Supplementary Figure and Table Captions

Table S1: Overall statistics of species' sequences, SSRs, and indels across the genomes analyzed. Columns give the species abbreviation, name, number of fully-sequenced strains analyzed, the total length of homologous coding sequence that was analyzed (the total length of all alignments), the number of SSR loci identified, the total number of SSR-associated indels, and the number of indels per SSR locus.

Figure S1: Observed vs. expected SSR frequency statistics for different species and motif types. (A) Ratio of observed vs. expected SSR frequency estimated by global synonymous codon shuffling. Each square represents the value of the ratio for a given SSR type and species. Over- and under-represented types are shown respectively in yellow and blue shades. (B) Cumulative statistics for each SSR type. See Table 1 for the prokaryotic species names.

Figure S2A and S2B: Symmetric/anti-symmetric decomposition of SSR density. (A) Symmetric components of original SSR density (blue curves) and local synonymous codon-shuffled SSR density (red curves). (B) Anti-symmetric components of original SSR density (blue curves) and local synonymous codon-shuffled SSR density (red curves). Note that for each species, symmetric and anti-symmetric component curves are all normalized by average SSR frequency of local shuffled genome.

Figure S3: Positional dependence of SSR density, using stringent length cutoff (see Supplementary Methods). For each plot, the x-axis represents binned positions from N–>C termini, and the y-axis indicates the SSR density (# SSRs / kb). Blue curves: original SSR density. Green curves: SSR density in global synonymous codon shufflings. Red curves: SSR density in local synonymous codon shufflings. Error bars represent the standard error of the estimated SSR density observed in the shuffled sequences.

Figure S4A and S4B: Symmetric/anti-symmetric decomposition of SSR density, using stringent length cutoff (see Supplementary Methods). Same color scheme was used as in Figure S2.

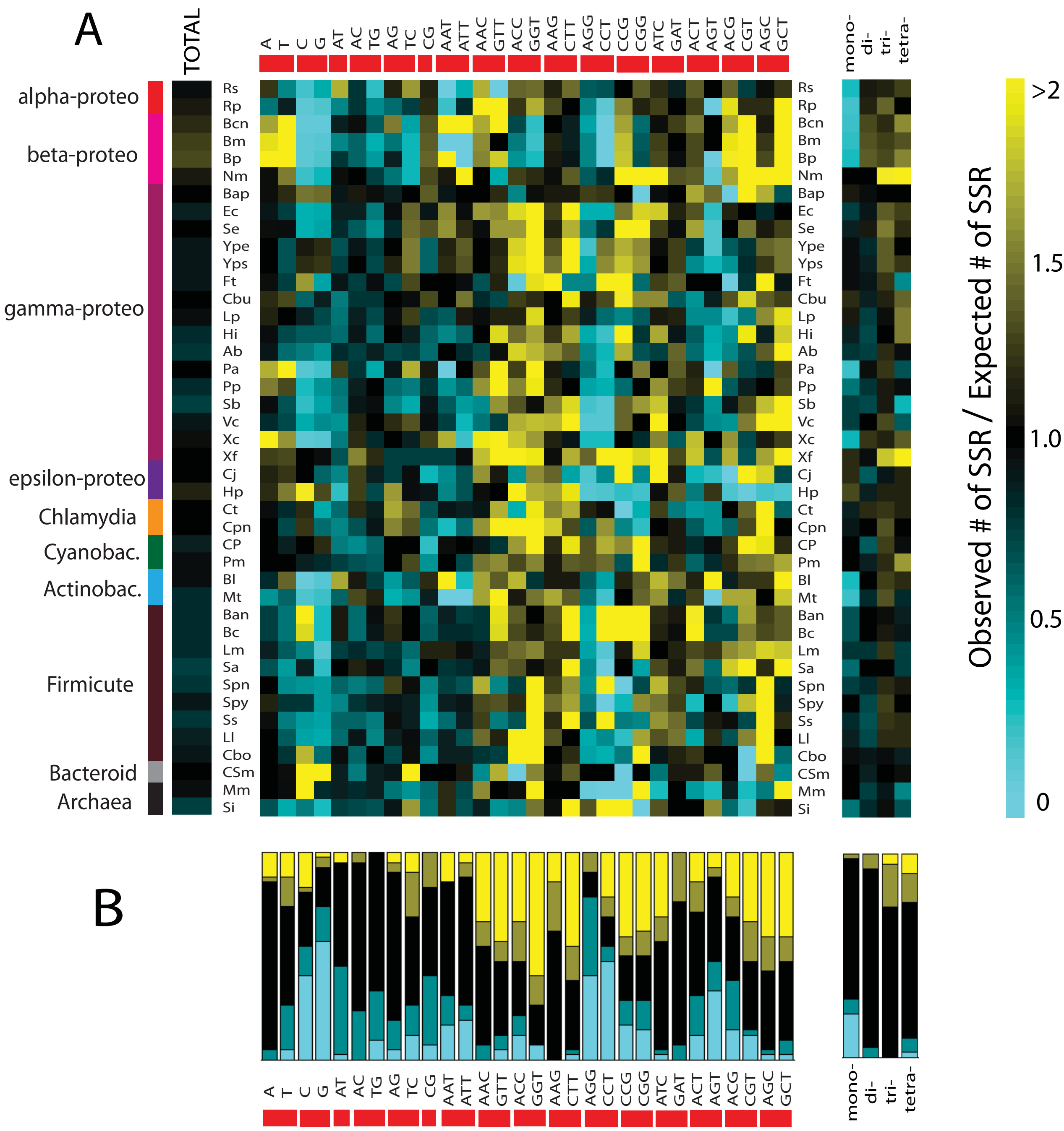
Figure S5: Flow chart describing procedure to generate local and global codon shuffling of CDS, used as null models for testing selective pressures on SSR density.

Table S1

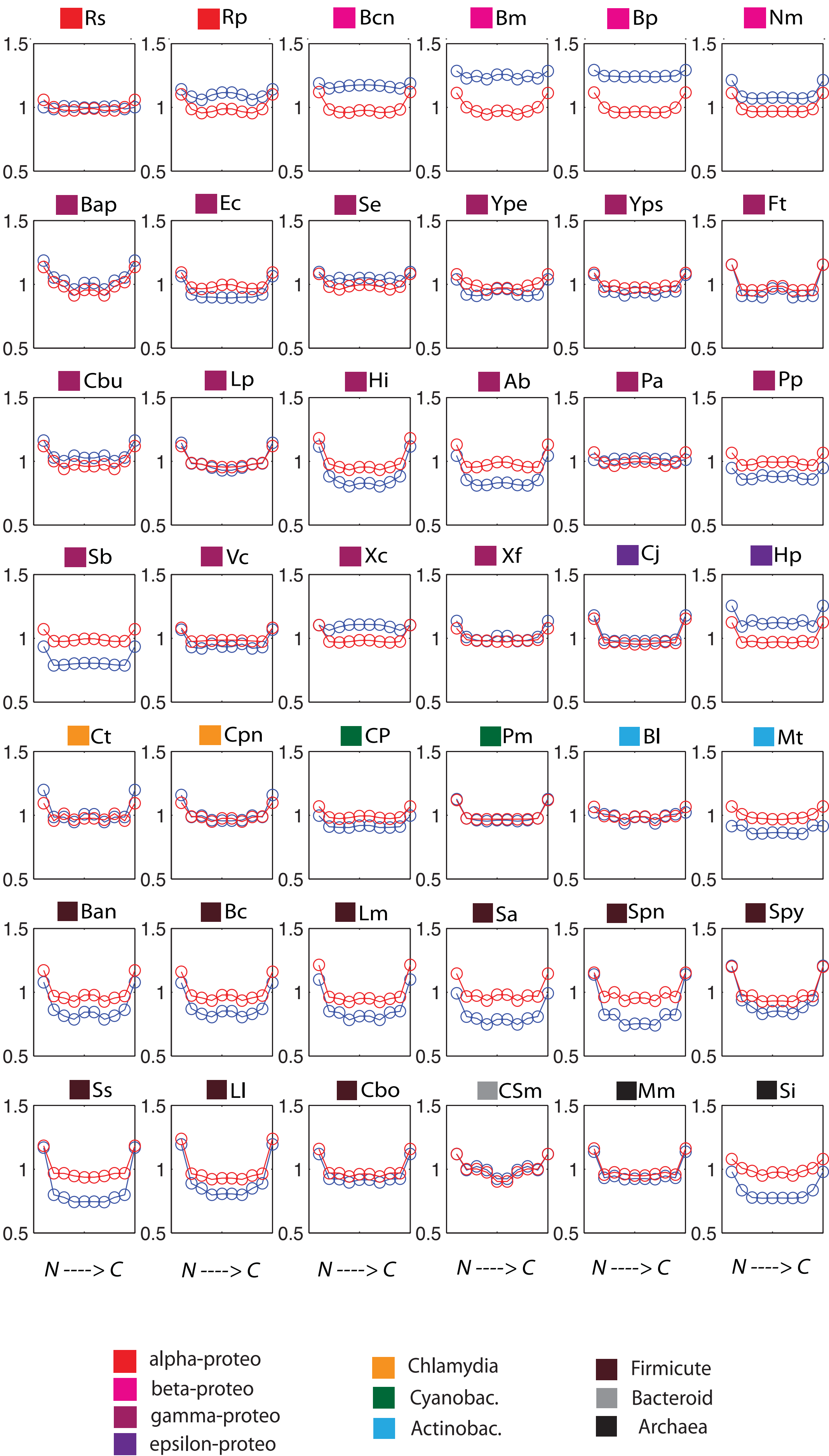
Abbr.	Species name	# of strains analyzed	homologous CDS length (bp)	# of SSR loci	# of indel events	indel events per SSR locus	Taxonomy
Ab	*Acinetobacter baumannii*	6	3583306	13512	208	0.0154	gamma-proteo.; Pseudomonadales
Ban	*Bacillus anthracis*	6	4765573	23521	297	0.0126	Firmicutes; Bacillales
Bc	*Bacillus cereus*	10	5585495	49981	972	0.0194	Firmicutes; Bacillales
Bl	*Bifidobacterium longum*	4	2211423	7382	86	0.0116	Actinobacteria; Bifidobacteriales
Bap	*Buchnera aphidicola*	5	575267	14912	651	0.0437	gamma-proteo.; Enterobacteriales
Bcn	*Burkhoderia cenocepacia*	4	6776765	53827	424	0.0079	beta-proteo.; Burkholderiales
Bm	*Burkhoderia mallei*	4	5028909	47444	342	0.0072	beta-proteo.; Burkholderiales
Bp	*Burkhoderia pseudomallei*	5	6599508	63839	1042	0.0163	beta-proteo.; Burkholderiales
Cj	*Campylobacter jejuni*	5	1638518	15145	194	0.0128	epsion-proteo. ;Campylobacteria
CSm	*Candidatus Sulcia muelleri*	4	262508	7376	536	0.0727	Bacteroidetes; flavobacteria
Ct	*Chlamydia trachomatis*	6	946013	5949	39	0.0066	Chlamydiae
Cpn	*Chlamydophia pneumoniae*	4	1112526	5348	8	0.0015	Chlamydiae
Cbo	*Clostridium botulinum*	10	5113611	58030	877	0.0151	Firmicutes; Clostridiales
Cbu	*Coxiella burnetii*	5	1790675	9577	151	0.0158	gamma-proteo.; Lagionellales
CP	*Cynothece PCC*	5	5658602	41385	516	0.0125	Cyanobacteria
Ec	*Escherichia coli*	30	6736504	36432	2351	0.0645	gamma-proteo.; Enterobacteriales
Ft	*Francisella tularensis*	8	1643566	10461	180	0.0172	gamma-proteo.; Thiotrichales
Hi	*Haemophilus influenzae*	4	1649625	7616	154	0.0202	gamma-proteo.; Pateurellales
Hp	*Helicobacter pylori*	10	1662741	18056	683	0.0378	epsion-proteo. ;Campylobacteria
Ll	*Lactococcus lactis*	4	2114417	13614	165	0.0121	Firmicutes; Lactobacillales
Lp	*Legionella pneumophia*	5	3242761	17885	108	0.0060	gamma-proteo.; Lagionellales
Lm	*Listeria monocytogenes*	6	2839101	16074	106	0.0066	Firmicutes; Bacillales

Mm	*Methanococcus maripaludis*	4	1507985	15206	132	0.0087	Archaea
Mt	*Mycobacterium tuberculosis*	5	4118934	13388	45	0.0034	Actinobacteria; Actinomycetales
Nm	*Neisseria meningitidis*	5	1843651	9445	133	0.0141	beta-proteo.; Nesseriales
Pm	*Prochlorococcus marinus*	12	3257772	42443	1067	0.0251	Cyanobacteria
Pa	*Pseudomonas aeruginosa*	4	5758814	28673	204	0.0071	gamma-proteo.; Pseudomonadales
Pp	*Pseudomonas putida*	4	5150324	25859	176	0.0068	gamma-proteo.; Pseudomonadales
Rs	*Rhodobacter sphaeroides*	4	4077709	30677	360	0.0117	alpha-proteo.; Rhodobacteriales
Rp	*Rhodopseudomonas palustris*	6	5470313	55145	2029	0.0368	alpha-proteo.; Rhizobiales
Se	*Salmonella entericia*	15	4995231	25638	338	0.0132	gamma-proteo.; Enterobacteriales
Sb	*Shewanella baltica*	4	4402143	13246	79	0.0060	gamma-proteo.; Alteromonadales
Sa	*Staphylococcus aureus*	15	2704147	14565	197	0.0135	Firmicutes; Bacillales
Spn	*Streptococcus pneumoniae*	14	2281593	10203	233	0.0228	Firmicutes; Lactobacillales
Spy	*Streptococcus pyogenes*	13	1901778	9242	193	0.0209	Firmicutes; Lactobacillales
Ss	*Streptococcus suis*	5	1869382	5824	25	0.0043	Firmicutes; Lactobacillales
Si	*Sulfolobus islandicus*	7	2545336	12544	169	0.0135	Archaea
Vc	*Vibrio cholerae*	4	3595367	11497	25	0.0022	gamma-proteo.; Vibrionales
Xc	*Xanthomonas campestris*	4	4486353	30687	334	0.0109	gamma-proteo.; Xanthomonadales
Xf	*Xylella fastidiosa*	4	2081769	8333	158	0.0190	gamma-proteo.; Xanthomonadales
Ype	*Yersinia pestis*	8	4032200	12564	125	0.0099	gamma-proteo.; Enterobacteriales
Yps	*Yersinia pseudotuberculosis*	4	3997610	13211	84	0.0064	gamma-proteo.; Enterobacteriales

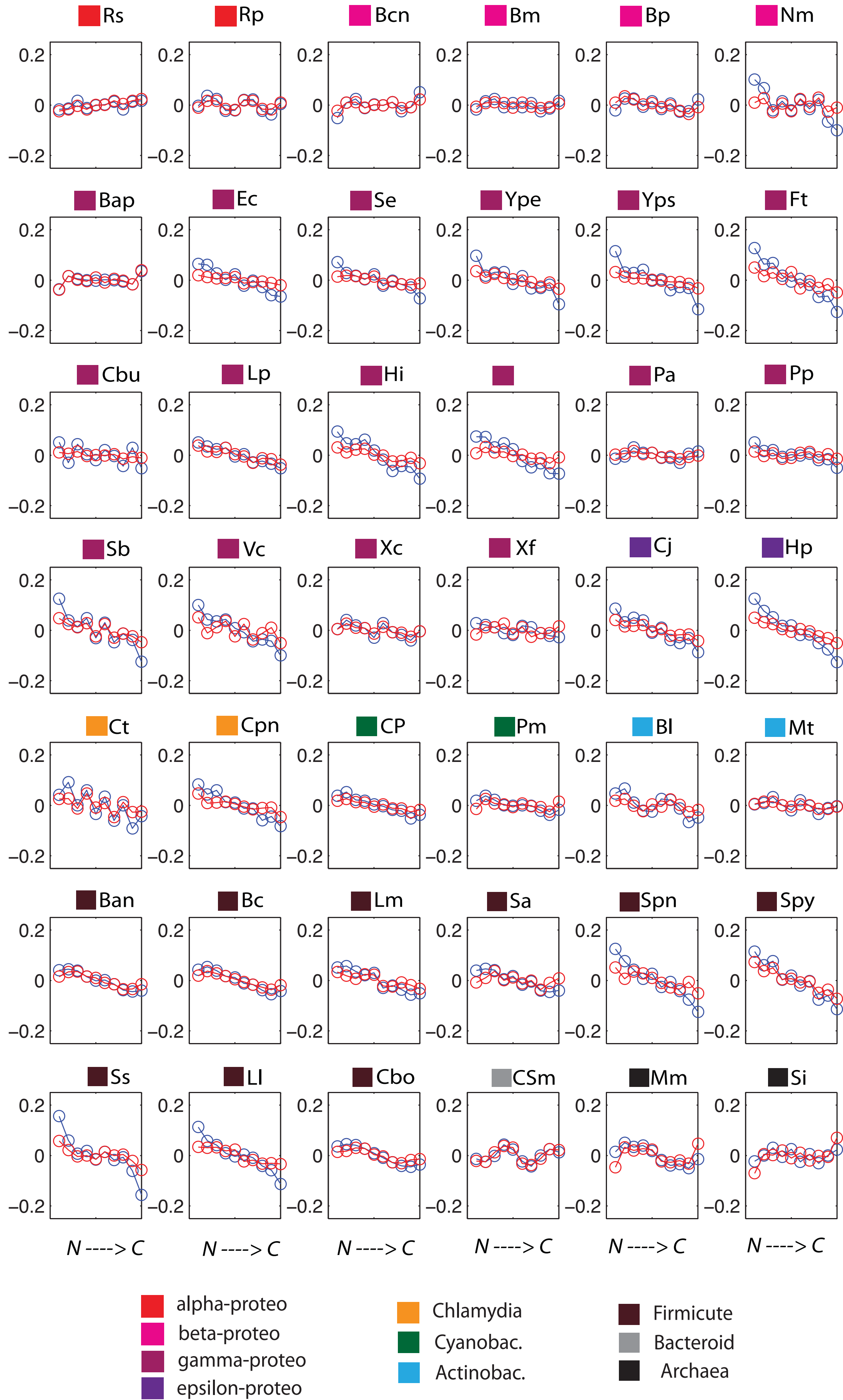
Supplemental Figure 1



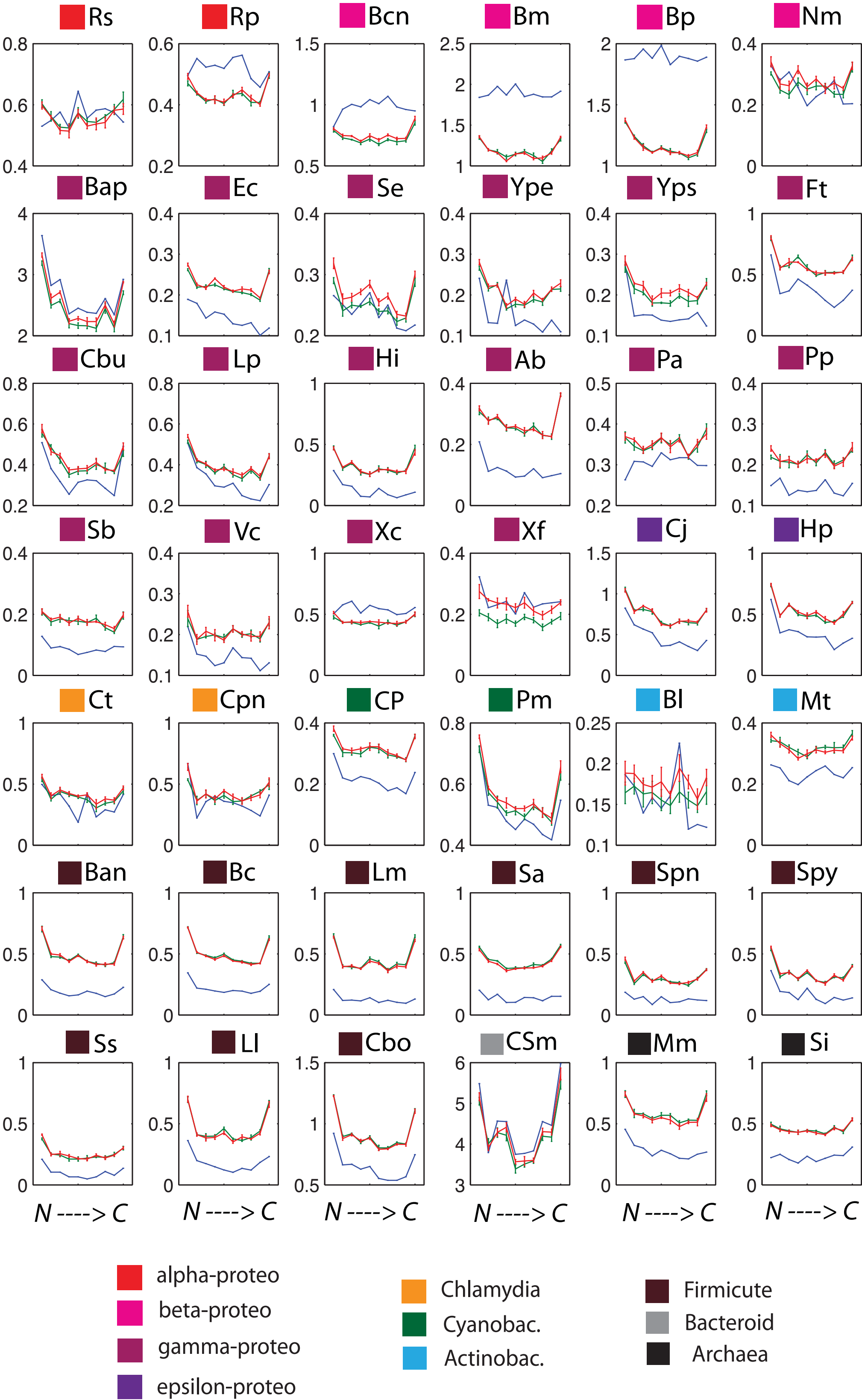
Supplemental Figure 2A



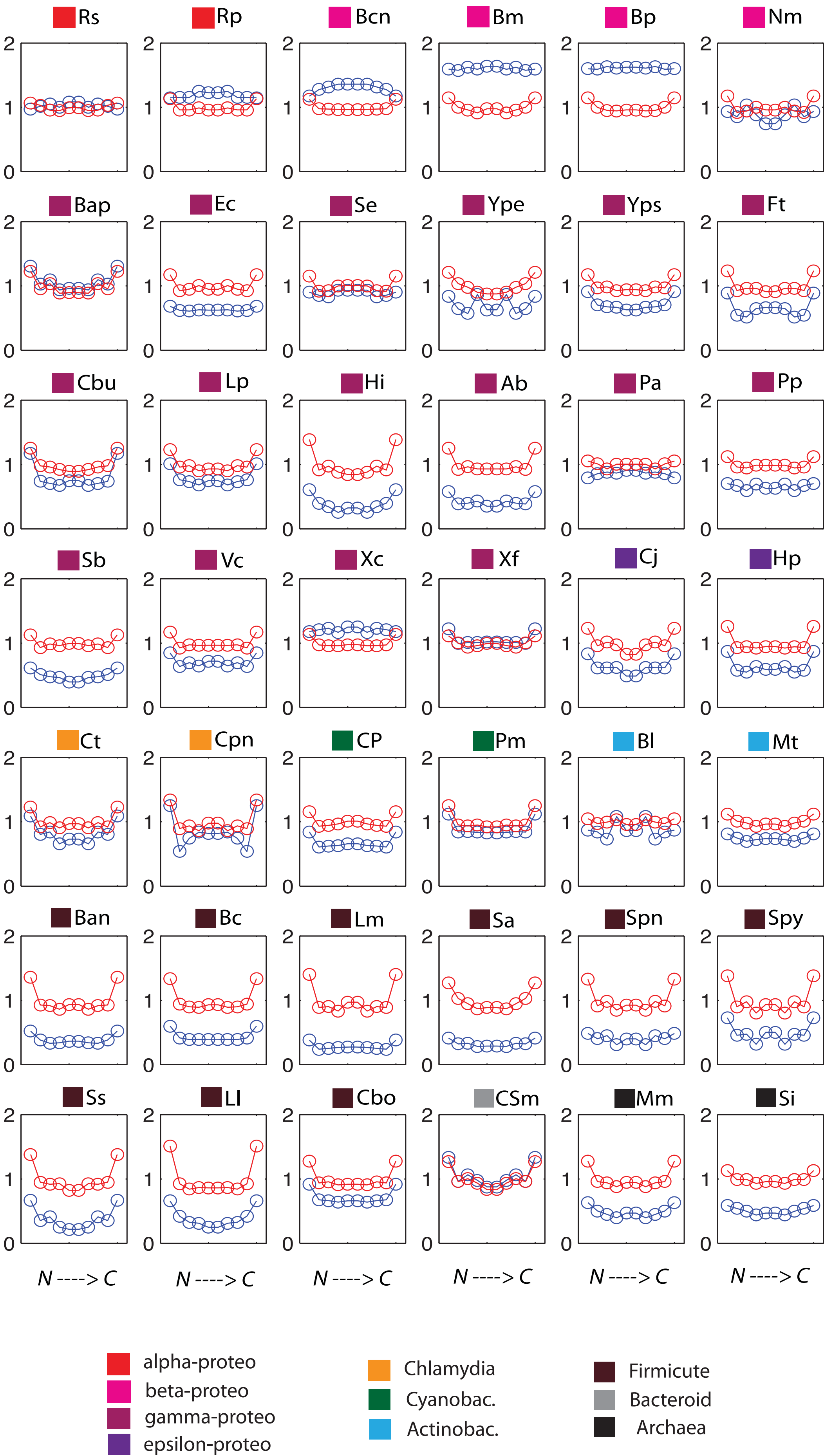
Supplemental Figure 2B



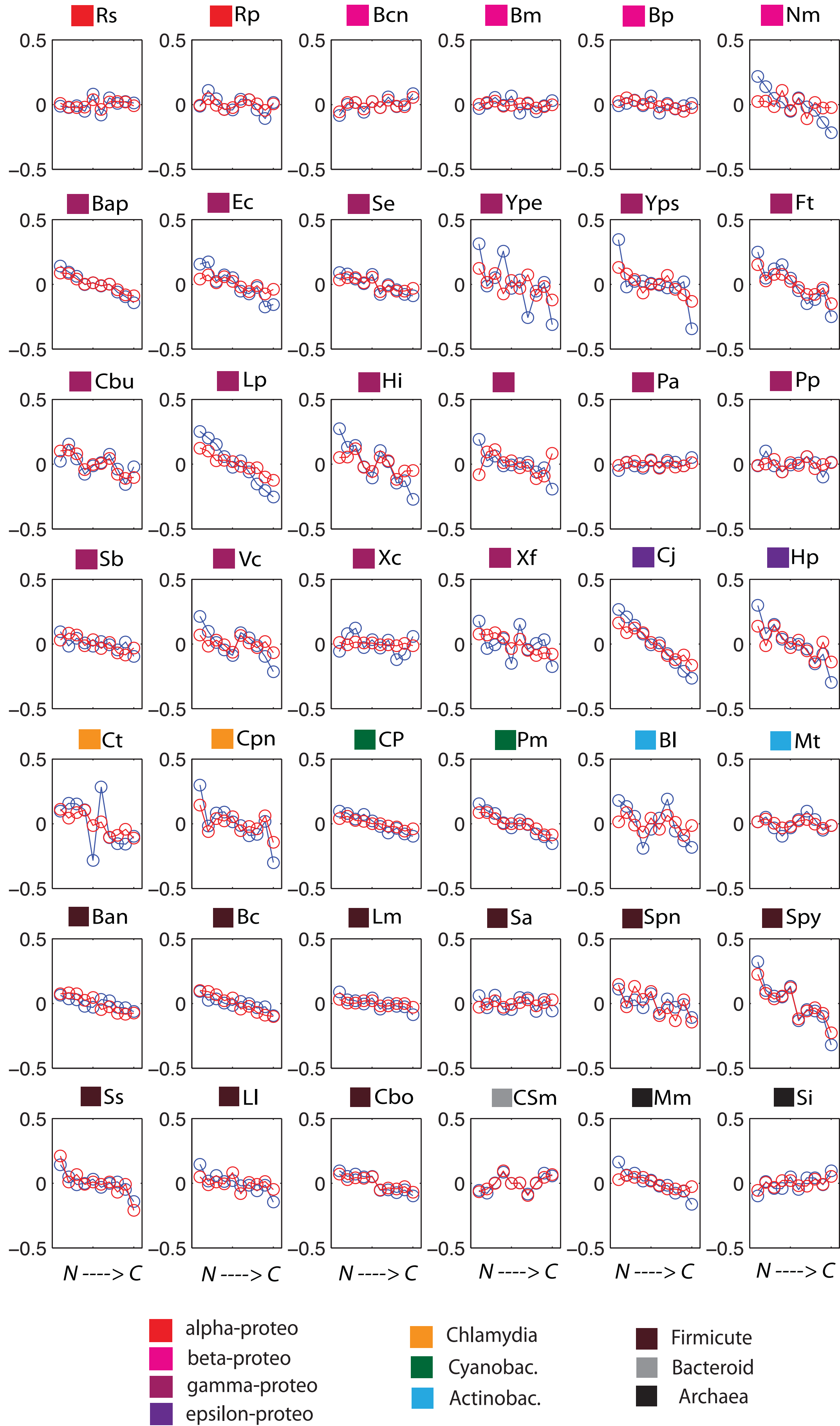
Supplemental Figure 3

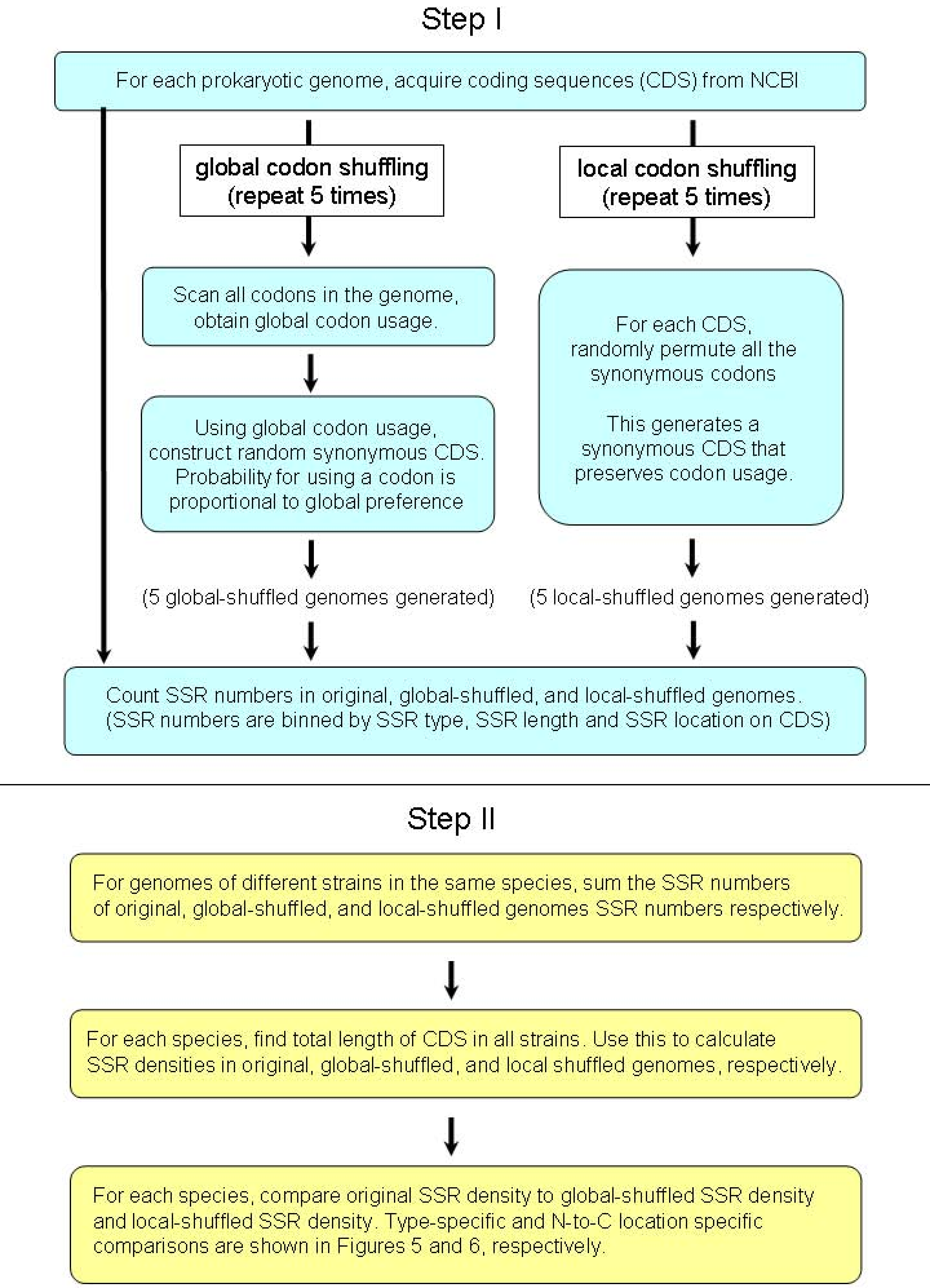


Supplemental Figure 4A



Supplemental Figure 4B




Step I
For each prokaryotic genome, acquire coding sequences (CDS) from NCBI
global codon shuffling (repeat 5 times)
local codon shuffling (repeat 5 times)
Scan all codons in the genome, obtain global codon usage.
Using global codon usage, construct random synonymous CDS. Probability for using a codon is proportional to global preference
For each CDS, randomly permute all the synonymous codons
This generates a synonymous CDS that preserves codon usage.
(5 global-shuffled genomes generated)
(5 local-shuffled genomes generated)
Count SSR numbers in original, global-shuffled, and local-shuffled genomes. (SSR numbers are binned by SSR type, SSR length and SSR location on CDS)
Step II
For genomes of different strains in the same species, sum the SSR numbers of original, global-shuffled, and local-shuffled genomes SSR numbers respectively.
For each species, find total length of CDS in all strains. Use this to calculate SSR densities in original, global-shuffled, and local shuffled genomes, respectively.
For each species, compare original SSR density to global-shuffled SSR density and local-shuffled SSR density. Type-specific and N-to-C location specific comparisons are shown in Figures 5 and 6, respectively.