Table S1. Performance summary for four computational methods using different reference genome sets on LQG dataset

Method	Variant	ALL	BAAC	BAS	BAC	GAMMA	BANR
PPM	BPPM	0.64	0.66	0.65	0.58	0.53	0.64
	SPPM	0.71	0.72	0.7	0.68	0.72	0.68
MDM	MDM	0.72	0.72	0.7	0.69	0.7	0.65
Mirrortree	Mirrortree	NA	NA	0.62	0.65	0.63	0.61
	Tol-mirrortree	NA	NA	0.64	0.66	0.63	0.6
	GD-Mirrortree	NA	NA	0.66	0.66	0.65	0.6

Notes: The performance summary of protein-protein prediction methods measured as Area Under the ROC Curve (AUC). BPPM stands for Binary Phylogenetic Profile Method; SPPM stands for Sequence Similarity (Bits scores) based Phylogenetic Profiling Method; GCM is Gene Cluster Method, MDM is gene neighborhood based Minimum Distance Method; GD is genome distance; NA stands for sets that are not analyzed for corresponding methods. ALL, BAAC, BAS, BAC, GAMMA and BANR are reference genome sets whose compositions is given in Table 1.