Supplementary Material

Raw datasets

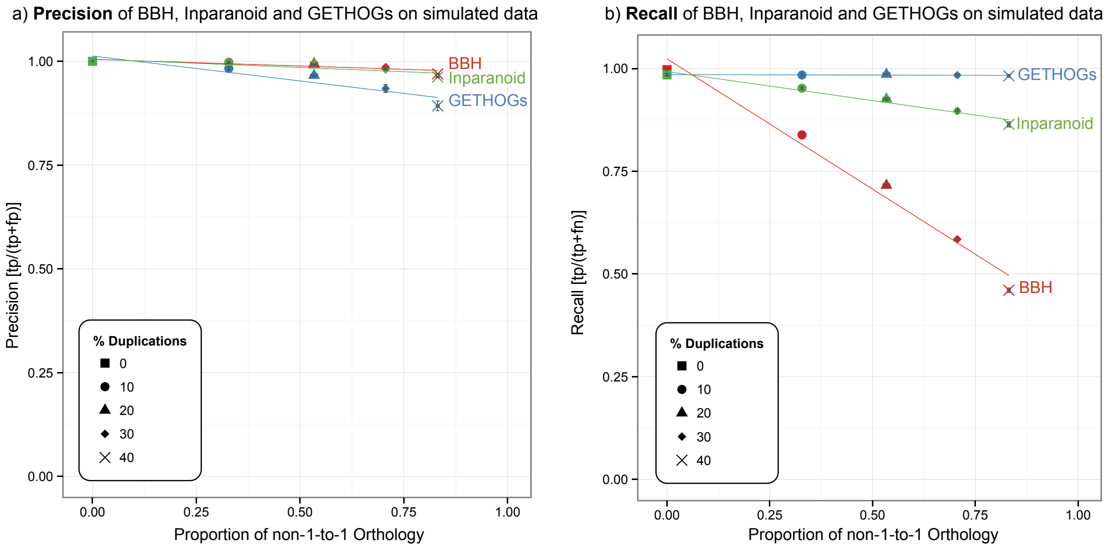
All raw datasets can be downloaded form the following URL:

- `http://lab.dessimoz.org/13_bbh`

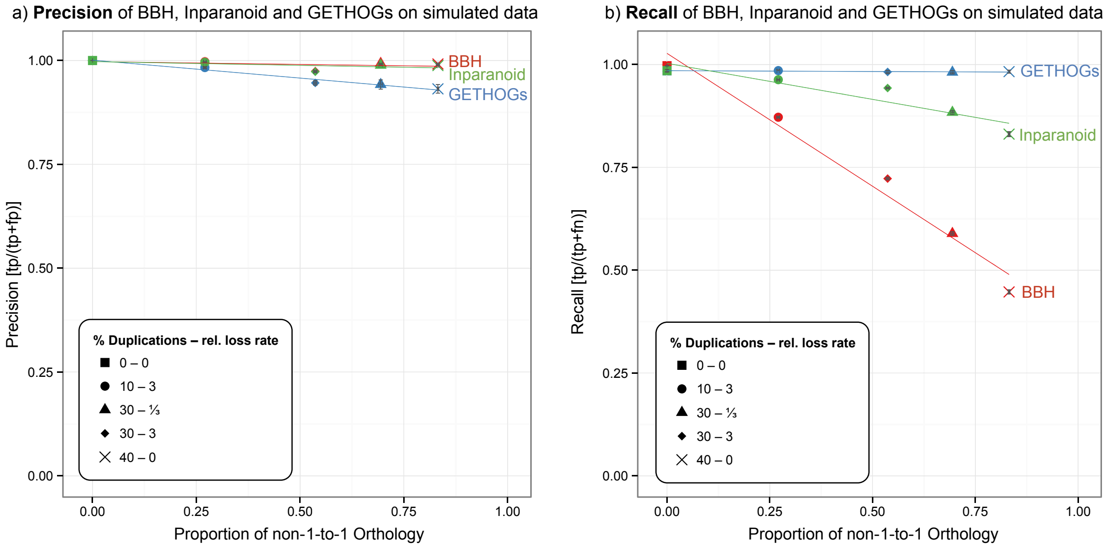
Additional experiments on simulated data (bacteria tree)

	% Duplications						
	equal loss rate			**varying relative loss rate**			
	20	**30**	**40**	**10**	**30**	**30**	**40**
parameters values							
# of sequences	1000						
distr. of seq. length	$\Gamma(k = 2.4, \theta = 133.8)$						
min. sequence length	50						
substitution model	WAG						
insertion and deletion rate	0.000125						
gene duplication rate	0.006	0.0105	0.017	0.0026	0.0087	0.009	0.0125
gene loss rate	0.006	0.0105	0.017	0.0078	0.0029	0.027	0
# of species	30						
key statistics							
seq. length - mean	321.6	324.4	323.7	323.1	317.9	320.9	324.6
seq. length - stderr	205.5	210.9	207.1	208.9	201.2	205.8	207.2
avg. % gap chars in MSA	22.5	21.44	20.35	19.54	26.48	12.08	31.11
variance of % gap chars	68.4	79.4	90.7	65.2	74.6	56.8	74.9
total tree length	763.6						
minimum tree height	31.70						
maximum tree height	77.80						
average tree height	41.36						
average pairwise distance	72.60						

Table 3: Key statistics for datasets of additional experiments (bacteria tree)

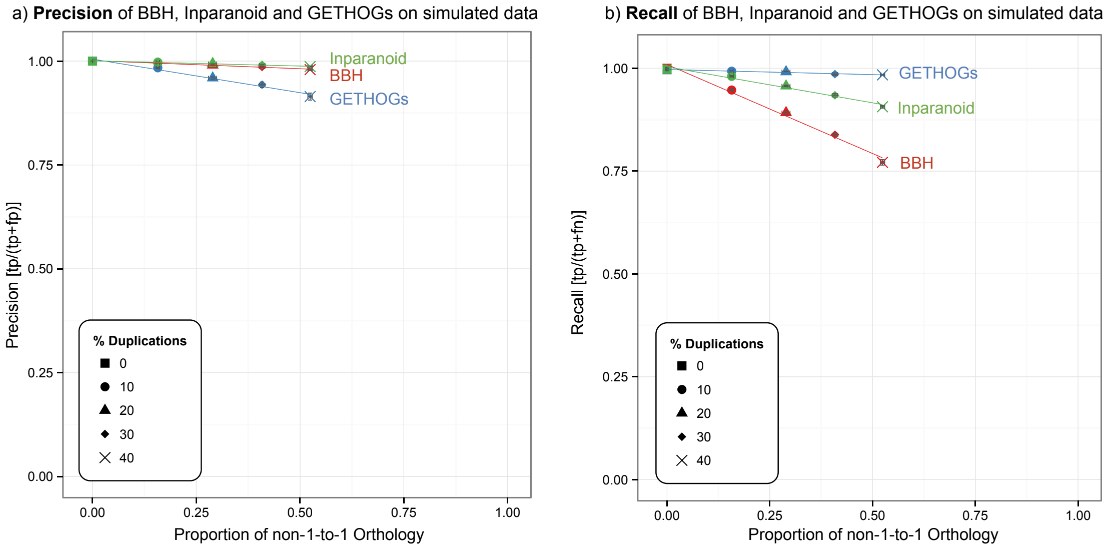


Supplementary Figure 1: Equal duplication and loss rates. Relationship between the proportion of non-1-to-1 orthology and precision/recall for BBH (in red) on simulated datasets with different proportions of genes with a history of duplications. Loss rates were equal to duplication rates. Results for Inparanoid (green) and OMA/GETHOGs (blue) are given for comparison. Each point corresponds to the mean value of five replicates. Error bars give the 95% confidence interval of the mean values in both dimensions.

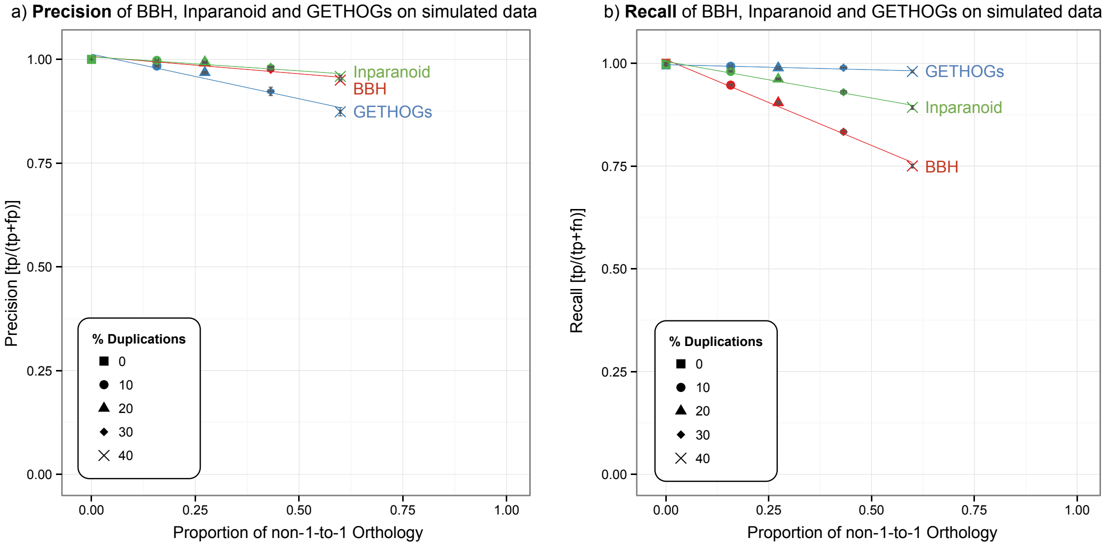


Supplementary Figure 2: Varying relative loss rates. Relationship between the proportion of non-1-to-1 orthology and precision/recall for BBH (in red) on simulated datasets with different proportions of genes with a history of duplications. Datasets with 10% duplications had a loss rate that was three times the duplication rate. For the datasets with 30% duplications, loss rate was either a third of or three times the duplication rate. For datasets with 40% duplications, loss rate was set to 0. Results for Inparanoid (green) and OMA/GETHOGs (blue) are given for comparison. Each point corresponds to the mean value of five replicates. Error bars give the 95% confidence interval of the mean values in both dimensions.

Additional experiments on simulated data (mammalia-like tree)



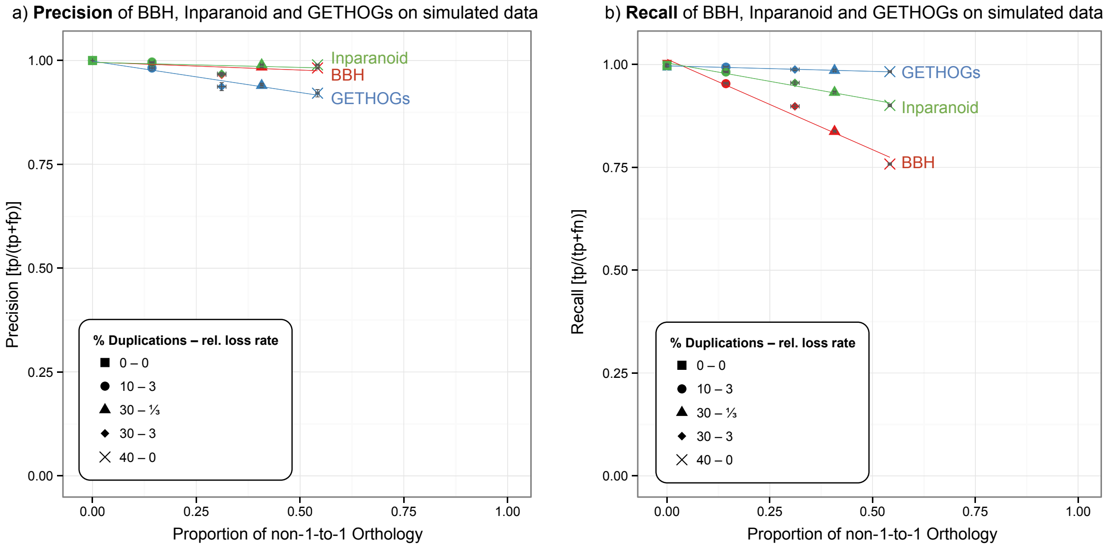
Supplementary Figure 3: Constant loss rates. Relationship between the proportion of non-1-to-1 orthology and precision/recall for BBH (in red) on simulated datasets with different proportions of genes with a history of duplications. Loss rate for all datasets was equal to the duplication rate of datasets with 10% duplication. Results for Inparanoid (green) and OMA/GETHOGs (blue) are given for comparison. Each point corresponds to the mean value of five replicates. Error bars give the 95% confidence interval of the mean values in both dimensions.



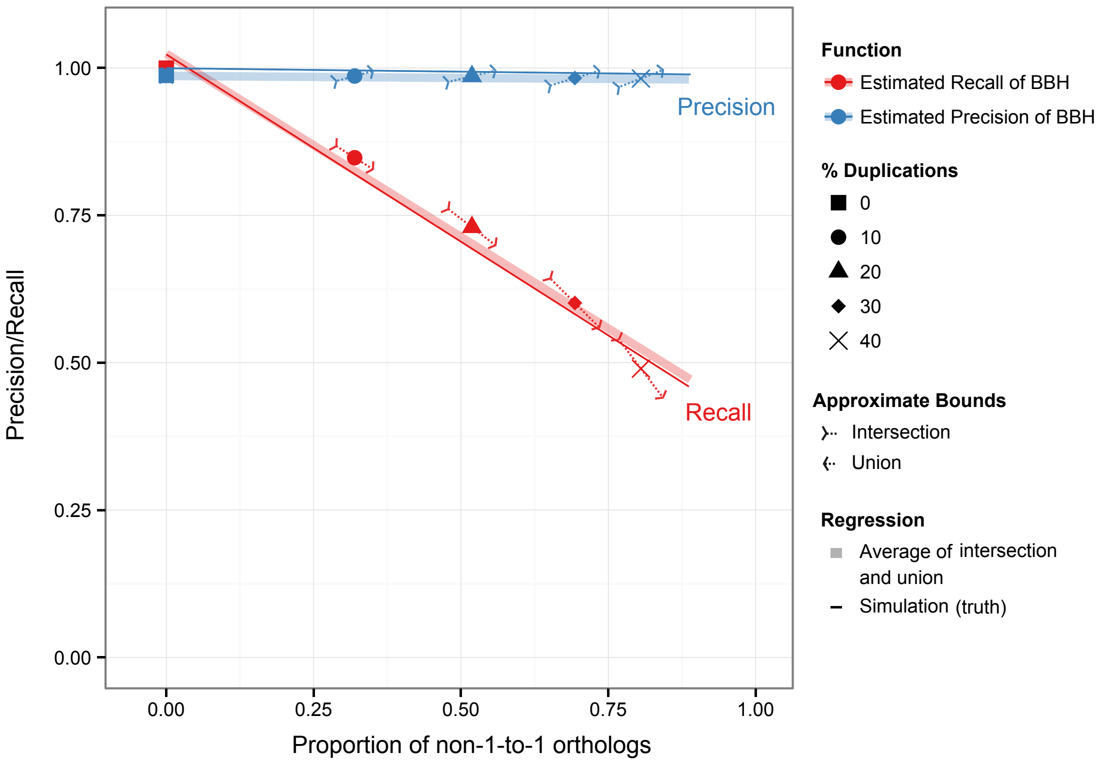
Supplementary Figure 4: Equal duplication and loss rates. Relationship between the proportion of non-1-to-1 orthology and precision/recall for BBH (in red) on simulated datasets with different proportions of genes with a history of duplications. Loss rates were equal to duplication rates. Results for Inparanoid (green) and OMA/GETHOGs (blue) are given for comparison. Each point corresponds to the mean value of five replicates. Error bars give the 95% confidence interval of the mean values in both dimensions.

	% Duplications											
		const. loss rate				equal loss rate			varying relative loss rate			
	0	**10**	**20**	**30**	**40**	**20**	**30**	**40**	**10**	**30**	**30**	**40**
parameters values												
# of sequences	1000											
distr. of seq. length	$\Gamma(k = 1.8, \theta = 274.1)$											
min. sequence length	50											
substitution model	WAG											
insertion and deletion rate	0.000125											
gene duplication rate	0	0.0065	0.013	0.0205	0.0295	0.013	0.025	0.0455	0.0065	0.0201	0.021	0.03
gene loss rate	0	0.0065				0.013	0.025	0.045	0.0195	0.0067	0.063	0
# of species	20											
key statistics												
seq. length - mean	487.6	485.8	483.0	487.6	482.8	488.6	488.1	495.5	483.9	483.9	484.2	481.1
seq. length - stderr	363.2	363.0	363.6	373.1	350.9	367.9	362.5	369.5	358.5	364.4	365.5	356.8
avg. % gap chars in MSA	4.0	3.87	4.27	4.74	5.29	3.8	3.86	3.75	3.26	4.72	2.35	5.89
variance of % gap chars	12.6	13.4	15.2	18.6	21.9	15.1	16.4	18.6	12.0	18.7	9.7	23.2
total tree length	101.2											
minimum tree height	14.70											
maximum tree height	19.18											
average tree height	17.48											
average pairwise distance	14.50											

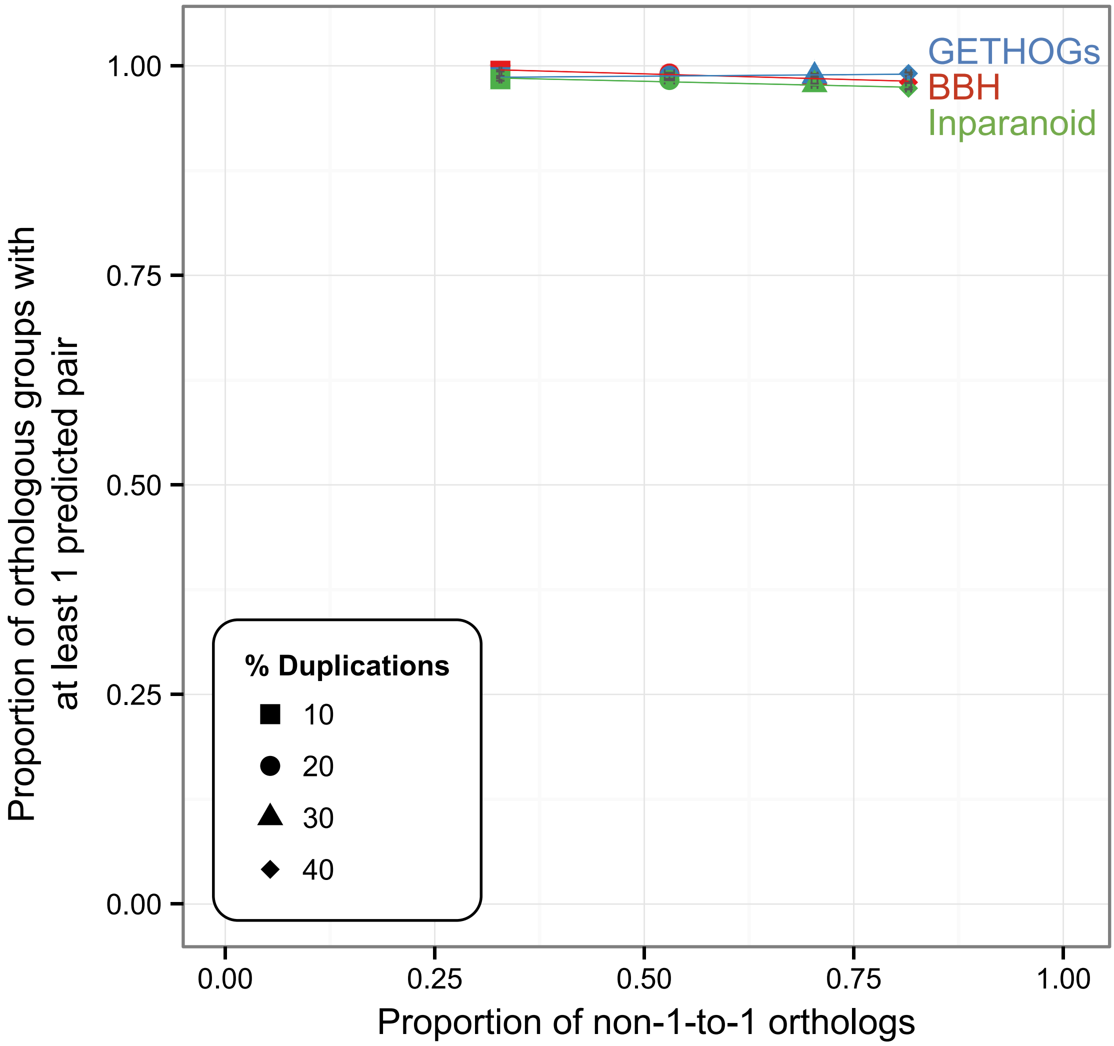
Table 4: Key statistics for datasets of additional experiments (mammalia-like tree)



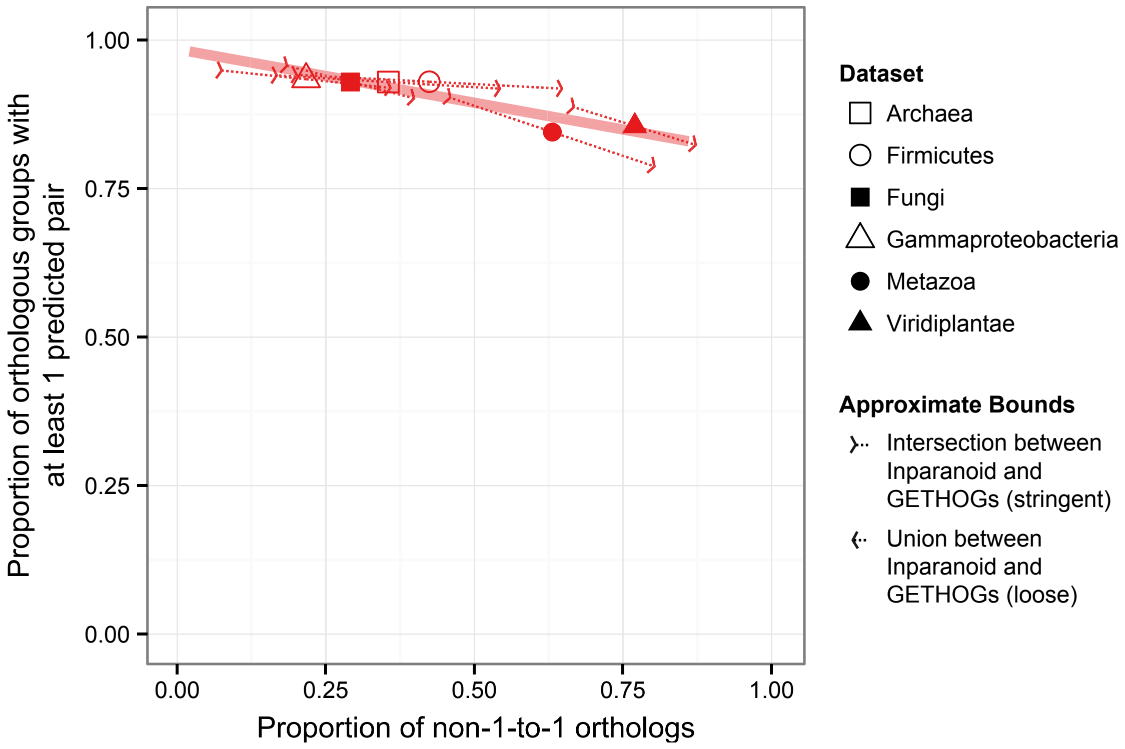
Supplementary Figure 5: Varying relative loss rates. Relationship between the proportion of non-1-to-1 orthology and precision/recall for BBH (in red) on simulated datasets with different proportions of genes with a history of duplications. Datasets with 10% duplications had a loss rate that was three times the duplication rate. For the datasets with 30% duplications, loss rate was either a third of or three times the duplication rate. For datasets with 40% duplications, loss rate was set to 0. Results for Inparanoid (green) and OMA/GETHOGs (blue) are given for comparison. Each point corresponds to the mean value of five replicates. Error bars give the 95% confidence interval of the mean values in both dimensions.



Supplementary Figure 6: Testing the intersection/union-bound approach on simulated data Precision and recall of BBH on simulated biological datasets, estimated from the intersection and union sets of orthologs inferred by Inparanoid and GETHOGs—the intersection yielding a lower bound and the union yielding an upper bound for precision and recall. Visually, the estimated trendlines (thick lines) coincide well with the true trendlines (slim solid lines).



Supplementary Figure 7: Suitability of BBH to identify ortholog seeds (simulation). Proportion of orthologous groups—defined here as sets of n:m orthologs—for which at least one pair is predicted by BBH (red), as a function of the proportion of non-1-to-1 orthologs. Analysis on simulated datasets, with different proportions of genes with a history of duplications. Loss rate for all datasets was equal to the duplication rate of datasets with 10% duplication. Results for Inparanoid (green) and OMA/GETHOGs (blue) are given for comparison.



Supplementary Figure 8: Suitability of BBH to identify ortholog seeds (real data). Proportion of orthologous groups—defined here as sets of n:m orthologs—for which at least one pair is predicted by BBH, as a function of the proportion of non-1-to-1 orthologs. Analysis on real biological datasets, estimated from the intersection and union sets of orthologs inferred by Inparanoid and GETHOGs.

List of Species

species	# of sequences
Aciduliprofundum boonei (strain DSM 19572)	1539
Aeropyrum pernix (strain ATCC 700893)	1699
Ferroglobus placidus (strain DSM 10642)	2463
Haloferax volcanii (strain ATCC 29605)	3986
Ignisphaera aggregans (strain DSM 17230)	1929
Ignicoccus hospitalis (strain KIN4/I)	1434
Korarchaeum cryptofilum (strain OPF8)	1602
Methanosarcina acetivorans (strain ATCC 35395)	4463
Methanopyrus kandleri (strain AV19)	1690
Methanococcus maripaludis	1849
Methanoplanus petrolearius (strain DSM 11571)	2779
Methanothermobacter marburgensis (strain DSM 2133)	1755
Nanoarchaeum equitans (strain Kin4-M)	536
Nitrosopumilus maritimus (strain SCM1)	1795
Picrophilus torridus (strain ATCC 700027)	1535
Pyrococcus furiosus (strain ATCC 43587)	2052
Staphylothermus marinus (strain ATCC 43588)	1570
Sulfolobus islandicus (strain M.16.4)	2729
Thermofilum pendens (strain Hrk 5)	1876
Vulcanisaeta distributa (strain DSM 14429)	2492

Supplementary Table 1: List of species in archaea dataset.

species	**# of sequences**
Alicyclobacillus acidocaldarius subsp. acidocaldarius (strain ATCC 27009)	3059
Alkaliphilus metalliredigens (strain QYMF)	4463
Anaerococcus prevotii (strain ATCC 9321)	1795
Caldicellulosiruptor saccharolyticus (strain ATCC 43494)	2619
Clostridium beijerinckii (strain ATCC 51743)	5003
Clostridium cellulolyticum (strain ATCC 35319)	3286
Clostridium phytofermentans (strain ATCC 700394)	3889
Coprothermobacter proteolyticus (strain ATCC 35245)	1481
Desulfitobacterium hafniense (strain Y51)	5015
Desulfotomaculum reducens (strain MI-1)	3214
Finegoldia magna (strain ATCC 29328)	1813
Geobacillus sp. (strain Y412MC10)	6237
Halothermothrix orenii (strain H 168)	2324
Lactobacillus acidophilus (strain ATCC 700396)	1860
Leuconostoc mesenteroides subsp. mesenteroides (strain ATCC 8293)	2002
Moorella thermoacetica (strain ATCC 39073)	2450
Natranaerobius thermophilus (strain ATCC BAA-1301)	2836
Symbiobacterium thermophilum (strain T)	3312
Syntrophomonas wolfei subsp. wolfei (strain Goettingen)	2471
Thermoanaerobacter sp. (strain X514)	2322

Supplementary Table 2: List of species in firmicutes dataset.

species	**# of sequences**
Agaricus bisporus	10366
Ashbya gossypii (strain ATCC 10895)	4707
Neosartorya fumigata (strain ATCC MYA-4609)	9832
Botryotinia fuckeliana (strain B05.10)	16299
Candida albicans (strain WO-1)	5695
Candida glabrata (strain ATCC 2001)	4752
Cryptococcus neoformans	6231
Debaryomyces hansenii (strain ATCC 36239)	6263
Emericella nidulans (strain FGSC A4)	10510
Encephalitozoon cuniculi (strain GB-M1)	1895
Kluyveromyces lactis (strain ATCC 8585)	5230
Laccaria bicolor	20052
Lodderomyces elongisporus (strain ATCC 11503)	5772
Magnaporthe grisea	12637
Mycosphaerella graminicola	10932
Neurospora crassa (strain ATCC 24698)	7569
Phaeosphaeria nodorum (strain SN15)	16489
Scheffersomyces stipitis (strain ATCC 58785)	5799
Puccinia graminis f. sp. tritici (strain CRL 75-36-700-3)	20364
Schizosaccharomyces pombe (strain 972)	4958
Ustilago maydis (strain 521)	6510
Yarrowia lipolytica (strain CLIB 122)	6603
Saccharomyces cerevisiae (strain ATCC 204508)	6328

Supplementary Table 3: List of species in fungi dataset.

species	**# of sequences**
Acinetobacter baumannii (strain AYE)	3715
Acidithiobacillus ferrooxidans (strain ATCC 23270)	3119
Alcanivorax borkumensis (strain SK2)	2752
Baumannia cicadellinicola subsp. Homalodisca coagulata	595
Buchnera aphidicola subsp. Acyrthosiphon pisum (strain APS)	571
Buchnera aphidicola subsp. Baizongia pistaciae (strain Bp)	507
Buchnera aphidicola subsp. Cinara cedri (strain Cc)	365
Carsonella ruddii (strain PV)	182
Coxiella burnetii (strain Dugway 5J108-111)	2110
Dichelobacter nodosus (strain VCS1703A)	1273
Francisella tularensis subsp. novicida (strain U112)	1719
Legionella pneumophila (strain Paris)	3059
Marinomonas sp. (strain MWYL1)	4415
Pseudomonas aeruginosa (strain UCBPP-PA14)	5886
Ruthia magnifica subsp. Calyptogena magnifica	976
Shewanella baltica (strain OS195)	4616
Thiomicrospira crunogena (strain XCL-2)	2183
Thioalkalivibrio sp. (strain HL-EbGR7)	3272
Wigglesworthia glossinidia brevipalpis	612
Xanthomonas campestris pv. campestris (strain ATCC 33913)	4123

Supplementary Table 4: List of species in Gammaproteobacteria dataset.

species	# of sequences
Acyrthosiphon pisum	33992
Aedes aegypti	15129
Apis mellifera	10378
Bombyx mori	14577
Branchiostoma floridae	28464
Caenorhabditis remanei	31096
Capitella sp. 1	31562
Ciona intestinalis	15464
Daphnia pulex	30088
Drosophila virilis	14351
Gasterosteus aculeatus	21832
Helobdella robusta	23263
Ixodes scapularis	20454
Lottia gigantea	23514
Nematostella vectensis	26036
Pediculus humanus subsp. corporis	10733
Pristionchus pacificus	29363
Schistosoma mansoni	11404
Strongylocentrotus purpuratus	26882
Trichoplax adhaerens	11466

Supplementary Table 5: List of species in Metazoa dataset.

species	# of sequences
Arabidopsis lyrata	32085
Arabidopsis thaliana	27489
Chlamydomonas reinhardtii	16900
Hordeum vulgare var. distichum	29214
Zea mays	107258
Manihot esculenta	32737
Oryza sativa subsp. japonica	58246
Ostreococcus lucimarinus (strain CCE9901)	7410
Ostreococcus tauri	7860
Populus trichocarpa	41297
Sorghum bicolor	35447
Vitis vinifera	26207

Supplementary Table 6: List of species in Viridiplantae dataset.