Supporting Information

Yoffe *et al.* 10.1073/pnas.0808089105

SI Text

Calculation of Z scores. Based on our analysis of random ssRNAs of lengths 2,500 through 7,000 nt with RNAsubopt, we find that $Log(\overline{\langle MLD \rangle})$ and the log of the standard deviation of $\overline{\langle MLD \rangle}$, $Log(\sigma(\overline{\langle MLD \rangle}))$, both scale linearly with sequence length (N) over this range ($R^2 = 0.993$ and 0.971, respectively), yielding the following predictive equations for these values as a function of N:

$$\overline{\langle MLD \rangle} = 1.37\ N^{0.67} \quad [1]$$

$$\sigma(\overline{\langle MLD \rangle}) = 0.122\ N^{0.71} \quad [2]$$

With these, one can calculate predicted values of $\overline{\langle MLD \rangle}$ and $\sigma(\overline{\langle MLD \rangle})$ for ssRNAs of arbitrary length. Thus, for each individual viral ssRNA, its $\langle MLD \rangle$ was compared with the $\overline{\langle MLD \rangle}$ predicted for random sequences of the same length. This difference was then divided by the value of $\sigma(\overline{\langle MLD \rangle})$ predicted for random sequences of that length, yielding a Z score for that individual viral sequence. The Z scores of all viral ssRNAs within each group were then averaged and presented in Table 1.

The same approach was used to determine the Z scores of the $\langle ALD \rangle$ values calculated with RNAfold, also presented in Table 1. Here, the following predictive equations were used:

$$\overline{\langle ALD \rangle} = 0.485\ N^{0.68} \quad [3]$$

$$\sigma(\overline{\langle ALD \rangle}) = 0.0373\ N^{0.72} \quad [4]$$

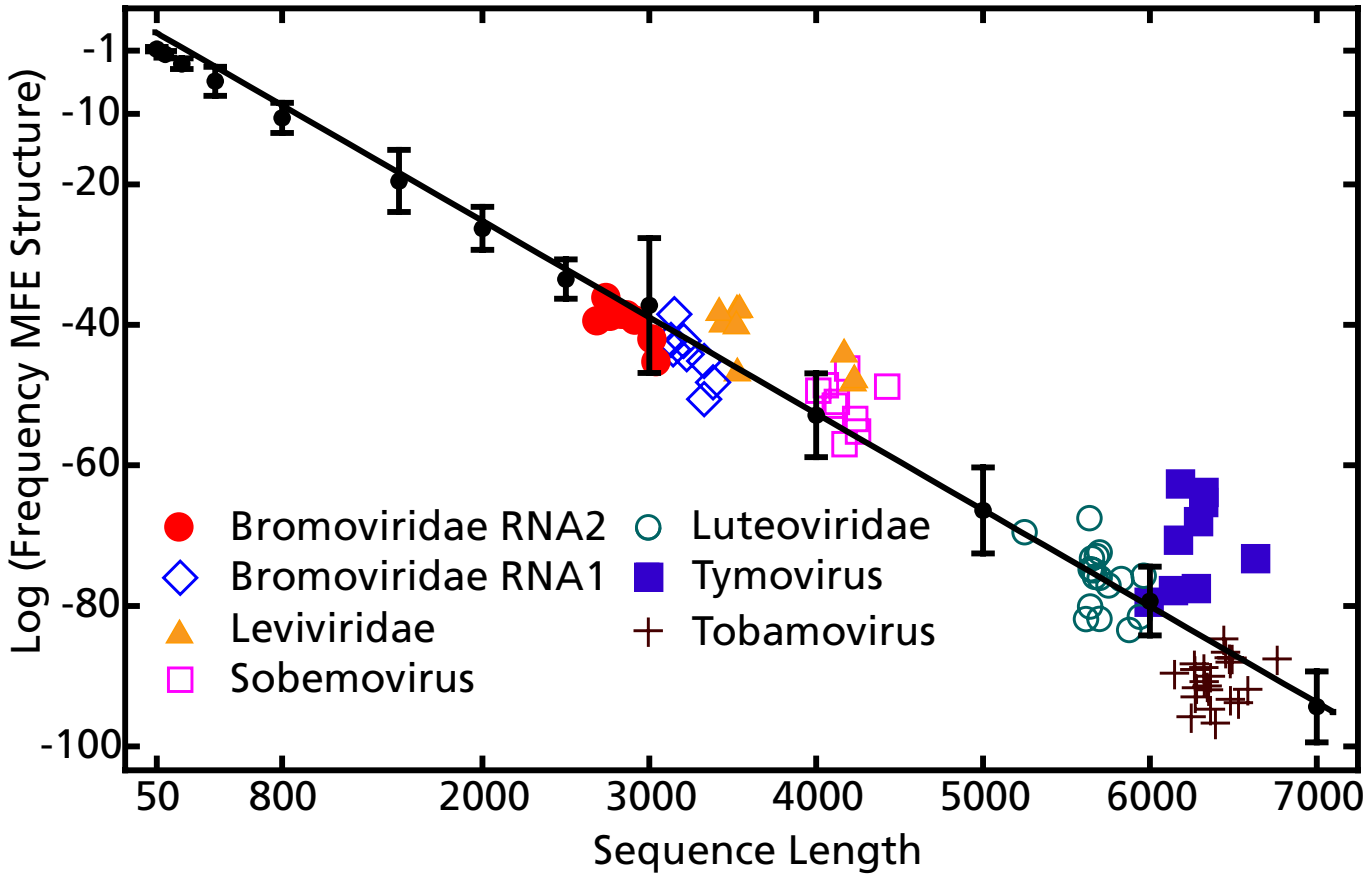

Fig. S1. Log (frequency MFE structure) vs. sequence length, for viral and randomly permuted ssRNAs. 'Frequency MFE structure' is the Boltzmann-weighted probability of the occurrence of the MFE (minimum free energy) structure within the ensemble. The viral ssRNAs are identified by the symbols listed in the *Inset*. The Bromoviridae analyzed here are from the Bromovirus and Cucomovirus genera. The straight line is a least-squares fit to the average values computed for random sequences 50-7,000 nt in length; the vertical lines show the standard deviations. The slope of the regression line indicates that the frequency of occurrence of the MFE structure $\sim 10^{-0.01N}$. Values were calculated with RNAfold.

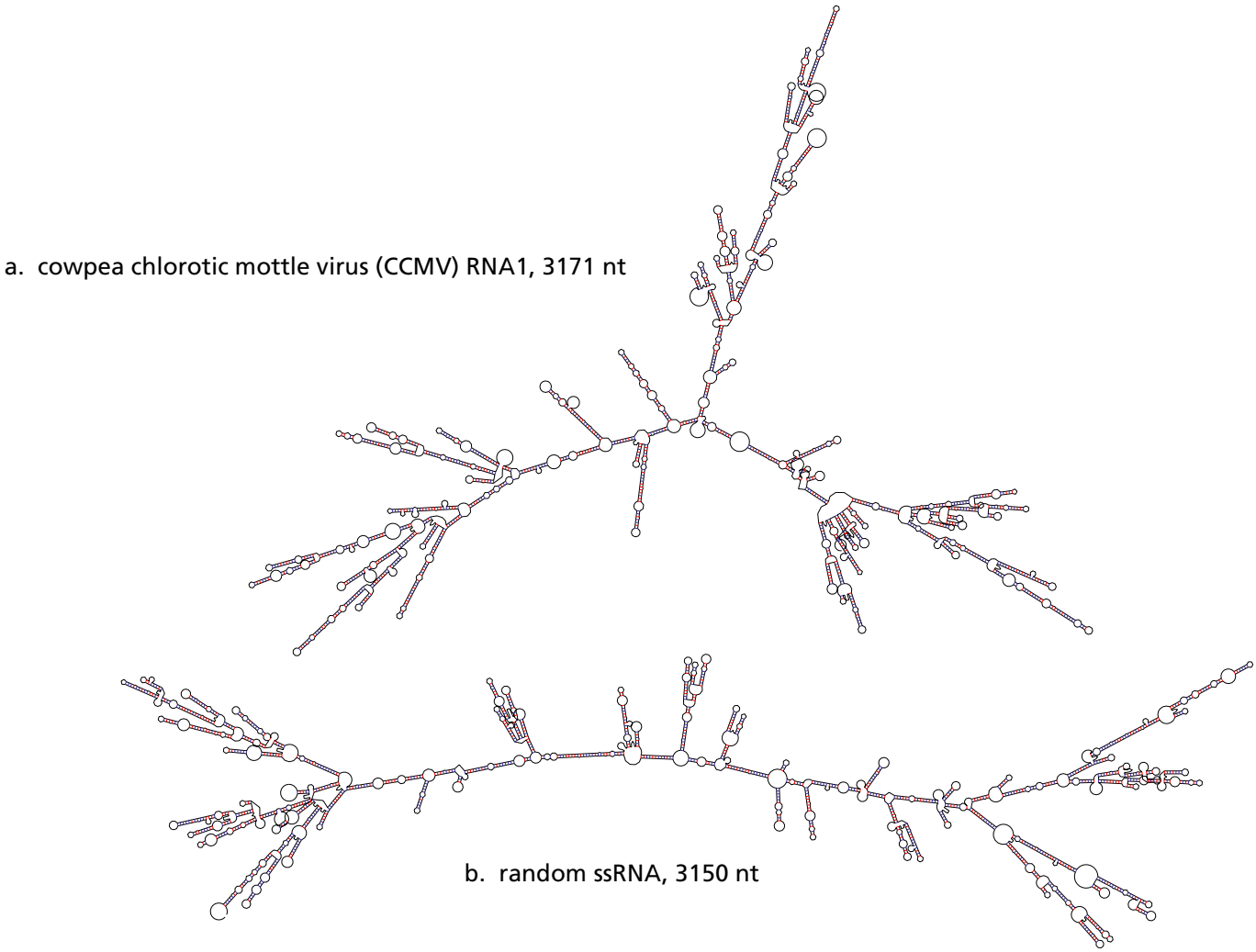

Fig. S2. Secondary structures of representative approximately equal-length viral and random ssRNAs shown to the same scale. (*A*) RNA1 of cowpea chlorotic mottle virus. The MLD of this secondary structure is equal to the ⟨MLD⟩ for this RNA (246), giving it a *Z* score of −1.8; the average *Z* score for all of the Bromoviridae RNA1 ssRNAs is −1.4. (*B*) Randomly permuted ssRNA. The MLD of this secondary structure is equal to the ⟨MLD⟩ predicted for random ssRNAs of that length (313). ⟨MLD⟩ values were calculated with RNAsubopt and figures were drawn with mfold.

a. ononis yellow mosaic virus ssRNA, 6211 nt

b. random ssRNA, 6250 nt

c. streptocarpus flower break virus ssRNA, 6279 nt

Fig. S3. Secondary structures of representative approximately equal-length viral and random ssRNAs, shown to the same scale. (*A*) Ononis yellow mosaic virus, a Tymovirus. The MLD of this secondary structure (308) is almost equal to the ⟨MLD⟩ for this RNA (310), giving it a *Z* score of −2.9; the average *Z* score for all Tymovirus ssRNAs is −2.8. (*B*) Randomly permuted ssRNA. The MLD of this secondary structure is equal to the $\overline{\langle \mathrm{MLD} \rangle}$ predicted for random ssRNAs of that length (497). (*C*) Streptocarpus flower break virus, a Tobamovirus. The MLD of this secondary structure (539) is almost equal to the ⟨MLD⟩ for this RNA (541), giving it a *Z* score of +0.7; the average *Z* score for all Tobamovirus ssRNAs is +0.6. Note the striking difference in extendedness between the secondary structures of the Tymovirus and random ssRNAs, as contrasted with the similarity between the Tobamovirus and random ssRNAs. Tymovirus ssRNAs fit into icosahedral capsids of fixed size, while Tobamovirus ssRNAs fit into rod-shaped capsids of variable length. ⟨MLD⟩ values were calculated with RNAsubopt, and figures were drawn with mfold.

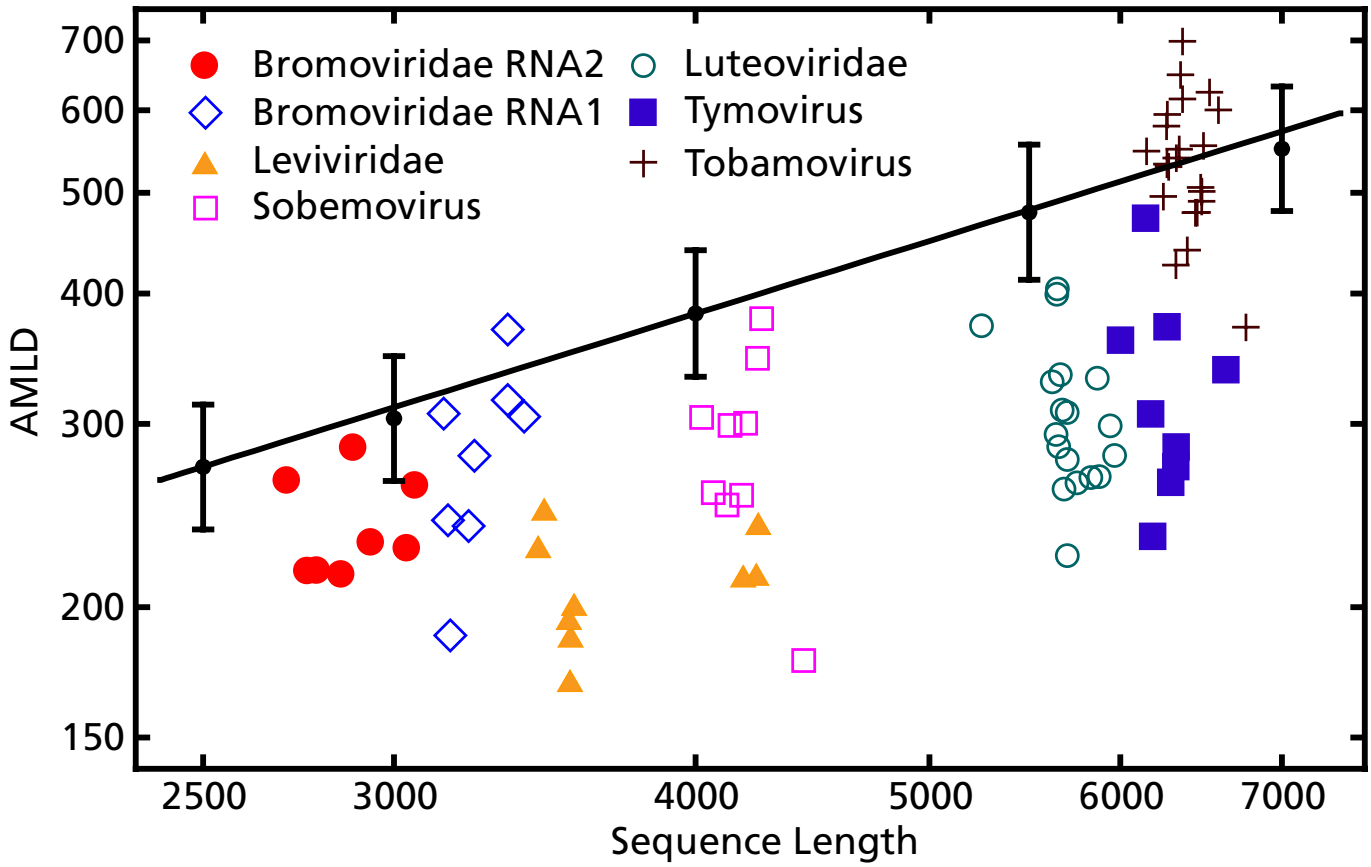

Fig. S4. Log–log plot of AMLD (calculated with mfold) vs. sequence length for viral and randomly permuted ssRNAs. The viral ssRNAs are identified by the symbols listed in the *Inset*. The Bromoviridae analyzed here are from the Bromovirus and Cucomovirus genera. The straight line is a least-squares fit to the $\overline{\text{AMLD}}$ values computed for random sequences 2,500, 3,000, 4,000, 5,500, and 7,000 nt in length; from its slope, we obtain $\overline{\text{AMLD}} \sim N^{0.74 \pm 0.01}$ over this range, where the uncertainty in the exponent is the standard deviation. The vertical lines show the standard deviations in the $\overline{\text{AMLD}}$ values.

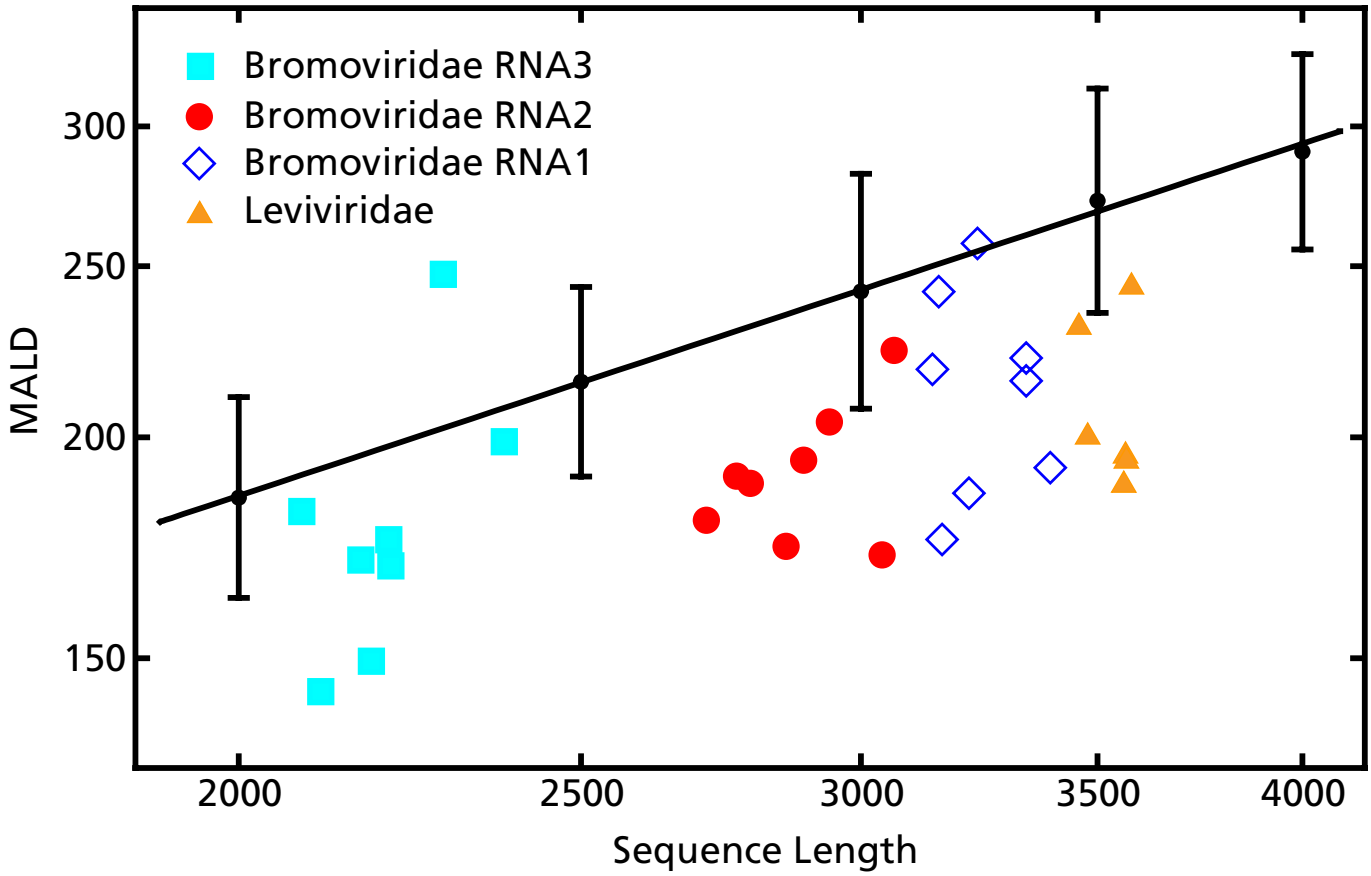

Fig. S5. Log–log plot of MALD (calculated with the simplified RNA folding program) vs. sequence length for viral and randomly permuted ssRNAs. The viral ssRNAs are identified by the symbols listed in the *Inset*. The Bromoviridae analyzed here are from the Bromovirus and Cucomovirus genera. The straight line is a least-squares fit to the MALD values computed for random sequences 2,000, 2,500, 3,000, 3,500, and 4,000 nt in length; from its slope, we obtain MALD $\sim N^{0.66\pm0.02}$ over this range. The vertical lines show the standard deviations.

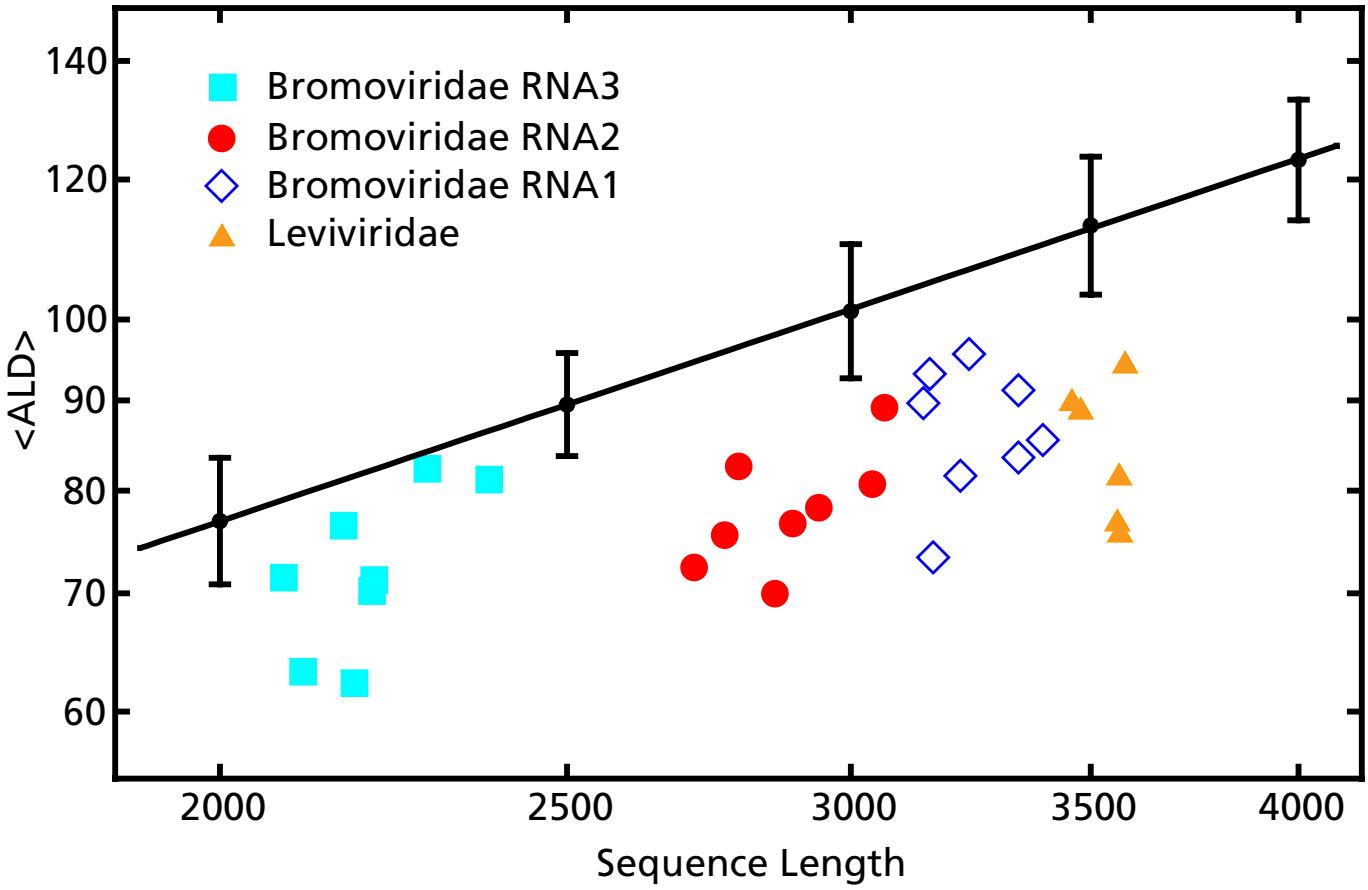

Fig. S6. Log–log plot of ⟨ALD⟩ (calculated with the simplified RNA folding program) vs. sequence length (same as Fig. S5, but with ⟨ALD⟩ replacing MALD). A least-squares fit yields $\overline{\langle ALD \rangle} \sim N^{0.68 \pm 0.01}$ for random sequences 2,000-4,000 nt in length.