

Supporting Information

Cui et al. 10.1073/pnas.0905818106

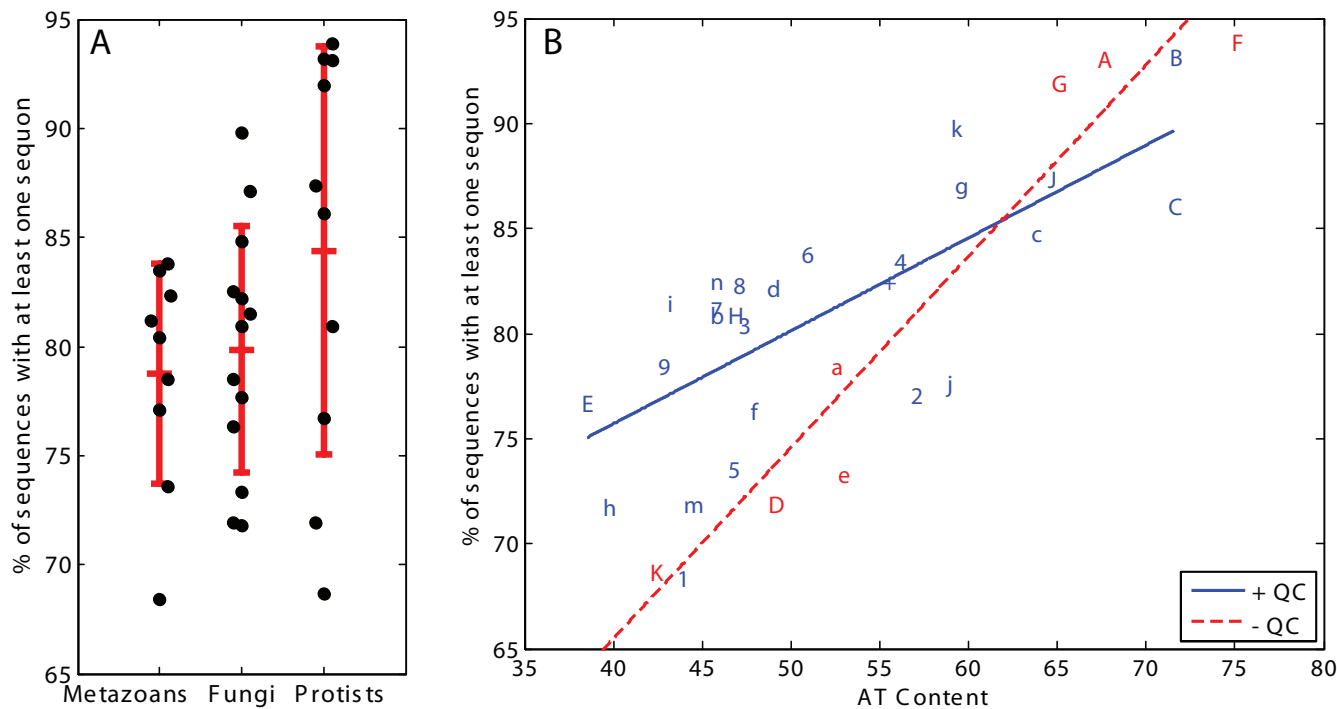


Fig. S1. This figure, which complements Fig. 1, shows the percentage of secreted proteins with at least one sequon is positively correlated with AT content. (A) The percentage of secreted and membrane proteins with at least 1 sequon are fairly similar among metazoans and fungi but are much more variant among protists. (B) The percentage of secreted and membrane proteins with at least 1 sequon is positively correlated with the AT content of these proteins in all eukaryotes, which are abbreviated as in Table S1. In organisms without N-glycan-dependent QC of folding (colored red), where there is no apparent positive selection for sequons (see Fig. 2), this relationship closely fits a line with a slope of 1. For these organisms, a 10% increase in AT content predicts a 10% increase in the number of secreted and membrane proteins with at least 1 sequon. In contrast, in organisms with N-glycan-dependent QC of protein folding (colored blue), where there is positive selection for sequons (again, see Fig. 2), the relationship between AT and percentage of secreted proteins with at least 1 sequon is not so linear, and the slope is less steep.

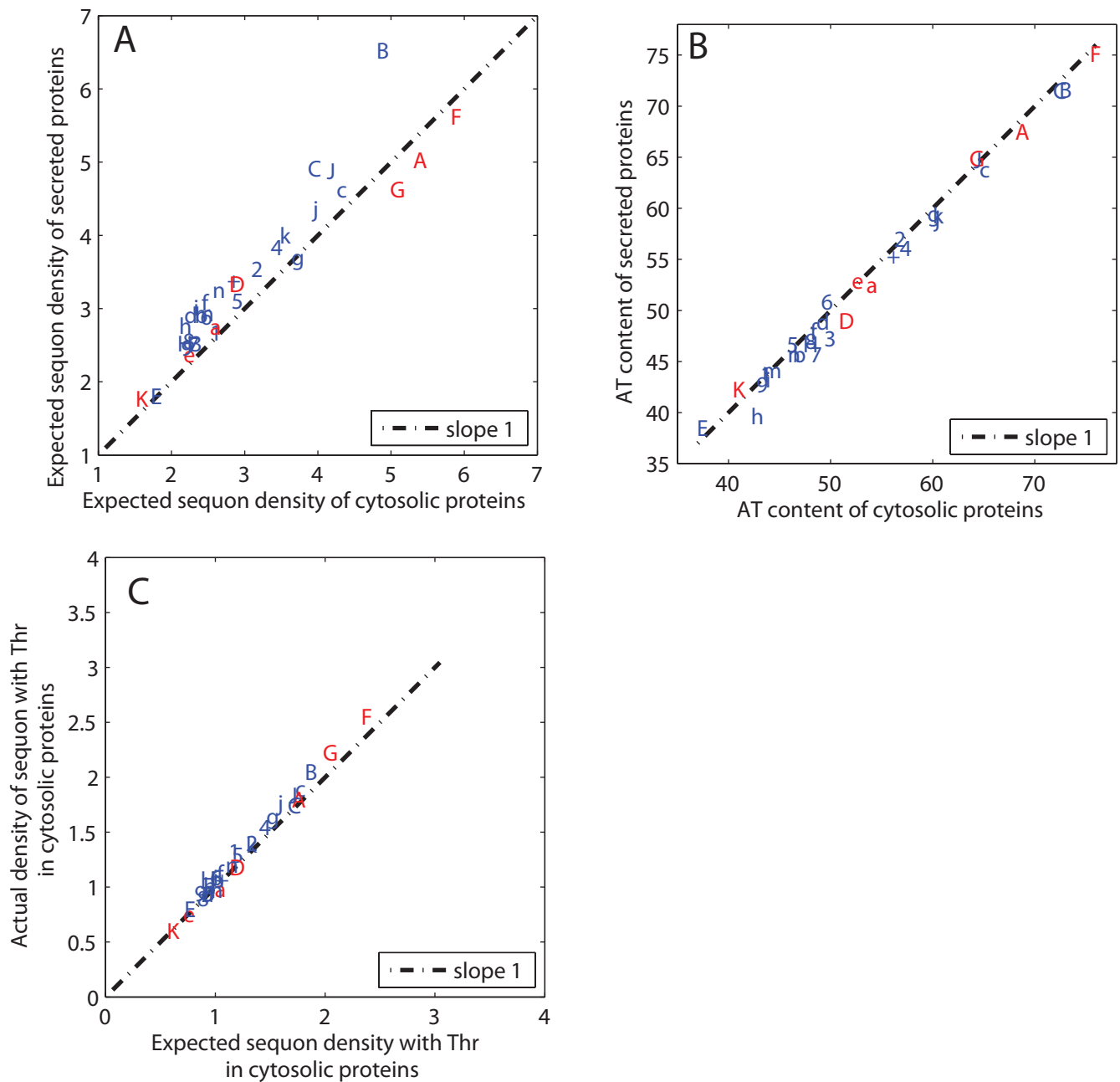


Fig. S2. This figure, which complements Fig. 2, shows that amino acid bias and AT content contribute little to positive selection for sequons with Thr. (A) There is little contribution of amino acid composition bias, which is determined by comparing the expected sequon densities of secreted and cytosolic proteins, to positive selection for sequons with Thr in organisms with N-glycan-dependent QC (colored blue and abbreviated as in Table S1). Therefore, all of the points fall on the dotted line with the slope of 1. Similarly, there is no amino acid composition bias in secreted proteins of organisms that lack N-glycan-dependent QC (colored red and abbreviated as in Table S1). (B) There is no difference in the AT content of the secreted and cytosolic proteins of any eukaryote, so that all of the points fall on the dotted line with a slope of 1. (C) The actual and expected sequon densities with Thr are the same for cytosolic proteins of all eukaryotes. This is a negative control for Fig. 2C.

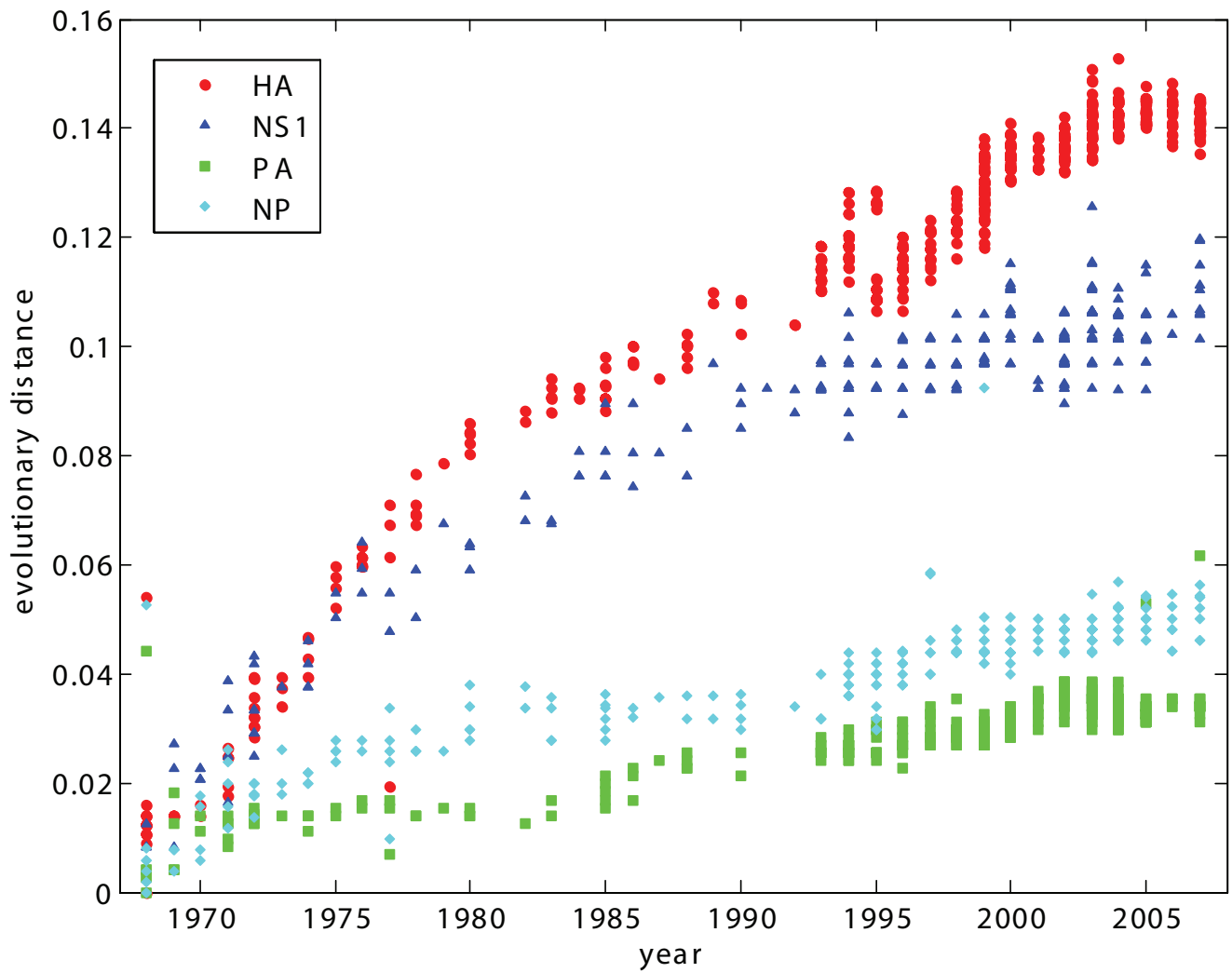


Fig. S3. This figure, which complements Fig. 3, shows linear changes in amino acid sequences of influenza virus A/H3N2 proteins. Here, the rate of change is greatest for HA, which in turn is greater than that for NS1 (a nonstructural protein), PA (a component of the viral polymerase), and NP (nucleoprotein). The evolutionary distance was calculated based on Jones–Taylor–Thornton matrix and 1 category of substitution rates.

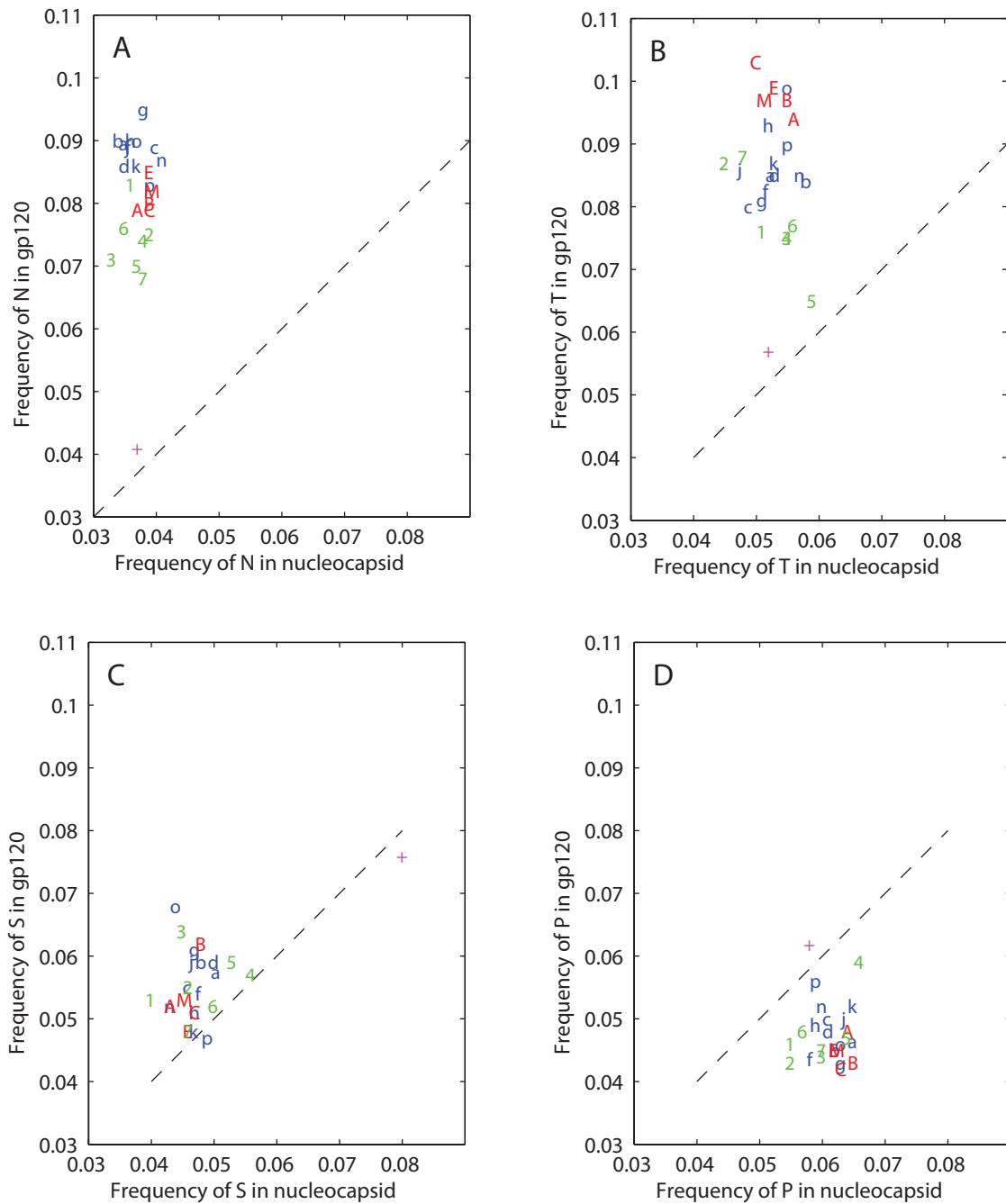


Fig. S4. Amino acid composition bias, which contributes to the high density of sequons in gp120 of HIV and other retroviruses (abbreviated as in Fig. 4), is broken down here for individual amino acids. Amino acid composition bias (determined by comparing the frequency of each amino acid in gp120 vs. capsid proteins and other viral enzymes) is greatest for Asn (A) and Thr (B), so that most points are well above the dotted line with the slope of 1. In contrast, this bias is less for Ser (C), whereas there is negative selection against Pro (D), which cannot be in sequons.

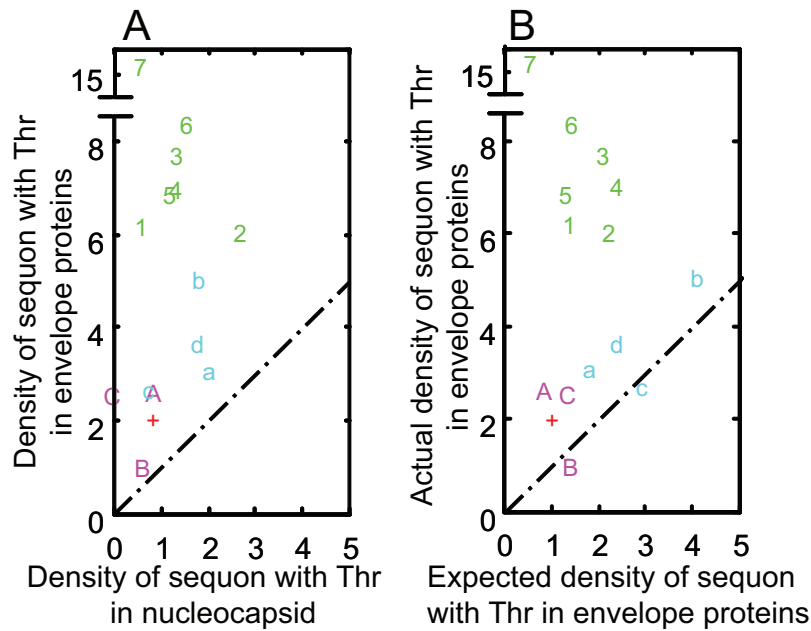


Fig. S6. Viral pathogens show a broad range of sequon densities in their envelope glycoproteins and multiple mechanisms for selecting sequons (human host proteins are marked with a plus sign). (A) Viruses that have strong selection for sequons in their envelope glycoproteins are marked in green numbers. These include HCV (1), SARS virus (2), influenza A/H3N2 virus (3), Ebola Reston (4), Ebola Sudan (5), Ebola Zaire (6), and HIV subtype B (7). Viruses that have moderate selection for N-glycans in their envelope glycoproteins are marked in light blue, lowercase letters: hepatitis b virus (a), human respiratory syncytial virus strain B1 (b), mumps virus (c), and Vaccinia virus WR (d). Viruses with little or no selection for N-glycans in their envelope glycoproteins are marked in pink, uppercase letters: human Herpes virus 1 (A), measles viruses (B), and yellow fever virus (C). Human host proteins are marked with an asterisk. (B) For viruses with the strongest selection for sequons in their envelope proteins (colored green), there is an increased likelihood that Asn, Ser, and Thr will be present in sequons rather than elsewhere. Because of this selection, the actual density of sequons is much greater than the calculated density of sequons for these viral envelope proteins.

Other Supporting Information Files

[Table S1 \(PDF\)](#)