

Supplementary material 1: List of genomes used in the study

(names are as in the KEGG database)

1. *Acaryochloris marina*
2. *Acholeplasma laidlawii*
3. *Acidiphilium cryptum* JF-5
4. *Acidithiobacillus ferrooxidans* ATCC 23270
5. *Acidithiobacillus ferrooxidans* ATCC 53993
6. *Acidobacteria bacterium*
7. *Acidobacterium capsulatum*
8. *Acidothermus cellulolyticus*
9. *Acidovorax avenae*
10. *Acidovorax* sp. JS42
11. *Acinetobacter baumannii* AB0057
12. *Acinetobacter baumannii* AB307-0294
13. *Acinetobacter baumannii* ACICU
14. *Acinetobacter baumannii* ATCC 17978
15. *Acinetobacter baumannii* AYE
16. *Acinetobacter baumannii* SDF
17. *Acinetobacter* sp. ADP1
18. *Actinobacillus pleuropneumoniae* AP76 (serotype 7)
19. *Actinobacillus pleuropneumoniae* JL03 (serotype 3)
20. *Actinobacillus pleuropneumoniae* L20 (serotype 5b)
21. *Actinobacillus succinogenes*
22. *Aeromonas hydrophila*
23. *Aeromonas salmonicida*
24. *Aeropyrum pernix*
25. *Agrobacterium radiobacter* K84
26. *Agrobacterium tumefaciens* C58
27. *Agrobacterium tumefaciens* C58 (Cereon)
28. *Agrobacterium vitis* S4
29. *Akkermansia muciniphila*
30. *Alcanivorax borkumensis*
31. *Aliivibrio salmonicida* LFI1238
32. *Alkalilimnicola ehrlichei*
33. *Alkaliphilus metalliredigens*
34. *Alkaliphilus oremlandii*
35. *Alteromonas macleodii*
36. *Anabaena* sp. PCC7120
37. *Anabaena variabilis*
38. *Anaerocellum thermophilum*
39. *Anaeromyxobacter dehalogenans* 2CP-1
40. *Anaeromyxobacter dehalogenans* 2CP-C
41. *Anaeromyxobacter* sp. Fw109-5
42. *Anaeromyxobacter* sp. K
43. *Anaplasma marginale* Florida
44. *Anaplasma marginale* St. Maries
45. *Anaplasma phagocytophilum*
46. *Anoxybacillus flavithermus*
47. *Aquifex aeolicus*
48. *Archaeoglobus fulgidus*
49. *Arcobacter butzleri*
50. *Aromatoleum aromaticum* EbN1
51. *Arthrobacter aurescens*
52. *Arthrobacter chlorophenolicus*
53. *Arthrobacter* sp. FB24
54. *Azoarcus* sp. BH72
55. *Azorhizobium caulinodans*
56. *Azotobacter vinelandii*
57. *Bacillus amyloliquefaciens*
58. *Bacillus anthracis* A2012
59. *Bacillus anthracis* Ames
60. *Bacillus anthracis* Ames 0581
61. *Bacillus anthracis* Sterne
62. *Bacillus cereus* 03BB102
63. *Bacillus cereus* AH187
64. *Bacillus cereus* AH820
65. *Bacillus cereus* ATCC 10987
66. *Bacillus cereus* ATCC 14579
67. *Bacillus cereus* B4264
68. *Bacillus cereus* G9842
69. *Bacillus cereus* Q1
70. *Bacillus cereus* ZK
71. *Bacillus cereus* subsp. cytotoxis NVH 391-98
72. *Bacillus clausii*
73. *Bacillus halodurans*
74. *Bacillus licheniformis* ATCC 14580
75. *Bacillus licheniformis* DSM13
76. *Bacillus pumilus*
77. *Bacillus subtilis*
78. *Bacillus thuringiensis* 97-27
79. *Bacillus thuringiensis* Al Hakam
80. *Bacillus weihenstephanensis*
81. *Bacteroides fragilis* NCTC9343
82. *Bacteroides fragilis* YCH46
83. *Bacteroides thetaiotaomicron*
84. *Bacteroides vulgatus*
85. *Bartonella bacilliformis*
86. *Bartonella henselae*
87. *Bartonella quintana*
88. *Bartonella tribocorum*
89. *Baumannia cicadellinicola*
90. *Bdellovibrio bacteriovorus*
91. *Beijerinckia indica*
92. *Bifidobacterium adolescentis*
93. *Bifidobacterium animalis* subsp. lactis AD011
94. *Bifidobacterium longum* DJO10A
95. *Bifidobacterium longum* NCC2705
96. *Bifidobacterium longum* subsp. infantis ATCC 15697
97. *Bordetella avium*
98. *Bordetella bronchiseptica*
99. *Bordetella parapertussis*
100. *Bordetella pertussis*
101. *Bordetella petrii*
102. *Borrelia afzelii*
103. *Borrelia burgdorferi*
104. *Borrelia burgdorferi* ZS7
105. *Borrelia duttonii*
106. *Borrelia garinii*
107. *Borrelia hermsii*
108. *Borrelia recurrentis*
109. *Borrelia turicatae*
110. *Brachyspira hyodysenteriae*
111. *Bradyrhizobium japonicum*
112. *Bradyrhizobium* sp. BTAi1
113. *Bradyrhizobium* sp. ORS278
114. *Brevibacillus brevis*
115. *Brucella abortus* 9-941
116. *Brucella abortus* S19
117. *Brucella canis*
118. *Brucella melitensis* 16M
119. *Brucella melitensis* ATCC 23457
120. *Brucella melitensis* biovar Abortus
121. *Brucella ovis*
122. *Brucella suis* 1330
123. *Brucella suis* ATCC 23445
124. *Buchnera aphidicola* 5A
125. *Buchnera aphidicola* APS
126. *Buchnera aphidicola* Bp
127. *Buchnera aphidicola* Cc
128. *Buchnera aphidicola* Sg
129. *Buchnera aphidicola* Tuc7
130. *Burkholderia ambifaria* MC40-6
131. *Burkholderia cenocepacia* AU1054
132. *Burkholderia cenocepacia* HI2424
133. *Burkholderia cenocepacia* J2315
134. *Burkholderia cenocepacia* MC0-3
135. *Burkholderia cepacia*
136. *Burkholderia mallei* ATCC 23344
137. *Burkholderia mallei* NCTC 10229
138. *Burkholderia mallei* NCTC 10247

139. *Burkholderia mallei* SAVP1
140. *Burkholderia multivorans* ATCC 17616 (JGI)
141. *Burkholderia multivorans* ATCC 17616 (Tohoku)
142. *Burkholderia phymatum*
143. *Burkholderia phytofirmans*
144. *Burkholderia pseudomallei* 1106a
145. *Burkholderia pseudomallei* 1710b
146. *Burkholderia pseudomallei* 668
147. *Burkholderia pseudomallei* K96243
148. *Burkholderia* sp. 383
149. *Burkholderia thailandensis*
150. *Burkholderia vietnamiensis*
151. *Burkholderia xenovorans*
152. *Caldicellulosiruptor saccharolyticus*
153. *Caldivirga maquilingensis*
154. *Campylobacter concisus* 13826
155. *Campylobacter curvus*
156. *Campylobacter fetus*
157. *Campylobacter hominis* ATCC BAA-381
158. *Campylobacter jejuni* 81-176
159. *Campylobacter jejuni* 81116
160. *Campylobacter jejuni* NCTC11168
161. *Campylobacter jejuni* RM1221
162. *Campylobacter jejuni* subsp. *doylei* 269.97
163. *Campylobacter lari*
164. *Candidatus Amoebophilus asiaticus*
165. *Candidatus Azobacteroides pseudotrichonymphae*
166. *Candidatus Blochmannia floridanus*
167. *Candidatus Blochmannia pennsylvanicus*
168. *Candidatus Carsonella ruddii*
169. *Candidatus Desulfococcus oleovorans*
170. *Candidatus Desulfurudis audaxviator*
171. *Candidatus Methanoregula boonei*
172. *Candidatus Methanosphaerula palustris*
173. *Candidatus Pelagibacter ubique*
174. *Candidatus Phytoplasma australiense*
175. *Candidatus Phytoplasma mali*
176. *Candidatus Protochlamydia amoebophila*
177. *Candidatus Ruthia magnifica*
178. *Candidatus Sulcia muelleri*
179. *Candidatus Vesicomysocius okutanii*
180. *Carboxydotherrmus hydrogenoformans*
181. *Caulobacter crescentus* CB15
182. *Caulobacter crescentus* NA1000
183. *Caulobacter* sp. K31
184. *Cellvibrio japonicus*
185. *Chlamydia muridarum*
186. *Chlamydia trachomatis* 434/Bu
187. *Chlamydia trachomatis* A/HAR-13 (serovar A)
188. *Chlamydia trachomatis* D/UW-3/CX (serovar D)
189. *Chlamydia trachomatis* L2b/UCH-1/proctitis
190. *Chlamydophila abortus*
191. *Chlamydophila caviae*
192. *Chlamydophila felis*
193. *Chlamydophila pneumoniae* AR39
194. *Chlamydophila pneumoniae* CWL029
195. *Chlamydophila pneumoniae* J138
196. *Chlamydophila pneumoniae* TW183
197. *Chlorobaculum parvum* NCIB 8327
198. *Chlorobaculum tepidum*
199. *Chlorobium chlorochromatii*
200. *Chlorobium limicola*
201. *Chlorobium phaeobacteroides* BS1
202. *Chlorobium phaeobacteroides* DSM 266
203. *Chloroflexus aggregans*
204. *Chloroflexus aurantiacus*
205. *Chloroflexus* sp. Y-400-fl
206. *Chloroherpeton thalassium*
207. *Chromobacterium violaceum*
208. *Chromohalobacter salexigens*
209. *Citrobacter koseri* ATCC BAA-895
210. *Clavibacter michiganensis* subsp. *michiganensis*
211. *Clavibacter michiganensis* subsp. *sepedonicus*
212. *Clostridium acetobutylicum*
213. *Clostridium beijerinckii*
214. *Clostridium botulinum* A ATCC 19397
215. *Clostridium botulinum* A ATCC 3502
216. *Clostridium botulinum* A Hall
217. *Clostridium botulinum* A2
218. *Clostridium botulinum* A3 Loch Maree
219. *Clostridium botulinum* B Eklund 17B
220. *Clostridium botulinum* B1 Okra
221. *Clostridium botulinum* E3
222. *Clostridium botulinum* F Langeland
223. *Clostridium cellulolyticum*
224. *Clostridium difficile*
225. *Clostridium kluyveri* DSM 555
226. *Clostridium kluyveri* NBRC 12016
227. *Clostridium novyi*
228. *Clostridium perfringens* 13
229. *Clostridium perfringens* ATCC 13124
230. *Clostridium perfringens* SM101
231. *Clostridium phytofermentans*
232. *Clostridium tetani* E88
233. *Clostridium thermocellum*
234. *Colwellia psychrerythraea*
235. *Coprothermobacter proteolyticus*
236. *Corynebacterium diphtheriae*
237. *Corynebacterium efficiens*
238. *Corynebacterium glutamicum* ATCC 13032 (Bielefeld)
239. *Corynebacterium glutamicum* ATCC 13032 (Kyowa Hakko)
240. *Corynebacterium glutamicum* R
241. *Corynebacterium jeikeium*
242. *Corynebacterium urealyticum*
243. *Coxiella burnetii* CbuG_Q212
244. *Coxiella burnetii* CbuK_Q154
245. *Coxiella burnetii* Dugway 5J108-111
246. *Coxiella burnetii* RSA 331
247. *Coxiella burnetii* RSA 493
248. *Cupriavidus taiwanensis*
249. *Cyanobacteria* Yellowstone A-Prime
250. *Cyanobacteria* Yellowstone B-Prime
251. *Cyanothece* sp. ATCC 51142
252. *Cyanothece* sp. PCC 7424
253. *Cyanothece* sp. PCC 7425
254. *Cyanothece* sp. PCC 8801
255. *Cytophaga hutchinsonii*
256. *Dechloromonas aromatica*
257. *Dehalococcoides ethenogenes*
258. *Dehalococcoides* sp. BAV1
259. *Dehalococcoides* sp. CBDB1
260. *Deinococcus deserti*
261. *Deinococcus geothermalis*
262. *Deinococcus radiodurans*
263. *Delftia acidovorans*
264. *Desulfatibacillum alkenivorans*
265. *Desulfotobacterium hafniense* DCB-2
266. *Desulfotobacterium hafniense* Y51
267. *Desulfobacterium autotrophicum*
268. *Desulfotalea psychrophila*
269. *Desulfotomaculum reducens*
270. *Desulfovibrio desulfuricans* ATCC 27774
271. *Desulfovibrio desulfuricans* G20
272. *Desulfovibrio vulgaris* DP4
273. *Desulfovibrio vulgaris* Hildenborough
274. *Desulfovibrio vulgaris* Miyazaki F
275. *Desulfurococcus kamchatkensis*
276. *Diaphorobacter* sp. TPSY
277. *Dichelobacter nodosus*
278. *Dictyoglomus thermophilum*
279. *Dictyoglomus turgidum*
280. *Dinoroseobacter shibae*
281. *Ehrlichia canis*

282. *Ehrlichia chaffeensis*
283. *Ehrlichia ruminantium* Gardel
284. *Ehrlichia ruminantium* Welgevonden (France)
285. *Ehrlichia ruminantium* Welgevonden (South Africa)
286. *Elusimicrobium minutum*
287. *Enterobacter sakazakii*
288. *Enterobacter* sp. 638
289. *Enterococcus faecalis*
290. *Erwinia carotovora*
291. *Erwinia tasmaniensis*
292. *Erythrobacter litoralis*
293. *Escherichia coli* 536 (UPEC)
294. *Escherichia coli* 55989
295. *Escherichia coli* APEC O1
296. *Escherichia coli* ATCC 8739
297. *Escherichia coli* CFT073 (UPEC)
298. *Escherichia coli* E24377A
299. *Escherichia coli* ED1a
300. *Escherichia coli* HS
301. *Escherichia coli* IAI1
302. *Escherichia coli* IAI39
303. *Escherichia coli* K-12 DH10B
304. *Escherichia coli* K-12 MG1655
305. *Escherichia coli* K-12 W3110
306. *Escherichia coli* LF82
307. *Escherichia coli* O127
308. *Escherichia coli* O157 EC4115 (EHEC)
309. *Escherichia coli* O157 EDL933 (EHEC)
310. *Escherichia coli* O157 Sakai (EHEC)
311. *Escherichia coli* S88
312. *Escherichia coli* SE11
313. *Escherichia coli* SECEC
314. *Escherichia coli* UMN026
315. *Escherichia coli* UTI89 (UPEC)
316. *Escherichia fergusonii*
317. *Exiguobacterium sibiricum*
318. *Fervidobacterium nodosum*
319. *Finegoldia magna*
320. *Flavobacterium johnsoniae*
321. *Flavobacterium psychrophilum*
322. *Francisella philomiragia*
323. *Francisella tularensis* subsp. *holarctica* FTNF002-00
324. *Francisella tularensis* subsp. *holarctica* LVS
325. *Francisella tularensis* subsp. *holarctica* OSU18
326. *Francisella tularensis* subsp. *mediasiatica* FSC147
327. *Francisella tularensis* subsp. *novicida* U112
328. *Francisella tularensis* subsp. *tularensis* FSC 198
329. *Francisella tularensis* subsp. *tularensis* SCHU S4
330. *Francisella tularensis* subsp. *tularensis* WY96-3418
331. *Frankia alni*
332. *Frankia* sp. CcI3
333. *Frankia* sp. EAN1pec
334. *Fusobacterium nucleatum*
335. *Gemmatimonas aurantiaca*
336. *Geobacillus kaustophilus*
337. *Geobacillus thermodenitrificans*
338. *Geobacter bemidjiensis*
339. *Geobacter lovleyi*
340. *Geobacter metallireducens*
341. *Geobacter* sp. FRC-32
342. *Geobacter sulfurreducens*
343. *Geobacter uraniumreducens*
344. *Gloeobacter violaceus*
345. *Gluconacetobacter diazotrophicus* PAI 5 (Brazil)
346. *Gluconacetobacter diazotrophicus* PAI 5 (JGI)
347. *Gluconobacter oxydans*
348. *Gramella forsetii*
349. *Granulobacter bethedensis*
350. *Haemophilus ducreyi*
351. *Haemophilus influenzae* 86-028NP (nontypeable)
352. *Haemophilus influenzae* PittEE
353. *Haemophilus influenzae* PittGG
354. *Haemophilus influenzae* Rd KW20 (serotype d)
355. *Haemophilus parasuis*
356. *Haemophilus somnus* 129PT
357. *Haemophilus somnus* 2336
358. *Hahella chejuensis*
359. *Haloarcula marismortui*
360. *Halobacterium salinarum* R1
361. *Halobacterium* sp. NRC-1
362. *Haloquadratum walsbyi*
363. *Halorhodospira halophila*
364. *Halorubrum lacusprofundi*
365. *Halothermothrix orenii*
366. *Helicobacter acinonychis*
367. *Helicobacter hepaticus*
368. *Helicobacter pylori* 26695
369. *Helicobacter pylori* G27
370. *Helicobacter pylori* HPAG1
371. *Helicobacter pylori* J99
372. *Helicobacter pylori* P12
373. *Helicobacter pylori* Shi470
374. *Heliobacterium modesticaldum*
375. *Hermiinimonas arsenicoxydans*
376. *Herpetosiphon aurantiacus*
377. *Hydrogenobaculum* sp. Y04AAS1
378. *Hyperthermus butylicus*
379. *Hyphomonas neptunium*
380. *Idiomarina loihiensis*
381. *Ignicoccus hospitalis*
382. *Jannaschia* sp. CCS1
383. *Kineococcus radiotolerans*
384. *Klebsiella pneumoniae*
385. *Klebsiella pneumoniae* 342
386. *Kocuria rhizophila*
387. *Lactobacillus acidophilus*
388. *Lactobacillus brevis*
389. *Lactobacillus casei* ATCC 334
390. *Lactobacillus casei* BL23
391. *Lactobacillus delbrueckii* ATCC 11842
392. *Lactobacillus delbrueckii* ATCC BAA-365
393. *Lactobacillus fermentum*
394. *Lactobacillus gasserii*
395. *Lactobacillus helveticus*
396. *Lactobacillus johnsonii*
397. *Lactobacillus plantarum*
398. *Lactobacillus reuteri* DSM 20016
399. *Lactobacillus reuteri* JCM 1112
400. *Lactobacillus sakei*
401. *Lactobacillus salivarius*
402. *Lactococcus lactis* subsp. *cremoris* MG1363
403. *Lactococcus lactis* subsp. *cremoris* SK11
404. *Lactococcus lactis* subsp. *lactis* IL1403
405. *Laribacter hongkongensis*
406. *Lawsonia intracellularis*
407. *Legionella pneumophila* Corby
408. *Legionella pneumophila* Lens
409. *Legionella pneumophila* Paris
410. *Legionella pneumophila* Philadelphia 1
411. *Leifsonia xyli xyli* CTCB07
412. *Leptospira biflexa* serovar Patoc Patoc 1 (Ames)
413. *Leptospira biflexa* serovar Patoc Patoc 1 (Paris)
414. *Leptospira borgpetersenii* JB197
415. *Leptospira borgpetersenii* L550
416. *Leptospira interrogans* serovar Copenhageni
417. *Leptospira interrogans* serovar lai
418. *Leptothrix cholodnii*
419. *Leuconostoc citreum*
420. *Leuconostoc mesenteroides*
421. *Listeria innocua*
422. *Listeria monocytogenes* EGD-e
423. *Listeria monocytogenes* F2365
424. *Listeria monocytogenes* HCC23
425. *Listeria welshimeri* SLCC5334

426.	<i>Lysinibacillus sphaericus</i>	498.	<i>Mycoplasma hyopneumoniae</i> J
427.	<i>Macrococcus caseolyticus</i>	499.	<i>Mycoplasma mobile</i>
428.	<i>Magnetococcus</i> sp. MC-1	500.	<i>Mycoplasma mycoides</i>
429.	<i>Magnetospirillum magneticum</i>	501.	<i>Mycoplasma penetrans</i>
430.	<i>Mannheimia succiniciproducens</i>	502.	<i>Mycoplasma pneumoniae</i>
431.	<i>Maricaulis maris</i>	503.	<i>Mycoplasma pulmonis</i>
432.	<i>Marinobacter aquaeolei</i>	504.	<i>Mycoplasma synoviae</i>
433.	<i>Marinomonas</i> sp. MWYL1	505.	<i>Myxococcus xanthus</i>
434.	<i>Mesoplasma florum</i>	506.	<i>Nanoarchaeum equitans</i>
435.	<i>Mesorhizobium loti</i>	507.	<i>Natronaerobius thermophilus</i>
436.	<i>Mesorhizobium</i> sp. BNC1	508.	<i>Natronomonas pharaonis</i>
437.	<i>Metallosphaera sedula</i>	509.	<i>Nautilia profundicola</i>
438.	<i>Methanobacterium thermoautotrophicum</i>	510.	<i>Neisseria gonorrhoeae</i>
439.	<i>Methanobrevibacter smithii</i> ATCC 35061	511.	<i>Neisseria gonorrhoeae</i> NCCP11945
440.	<i>Methanococcoides burtonii</i>	512.	<i>Neisseria meningitidis</i> 053442 (serogroup C)
441.	<i>Methanococcus aeolicus</i>	513.	<i>Neisseria meningitidis</i> FAM18 (serogroup C)
442.	<i>Methanococcus jannaschii</i>	514.	<i>Neisseria meningitidis</i> MC58 (serogroup B)
443.	<i>Methanococcus maripaludis</i> C5	515.	<i>Neisseria meningitidis</i> Z2491 (serogroup A)
444.	<i>Methanococcus maripaludis</i> C6	516.	<i>Neorickettsia sennetsu</i>
445.	<i>Methanococcus maripaludis</i> C7	517.	<i>Nitratiruptor</i> sp. SB155-2
446.	<i>Methanococcus maripaludis</i> S2	518.	<i>Nitrobacter hamburgensis</i>
447.	<i>Methanococcus vannieli</i>	519.	<i>Nitrobacter winogradskyi</i>
448.	<i>Methanocorpusculum labreanum</i>	520.	<i>Nitrosococcus oceani</i>
449.	<i>Methanoculleus marisnigri</i>	521.	<i>Nitrosomonas europaea</i>
450.	<i>Methanopyrus kandleri</i>	522.	<i>Nitrosomonas eutropha</i>
451.	<i>Methanosaeta thermophila</i>	523.	<i>Nitrosopumilus maritimus</i>
452.	<i>Methanosarcina acetivorans</i>	524.	<i>Nitrospira multiformis</i>
453.	<i>Methanosarcina barkeri</i>	525.	<i>Nocardia farcinica</i>
454.	<i>Methanosarcina mazei</i>	526.	<i>Nocardioides</i> sp. JS614
455.	<i>Methanosphaera stadtmanae</i>	527.	<i>Nostoc punctiforme</i>
456.	<i>Methanospirillum hungatei</i>	528.	<i>Novosphingobium aromaticivorans</i>
457.	<i>Methylococcus thermophilus</i>	529.	<i>Oceanobacillus ihayensis</i>
458.	<i>Methylobium petroleiphilum</i>	530.	<i>Ochrobactrum anthropi</i>
459.	<i>Methylobacillus flagellatus</i>	531.	<i>Oenococcus oeni</i>
460.	<i>Methylobacterium chloromethanicum</i>	532.	<i>Oligotropha carboxidovorans</i>
461.	<i>Methylobacterium extorquens</i>	533.	<i>Opitutus terrae</i>
462.	<i>Methylobacterium nodulans</i>	534.	<i>Orientia tsutsugamushi</i> Boryong
463.	<i>Methylobacterium populi</i>	535.	<i>Orientia tsutsugamushi</i> Ikeda
464.	<i>Methylobacterium radiotolerans</i>	536.	<i>Parabacteroides distasonis</i>
465.	<i>Methylobacterium</i> sp. 4-46	537.	<i>Paracoccus denitrificans</i>
466.	<i>Methylocella silvestris</i>	538.	<i>Parvibaculum lavamentivorans</i>
467.	<i>Methylococcus capsulatus</i>	539.	<i>Pasteurella multocida</i>
468.	<i>Microcystis aeruginosa</i>	540.	<i>Pediococcus pentosaceus</i>
469.	<i>Minibacterium massiliensis</i>	541.	<i>Pelobacter carbinolicus</i>
470.	<i>Moorella thermoacetica</i>	542.	<i>Pelobacter propionicus</i>
471.	<i>Mycobacterium abscessus</i> ATCC 19977T	543.	<i>Pelodictyon luteolum</i>
472.	<i>Mycobacterium avium</i> 104	544.	<i>Pelodictyon phaeoclathratiforme</i>
473.	<i>Mycobacterium avium</i> paratuberculosis	545.	<i>Pelotomaculum thermopropionicum</i>
474.	<i>Mycobacterium bovis</i> AF2122/97	546.	<i>Persephonella marina</i>
475.	<i>Mycobacterium bovis</i> BCG Pasteur 1173P2	547.	<i>Petrogoba mobilis</i>
476.	<i>Mycobacterium bovis</i> BCG Tokyo 172	548.	<i>Phenylobacterium zucineum</i>
477.	<i>Mycobacterium gilvum</i>	549.	<i>Photobacterium profundum</i>
478.	<i>Mycobacterium leprae</i> Br4923	550.	<i>Photorhabdus luminescens</i>
479.	<i>Mycobacterium leprae</i> TN	551.	<i>Phytoplasma AYWB</i>
480.	<i>Mycobacterium marinum</i> M	552.	<i>Phytoplasma OY</i>
481.	<i>Mycobacterium smegmatis</i>	553.	<i>Picrophilus torridus</i>
482.	<i>Mycobacterium</i> sp. JLS	554.	<i>Polaromonas naphthalenivorans</i>
483.	<i>Mycobacterium</i> sp. KMS	555.	<i>Polaromonas</i> sp. JS666
484.	<i>Mycobacterium</i> sp. MCS	556.	<i>Polynucleobacter necessarius</i>
485.	<i>Mycobacterium tuberculosis</i> CDC1551	557.	<i>Polynucleobacter</i> sp. QLV-P1DMWA-1
486.	<i>Mycobacterium tuberculosis</i> F11	558.	<i>Porphyromonas gingivalis</i> ATCC 33277
487.	<i>Mycobacterium tuberculosis</i> H37Ra	559.	<i>Porphyromonas gingivalis</i> W83
488.	<i>Mycobacterium tuberculosis</i> H37Rv	560.	<i>Prochlorococcus marinus</i> AS9601
489.	<i>Mycobacterium ulcerans</i>	561.	<i>Prochlorococcus marinus</i> MED4
490.	<i>Mycobacterium vanbaalenii</i>	562.	<i>Prochlorococcus marinus</i> MIT 9211
491.	<i>Mycoplasma agalactiae</i>	563.	<i>Prochlorococcus marinus</i> MIT 9215
492.	<i>Mycoplasma arthritidis</i>	564.	<i>Prochlorococcus marinus</i> MIT 9301
493.	<i>Mycoplasma capricolum</i>	565.	<i>Prochlorococcus marinus</i> MIT 9303
494.	<i>Mycoplasma gallisepticum</i>	566.	<i>Prochlorococcus marinus</i> MIT 9515
495.	<i>Mycoplasma genitalium</i>	567.	<i>Prochlorococcus marinus</i> MIT9312
496.	<i>Mycoplasma hyopneumoniae</i> 232	568.	<i>Prochlorococcus marinus</i> MIT9313
497.	<i>Mycoplasma hyopneumoniae</i> 7448	569.	<i>Prochlorococcus marinus</i> NATL1A

570. *Prochlorococcus marinus* NATL2A
571. *Prochlorococcus marinus* SS120
572. *Propionibacterium acnes*
573. *Prosthecochloris aestuarii*
574. *Prosthecochloris vibrioformis*
575. *Proteus mirabilis*
576. *Pseudoalteromonas atlantica*
577. *Pseudoalteromonas haloplanktis*
578. *Pseudomonas aeruginosa* LESB58
579. *Pseudomonas aeruginosa* PA7
580. *Pseudomonas aeruginosa* PAO1
581. *Pseudomonas aeruginosa* UCBPP-PA14
582. *Pseudomonas entomophila*
583. *Pseudomonas fluorescens* Pf-5
584. *Pseudomonas fluorescens* Pf0-1
585. *Pseudomonas mendocina*
586. *Pseudomonas putida* F1
587. *Pseudomonas putida* GB-1
588. *Pseudomonas putida* KT2440
589. *Pseudomonas putida* W619
590. *Pseudomonas stutzeri*
591. *Pseudomonas syringae* pv. *phaseolicola* 1448A
592. *Pseudomonas syringae* pv. *syringae* B728a
593. *Pseudomonas syringae* pv. *tomato* DC3000
594. *Psychrobacter arcticum*
595. *Psychrobacter cryohalolentis*
596. *Psychrobacter* sp. PRwf-1
597. *Psychromonas ingrahamii*
598. *Pyrobaculum aerophilum*
599. *Pyrobaculum arsenaticum*
600. *Pyrobaculum caldifontis*
601. *Pyrobaculum islandicum*
602. *Pyrococcus abyssi*
603. *Pyrococcus furiosus*
604. *Pyrococcus horikoshii*
605. *Ralstonia eutropha* H16
606. *Ralstonia eutropha* JMP134
607. *Ralstonia metallidurans*
608. *Ralstonia pickettii* 12J
609. *Ralstonia solanacearum*
610. *Renibacterium salmoninarum*
611. *Rhizobium etli* CFN 42
612. *Rhizobium etli* CIAT 652
613. *Rhizobium leguminosarum*
614. *Rhizobium leguminosarum* bv. *trifolii* WSM2304
615. *Rhodobacter sphaeroides* 2.4.1
616. *Rhodobacter sphaeroides* ATCC 17025
617. *Rhodobacter sphaeroides* ATCC 17029
618. *Rhodobacter sphaeroides* KD131
619. *Rhodococcus erythropolis*
620. *Rhodococcus opacus*
621. *Rhodococcus* sp. RHA1
622. *Rhodoferrax ferrireducens*
623. *Rhodopirellula baltica*
624. *Rhodopseudomonas palustris* BisA53
625. *Rhodopseudomonas palustris* BisB18
626. *Rhodopseudomonas palustris* BisB5
627. *Rhodopseudomonas palustris* CGA009
628. *Rhodopseudomonas palustris* HaA2
629. *Rhodopseudomonas palustris* TIE-1
630. *Rhodospirillum centenum*
631. *Rhodospirillum rubrum*
632. *Rickettsia akari*
633. *Rickettsia bellii* OSU 85-389
634. *Rickettsia bellii* RML369-C
635. *Rickettsia canadensis*
636. *Rickettsia conorii*
637. *Rickettsia felis*
638. *Rickettsia massiliae*
639. *Rickettsia prowazekii*
640. *Rickettsia rickettsii* Iowa
641. *Rickettsia rickettsii* Sheila Smith
642. *Rickettsia typhi*
643. *Roseiflexus castenholzii* DSM13941
644. *Roseiflexus* sp. RS-1
645. *Roseobacter denitrificans*
646. *Rubrobacter xylanophilus*
647. *Ruegeria pomeroyi*
648. *Saccharophagus degradans*
649. *Saccharopolyspora erythraea*
650. *Salinibacter ruber*
651. *Salinispora arenicola*
652. *Salinispora tropica*
653. *Salmonella enterica* subsp. *arizonae*
654. *Salmonella enterica* subsp. *enterica* serovar *Agona*
655. *Salmonella enterica* subsp. *enterica* serovar *Choleraesuis*
656. *Salmonella enterica* subsp. *enterica* serovar *Dublin*
657. *Salmonella enterica* subsp. *enterica* serovar *Enteritidis*
658. *Salmonella enterica* subsp. *enterica* serovar *Gallinarum*
659. *Salmonella enterica* subsp. *enterica* serovar *Heidelberg*
660. *Salmonella enterica* subsp. *enterica* serovar *Newport*
661. *Salmonella enterica* subsp. *enterica* serovar *Paratyphi A*
662. *Salmonella enterica* subsp. *enterica* serovar *Paratyphi A* AKU12601
663. *Salmonella enterica* subsp. *enterica* serovar *Paratyphi B*
664. *Salmonella enterica* subsp. *enterica* serovar *Paratyphi C*
665. *Salmonella enterica* subsp. *enterica* serovar *Schwarzengrund*
666. *Salmonella enterica* subsp. *enterica* serovar *Typhi* CT18
667. *Salmonella enterica* subsp. *enterica* serovar *Typhi* Ty2
668. *Salmonella typhimurium* LT2
669. *Serratia proteamaculans*
670. *Shewanella amazonensis*
671. *Shewanella baltica* OS155
672. *Shewanella baltica* OS185
673. *Shewanella baltica* OS195
674. *Shewanella baltica* OS223
675. *Shewanella denitrificans*
676. *Shewanella frigidimarina*
677. *Shewanella halifaxensis*
678. *Shewanella loihica*
679. *Shewanella oneidensis*
680. *Shewanella pealeana*
681. *Shewanella piezotolerans* WP3
682. *Shewanella putrefaciens*
683. *Shewanella sediminis*
684. *Shewanella* sp. ANA-3
685. *Shewanella* sp. MR-4
686. *Shewanella* sp. MR-7
687. *Shewanella* sp. W3-18-1
688. *Shewanella woodyi* ATCC 51908
689. *Shigella boydii* CDC 3083-94
690. *Shigella boydii* Sb227
691. *Shigella dysenteriae*
692. *Shigella flexneri* 2457T (serotype 2a)
693. *Shigella flexneri* 301 (serotype 2a)
694. *Shigella flexneri* 8401 (serotype 5b)
695. *Shigella sonnei*
696. *Silicibacter* sp. TM1040
697. *Sinorhizobium medicae*
698. *Sinorhizobium meliloti*
699. *Sodalis glossinidius*
700. *Solibacter usitatus*
701. *Sorangium cellulosum*
702. *Sphingomonas wittichii*
703. *Sphingopyxis alaskensis*
704. *Staphylococcus aureus* COL
705. *Staphylococcus aureus* JH1
706. *Staphylococcus aureus* JH9
707. *Staphylococcus aureus* MRSA252
708. *Staphylococcus aureus* MSSA476
709. *Staphylococcus aureus* MW2
710. *Staphylococcus aureus* Mu3
711. *Staphylococcus aureus* Mu50

712. *Staphylococcus aureus* N315
713. *Staphylococcus aureus* NCTC8325
714. *Staphylococcus aureus* Newman
715. *Staphylococcus aureus* RF122
716. *Staphylococcus aureus* USA300
717. *Staphylococcus aureus* USA300 TCH1516
718. *Staphylococcus carnosus*
719. *Staphylococcus epidermidis* ATCC 12228
720. *Staphylococcus epidermidis* RP62A
721. *Staphylococcus haemolyticus*
722. *Staphylococcus saprophyticus*
723. *Staphylothermus marinus*
724. *Stenotrophomonas maltophilia*
725. *Stenotrophomonas maltophilia* R551-3
726. *Streptococcus agalactiae* 2603 (serotype V)
727. *Streptococcus agalactiae* A909 (serotype Ia)
728. *Streptococcus agalactiae* NEM316 (serotype III)
729. *Streptococcus equi* subsp. *equi* 4047
730. *Streptococcus equi* subsp. *zooepidemicus*
731. *Streptococcus equi* subsp. *zooepidemicus* MGCS10565
732. *Streptococcus gordonii*
733. *Streptococcus mutans*
734. *Streptococcus pneumoniae* 70585
735. *Streptococcus pneumoniae* ATCC 700669
736. *Streptococcus pneumoniae* CGSP14
737. *Streptococcus pneumoniae* D39
738. *Streptococcus pneumoniae* G54
739. *Streptococcus pneumoniae* Hungary19A 6
740. *Streptococcus pneumoniae* JJA
741. *Streptococcus pneumoniae* P1031
742. *Streptococcus pneumoniae* R6
743. *Streptococcus pneumoniae* TIGR4
744. *Streptococcus pneumoniae* Taiwan19F-14
745. *Streptococcus pyogenes* MGAS10270 (serotype M2)
746. *Streptococcus pyogenes* MGAS10394 (serotype M6)
747. *Streptococcus pyogenes* MGAS10750 (serotype M4)
748. *Streptococcus pyogenes* MGAS2096 (serotype M12)
749. *Streptococcus pyogenes* MGAS315 (serotype M3)
750. *Streptococcus pyogenes* MGAS5005 (serotype M1)
751. *Streptococcus pyogenes* MGAS6180 (serotype M28)
752. *Streptococcus pyogenes* MGAS8232 (serotype M18)
753. *Streptococcus pyogenes* MGAS9429 (serotype M12)
754. *Streptococcus pyogenes* Manfredo (serotype M5)
755. *Streptococcus pyogenes* NZ131 (serotype M49)
756. *Streptococcus pyogenes* SF370 (serotype M1)
757. *Streptococcus pyogenes* SSI-1 (serotype M3)
758. *Streptococcus sanguinis*
759. *Streptococcus suis* 05ZYH33
760. *Streptococcus suis* 98HAH33
761. *Streptococcus thermophilus* CNRZ1066
762. *Streptococcus thermophilus* LMD-9
763. *Streptococcus thermophilus* LMG18311
764. *Streptococcus uberis*
765. *Streptomyces avermitilis*
766. *Streptomyces coelicolor*
767. *Streptomyces griseus*
768. *Sulfolobus acidocaldarius*
769. *Sulfolobus solfataricus*
770. *Sulfolobus tokodaii*
771. *Sulfurihydrogenibium azorense*
772. *Sulfurihydrogenibium* sp. YO3AOP1
773. *Sulfurovum* sp. NBC37-1
774. *Symbiobacterium thermophilum*
775. *Synechococcus elongatus* PCC6301
776. *Synechococcus elongatus* PCC7942
777. *Synechococcus* sp. CC9311
778. *Synechococcus* sp. CC9605
779. *Synechococcus* sp. CC9902
780. *Synechococcus* sp. PCC7002
781. *Synechococcus* sp. RCC307
782. *Synechococcus* sp. WH7803
783. *Synechococcus* sp. WH8102
784. *Synechocystis* sp. PCC6803
785. *Syntrophobacter fumaroxidans*
786. *Syntrophomonas wolfei*
787. *Syntrophus aciditrophicus*
788. *Thauera* sp. MZ1T
789. *Thermoanaerobacter pseudethanolicus*
790. *Thermoanaerobacter* sp. X514
791. *Thermoanaerobacter tengcongensis*
792. *Thermobifida fusca*
793. *Thermococcus kodakaraensis*
794. *Thermococcus onnurineus*
795. *Thermodesulfovibrio yellowstonii*
796. *Thermofilum pendens*
797. *Thermomicrobium roseum*
798. *Thermoplasma acidophilum*
799. *Thermoplasma volcanium*
800. *Thermoproteus neutrophilus*
801. *Thermosipho africanus*
802. *Thermosipho melanesiensis*
803. *Thermosynechococcus elongatus*
804. *Thermotoga lettingae*
805. *Thermotoga maritima*
806. *Thermotoga neapolitana*
807. *Thermotoga petrophila*
808. *Thermotoga* sp. RQ2
809. *Thermus thermophilus* HB27
810. *Thermus thermophilus* HB8
811. *Thioalkalivibrio* sp. HL-EbGR7
812. *Thiobacillus denitrificans*
813. *Thiomicrospira crunigena*
814. *Thiomicrospira denitrificans*
815. *Treponema denticola*
816. *Treponema pallidum* subsp. *pallidum* Nichols
817. *Treponema pallidum* subsp. *pallidum* SS14
818. *Trichodesmium erythraeum*
819. *Tropheryma whippelii* TW08/27
820. *Tropheryma whippelii* Twist
821. Uncultured Termite group 1 bacterium phylotype Rs-D17
822. Uncultured methanogenic archaeon RC-I
823. *Ureaplasma parvum* serovar 3 ATCC 27815
824. *Ureaplasma parvum* serovar 3 ATCC 700970
825. *Ureaplasma urealyticum* serovar 10 ATCC 33699
826. *Verminephrobacter eiseniae*
827. *Vibrio cholerae* O1
828. *Vibrio cholerae* O395
829. *Vibrio fischeri*
830. *Vibrio fischeri* MJ11
831. *Vibrio harveyi*
832. *Vibrio parahaemolyticus*
833. *Vibrio splendidus*
834. *Vibrio vulnificus* CMCP6
835. *Vibrio vulnificus* YJ016
836. *Wigglesworthia glossinidia*
837. *Wolbachia pipientis*
838. *Wolbachia* sp. wRi
839. *Wolbachia* wBm
840. *Wolbachia* wMel
841. *Wolinella succinogenes*
842. *Xanthobacter autotrophicus*
843. *Xanthomonas axonopodis*
844. *Xanthomonas campestris* pv. *campestris* 8004
845. *Xanthomonas campestris* pv. *campestris* ATCC 33913
846. *Xanthomonas campestris* pv. *campestris* B100
847. *Xanthomonas campestris* pv. *vesicatoria*
848. *Xanthomonas oryzae* KACC10331
849. *Xanthomonas oryzae* MAFF311018
850. *Xanthomonas oryzae* PXO99A
851. *Xylella fastidiosa* 9a5c
852. *Xylella fastidiosa* M12
853. *Xylella fastidiosa* M23
854. *Xylella fastidiosa* Temecula1

- 855. *Yersinia enterocolitica*
- 856. *Yersinia pestis* Angola
- 857. *Yersinia pestis* Antiqua
- 858. *Yersinia pestis* CO92
- 859. *Yersinia pestis* KIM
- 860. *Yersinia pestis* Mediaevails
- 861. *Yersinia pestis* Nepal516
- 862. *Yersinia pestis* Pestoides
- 863. *Yersinia pseudotuberculosis* IP31758
- 864. *Yersinia pseudotuberculosis* IP32953
- 865. *Yersinia pseudotuberculosis* PB1/+
- 866. *Yersinia pseudotuberculosis* YPIII
- 867. *Zymomonas mobilis*

Supplementary Material 2: PFAMs used in the study

PFAMs used to identify TFs

PFAM name	PFAM ID	Description
Arc	PF03869	Arc-like DNA binding domain
Arg_repressor	PF01316	Arginine repressor, DNA binding domain
AsnC_trans_reg	PF01037	AsnC family
Bac_DnaA_C	PF08299	Bacterial dnaA protein helix-turn-helix domain
Bac_DNA_binding	PF00216	Bacterial DNA-binding protein
CoiA	PF06054	Competence protein CoiA-like family
Crl	PF07417	Transcriptional regulator Crl
Crp	PF00325	Bacterial regulatory proteins, crp family
CtsR	PF05848	Firmicute transcriptional repressor of class III stress genes (CtsR)
FeoC	PF09012	FeoC like transcriptional regulator
Fe_dep_repress	PF01325	Iron dependent repressor, N-terminal DNA binding domain
FlhC	PF05280	Flagellar transcriptional activator (FlhC)
FlhD	PF05247	Flagellar transcriptional activator (FlhD)
FUR	PF01475	Ferric uptake regulator family
GcrA	PF07750	GcrA cell cycle regulator
GerE	PF00196	Bacterial regulatory proteins, luxR family
GntR	PF00392	Bacterial regulatory proteins, gntR family
GutM	PF06923	Glucitol operon activator protein (GutM)
HHA	PF05321	Haemolysin expression modulating protein
Histone_HNS	PF00816	H-NS histone family
HTH_1	PF00126	Bacterial regulatory helix-turn-helix protein, lysR family
HTH_10	PF04967	HTH DNA binding domain
HTH_11	PF08279	HTH domain
HTH_3	PF01381	Helix-turn-helix
HTH_5	PF01022	Bacterial regulatory protein, arsR family
HTH_6	PF01418	Helix-turn-helix domain, rpiR family
HTH_8	PF02954	Bacterial regulatory protein, Fis family
HTH_AraC	PF00165	Bacterial regulatory helix-turn-helix proteins, AraC family
HTH_CodY	PF08222	CodY helix-turn-helix domain
HTH_DeoR	PF08220	DeoR-like helix-turn-helix domain
HTH_lclR	PF09339	lclR helix-turn-helix domain
HTH_Mga	PF08280	M protein trans-acting positive regulator (MGA) HTH domain
HTH_psq	PF05225	helix-turn-helix, Psq domain
HTH_WhiA	PF02650	Sporulation Regulator WhiA C terminal domain
KilA-N	PF04383	KilA-N domain
KorB	PF08535	KorB domain
LacI	PF00356	Bacterial regulatory proteins, lacI family
LexA_DNA_bind	PF01726	LexA DNA binding domain
LytTR	PF04397	LytTr DNA-binding domain
MarR	PF01047	MarR family
MerR	PF00376	MerR family regulatory protein
MetJ	PF01340	Met Apo-repressor, MetJ
MtlR	PF05068	Mannitol repressor
Ogr_Delta	PF04606	Ogr/Delta-like zinc finger
PaaX	PF07848	PaaX-like protein
PadR	PF03551	Transcriptional regulator PadR-like family
PerC	PF06069	PerC transcriptional activator
Phage_Alpa	PF05930	Prophage CP4-57 regulatory protein (AlpA)
Phage_Treg	PF04761	Lactococcus bacteriophage putative transcription regulator
PHB_acc_N	PF07879	PHB/PHA accumulation regulator DNA-binding domain
PSK_trans_fac	PF07704	Rv0623-like transcription factor
PspC	PF04024	PspC domain
PuR_N	PF09182	Bacterial purine repressor, N-terminal
RinB	PF06116	Transcriptional activator RinB
ROS_MUCR	PF05443	ROS/MUCR transcriptional regulator protein
SfsA	PF03749	Sugar fermentation stimulation protein
SirB	PF04247	Invasion gene expression up-regulator, SirB
SpoVT_AbrB	PF04014	SpoVT / AbrB like domain
TetR_N	PF00440	Bacterial regulatory proteins, tetR family
Trans_reg_C	PF00486	Transcriptional regulatory protein, C terminal
TrmB	PF01978	Sugar-specific transcriptional regulator TrmB
Trp_repressor	PF01371	Trp repressor protein
UPF0122	PF04297	Putative helix-turn-helix protein, YlxM / p13 like

PFAMs used to identify HisKs and RECs

PFAM name	PFAM ID	Description
H-kinase_dim	PF02895	Signal transducing histidine kinase, homodimeric domain
HisKA	PF00512	His Kinase A (phosphoacceptor) domain
HisKA_2	PF07568	Histidine kinase
HisKA_3	PF07730	Histidine kinase
His_kinase	PF06580	Histidine kinase
Hpt	PF01627	Hpt domain
HWE_HK	PF07536	HWE histidine kinase

The PFAM Response_reg (PF00072; Response regulator receiver domain) was used to identify RECs.

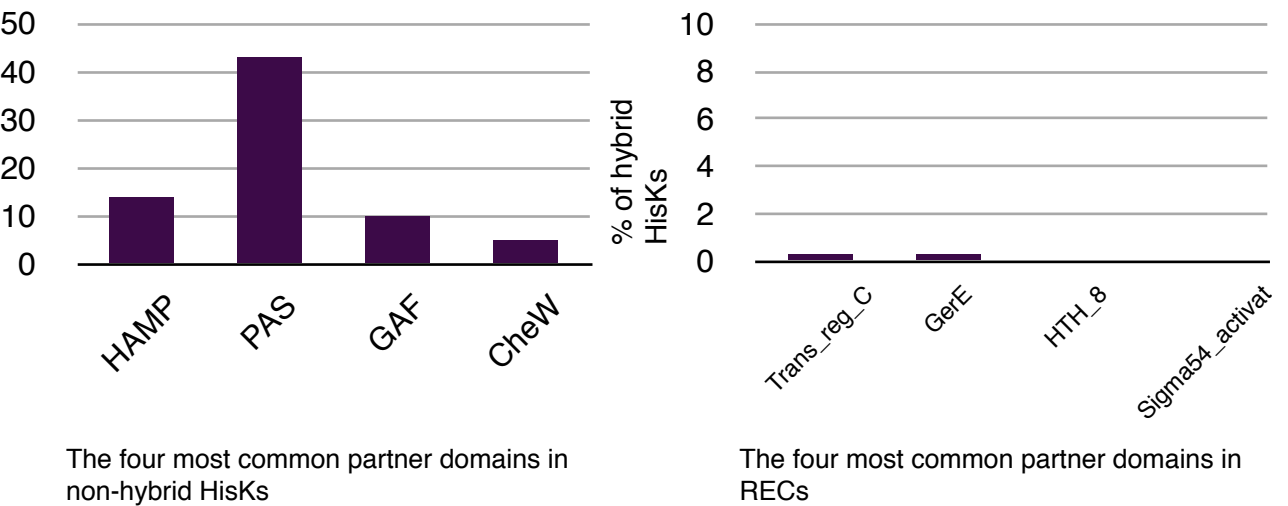
PFAMs used to identify Small-molecule-binding domains

PFAM name	PFAM ID	Description
ACBP	PF00887	Acyl CoA binding protein
ACT	PF01842	ACT domain
Ada_Zn_binding	PF02805	Metal binding domain of Ada
ANF_receptor	PF01094	Receptor family ligand binding region
AraC_binding	PF02311	AraC-like ligand binding domain
AraC_E_bind	PF06445	Bacterial transcription activator, effector binding domain
Autoind_bind	PF03472	Autoinducer binding domain
Cache_1	PF02743	Cache domain
Cache_2	PF08269	Cache domain
CHASE	PF03924	CHASE domain
cNMP_binding	PF00027	Cyclic nucleotide-binding domain
CodY	PF06018	CodY GAF-like domain
DeoR	PF00455	Bacterial regulatory proteins, deoR family
Diacid_rec	PF05651	Putative sugar diacid recognition
FCD	PF07729	FCD domain
FeoA	PF04023	FeoA domain
Fe_dep_repr_C	PF02742	Iron dependent repressor, metal binding and dimerisation domain
FHA	PF00498	FHA domain
FIST	PF08495	FIST N domain
GAF	PF01590	GAF domain
HMA	PF00403	Heavy-metal-associated domain
Hormone_recep	PF00104	Ligand-binding domain of nuclear hormone receptor
IclR	PF01614	Bacterial transcriptional regulator
LysR_substrate	PF03466	LysR substrate binding domain
MCPsignal	PF00015	Methyl-accepting chemotaxis protein (MCP) signaling domain
NIT	PF08376	Nitrate and nitrite sensing
PAS	PF00989	PAS fold
PAS_2	PF08446	PAS fold
PAS_3	PF08447	PAS fold
PAS_4	PF08448	PAS fold
Peripla_BP_1	PF00532	Periplasmic binding proteins and sugar binding domain of the LacI family
Peripla_BP_2	PF01497	Periplasmic binding protein
PilZ	PF07238	PilZ domain
SBP_bac_3	PF00497	Bacterial extracellular solute-binding proteins, family 3
SIS	PF01380	SIS domain
STAS	PF01740	STAS domain
Sugar-bind	PF04198	Putative sugar-binding domain
TetR_C	PF02909	Tetracyclin repressor, C-terminal all-alpha domain
TOBE	PF03459	TOBE domain
UTRA	PF07702	UTRA domain
V4R	PF02830	V4R domain

Note: These are not described in the main text, but used for the analysis presented in Supplementary Material 14. This list is based on data available in the Microbial Signal Transduction (MiST) database (<http://mistdb.com/>, version 1).

Supplementary Material 3: Comparison of predominant partner domains in hybrid HisKs with those of non-hybrid HisKs and RECs

The four most prevalent partner domains of non-hybrid HisKs are the most commonly seen partner domains in hybrid HisKs. On the other hand, the most common partner domains of RECs are hardly observed in hybrid HisKs.



Supplementary Material 4: Benchmarks

Comparison with manually-curated datasets

In order to test the coverage of our dataset we benchmarked our lists of regulatory proteins against databases covering the best characterised bacterium *Escherichia coli* K12 MG1655 (henceforth referred to as *E. coli*). First we used RegulonDB, which contains a list of 168 ORF identifiers corresponding to genes whose protein products have experimentally-verified roles in transcriptional regulation. Our list of 249 *E. coli* TFs includes 143 of these, thus giving a coverage of 85%. Several TFs not included in our dataset, including multifunctional enzymes, do not contain any PFAM domain that can be associated with DNA binding. Some other known TFs contain domains such as CSD (Cold Shock Domain) and RHH (Ribbon-Helix-Helix), whose inclusion might lead to higher false-positive rates as these cannot be unambiguously assigned to TFs.

For HisKs, we compared our data for *E. coli* to the semi-manually-curated list of signalling proteins assembled by Galperin. We identify 28 of the 30 HisKs in the above list. In the two HisKs we miss (DcuS and CitA) we can detect only the HATPase_c domain, which is a promiscuous domain present in a range of ATP-binding proteins including DNA topoisomerases. In general, our decision not to use this domain in identifying HisKs results in a certain loss of sensitivity. However this does not significantly affect our conclusions (see below).

Finally, we identified all 29 experimentally-characterised GGDEF/EAL proteins in *E. coli*.

Comparison with RPS-BLAST results

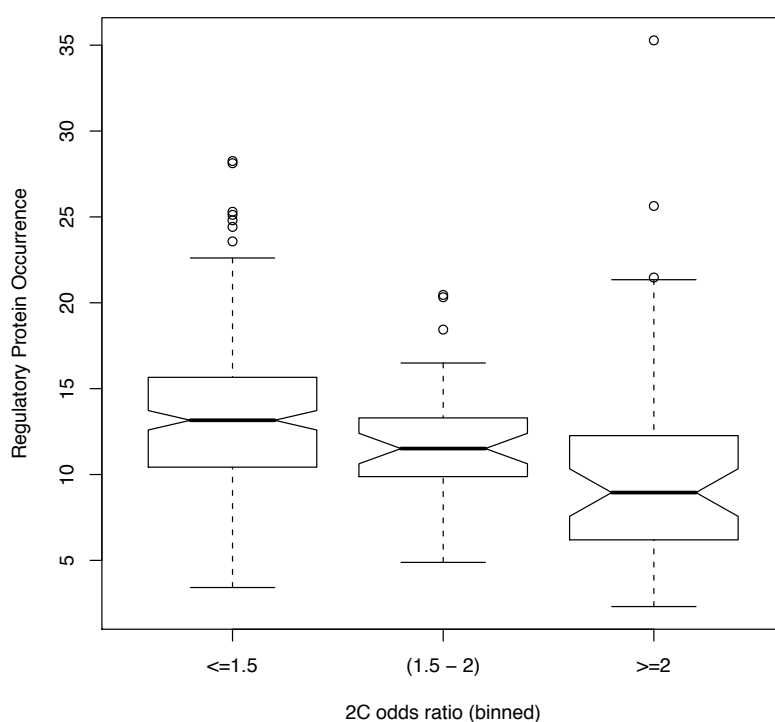
To perform a more systematic comparison with alternative search procedures, we scanned all 867 prokaryotic genomes for TFs and HisKs using reverse psi-blast, which is a significantly different alternative to HMMER for protein family identification. At an RPS-BLAST E-value cutoff of 1×10^{-6} , we obtain the following results, which indicate that our results are not substantially affected by the choice of search procedure.

Type	#HMMER and RPS-BLAST	#HMMER- only	#RPS-BLAST only	Sensitivity (%) ^a	Specificity (%) ^a
TF	114,244	11,875	2,151	98.2	90.6
HisK	20,742	830	325	98.5	96.2

^aSensitivity and specificity are calculated for the HMMER-based results, using RPS-BLAST results at $E < 1 \times 10^{-6}$ as benchmark.

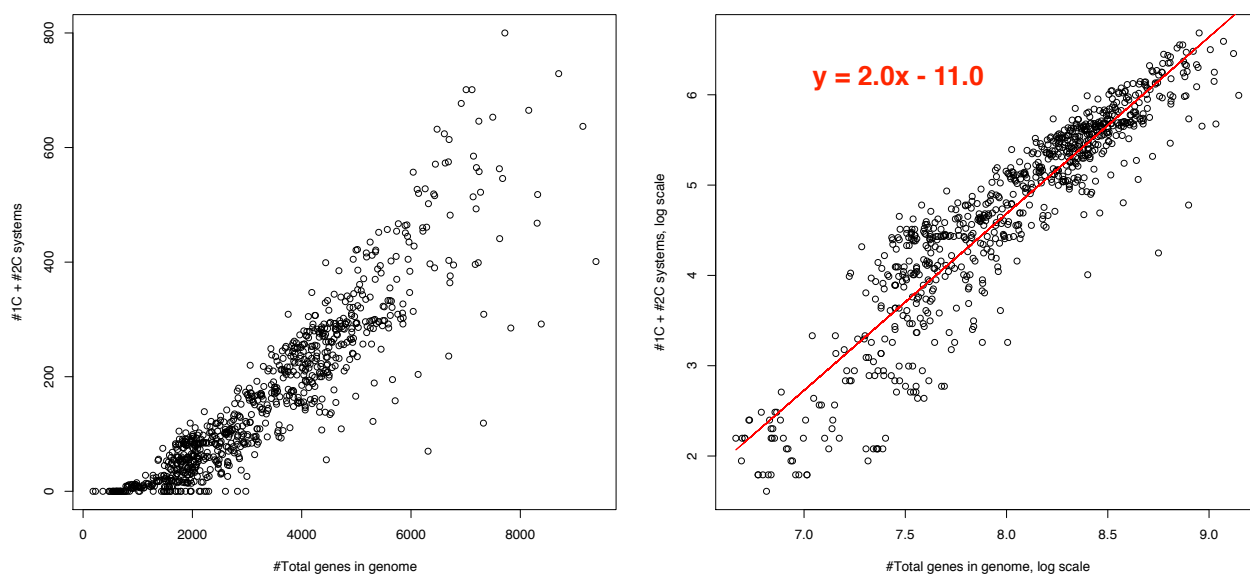
Effect of HATPase_c on our results

To assess whether the inclusion of the HATPase_c domain to our dataset would affect our results, we identified proteins with the HATPase_c domain. To minimise false positives, we selected only HATPase_c proteins which had a second domain typically associated with HisKs (PAS, GAF, GGDEF, EAL, CheW, CHASE, Cache, HAMP, Response_reg for hybrid HisKs). This list, when added to the dataset of HisKs used in our work contains 23,030 HisKs (7% increase). This increase in the number of HisKs, due to possible biases in the phylogenetic distribution of these newly included HisKs, might particularly affect the differences between different groups of genomes in their regulatory protein occurrence. However, comparison of trends observed in our original dataset with the new data shows that the RPO trend reported in the main text is not affected by the new data. Therefore, our results are not affected by our decision to neglect possible HisKs which are predicted to contain the HATPase_c domain, but not other domains for HisK activity.



Distribution of regulatory protein occurrence scores for genomes with 2C odds ratios ≤ 1.5 or $1.5 - 2$ or ≥ 2 . Here the RPO calculation *includes* predicted HATPase_c-containing HisKs.

Supplementary Material 5: Relationship between number of regulatory proteins and the total number of genes in a genome

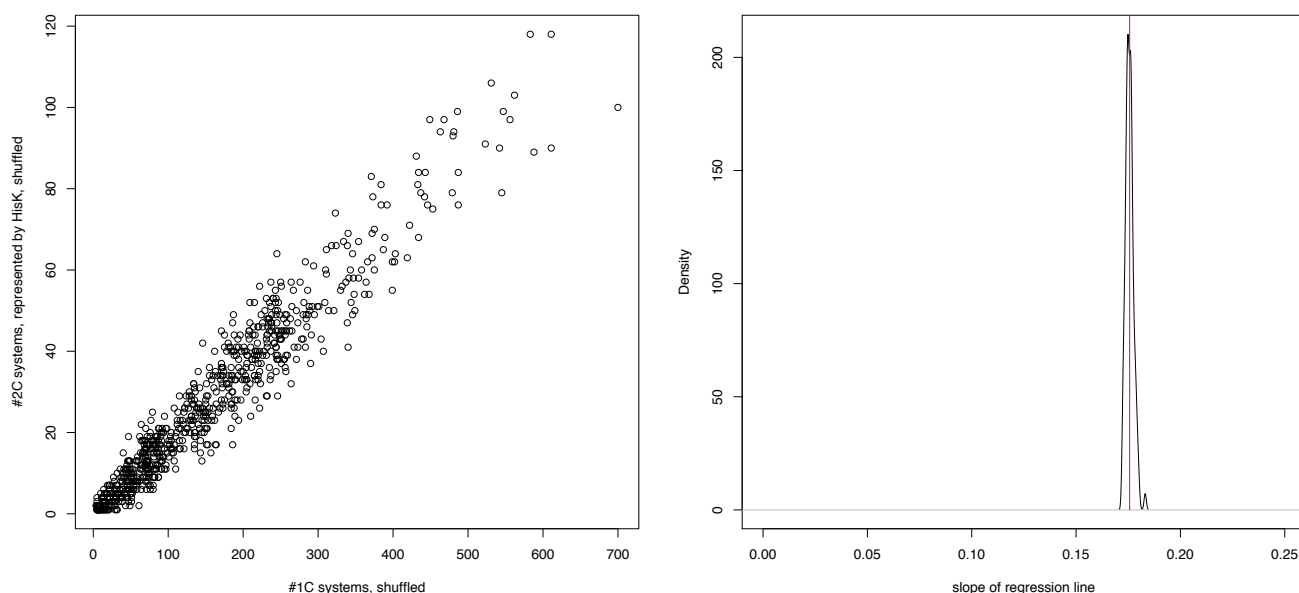


The left panel shows a scatter plot of the sum of the number of OC (represented by the sum of non-2C TFs and GGDEF and EAL domain proteins) and 2C (represented by HKs) systems against the total number of genes encoded in the genome. The right panel shows the same in the log-log scale; the red line is the regression line.

Supplementary Material 6: Calculation and properties of 2C odds ratio

A scatter plot of the numbers of 2C and 1C systems shows a significant correlation, but a high dispersion (see figure A on the following page). Therefore, we calculated a 2C odds ratio to identify genomes with a preference for encoding 2C systems.

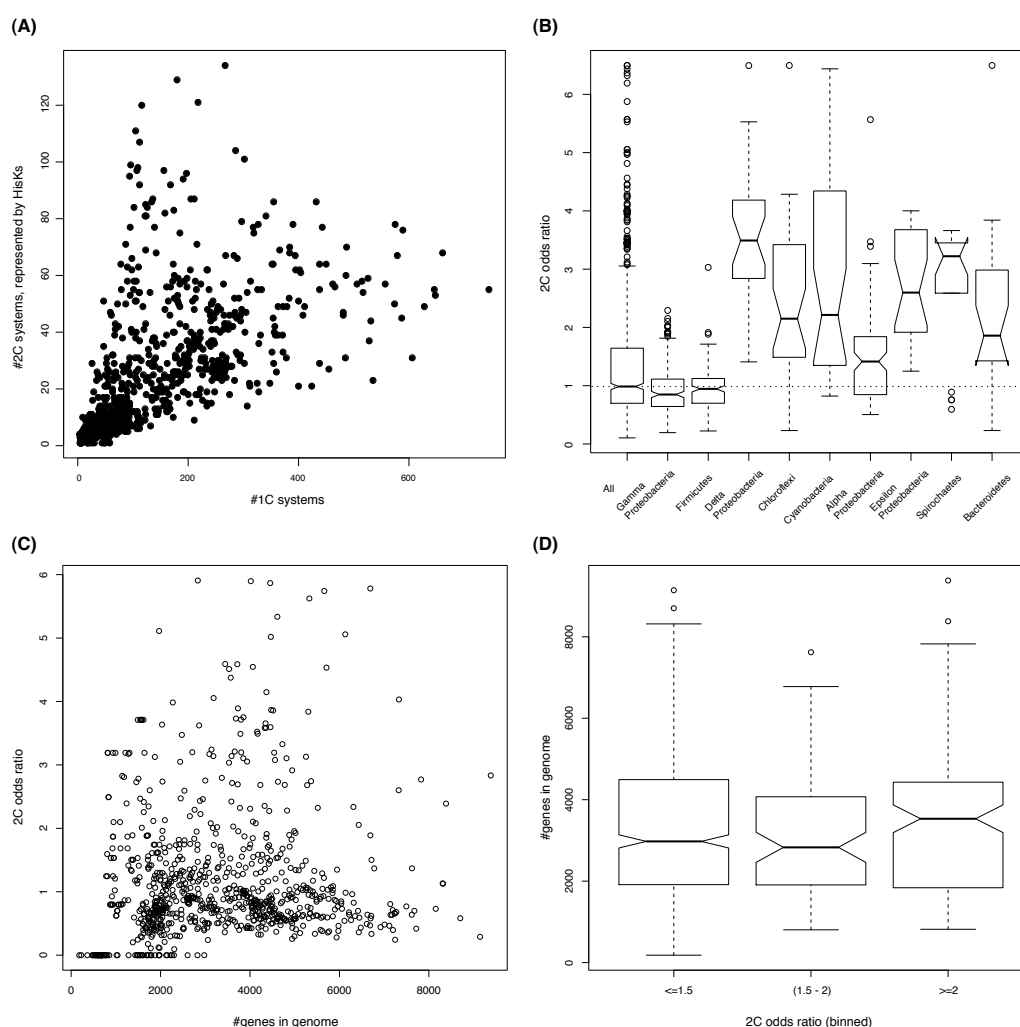
2C odds ratio was calculated given the null hypothesis that there is a near-perfect association between the numbers of 1C (defined by TFs and GGDEF and EAL proteins without the REC domain) and 2C (defined by HisKs) systems in prokaryotic genomes. The number of 2C systems expected, given a certain number of 1C systems, was calculated based on the following simulations. For each genome, the sum of the numbers of 1C and 2C systems was maintained; however, the number of each of 1C and 2C systems in each genome was randomly varied, such that the total number of 2C and 1C systems across all genomes was maintained. As a result of this, a near-perfect linear relationship between the numbers of 2C and 1C systems was obtained. This computation was repeated 1,000 times, and the average slope (m) and intercept (c) for the relationship across these simulations were calculated. Then, for each genome, the expected number of 2C systems ($\#HisK_{exp}$) was set to $mx+c$ where x is the number of 1C systems encoded by the genome. 2C odds ratio was equal to $\#HisK_{obs}/\#HisK_{exp}$.



The **left panel** shows the near-perfect correlation between #2C and #1C systems for one of the 1,000 random simulations. The **right panel** shows the distribution of slopes obtained across the 1,000 simulations. The average value is marked by the red vertical line.

Contrary to popular expectation that larger genomes should show a preference for 2C systems, we find little association between 2C odds ratio and genome size (see C, D below). For example, the number of HisKs encoded in the large genome of *Pseudomonas aeruginosa* (n=62) is approximately equal to the number expected (n0=69) for a genome encoding 395 1C systems. On the other hand, the 2C odds ratio of *Myxococcus xanthus* is over four.

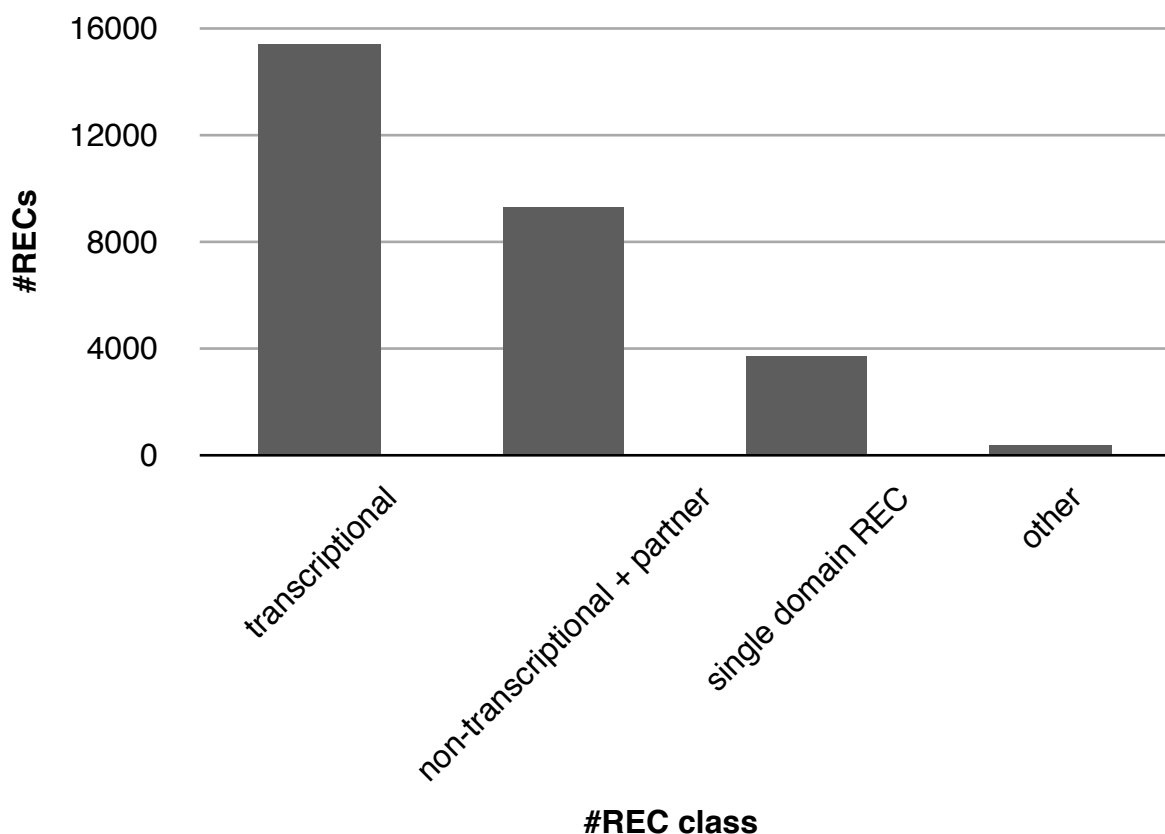
Certain phylogenetic families of bacteria display higher 2C odds ratios than others (see B below). We can classify most of these families into two major groups: (a) δ -proteobacteria, Cyanobacteria and Chloroflexi, and (b) α -proteobacteria, ϵ -proteobacteria and Spirochaetes. The former group, which has been well-studied in the present context using comparative genomics approaches, comprises large genomes from organisms with complex lifestyles; examples include *M. xanthus*, a model organism for starvation-dependent fruiting body formation, and *Anabaena variabilis*, a photosynthetic cyanobacterium that forms developmental structures called heterocysts. Among organisms belonging to the latter set of phylogenetic families, those with high 2C odds ratios tend to be pathogens (22 of 26 against 69 of 124). A final family of bacteria showing an elevated 2C odds ratio is Bacteroidetes which form one of the most prominent component of the human gut microbiota.



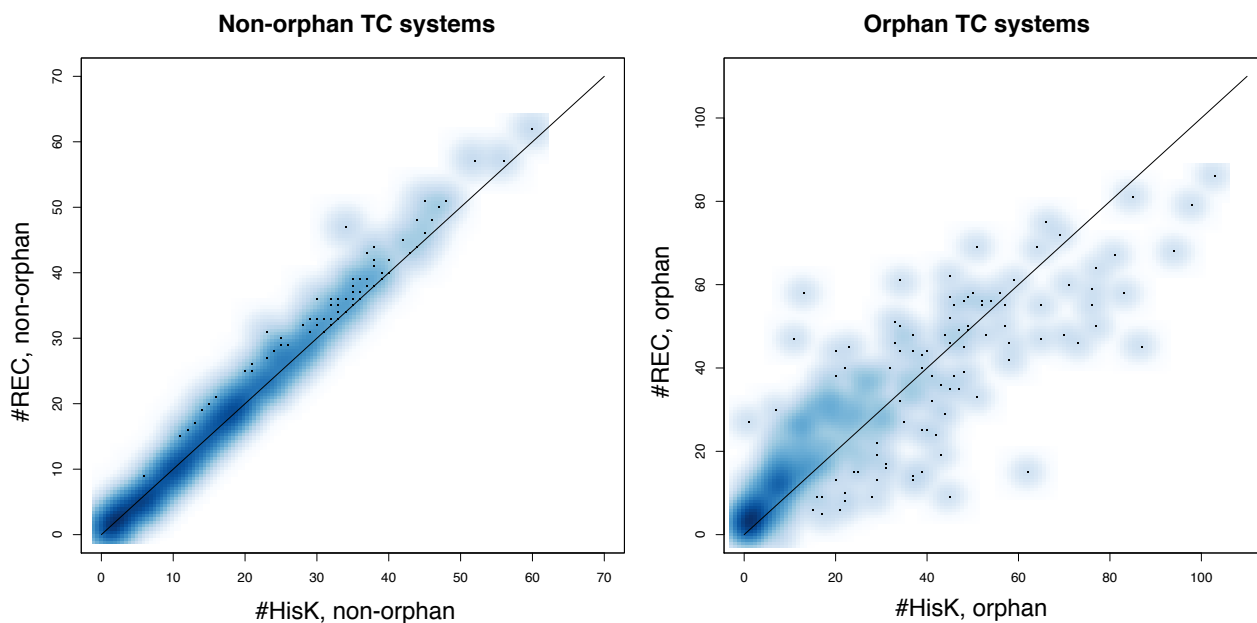
2C odds ratios in prokaryotic genomes: (A) Scatter plot of the numbers of 1C and 2C systems in prokaryotic genomes. (B) Boxplots showing the distribution of 2C odds ratios in selected phylogenetic classes of bacteria. (C) Scatter plot of 2C odds ratio and the total number of genes encoded in a genome. (D) Boxplots showing the distributions of the number of genes encoded in genomes belonging to three classes of organisms, binned on the basis of their 2C odds ratios (≤ 1.5 , $1.5 - 2$, ≥ 2).

Supplementary Material 7: Non-transcriptional RECs

Over 62% of all RECs are classifiable as transcriptional. Among the 4,061 RECs without recognisable partner domains, 3,700 RECs are classified as single domain RECs as the longest stretch of amino acids without a domain assignment is less than 50.

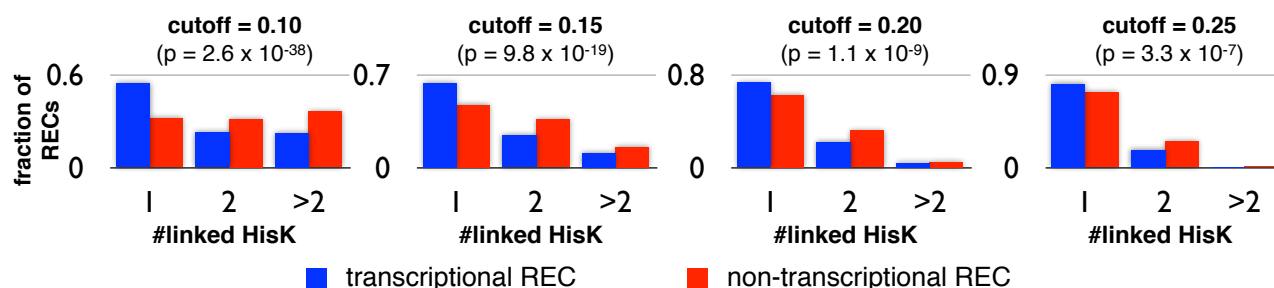


Supplementary Material 8: Correlations between numbers of orphan and non-orphan HisKs and RECs



The above are plots of the numbers of HisKs and RECs for non-orphan (left panel) and orphan (right panel) systems. The solid line represents $x=y$.

Supplementary Material 9: HisK-REC networks



The above figure shows distributions of the number of HisKs linked to transcriptional and non-transcriptional RECs in the Burgen and van Nimwegen prediction dataset. Different probability score cutoffs (0.1, 0.15, 0.2 and 0.25) were used to define an interaction between a HisK and a REC; results obtained at a cutoff of 0.15 (where 80% of typical transcriptional RECs have at least one predicted HisK) are discussed in the paper. Also shown are p-values from the Wilcoxon test comparing the degree distributions for the two classes of RECs being compared at each cutoff. The difference in degree distributions between the two classes of RECs is significant at each of the four cutoffs.

Supplementary Material 10: Computation of signalling interaction occurrence and its relationship with genome size

The signalling interaction occurrence (SIO) was calculated using the following formula:

$$\text{SIO} = (\text{N}_{\text{OC}} + \text{N}_{\text{non-orphan-REC}} + \text{N}_{\text{SDREC}} + \lambda \text{N}_{\text{orphan-REC}}) / (\text{N}_{\text{genes}} \times \text{N}_{\text{genes}})$$

N_{OC} = Number of one-component systems

N_{SDREC} = Number of single domain RECs

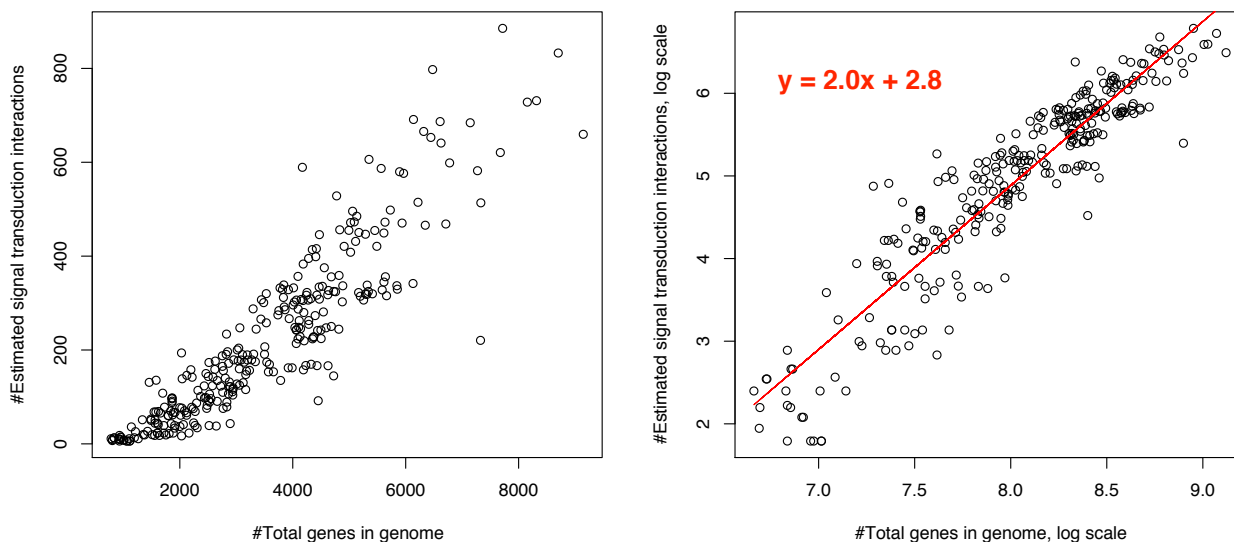
$\text{N}_{\text{orphan-REC}}$ = Number of orphan RECs

$\text{N}_{\text{non-orphan-REC}}$ = Number of non-orphan RECs.

N_{genes} = Total number of genes in genome

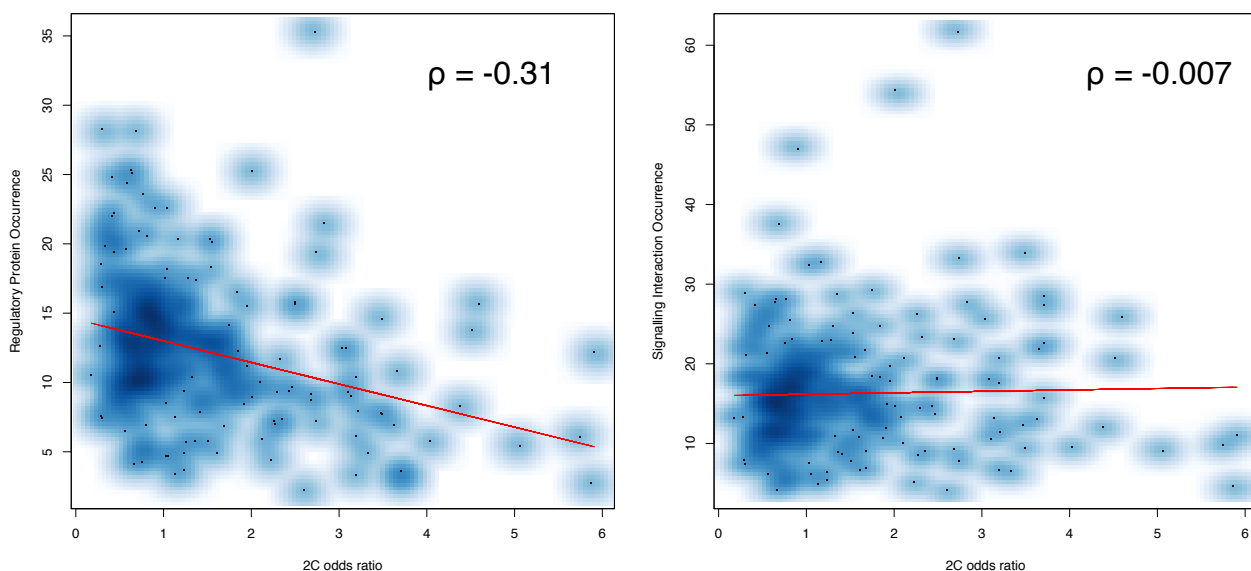
λ is calculated as follows: for each REC in the Burgen and van Nimwegen dataset, the highest score ($\text{P}_{\text{REC:genome}}$) among the probabilities assigned to its interactions with HKs in the genome was computed. Then, the average of the above scores ($\text{avg}_{\text{P}_{\text{REC:genome}}}$) was computed for each genome. Then, the number of interactions per orphan REC, λ , was calculated as $1 / \text{avg}_{\text{P}_{\text{REC:genome}}}$.

The square of the total number of genes was used as the normalisation factor because the best fit for the relationship between the number of signal transduction interactions and the total number of genes in a genome is quadratic (with exponent 2.0) as shown below.



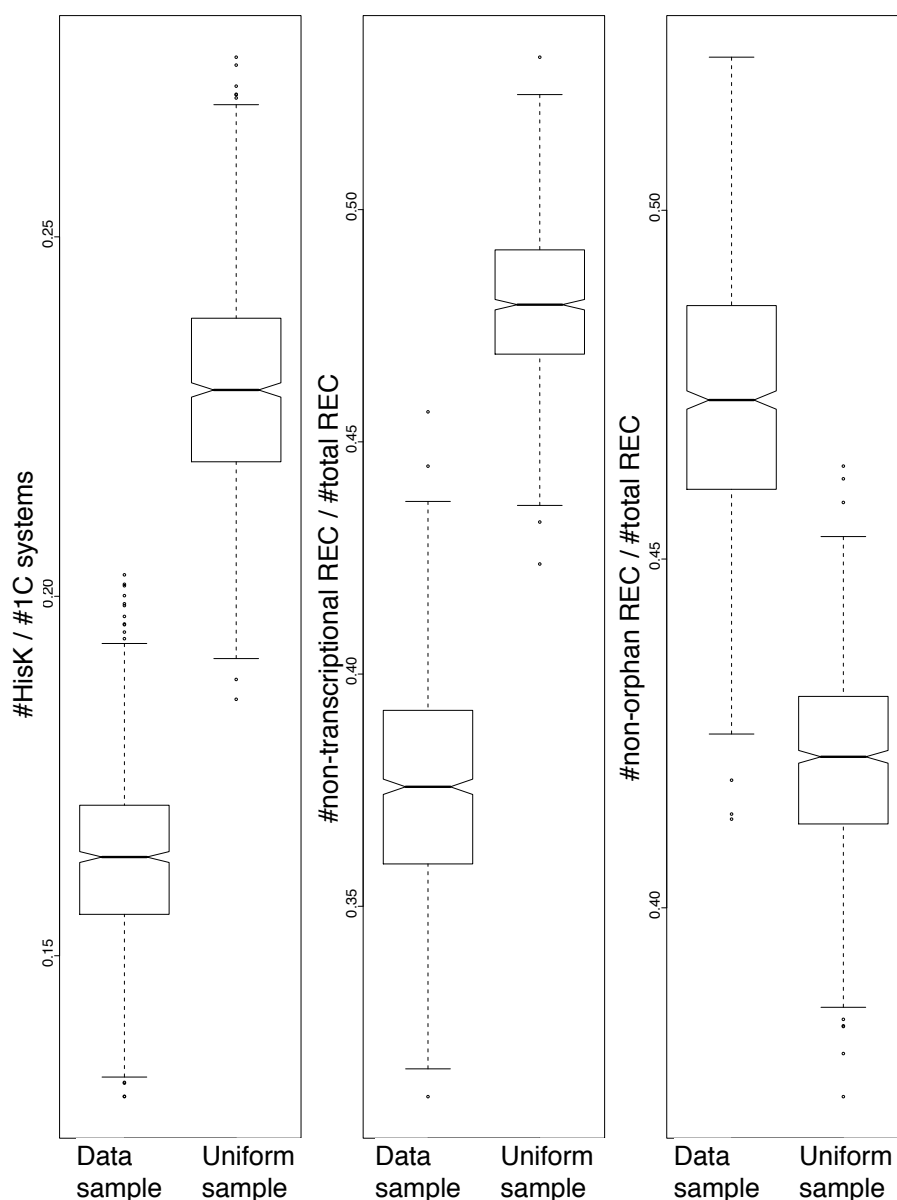
Supplementary Material 11: Regulatory Protein Occurrence, Signalling Interaction Occurrence and 2C odds ratio

Regulatory Protein Interaction decreases with increasing 2C odds ratio (left panel), whereas signalling interaction occurrence remains relatively constant (right panel). This is shown as smoothed scatter plots below. Spearman correlation coefficients are quoted.



Supplementary Material 12: Comparison of currently available genomes with genome sets of uniform phylogenetic distribution

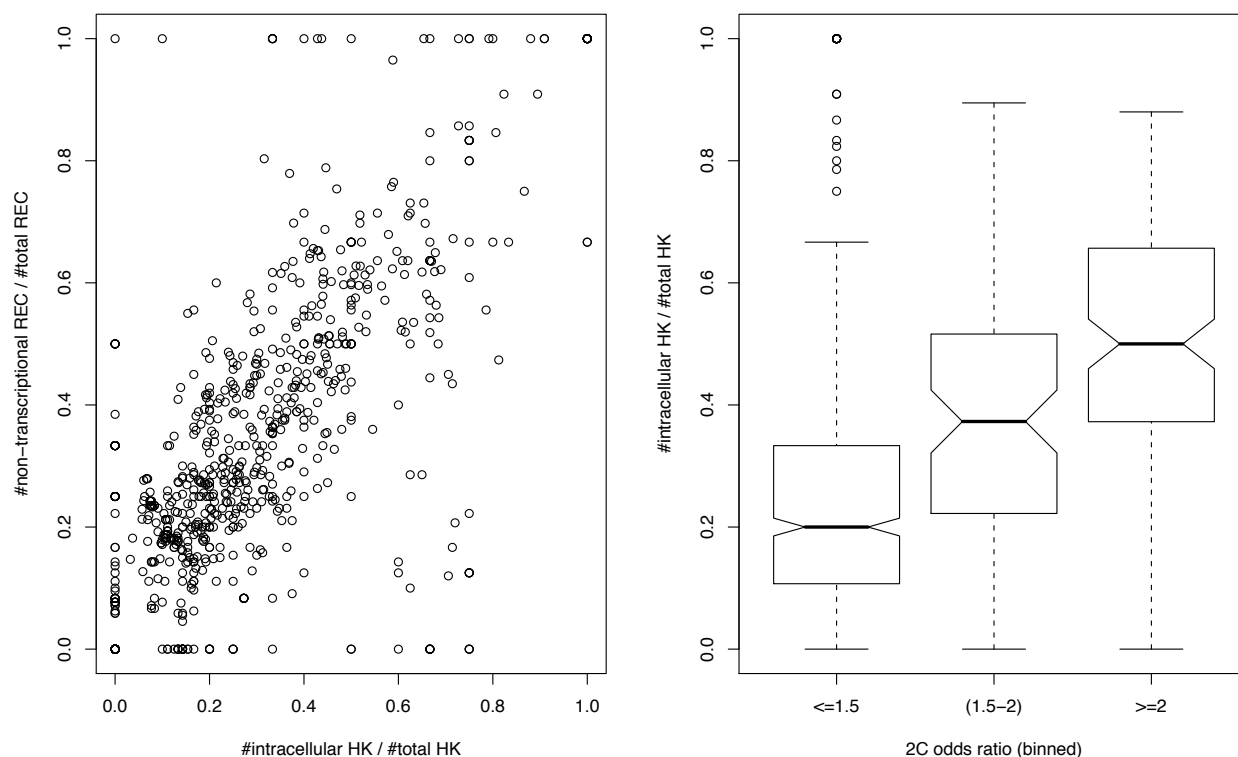
We generated 1,000 sets of 91 genomes of near-uniform phylogenetic distribution, such that no phylogenetic class was represented more than five times. This contrasts the phylogenetic distribution of the set of genomes used in our study, where 45% of all genomes fall within two classes namely Gamma Proteobacteria and Firmicutes. We compared the distributions of the following parameters between these phylogenetically uniform sets of genomes and random samples (n=91) from the complete set of genomes used in our study. In summary, in phylogenetically uniform sets, (a) the ratio of 2C and 1C systems, (b) the proportion of non-transcriptional RECs and, (c) the proportion of orphan 2C systems, are higher (by ~7-10 percentage points) than in the phylogenetically biased genome data.



The **left panel** shows the distributions of the ratios of the numbers of 2C (represented by HisK) and 1C systems. The **middle panel** shows the distributions of the proportions of non-transcriptional RECs. The **right panel** shows the distributions of the proportions of non-orphan RECs. In each panel, the left plot is for random samples from the genome dataset used in our study; the right panel is for phylogenetically uniform sets of genomes.

Supplementary Material 13: Association between the prevalence of intracellular HisKs and HisK preference

There is a significant correlation between the proportions of intracellular HisKs (as defined by not containing a TMHMM-predicted transmembrane helix) and that of non-transcriptional RECs. As might be expected from this association, genomes with high HisK preference tend to code for a higher proportion of intracellular HisKs.



The **left panel** shows a scatter plot of the relationship between the proportions of intracellular HisKs and non-transcriptional RECs. The **right panel** shows the distributions of the proportions of intracellular HisKs for genomes falling in different bins of HisK preference.

Supplementary Material 14: Signal-sensing partner domains of HisKs and TFs

The identity of the most prevalent signal-sensing partner domains is different between HisKs and TFs. Whereas the most common signal sensing domains in TFs - namely LysR_substrate, FCD, LacI and Peripla_BP_1 (not including RECs) - are rarely seen in HisKs, the PAS and GAF domains which are prevalent in HisKs are relatively uncommon in TFs. These numbers are given below.

Domains for	Partner domain	#TFs with domain	#HisKs with domain
TFs	LysR_substrate	17,847 (14%)	0
	FCD	4,100 (3%)	0
	LacI	3,925 (3%)	0
	Peripla_BP_1	3,584 (3%)	12 (<1%)
HisKs	PAS	758 (<1%)	5,366 (25%)
	GAF	797 (<1%)	1,665 (8%)