Supporting Information

Tamas *et al.* 10.1073/pnas.0806554105

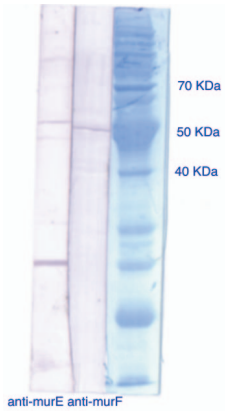

Fig. S1. Western blot of the total protein content of *Buchnera aphidicola* cells extracted from adult *Rhopalosiphum padi* aphids and probed with anti-murE (left lane) and anti-murF (middle lane) antibodies.

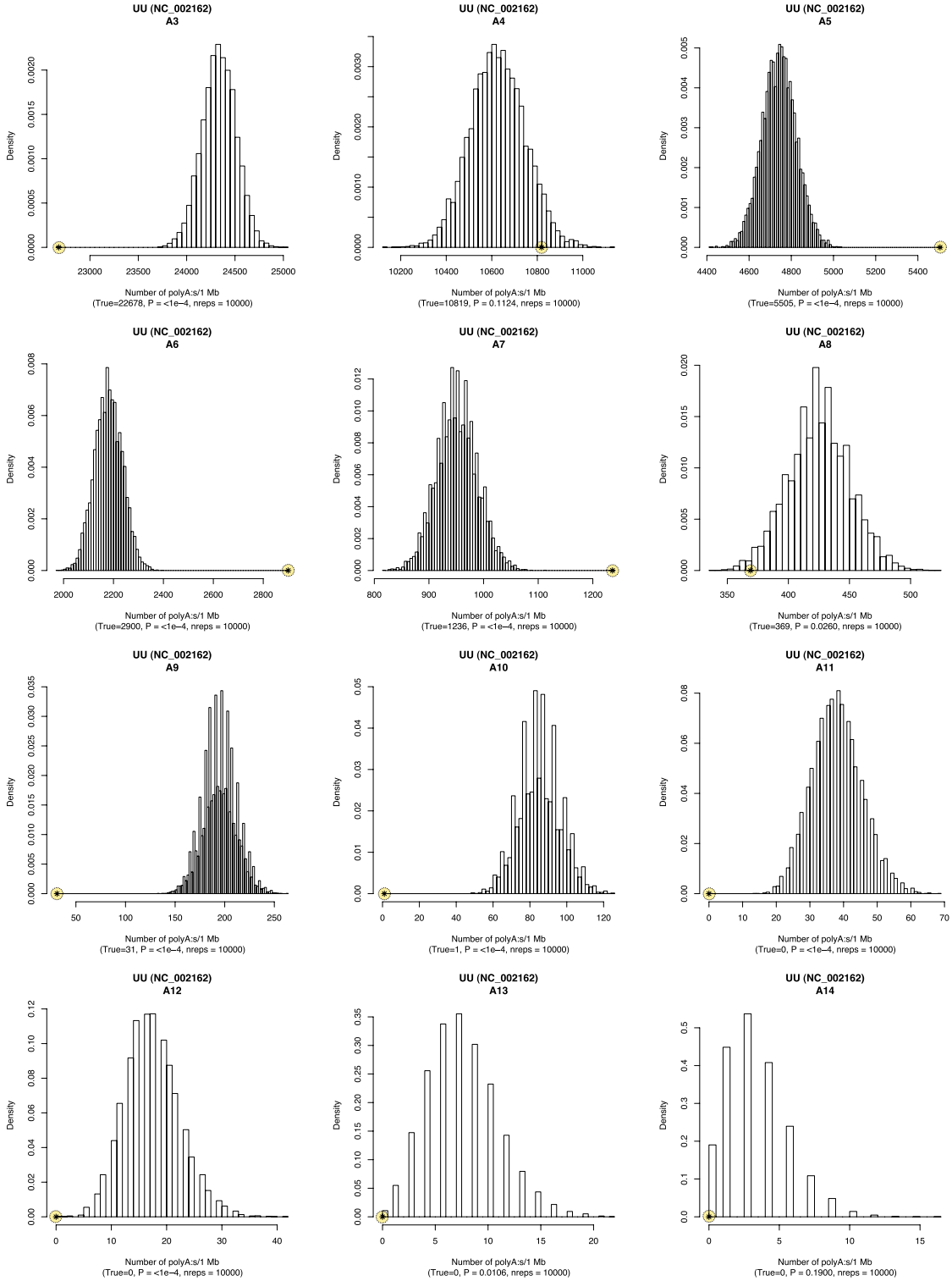

Fig. S2. The number of poly(A) tracts in real genomes (stars surrounded by dotted circles) per Mb of sequence data compared to the distribution profiles generated for 10,000 simulated genomes (histograms) for *Ureaplasma urealyticum* (UU), *Buchnera aphidicola* (Sg) (BA_Sg), *Buchnera aphidicola* (Cc) (BA_Cc), *Blochmannia pennsylvanicus* (BP), *Wigglesworthia glossinidia* (WG) and *Escherichia coli* str. K12 (EC_K12). The accession numbers and the lengths of the A runs (3–14) are indicated above each plot. The gene lengths and the overall codon frequency were preserved in each of the simulated genomes compared to the corresponding real genome. The *P* values refer to the two-sided permutation test that the true number of poly(A) tracts of a given length in a genome differs from the expectation value for a corresponding simulated genome.

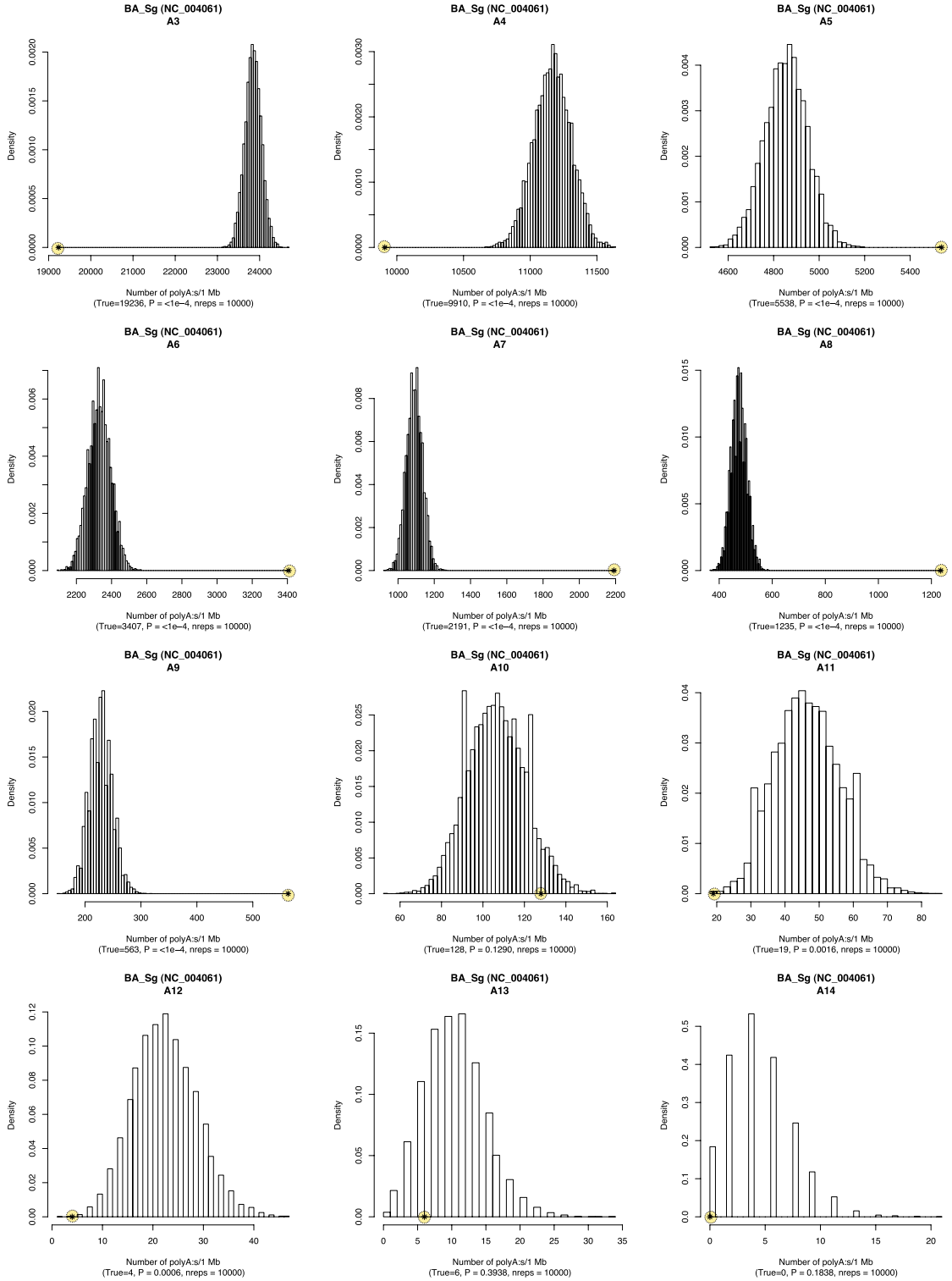

Fig. S2. (Continued.)

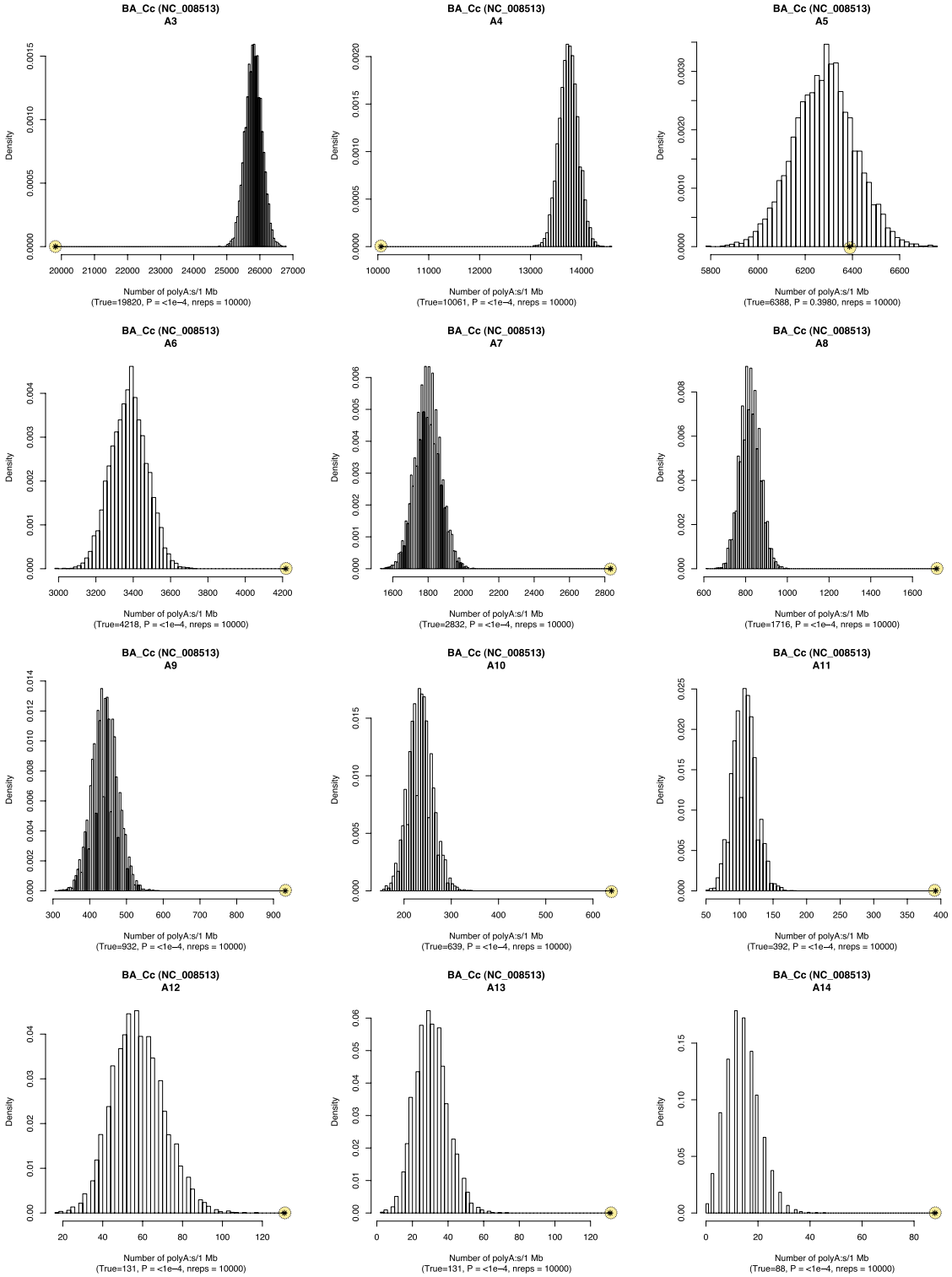

Fig. S2. (Continued.)

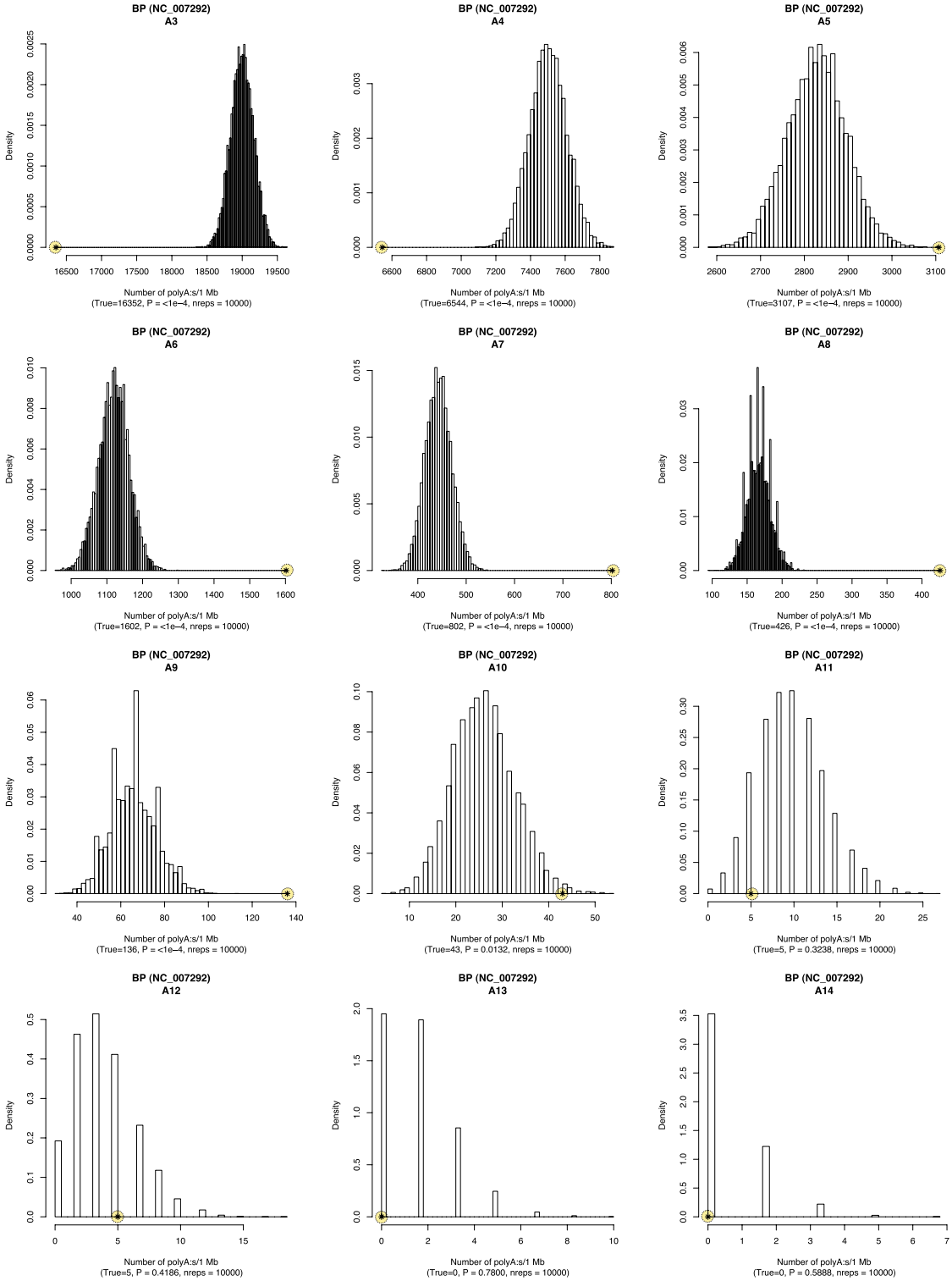

Fig. S2. (Continued.)

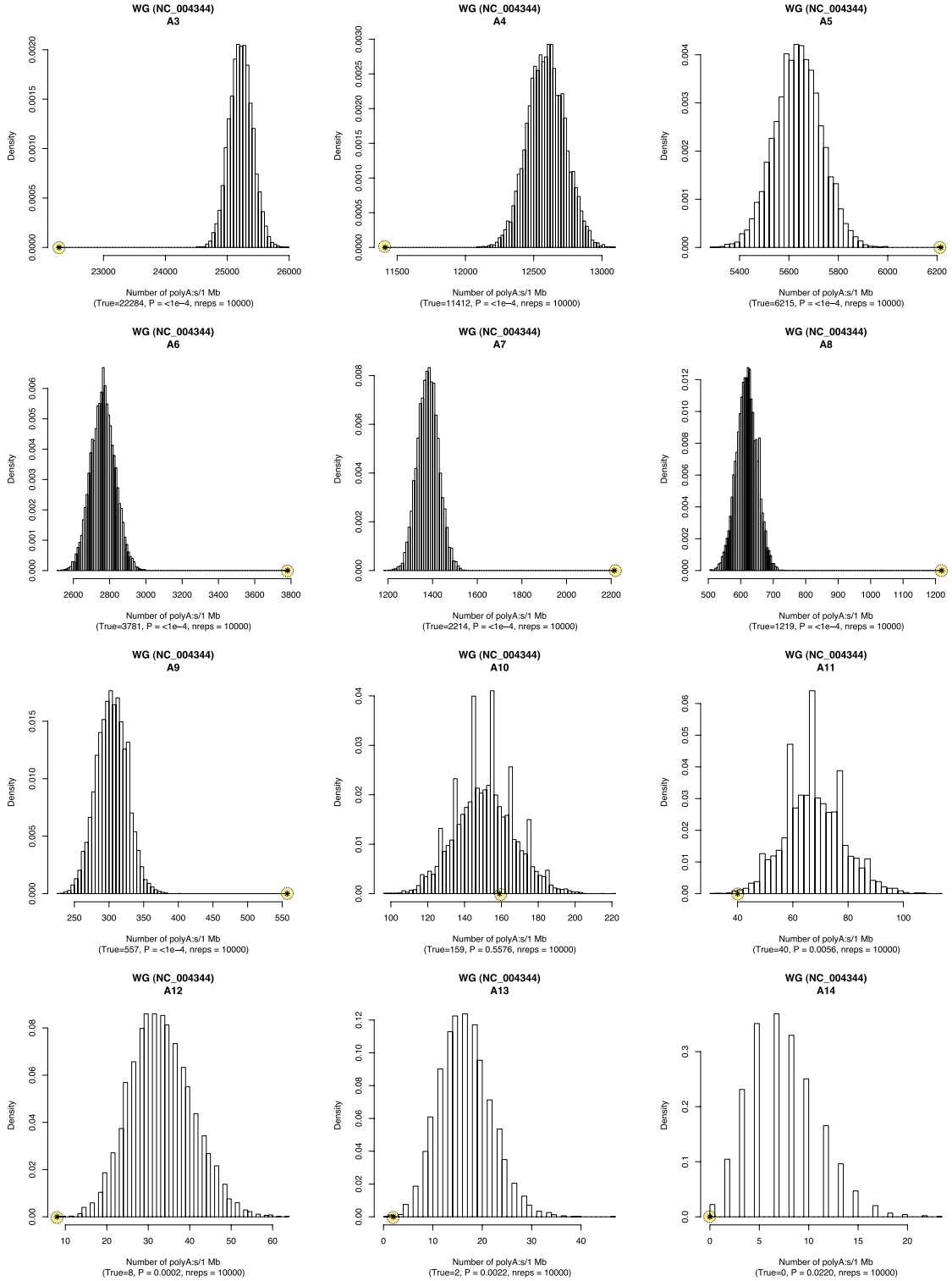

Fig. S2. (Continued.)

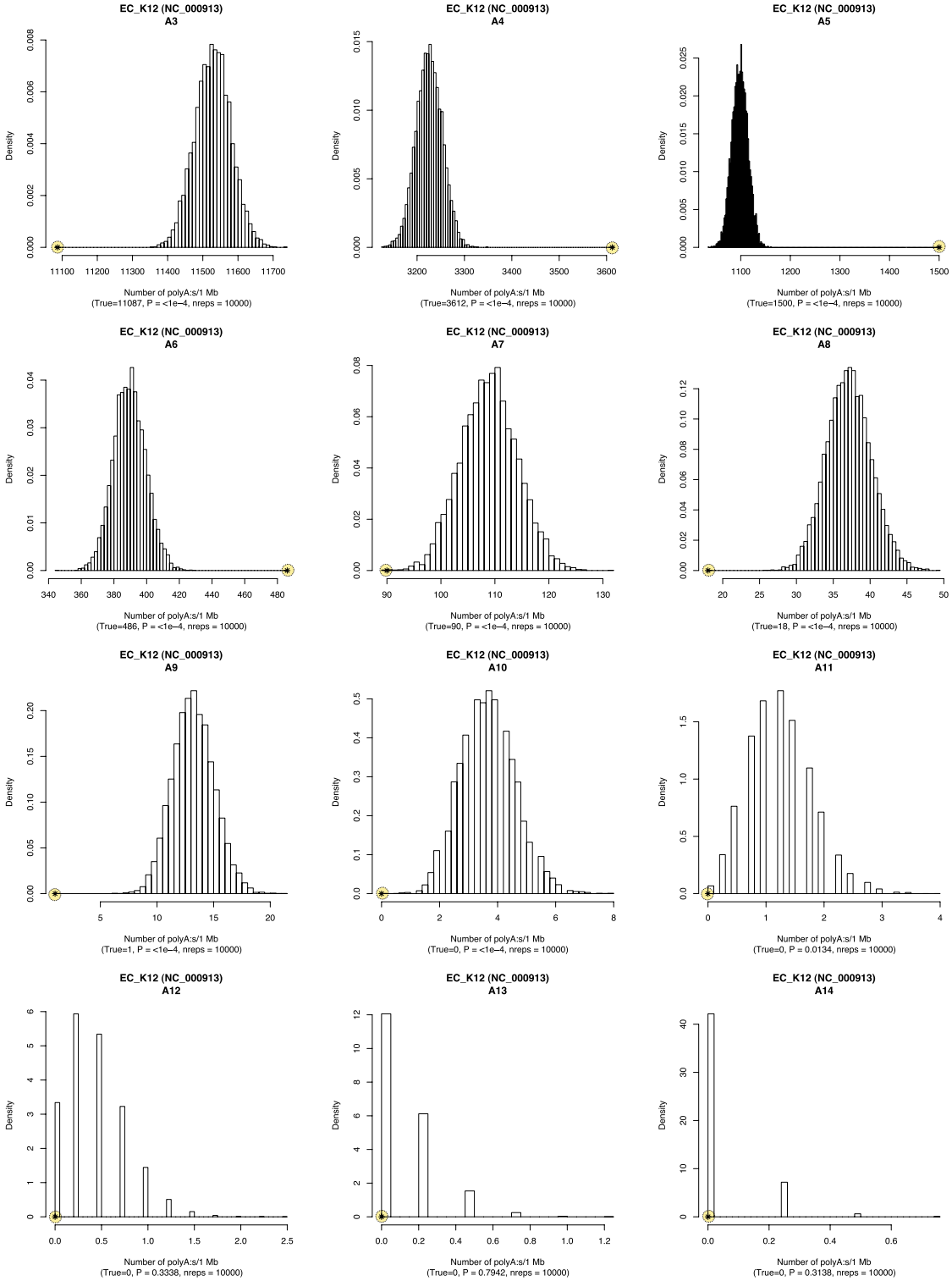

Fig. S2. (Continued.)

Table S1. Nonsynonymous (dN) and synonymous substitutions (dS) per site for genes in the *dcw* cluster in *B. aphidicola* are shown above and below the diagonal, respectively, with the ratio dN/dS shown below the substitution frequency matrix for each gene

	Sg	Rp	As	Ap	Mt	Dn
murE						
Sg	NA	0.06 ± 0.008	0.17 ± 0.015	0.20 ± 0.016	0.22 ± 0.018	0.21 ± 0.017
Rp	0.98 ± 0.35	NA	0.15 ± 0.014	0.19 ± 0.016	0.21 ± 0.017	0.22 ± 0.017
As	2.96 ± 1.97	1.96 ± 0.54	NA	0.23 ± 0.018	0.23 ± 0.018	0.23 ± 0.018
Ap	2.10 ± 0.64	2.02 ± 0.57	3.72 ± 5.21	NA	0.14 ± 0.014	0.18 ± 0.015
Mt	1.92 ± 0.66	2.47 ± 1.06	3.65 ± 4.96	1.54 ± 0.32	NA	0.19 ± 0.016
Dn	3.65 ± 4.96	3.68 ± 5.06	3.67 ± 5.04	2.33 ± 0.85	3.62 ± 4.86	NA
dN/dS						
Sg	NA					
Rp	0.061	NA				
As	0.056	0.078	NA			
Ap	0.094	0.092	0.062	NA		
Mt	0.011	0.084	0.064	0.093	NA	
Dn	0.058	0.059	0.063	0.076	0.053	NA
murF						
Sg	NA	0.074 ± 0.009	0.16 ± 0.013	0.19 ± 0.014	0.20 ± 0.015	0.18 ± 0.014
Rp	0.87 ± 0.20	NA	0.15 ± 0.013	0.19 ± 0.014	0.19 ± 0.014	0.18 ± 0.014
As	2.21 ± 0.62	1.52 ± 0.25	NA	0.23 ± 0.016	0.21 ± 0.015	0.23 ± 0.016
Ap	2.88 ± 1.44	1.51 ± 0.24	1.63 ± 0.43	NA	0.12 ± 0.011	0.15 ± 0.013
Mt	3.97 ± 6.13	1.97 ± 0.77	2.49 ± 0.91	1.32 ± 0.25	NA	0.17 ± 0.013
Dn	2.70 ± 1.12	1.99 ± 0.79	2.82 ± 1.35	1.76 ± 0.34	1.90 ± 0.41	NA
dN/dS						
Sg	NA					
Rp	0.085	NA				
As	0.072	0.010	NA			
Ap	0.066	0.012	0.014	NA		
Mt	0.050	0.097	0.086	0.091	NA	
Dn	0.068	0.091	0.081	0.086	0.087	NA
mraY						
Sg	NA	0.077 ± 0.010	0.17 ± 0.016	0.20 ± 0.017	0.21 ± 0.017	0.22 ± 0.018
Rp	0.80 ± 0.23	NA	0.16 ± 0.015	0.19 ± 0.017	0.19 ± 0.017	0.20 ± 0.017
As	1.57 ± 0.444	1.33 ± 0.32	NA	0.20 ± 0.018	0.22 ± 0.018	0.22 ± 0.018
Ap	1.40 ± 0.31	1.66 ± 0.53	1.91 ± 0.45	NA	0.12 ± 0.013	0.13 ± 0.013
Mt	1.25 ± 0.22	1.68 ± 0.41	1.49 ± 0.28	0.94 ± 0.22	NA	0.16 ± 0.015
Dn	3.45 ± 10.8	2.49 ± 0.98	3.15 ± 4.77	1.49 ± 0.35	1.36 ± 0.08	NA
dN/dS						
Sg	NA					
Rp	0.096	NA				
As	0.110	0.119	NA			
Ap	0.139	0.112	0.107	NA		
Mt	0.165	0.116	0.147	0.129	NA	
Dn	0.063	0.081	0.069	0.086	0.122	NA
murD						
Sg	NA	0.06 ± 0.008	0.18 ± 0.014	0.18 ± 0.014	0.20 ± 0.015	0.21 ± 0.015
Rp	1.23 ± 0.55	NA	0.18 ± 0.014	0.18 ± 0.014	0.20 ± 0.015	0.20 ± 0.015
As	2.60 ± 1.07	3.86 ± 5.69	NA	0.20 ± 0.015	0.23 ± 0.016	0.21 ± 0.016
Ap	3.90 ± 5.84	3.86 ± 5.70	2.94 ± 1.66	NA	0.15 ± 0.012	0.17 ± 0.014
Mt	3.86 ± 5.70	3.82 ± 5.55	2.96 ± 1.74	1.43 ± 0.25	NA	0.18 ± 0.014
Dn	3.89 ± 5.81	3.85 ± 5.68	3.94 ± 6.25	1.64 ± 0.31	1.77 ± 0.38	NA
dN/dS						
Sg	NA					
Rp	0.050	NA				
As	0.068	0.048	NA			
Ap	0.046	0.048	0.069	NA		
Mt	0.053	0.052	0.076	0.103	NA	
Dn	0.053	0.052	0.054	0.106	0.104	NA
ftsW						
Sg	NA	0.02 ± 0.005	0.08 ± 0.01	0.12 ± 0.013	0.14 ± 0.014	0.14 ± 0.014
Rp	0.65 ± 0.17	NA	0.08 ± 0.01	0.12 ± 0.013	0.15 ± 0.014	0.14 ± 0.014
As	1.61 ± 1.00	1.12 ± 0.21	NA	0.13 ± 0.013	0.15 ± 0.014	0.13 ± 0.014
Ap	1.21 ± 0.24	2.13 ± 2.78	1.35 ± 0.31	NA	0.10 ± 0.011	0.10 ± 0.012
Mt	1.43 ± 0.44	1.46 ± 0.37	2.75 ± 14.6	1.09 ± 0.24	NA	0.13 ± 0.013
Dn	1.91 ± 0.68	1.74 ± 0.54	1.98 ± 0.86	1.64 ± 0.76	1.24 ± 0.35	NA

	Sg	Rp	As	Ap	Mt	Dn
dN/dS						
Sg	NA					
Rp	0.037	NA				
As	0.051	0.067	NA			
Ap	0.100	0.058	0.095	NA		
Mt	0.100	0.100	0.054	0.088	NA	
Dn	0.071	0.081	0.067	0.064	0.104	NA
murG						
Sg	NA	0.05 ± 0.008	0.15 ± 0.014	0.18 ± 0.016	0.19 ± 0.016	0.18 ± 0.016
Rp	1.24 ± 0.39	NA	0.14 ± 0.013	0.17 ± 0.015	0.18 ± 0.016	0.17 ± 0.016
As	3.75 ± 5.30	3.70 ± 5.16	NA	0.20 ± 0.017	0.20 ± 0.017	0.21 ± 0.017
Ap	3.76 ± 5.35	3.72 ± 5.20	2.12 ± 0.65	NA	0.12 ± 0.013	0.17 ± 0.015
Mt	3.75 ± 5.30	3.70 ± 5.14	4.40 ± 12.9	2.40 ± 0.93	NA	0.18 ± 0.016
Dn	3.81 ± 5.54	3.77 ± 5.40	1.93 ± 0.57	2.40 ± 0.88	4.10 ± 8.32	NA
dN/dS						
Sg	NA					
Rp	0.044	NA				
As	0.039	0.037	NA			
Ap	0.049	0.047	0.096	NA		
Mt	0.052	0.049	0.047	0.052	NA	
Dn	0.046	0.047	0.107	0.071	0.043	NA
murC						
Sg	NA	0.06 ± 0.007	0.15 ± 0.012	0.21 ± 0.015	0.19 ± 0.014	0.20 ± 0.015
Rp	0.80 ± 0.26	NA	0.16 ± 0.013	0.21 ± 0.015	0.19 ± 0.014	0.21 ± 0.015
As	1.79 ± 0.37	1.52 ± 0.26	NA	0.22 ± 0.015	0.18 ± 0.014	0.21 ± 0.015
Ap	1.85 ± 0.40	2.03 ± 0.49	1.67 ± 0.31	NA	0.14 ± 0.011	0.18 ± 0.014
Mt	3.86 ± 5.72	3.90 ± 5.84	2.34 ± 0.76	2.20 ± 1.47	NA	0.18 ± 0.013
Dn	2.39 ± 0.78	2.54 ± 0.92	3.02 ± 1.73	2.04 ± 0.48	2.64 ± 1.09	NA
dN/dS						
Sg	NA					
Rp	0.072	NA				
As	0.086	0.105	NA			
Ap	0.112	0.102	0.129	NA		
Mt	0.050	0.050	0.079	0.062	NA	
Dn	0.086	0.081	0.071	0.090	0.067	NA
ddlB						
Sg	NA	0.069 ± 0.01	0.17 ± 0.016	0.18 ± 0.017	0.21 ± 0.018	0.22 ± 0.019
Rp	1.22 ± 0.74	NA	0.17 ± 0.016	0.19 ± 0.017	0.22 ± 0.019	0.22 ± 0.019
As	4.48 ± 15.86	3.59 ± 4.78	NA	0.20 ± 0.018	0.21 ± 0.018	0.22 ± 0.019
Ap	3.59 ± 4.79	3.62 ± 4.87	2.95 ± 2.07	NA	0.13 ± 0.014	0.19 ± 0.018
Mt	3.57 ± 4.71	3.60 ± 4.80	3.55 ± 4.65	2.87 ± 1.84	NA	0.20 ± 0.018
Dn	3.62 ± 4.86	3.64 ± 4.94	1.89 ± 0.52	1.94 ± 0.54	2.11 ± 0.68	NA
dN/dS						
Sg	NA					
Rp	0.056	NA				
As	0.038	0.049	NA			
Ap	0.051	0.052	0.067	NA		
Mt	0.059	0.062	0.060	0.045	NA	
Dn	0.060	0.061	0.116	0.098	0.095	NA

Sequences have been corrected for frameshift mutations prior to the calculations with PAML 3.14 under yn00 (1). In comparison to the synonymous substitutions, which are in most cases saturated, non-synonymous substitution frequencies are lower and similar across genes, irrespectively of whether the frame is shifted or not. Strain designations are as described in *Materials and Methods*. Gene names and gene products are as follows: *murE*, UDP-*N*-acetylmuramyl-L-alanyl-D-glutamate: 2,6-diaminopimelate ligase; *murF*, UDP–*N*-acetylmuramyl-L-alanyl-D-glutamyl-2,6-diaminopimelate: D-alanyl-D-alanineligase; *mraY*, phospho-*N*-acetylmuramylpentapeptidetransferase; *murD*,UDP-*N*-acetylmuramyl-L-alanine: D-glutamate ligase; *ftsW*, cell division protein; *murG*, UDP-*N*-acetylglucosamine: N-acetylmuramylpentapeptidepyrophosphoryl-undecaprenol-*N*-acetylglucosaminetransferase; *murC*,UDP-*N*-acetylmuramate:L-alanineligase;*ddlB*,D-alanine: D-alanineligase.

1. Yang Z, Nielsen R (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol Biol Evol* 17:32–43.

Table S2. Control experiments of the error frequencies at homopolymeric tracts of 10–12 As in the PCR and RT-PCR experiments

Template*	RNA-oligo	Re-seq	PCR	Re-streak	PCR
		Single Plasmid	Single Plasmid	Single *E. coli* strain	gDNA BA
Locus	*ytfM*-2	*ytfM*-2	*hisH*	*murE*	
<8 A	3	0	0	0	0
8 A	1	0	0	0	0
9 A	3	0	0	0	3
10 A	1	0	0	0	**59**
11 A	**56**	**8**	**10**	0	0
12 A	2	0	0	**14**	0
13 A	1	0	0	0	0
Total	67	8	10	14	62
Error, %	16	0	0	0	5

Numbers in bold = the number of As in the original genomic DNA.

*To control for slippage errors in M-MLV, we followed a previously suggested approach (1) and analyzed a synthetic RNA oligonucleotide of known sequence (Lane RNA-oligo): (GUCCACGAUGAGACCAAUUAAAAAAAAAAA-UACCAGCACAUCCAUCC, synthesized by Dharmacon).The region$CAAUUA_{11}U$ matches the genomic DNA sequence for *B. pennsylvanicus hisH*. The RNA oligomer was used as template for first-strand cDNA synthesis, using the same reverse transcription method that we applied to the ant RNA samples above (M-MLV reverse transcriptase and random decamers; RETROscript: Ambion). The single-stranded cDNA was PCR-amplified with complementing primers (*hisH*_oligo_F: GTCCACGATGAGACC and *hisH*_oligo_R: GGATGGATGTGCTGG). The resulting product was TA cloned into the pCR4-TOPO vector (Invitrogen). Multiple clones were sequenced as described above. Whereas endosymbiont cDNAs often showed slippage of +/- one bp, the few length variants in the RNA oligo control often show larger differences of 2 bp or more.

To control for errors during plasmid replication in *E. coli*, we streaked a single colony onto LB. From 14 resulting colonies, plasmid DNA was prepared and sequenced (Lane "Re-streak single *E. coli* strain"). As a control for polymerase errors in DNA sequencing, plasmid DNA from one of the *ytfM*-2 cDNA clones with an 11 A tract from *B. pennsylvanicus* was sequenced 8 times (Lane "Re-seq single plasmid"). To control for slippage by the Taq polymerase used in PCR, we performed 10 replicate PCR reactions of the *ytfM*-2 cDNA clone and sequenced these (Lane "PCR single plasmid"). In all of these cases, the original number of As was observed.

We also amplified a *murE* gene fragment with 10 A by PCR from genomic DNA of *B. aphidicola* and made a plasmid library with the same procedure as in the RT-PCR experiments except that no Reverse Transcriptase step was used in the experiment (Lane "PCR gDNA BA"). From the resulting 62 colonies, plasmid DNA was prepared and sequenced, of which 59 contained the original 10 A tract. Another fragment of the *murC* gene bearing a 9 A tract was also cloned into an expression vector and expressed in *Saccharomycescerevisiae*. We performed RT-PCR and sequenced cloned products in both directions. Consistent with previous observations, no slippage was observed in *S. cerevisiae* (2).

1. Larsen B, Wills NM, Nelson C, Atkins JF, Gesteland RF (2000) Nonlinearity in genetic decoding: Homologous DNA replicase genes use alternatives of transcriptional slippage or translational frameshifting. *Proc Natl Acad Sci USA* 97:1683–1688.
2. Wagner LA, Weiss RB, Driscoll R, Dunn DS, Gesteland RF (1990) Transcriptional slippage occurs during elongation at runs of adenine or thymine in *Escherichia coli*. *Nucleic Acids Res* 18:3529–3535.

Table S3. Poly(A) tracts of 10 nucleotides or more in protein coding genes

Species*	poly(A)/Mb†	% of all CDS‡	poly(A)/CDS§	% GC¶
BA (Cc)	1,491	57	2.6	20
WG	210	18	1.2	22
BA (Sg)	156	14	1.1	25
UU	1.5	<1	1.0	25
BP	48	5	1.0	29
EC (K12)	0	—	—	50

*Species abbreviations: BA (Cc), *B. aphidicola* (Cc); BA (Sg): *B. aphidicola* (Sg); WG, *W. glossinidia*; UU, *U. urealyticum*; BP: *B. pennsylvanicus*; EC (K12): *E. coli* (K12).

†Number of poly(A) tracts per million nucleotides of the proteome.

‡Fraction of all coding genes with at least one poly(A) tract.

§Average number of poly(A) tracts in genes containing at least one poly(A) tract.

¶Genomic G + C content.

Table S4. Functional categories for protein coding genes containing poly(A) tracts of 10 nucleotides or more

Functional category	BA (Sg)*		BA (Cc)*	
	Poly(A)†	Tot‡	Poly(A)†	Tot‡
Information storage				
[J] Translation, ribosome structure	14	89	40	93
[A] RNA processing/modification	0	1	0	0
[K] Transcription	3	16	5	10
[L] Replication, recombination	9	33	23	29
Cellular processes				
[D] Cell cycle control	1	8	5	6
[V] Defense mechanisms	1	3	2	2
[T] Signal transduction	0	4	3	3
[M] Cell wall/membrane	6	23	2	3
[N] Cell motility	1	23	9	12
[U] Intracellular trafficking	4	24	4	8
[O] Posttranslational modification	4	39	13	19
Metabolism				
[C] Energy production	5	41	14	28
[G] Carbohydrate transport/metabolism	2	22	8	16
[E] Amino acid transport/metabolism	5	55	15	25
[F] Nucleotide transport/metabolism	2	23	3	5
[H] Coenzyme transport/metabolism	1	30	4	5
[I] Lipid transport/metabolism	1	12	6	7
[P] Inorganic ion transport/metabolism	1	15	3	5
[Q] Secondary metabolites	0	2	1	1
Poorly characterized				
[R] General function prediction only	8	29	11	12
[S] Funtion unknown	3	14	7	12

*Species abbreviations: BA (Sg), *B. aphidicola* (Sg); BA (Cc), *B. aphidicola* (Cc).
†Poly(A) = Number of protein coding genes containing at least one poly(A) tract of 10 nucleotides or more.
‡Total number of protein coding genes.

Table S5. Functional categories for protein coding genes containing poly(A) tracts of 10 nucleotides or more

Functional category	BP*		WG*	
	Poly(A)†	Tot‡	Poly(A)†	Tot‡
Information storage				
[J] Translation, ribosome structure	5	89	14	82
[A] RNA processing/modification	0	1	0	1
[K] Transcription	1	21	2	20
[L] Replication, recombination	2	29	9	28
Cellular processes				
[D] Cell cycle control	0	10	3	9
[V] Defense mechanisms	0	2	3	5
[T] Signal transduction	0	3	0	3
[M] Cell wall/membrane	3	51	10	53
[N] Cell motility	0	0	4	32
[U] Intracellular trafficking	1	18	5	28
[O] Posttranslational modification	1	32	3	33
Metabolism				
[C] Energy production	1	46	3	34
[G] Carbohydrate transport/metabolism	0	25	1	16
[E] Amino acid transport/metabolism	2	68	8	30
[F] Nucleotide transport/metabolism	1	21	5	33
[H] Coenzyme transport/metabolism	4	46	14	54
[I] Lipid transport/metabolism	1	26	4	25
[P] Inorganic ion transport/metabolism	0	25	2	15
[Q] Secondary metabolites	0	3	0	4
Poorly characterized				
[R] General function prediction only	2	37	8	37
[S] Funtion unknown	2	16	1	18

*Species abbreviations: BP, *B. pennsylvanicus*; WG, *W. glossinidia*.
†Poly(A) = Number of protein coding genes containing at least one poly(A) tract of 10 nucleotides or more.
‡Total number of protein coding genes.

Table S6. Real and expected number of protein coding genes with poly(A) tracts of nine nucleotides or more in protein coding genes

Species*		Nb poly(A) tracts†	CDS:s with poly(A)‡		P§
			Real	Expected	
BA (Cc)	9+	857	253	288	**$<10^{-4}$**
	10+	525	203	242	**$<10^{-4}$**
BA (Sg)	9+	385	230	253	**7.00×10^{-4}**
	10+	83	74	75	0.41
BP	9+	108	89	96	0.02
	10+	29	28	28	0.6
WG	9+	464	280	299	**0.01**
	10+	127	109	111	0.34
UU	9+	22	21	21	0.52
	10+	1	1	1	1.0
EC (K12)	9+	3	3	3	1
	10+	0	—	—	—

*Species abbreviations: BA (Cc), *B. aphidicola* (Cc); BA (Sg): *B. aphidicola* (Sg); BP: *B. pennsylvanicus*; WG, *W. glossinidia*; UU, *U. urealyticum*; EC (K12): *E. coli* (K12).

†Total number of poly(A) tracts of nine nucleotides or more (9+) and 10 nucleotides or more (10+) in protein coding genes.

‡Real and expected number of protein coding genes (CDS) containing poly(A) tracts of nine nucleotides or more (9+) and 10 nucleotides or more (10+). The expected number is the median of 1000 simulated genomes with number of poly(A) tracts and gene lengths identical to the real proteome.

§*P* values refer to the one-sided permutation test that real genomes contain multiple poly(A) tracts in the same genes more often than expected by chance. Bold indicates $P < 0.01$. The *P* value was set to the fraction of simulated genomes with a lower fraction of genes with poly(A) tracts than the real genome.