Table S4. Performance summary for Gene Cluster Method (GCM) at various Intergenic Distance Cutoffs (IDC) on KEGG pathway associations as benchmark for six reference genome sets

IDC	ALL	BAAC	BAS	BAC	GAMMA	BANR
100	0.76	0.76	0.78	0.76	0.8	0.75
200	0.75	0.75	0.77	0.75	0.79	0.75
300	0.73	0.74	0.76	0.75	0.8	0.76
400	0.72	0.73	0.75	0.74	0.8	0.76
500	0.71	0.72	0.74	0.73	0.8	0.76

Notes: The performance of GCM is higher for GAMMA and BAC reference genome sets. IDC of 100 nucleotide bases is seems to be optimal distance between adjacent genes to define gene clusters in the reference genomes. The performance summary of GCM measured as Area Under the ROC Curve (AUC). ALL, BAAC, BAS, BAC, GAMMA and BANR are reference genome sets whose compositions is given in Table 1. The proteins that share at least one KEGG served as positives examples otherwise negative examples.