

Supplementary Materials

Analysis of stop codons within prokaryotic protein-coding genes suggests frequent readthrough events

132,817 total putative nonsense substitutions

104,457 after removal of 'plasma' genomes – known or suspected of stop codon reassignments

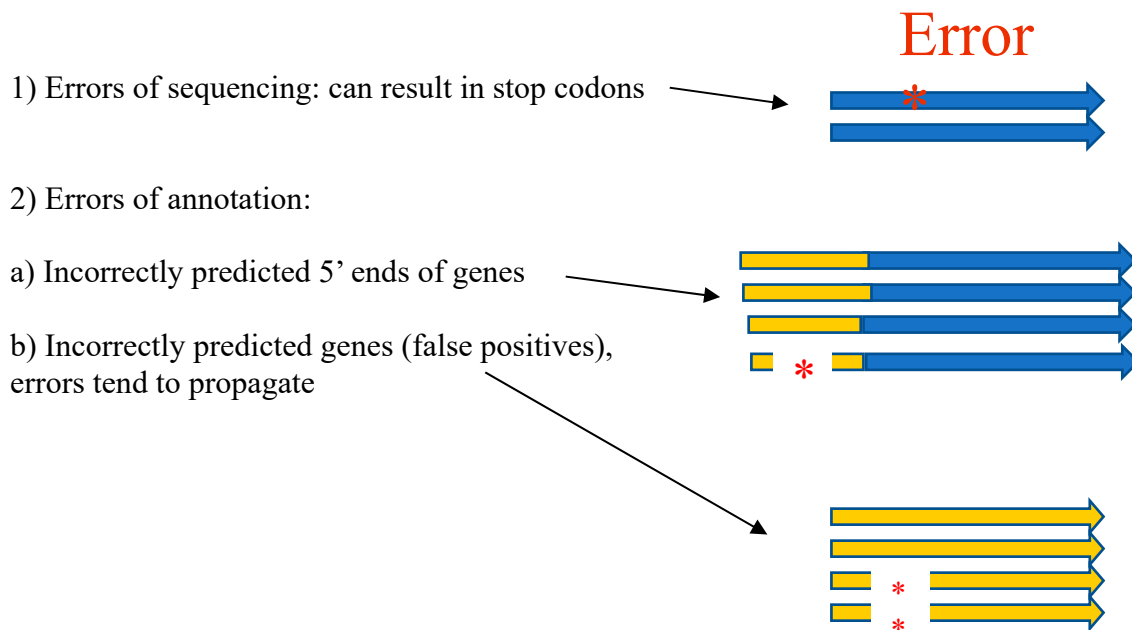
20,969 cases with at least two genomes with an orthologous nonsense substitution

15,695 cases where there are at least 40 bases upstream and 40 bases downstream to the stop (2,496 where both are on the + strand)

1,054 cases where the identity along the 83 positions is ≥ 50

Figure S1. Pipeline steps and the number of detected stop codons.

Figure S2. Major technical problems. Blue indicates real genes, yellow indicates errors of prediction.



3'-phosphoadenosine 5'-phosphosulfate sulfotransferase [Paenibacillus thiaminolyticus]

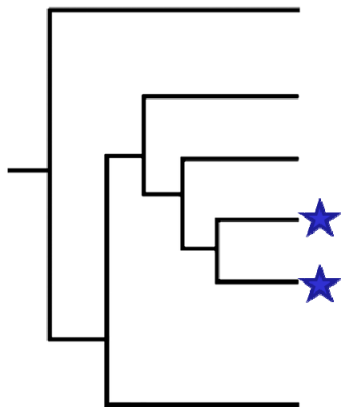
Sequence ID: WP_174818191.1 Length: 281 Number of Matches: 1

Alignment statistics for match #1

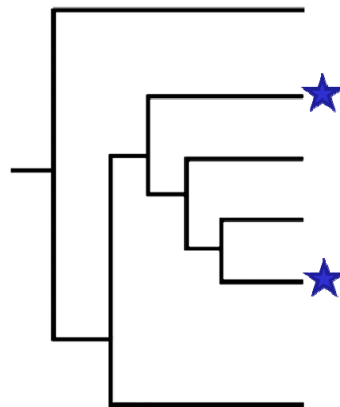
Score	Expect	Method	Identities	Positives	Gaps
204 bits (520)	2e-61	Composition-based stats.	58/251 (23%)	101/251 (40%)	12/251 (4%)

Query	2	RQAFKQKLIEIIPALQGRVYDVQPPSQTAEEPYAVMALGEEIWKSSWAGYRQVVRIKLYA	61
		R +Q+L++ IPA+ G++ D +PY V+ +G E ++ WAG + + YA	
Sbjct	11	RSVIRQRLLDRIPAVNGKILDAGTAEAGEAKPYLVLTIGSETVENDWAGSSCKIEVAPYA	70
Query	62	GQAGLAQADVWANALISGLHREPVTGAGEDTSAFTAHYLGVRDAEKLDVTGKAYRTLRF	121
		A L D A A+IS L R+ +T A S Y+G ++ +D R+++F	
Sbjct	71	PPAELPHVDSLAAAVISALDRQRLTDAISGKSILF-RYIGT-GSDTVDEKLKAVARSVQF	128
Query	122	GVYVPETEGDSAVPANGATQPEEWLAALVRWTQKQLGETWSIYANAWPAQPGRHAVLWRM	181
		V+ + + +AA+ WT+ + E + AW A+ WR+	
Sbjct	129	EVFSL-----GWMSHTPLDPPVAAMAAWTRARFPEIET-DPVAWNPSEDTPALYWRV	180
Query	182	SGCETRMAGASMYELRKRFIGHITAPDTTEENRAASALIEGFAAQIQLPLDQDKGRYMST	241
		+ + + + GH+ P +R + A Q+ + M	
Sbjct	181	AAIRSIQTVSQGAWIEATLRGHMLVPHAEARSRWLDLTVRQLALDGQVRM--LDQSPMLI	238
Query	242	AEASADLQADA	252
		SAD DA	
Sbjct	239	QSVSADGTQDA	249

Figure S3. PSI-BLAST (iteration#2, default parameters) output for putative 3'-phosphoadenosine 5'-phosphosulfate sulfotransferase proteins (Query: WP_058711956 and Sbjct: WP_036625970).



Phyletic index=2/2=1



Phyletic index=2/4=0.5

Figure S4. Distribution of nonsense substitution on species tree: phyletic index.

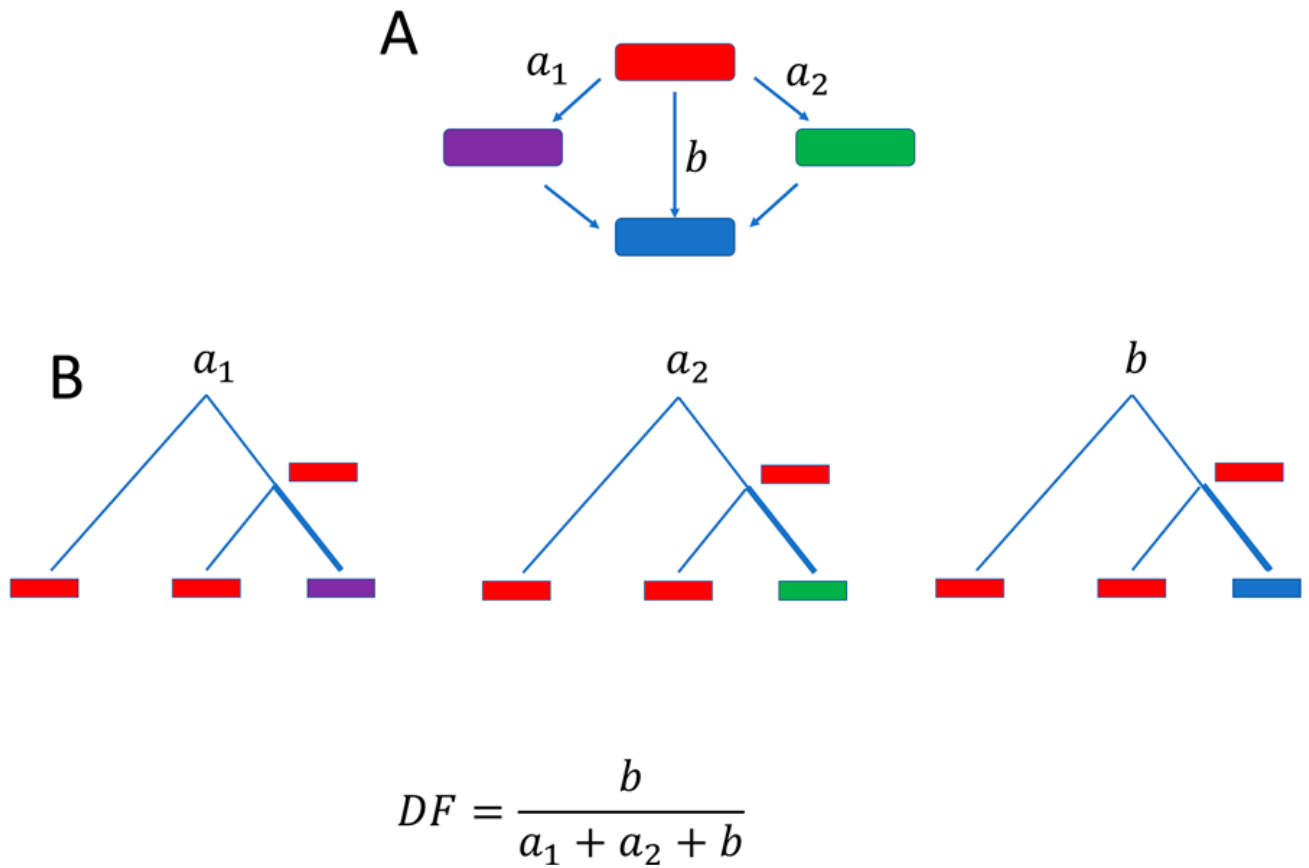


Figure S5. Conceptual scheme of double substitution analysis and the **double fraction (DF) measure**. (A) Point mutations are assumed to appear one at a time, such that observed double substitutions (b) occur through intermediate single substitution states. For each double substitution instance, there are two possible single substitution trajectories (a_1 and a_2). (B) Instances of single or double substitutions are inferred from the genomic data by construction of genomes triplets and relying on parsimony principle (see the “Methods” section). In brief, the parsimony principle implies that mutations occur along the thick branches in the trees shown in (B). The double fraction is defined as the ratio between the number of double substitution instances b and the sum of relevant single ($a_1 + a_2$) and double (b) substitution instances.

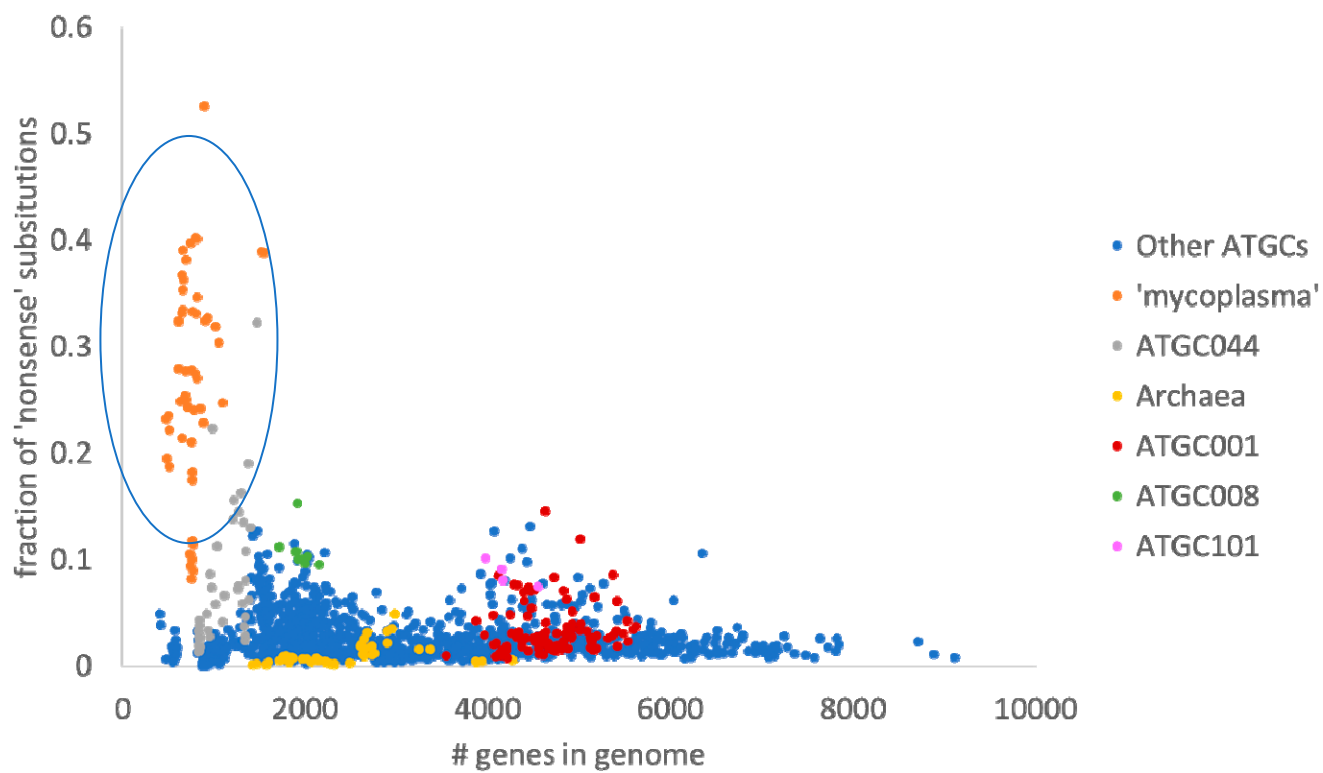


Figure S7. Fraction of nonsense substitution compared to the genome size. ATGC001 is the *Enterobacter* spp. group; ATGC008 is the *Streptococcus* spp. group; ATGC044 is the *Rickettsia* spp. group; ATGC101 is the *Dickeya* spp. group. The removed 'mycoplasma' group corresponds to ATGC32-ATGC40. "Other ATGCs" corresponds to all other ATGCs from the ATGC database used in this study. The blue ellipse shows obvious outliers.

Table S1. List of *Paenibacillus* species that are shown in Figures 1 and 2.

CP006941	<i>Paenibacillus polymyxa</i>
CP017967	<i>Paenibacillus polymyxa</i>
CP011512	<i>Paenibacillus peoriae</i>
CP017968	<i>Paenibacillus polymyxa</i>
CP000154	<i>Paenibacillus polymyxa</i>
CP011420	<i>Paenibacillus polymyxa</i>
CP015423	<i>Paenibacillus polymyxa</i>
CP034141	<i>Unidentified Paenibacillus sp.</i>
HE577054	<i>Paenibacillus polymyxa</i>
CP042272	<i>Paenibacillus polymyxa</i>
CP040829	<i>Paenibacillus polymyxa</i>
CP025957	<i>Paenibacillus polymyxa</i>
CP009909	<i>Paenibacillus polymyxa</i>
CP006872	<i>Paenibacillus polymyxa</i>
CP013203	<i>Unidentified Paenibacillus sp.</i>
CP003107	<i>Paenibacillus terrae</i>
CP020028	<i>Paenibacillus kribbensis</i>
CP045298	<i>Paenibacillus brasiliensis</i>

Table S2. Description of ATGC-COG functional categories where in-frame stop codons have been detected.

COG functional category	Number of orthologous genes containing in-frame stop codons	Number of singletons
[X] Mobilome: phages and transposons	1311	8578
[R] General function prediction only	610	5516
[G] Carbohydrate transport and metabolism	508	6689
[E] Amino acid transport and metabolism	472	7600
[S] Function unknown	348	2003
[T] Signal transduction mechanisms	371	5389
[K] Transcription	363	2836
[M] Cell wall/membrane/envelope biogenesis	335	3218
[L] Replication, recombination, and repair	306	3269
[P] Inorganic ion transport and metabolism	287	4027

Orthologous genes containing in-frame stop codons – in-frame stop codons that are present within orthologous protein coding genes shared by two or more bacterial species. Singletons – in-frame stop codons that were found only once (do not have orthologs). The most frequent categories are shown. ATGC-COG functional categories are ordered according to the “Number of orthologous genes containing in-frame stop codons” column.

Table S3. Statistics of double mutations and modes of selection

Codon change		Null model - double synonymous changes				Number of observed changes in codons		Fisher test	Class	mode of selection
		ancestral	de-derived	double	single					
		ancestral	de-derived	double	single	ancestral	double	single	ancestral	after Bonferoni
CAG	TCG	39	1023	14713	201	1288	630066	1.40E-18	XNnN	+
TGC	TCA	56	2067	184686	36	176	154212	1.02E-15	XNsN	+
TAT	GAA	23	759	13413	52	243	325159	4.22E-15	XNnN	+
TAC	GAA	22	711	13176	28	126	336894	1.71E-10	XNnN	+
TAC	AAA	14	524	13176	30	144	336894	2.62E-10	XNnN	+
TAT	AAA	47	828	13413	63	323	325159	1.75E-09	XNnN	+
TGG	GGA	329	7154	51165	38	252	350601	7.19E-09	XNsN	+
CAG	TAT	54	3107	72658	68	1514	630066	2.83E-07	XNnN	+
TCG	CAG	20	542	7574	160	1414	314455	3.22E-07	XNnN	+
TGC	TTA	9	1223	184686	9	150	154212	5.24E-05	XNnN	+
AAG	TCG	18	703	20988	91	1354	534568	9.50E-05	XNnN	+
TAT	GAG	17	641	13413	20	211	325159	1.99E-04	XNnN	+
TAT	AAG	22	691	13413	28	288	325159	2.10E-04	XNnN	+
TAC	AAG	16	477	13176	17	131	336894	3.73E-04	XNnN	+
TGG	AGA	242	6275	51165	12	95	350601	7.54E-04	XNsN	+
TAC	GAG	29	669	13176	15	113	336894	1.82E-03	XNnN	=
CAG	TAC	163	7212	72658	62	1844	630066	1.19E-02	XNnN	=
TAC	CAG	62	2089	13176	52	1113	336894	2.13E-02	XNnN	=
CAA	TAT	51	990	7534	70	879	270512	2.40E-02	XNnN	=
GAG	TCG	6	324	8129	129	3634	668983	1.46E-01	XNnN	=
TGC	GGA	120	4369	84580	11	263	154212	1.82E-01	XNnN	=
TAC	CAA	54	2130	13176	34	1095	336894	3.64E-01	XNnN	=
TGT	CGA	270	9300	100899	4	98	60699	5.38E-01	XNnN	=
AGA	TGC	100	3299	54066	5	161	94676	8.16E-01	XNnN	=
TGG	CGA	795	11823	51165	27	399	350601	9.19E-01	XNnN	=

TAC	TTA	2	134	8037	29	1566	336894	1.00E+00	XNn N	=
TTA	TAC	5	129	3439	25	610	269701	1.00E+00	XNn N	=
GGA	TGG	267	8630	56720	31	23114	186121	9.69E-113	XSN	-
AGA	TGG	252	5843	54066	7	7916	94676	1.08E-81	XSN	-
CGA	TGG	752	11369	53638	9	3856	38022	2.66E-78	XSN	-
GGA	TGC	131	3465	56720	5	8789	186121	5.25E-64	XSN	-
GGA	TGT	100	2699	56720	15	8775	186121	2.28E-46	XSN	-
CGA	TGC	302	7107	53638	8	2152	38022	2.08E-24	XSN	-
CGA	TGT	277	5907	53638	8	1619	38022	2.22E-19	XSN	-
TCA	TGC	16	212	4610	26	3930	122445	1.69E-10	XSN	-
GAA	TAT	24	548	11884	71	4927	747383	2.03E-05	XNn N	-
GAA	TAC	15	547	11884	42	5754	747383	8.62E-05	XNn N	-
CAA	TAC	63	1024	7534	36	1147	270512	1.39E-03	XNn N	=
AGA	TGT	241	3795	54066	2	154	94676	8.10E-03	XNn N	=
GAG	TAT	15	814	70692	37	4449	668983	1.23E-02	XNn N	=
TCA	TAC	3	179	4610	13	3917	122445	3.12E-02	XSN	=
AAA	TAC	30	983	20660	36	1819	574412	9.04E-02	XNn N	=
AAG	TAT	18	462	12030	28	1157	534568	1.37E-01	XNn N	=
GAG	TAC	29	3740	70692	30	5831	668983	1.40E-01	XNn N	=
AAA	TAT	58	1291	20660	79	2293	574412	1.47E-01	XNn N	=
TAT	CAA	59	1249	13413	63	1745	325159	1.60E-01	XNn N	=
AAG	TAC	18	558	12030	18	886	534568	1.71E-01	XNn N	=
TCG	GAG	11	271	7574	105	3891	314455	1.86E-01	XNn N	=
TCG	AAG	12	191	7574	65	1471	314455	2.75E-01	XNn N	=
TGC	AGA	108	3164	84580	8	355	154212	3.43E-01	XNn N	=
TAT	CAG	36	1114	13413	47	1729	325159	4.94E-01	XNn N	=
TGC	CGA	315	13019	84580	6	286	154212	1.00E+00	XNs N	=
TGT	GGA	179	5403	100899	3	117	60699	1.00E+00	XNs N	=
TTA	TGC	3	106	3439	16	601	269701	1.00E+00	XNn N	=

+, positive; -, negative; =, neutral selection cannot be rejected.