and the observed G+C content is based on direct

observation of genome contents. All data can be found in Supplementary Dataset 1: Tables 1–5.

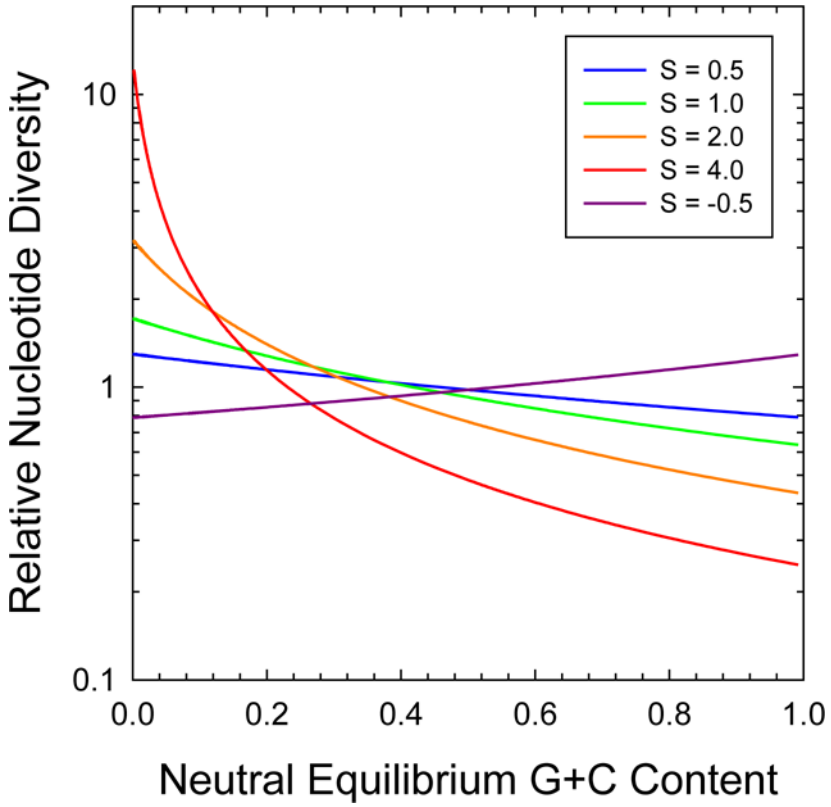

Figure 2. Expected equilibrium levels of within-population nucleotide diversity scaled by the neutral expectation

Derived from Equation (3) in the text, with various strengths of selection (*S*) color coded as in Figure 1.