Supplementary Material For

Carl Kingsford, Arthur Delcher, Steven L. Salzberg. A unified model explaining the offsets of overlapping and near-overlapping prokaryotic genes.

Estimation of p-values in section "Phase Bias in closely spaced genes"

The probability estimates of a particular phase ϕ being over-represented at random assume equal probabilities of $\frac{1}{3}$ for each of the 3 possible offset phases for a single tail-to-tail pair of genes. In a genome with k tail-to-tail pairs, the probability that strictly more pairs occur in phase ϕ than in either of the other two phases is strictly less than $\frac{1}{3}$ because ties are not counted. The exact probability depends on k, but for all values of k it is less than $\frac{1}{3}$. We therefore obtain a conservative upper bound on the probability that a out of n genomes have a strict maximum of pairs in phase ϕ, from the tail of a binomial distribution for a successes in n trials with success probability $\frac{1}{3}$. The same calculation applies for a strict minimum of pairs to occur in any given phase.

Similarly, the probability that a particular strict ordering of phase counts, say phase $0 >$ phase $1 >$ phase 2, occurs in a genome with k pairs is strictly less than $\frac{1}{6}$, because there are 6 equally likely orderings, but we have eliminated orderings with tie counts. Thus, we can bound the probability that a out of n genomes exhibit a particular ordering using a binomial distribution with success probability $\frac{1}{6}$.

Tables of phase distributions for subsets of organisms

The following tables give the percentage of overlaps and near-overlaps over subsets of the non-redundant set of 220 genomes described in the paper. The subsets were defined either taxonomically (using the NCBI taxonomy tree) or based on the GC content of the

genomes. In each of the three following tables, "Group" specifics a subset of organisms, the "Overlaps $-98\ldots-6$ column counts the percentage of overlaps with lengths between 98 and 6 that are in the given phase, the "Near-laps $0\ldots44$" column gives the percentage of non-overlapping, tail-to-tail pairs separated by ≤ 44 bases that are in each phase. Finally, the last two columns give the percentage of tail-to-tail pairs that overlap and the average position of the 3′-most reverse stop codon among tail-to-tail pairs from genomes in the given set.

In Table S1 below, the groups greedily divide the sorted non-redundant organisms into sets of organisms for which the GC content differs by less than 5%. The group name is the average GC content of the organisms in the group.

Table S1: Groups defined by GC content of 5% intervals

	Overlaps $-98\ldots-6$			Near-laps $0\ldots44$				Avg Last
Group	0	1	2	0	1	2	%Olaps	RC Stop
gc26	40.48	47.62	11.90	33.07	36.25	30.68	8.1	-40.8
gc30	39.16	46.99	13.86	35.42	34.19	30.38	16.4	-45.1
gc35	41.58	42.71	15.70	31.87	35.40	32.74	13.7	-48.3
gc40	32.48	49.58	17.94	35.27	31.85	32.88	8.8	-55.7
gc46	33.60	46.16	20.23	36.77	31.41	31.81	10.8	-65.5
gc51	28.59	49.67	21.74	37.39	33.63	28.98	20.5	-96.8
gc56	26.15	44.59	29.27	38.60	31.24	30.15	17.2	-125.3
gc62	20.06	44.55	35.40	43.62	31.27	25.11	12.6	-195.6
gc66	16.61	50.73	32.66	42.58	34.71	22.71	12.3	-271.1
gc71	16.85	53.05	30.11	40.31	37.24	22.45	19.0	-339.9
gc75	5.26	57.89	36.84	45.92	32.65	21.43	17.3	-461.3

In Table S2, the 220 organisms in the non-redundant set are divided into 10 groups of 22 organisms each. Again, the group name is based on the average GC content of the organisms in the group.

Table S2: Groups defined by deciles of GC content

	Overlaps $-98 \ldots -6$			Near-laps $0 \ldots 44$				Avg Last
Group	0	1	2	0	1	2	%Olaps	RC Stop
gc28	47.39	40.00	12.61	35.41	33.66	30.93	12.4	-44.4
gc34	35.77	47.75	16.47	31.47	36.08	32.45	12.6	-46.3
gc38	40.62	43.37	16.01	33.42	33.80	32.78	11.3	-52.4
gc42	31.68	49.92	18.39	36.29	31.60	32.11	9.9	-59.0
gc46	33.43	46.83	19.74	37.34	30.84	31.82	11.1	-64.1
gc50	29.46	48.91	21.63	36.72	33.60	29.68	18.2	-91.0
gc56	25.33	44.21	30.46	39.60	31.24	29.15	16.3	-132.3
gc61	21.27	44.41	34.31	42.56	31.24	26.20	12.3	-191.7
gc64	18.42	48.51	33.07	43.89	32.54	23.57	12.5	-240.4
gc69	14.88	52.63	32.49	42.19	36.19	21.61	14.5	-310.6

In Table S3 below, several taxonomic subsets of the organisms are considered, derived from the NCBI taxonomy tree. In addition, the set "Firm > 40% GC" considers only the Firmicutes with at least 40% GC content; "< 40% GC" and "> 60% GC" consider the organisms with less than 40% GC or more than 60% GC content respectively. Finally, the set "One / Genus" was selected by choosing a random member of each genus present in the non-redundant set (based on the first word of the species name). "All Non-redundant" gives the values for the complete set of 220 organisms.

Table S3: Taxonomic Groups

	Overlaps $-98 \ldots -6$			Near-laps $0 \ldots 44$				Avg Last
Group	0	1	2	0	1	2	%Olaps	RC Stop
Actinobacteria	17.69	52.31	30.00	40.43	35.86	23.70	17.7	-266.3
Archaea	35.78	47.86	16.36	37.99	33.46	28.55	24.2	-102.8
Bacteria	24.12	47.50	28.38	39.92	33.11	26.97	12.2	-156.6
Proteobacteria	22.77	47.09	30.14	41.14	32.53	26.34	10.7	-168.2
Firmicutes	35.87	46.32	17.81	33.46	33.35	33.19	6.3	-58.1
Firm > 40% GC	32.79	50.82	16.39	35.57	31.84	32.59	6.6	-72.1
< 40% GC	39.66	44.69	15.65	33.17	34.60	32.23	12.0	-48.6
> 60% GC	17.21	49.29	33.50	42.83	33.70	23.47	13.2	-253.0
One / Genus	25.72	47.71	26.57	39.96	33.06	26.99	14.0	-157.4
All Non-redundant	26.15	47.56	26.29	39.73	33.14	27.12	13.3	-152.6

Discussion of co-directed overlapping pairs

Although the main text is concerned with a phenomenon among tail-to-tail overlapping pairs, we briefly mention here that a similar effect is seen among co-directed gene pairs (tail-to-head pairs). Figure S1 below is the analogue of Figure 1 in the main text, but giving instead the distribution of overlap lengths and separation distances among co-directed ($\rightarrow\rightarrow$ or $\leftarrow\leftarrow$) pairs within the non-redundant set of genomes. Again, a phase bias in overlap lengths is clearly observed. In the extreme, in-frame (phase 0) overlaps are not possible. The large spike at -4 arises due to the `ATG` start codon ending with the `TG` beginning of the `TGA` stop codon.

Beyond overlaps of length 6, overlaps tend to be in phase 1 with fewer in phase 2. A similar bias is observed in the positions of the first (3′-most) stop codons (Figure S2 below, analogous to Figure 3 in the main text). Thus, it is likely that the bias distribution of overlap lengths in co-directed pairs is also significantly determined by the expected location of the stop codons.

In contrast to the tail-to-tail case, no bias is observed in the phase of the separation distances between closely spaced, co-directed genes. However, the position of start codons (especially those that do not cause overlaps) are a good deal less certain than the positions of stop codons, and there are many other factors (such as ribosomal binding sites and regulatory binding sites) constraining the separation of pairs of co-directed genes. Further, the extension of an upstream gene into a co-transcribed downstream gene may affect the transcription of the downstream gene in ways not applicable to tail-to-tail pairs.

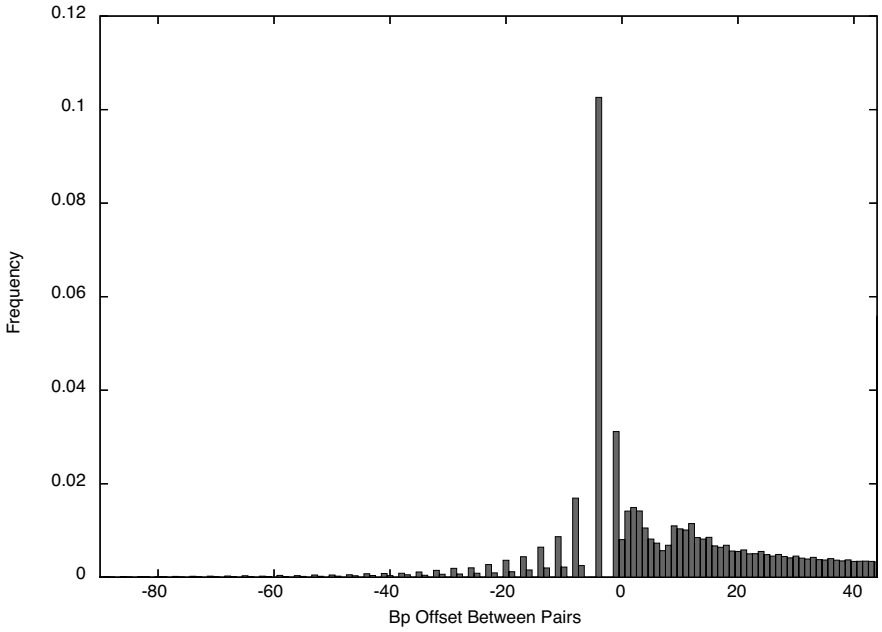

Figure S1: Distribution of overlaps and separations for co-directed gene pairs within the non-redundant genomes.

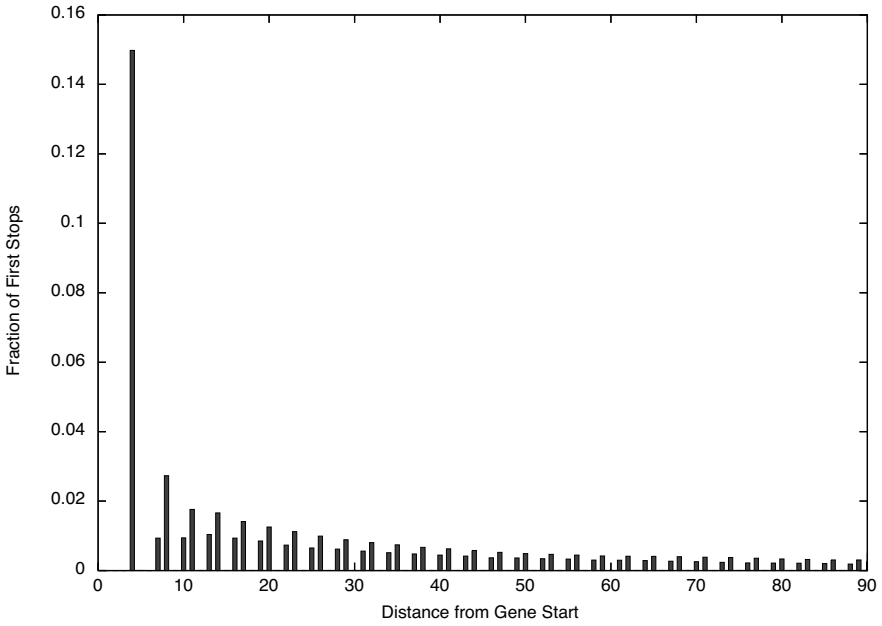

Figure S2: Distribution of the 5′-most stop-codons among all genes in the non-redundant set of genomes.