SUPPLEMENTAL:

Table S1
Distribution of reference organisms used sorted by family with list of accession numbers

Family	Species count	Accession List:
unclassified *Actinobacteria*	1	CP003219, CP000431, AP012204, CP011541, CP015235, CP000249, CP009110, CP011546, CP012184, CP003119, FO082843, CP003876, CP003275, CP011853, FO117623, CP003325, AP012326, AP012327, AP012331, AP012334, CP005287, HE804045, CP003696, CP011883, CP003788, CP013747, CP006259, CP007790, CP003924, CP004353, CP006272, CP004346, CP008889, CP004351, CP005080, AP012057, CP003170, CP006365, CP011491, CP011545, AP013105, CP006734, CP007699, CP006842, CP006841, CP006850, CP006996, CP011786, CP007155, CP007156, CP013292, CP007595, CP009302, CP011773, CP009211, CP011005, CP011311, CP010827, CP011868, CP011489, CP012070, CP012171, CP014869, CP012479, CP012697, LN831026, CP013743, CP014196, CP014475, CP015849, CP012949, CP000518, CP011341, CP008944, CP000509, AE014184, CP009896, CP008953, CP003078, CP000854, AP008957, CP012750, AE016958, CP002638, CP000386, CP000750, AE017283, CP000088, CP002385, CP000910, CP000474, CP000454, CP000820, CP002299, CR931997, AM849034, CT573213, CP000511, CP000481, CP011312, CP010407, CP000667, LN868938, AP009152, CP011340, CP000850, CP009215, CP009922, CP007514, CP002786, CP011492, CP001630, CP001964, CP002040, CP001819, CP001778, CP001821, CP013254, AP010968, CP001341, AP009493, CP004370, CP001874, CP001682, CP001854, CP001738, CP001618, CP001684, CP001706, CP001683, FO203431, CP001686, CP001737, CP001814, CP001700, CP001736, CP001726, CP003322, AP012211, AM942444, AP012319, AP012322, CP001721, CP001966, CP001631, CP001867, CP001802, CP012072, CP007490, CP001992, CP001849, CP001606, CP006835, CP006764, CP011112, CP011542, CP002666, CP002475, CP002665, CP008802, CP001958, CP002162, CP002045, CP001620, CP005929, CP002994, CP002801, CP002857, CP002593, FN554889, AP011540, CP001829, FN563149, CP002628, CP006936, CP003053, CP003169, CP002343, CP012069, CP002810, CP002047, CP002280, CP002734, CP015810, CP002917, CP012752, AP012333, CP002379, CP009220
Pseudonocardiales	11	
Actinomycetales	4	
Kineosporiales	1	
Nakamurellales	1	
Frankiales	5	
Coriobacteriales	4	
Micromonosporales	8	
Bifidobacteriales	11	
Eggerthellales	5	
Acidimicrobiales	2	
Rubrobacterales	2	
Catenulisporales	1	
Propionibacteriales	7	
Corynebacteriales	59	
Acidothermales	1	
Streptosporangiales	5	
Glycomycetales	1	
Solirubrobacterales	1	
Micrococcales	37	
Geodermatophilales	3	
Streptomycetales	19	

Table S2
Table of core genes found in the reference set listed by functional category (continued to pg. 6)

Model Classification	Count	Model Names
Amino acid biosynthesis	54	TIGR00032, TIGR00033, TIGR00036, TIGR00069, TIGR00070, TIGR00110, TIGR00112, TIGR00118, TIGR00119, TIGR00120, TIGR00170, TIGR00171, TIGR00191, TIGR00260, TIGR00262, TIGR00263, TIGR00338, TIGR00407, TIGR00465, TIGR00507, TIGR00564, TIGR00652, TIGR00653, TIGR00658, TIGR00674, TIGR00676, TIGR00735, TIGR00761, TIGR00838, TIGR00970, TIGR01027, TIGR01048, TIGR01088, TIGR01123, TIGR01127, TIGR01137, TIGR01141, TIGR01245, TIGR01296, TIGR01327, TIGR01356, TIGR01357, TIGR01358, TIGR01366, TIGR01392, TIGR01536, TIGR01795, TIGR01808, TIGR01850, TIGR01855, TIGR01900, TIGR02067, TIGR02082, TIGR03188
Biosynthesis of cofactors, prosthetic groups, and carriers	79	TIGR00018, TIGR00063, TIGR00078, TIGR00083, TIGR00097, TIGR00109, TIGR00114, TIGR00151, TIGR00152, TIGR00154, TIGR00173, TIGR00187, TIGR00190, TIGR00204, TIGR00212, TIGR00216, TIGR00222, TIGR00223, TIGR00243, TIGR00313, TIGR00343, TIGR00347, TIGR00379, TIGR00380, TIGR00433, TIGR00453, TIGR00482, TIGR00508, TIGR00510, TIGR00521, TIGR00525, TIGR00539, TIGR00550, TIGR00551, TIGR00554, TIGR00558, TIGR00562, TIGR00581, TIGR00612, TIGR00636, TIGR00693, TIGR00708, TIGR00713, TIGR00715, TIGR00751, TIGR01035, TIGR01378, TIGR01379, TIGR01464, TIGR01465, TIGR01469, TIGR01473, TIGR01475, TIGR01496, TIGR01498, TIGR01510, TIGR01513, TIGR01683, TIGR01819, TIGR01923, TIGR01929, TIGR01978, TIGR01980, TIGR01981, TIGR01994, TIGR02150, TIGR02257, TIGR02352, TIGR02476, TIGR02666, TIGR03160, TIGR03438, TIGR03442, TIGR03447, TIGR03448, TIGR03552, TIGR03699, TIGR03701, TIGR03800

Model Classification	Count	Model Names
Cell envelope	21	TIGR00031, TIGR00055, TIGR00067, TIGR00179, TIGR00219, TIGR00274, TIGR00445, TIGR00492, TIGR00753, TIGR01072, TIGR01082, TIGR01087, TIGR01099, TIGR01133, TIGR01181, TIGR01207, TIGR01214, TIGR01221, TIGR01695, TIGR03423, TIGR03426
Cellular processes	15	TIGR00065, TIGR00172, TIGR00220, TIGR00647, TIGR00685, TIGR00732, TIGR02210, TIGR02400, TIGR02614, TIGR02673, TIGR03137, TIGR03253, TIGR03445, TIGR03446, TIGR03451
Central intermediary metabolism	16	TIGR00101, TIGR00192, TIGR00193, TIGR00221, TIGR00455, TIGR00642, TIGR00700, TIGR01034, TIGR01135, TIGR01173, TIGR01217, TIGR01455, TIGR01792, TIGR02727, TIGR03383, TIGR03705
DNA metabolism	35	TIGR00042, TIGR00084, TIGR00194, TIGR00228, TIGR00237, TIGR00281, TIGR00362, TIGR00416, TIGR00575, TIGR00577, TIGR00580, TIGR00593, TIGR00595, TIGR00613, TIGR00615, TIGR00628, TIGR00630, TIGR00631, TIGR00634, TIGR00635, TIGR00643, TIGR00663, TIGR00665, TIGR01051, TIGR01059, TIGR01063, TIGR01073, TIGR01083, TIGR01128, TIGR01280, TIGR01389, TIGR01391, TIGR02012, TIGR02168, TIGR02225

Model Classification	Count	Model Names
Energy metabolism	86	TIGR00016, TIGR00021, TIGR00129, TIGR00131, TIGR00203, TIGR00209, TIGR00218, TIGR00232, TIGR00239, TIGR00273, TIGR00330, TIGR00419, TIGR00461, TIGR00518, TIGR00527, TIGR00528, TIGR00561, TIGR00651, TIGR00692, TIGR00720, TIGR00871, TIGR00872, TIGR00873, TIGR00876, TIGR00936, TIGR00962, TIGR00979, TIGR01039, TIGR01060, TIGR01064, TIGR01068, TIGR01078, TIGR01131, TIGR01132, TIGR01144, TIGR01145, TIGR01146, TIGR01163, TIGR01179, TIGR01198, TIGR01216, TIGR01224, TIGR01225, TIGR01228, TIGR01235, TIGR01236, TIGR01255, TIGR01260, TIGR01263, TIGR01266, TIGR01292, TIGR01311, TIGR01312, TIGR01341, TIGR01344, TIGR01515, TIGR01520, TIGR01534, TIGR01722, TIGR01771, TIGR01798, TIGR01814, TIGR01816, TIGR01828, TIGR01915, TIGR01959, TIGR01962, TIGR01973, TIGR02022, TIGR02152, TIGR02156, TIGR02157, TIGR02158, TIGR02159, TIGR02422, TIGR02423, TIGR02426, TIGR02427, TIGR02866, TIGR02891, TIGR02970, TIGR03036, TIGR03181, TIGR03377, TIGR04380, TIGR04382
Fatty acid and phospholipid metabolism	6	TIGR00189, TIGR00473, TIGR00517, TIGR00560, TIGR01830, TIGR03150
Hypothetical proteins	8	TIGR00247, TIGR00252, TIGR00370, TIGR00730, TIGR01777, TIGR02569, TIGR03084, TIGR03085
Protein fate	43	TIGR00064, TIGR00077, TIGR00079, TIGR00115, TIGR00121, TIGR00214, TIGR00357, TIGR00382, TIGR00401, TIGR00493, TIGR00500, TIGR00504, TIGR00534, TIGR00544, TIGR00546, TIGR00556, TIGR00706, TIGR00763, TIGR00810, TIGR00945, TIGR00959, TIGR00963, TIGR00964, TIGR00966, TIGR00967, TIGR01129, TIGR01410, TIGR02227, TIGR02348, TIGR02350, TIGR02493, TIGR03144, TIGR03346, TIGR03534, TIGR03686, TIGR03687, TIGR03688, TIGR03689, TIGR03690, TIGR03691, TIGR03919, TIGR03920, TIGR03921

Model Classification	Count	Model Names
Protein synthesis	115	Ribosomal_L10, Ribosomal_L23, Ribosomal_S14, Ribosomal_S8, Ribosomal_S9, TIGR00001, TIGR00008, TIGR00009, TIGR00011, TIGR00012, TIGR00019, TIGR00020, TIGR00029, TIGR00038, TIGR00043, TIGR00048, TIGR00057, TIGR00059, TIGR00060, TIGR00061, TIGR00062, TIGR00071, TIGR00086, TIGR00088, TIGR00090, TIGR00091, TIGR00095, TIGR00096, TIGR00105, TIGR00116, TIGR00132, TIGR00133, TIGR00135, TIGR00138, TIGR00150, TIGR00157, TIGR00158, TIGR00165, TIGR00166, TIGR00168, TIGR00174, TIGR00186, TIGR00233, TIGR00234, TIGR00256, TIGR00344, TIGR00392, TIGR00396, TIGR00398, TIGR00409, TIGR00414, TIGR00418, TIGR00420, TIGR00422, TIGR00430, TIGR00431, TIGR00435, TIGR00436, TIGR00442, TIGR00447, TIGR00456, TIGR00459, TIGR00460, TIGR00464, TIGR00467, TIGR00468, TIGR00472, TIGR00479, TIGR00484, TIGR00485, TIGR00487, TIGR00496, TIGR00499, TIGR00563, TIGR00731, TIGR00755, TIGR00855, TIGR00952, TIGR00981, TIGR01009, TIGR01011, TIGR01017, TIGR01021, TIGR01022, TIGR01023, TIGR01024, TIGR01029, TIGR01030, TIGR01031, TIGR01032, TIGR01044, TIGR01049, TIGR01050, TIGR01066, TIGR01067, TIGR01071, TIGR01079, TIGR01125, TIGR01164, TIGR01169, TIGR01171, TIGR01308, TIGR01575, TIGR01632, TIGR02692, TIGR02729, TIGR03594, TIGR03631, TIGR03632, TIGR03635, TIGR03654, TIGR03704, TIGR03723, TIGR03725, TIGR03953
Purines, pyrimidines, nucleosides, and nucleotides	43	TIGR00017, TIGR00041, TIGR00081, TIGR00126, TIGR00184, TIGR00302, TIGR00336, TIGR00337, TIGR00355, TIGR00639, TIGR00655, TIGR00670, TIGR00877, TIGR00878, TIGR00884, TIGR00928, TIGR01036, TIGR01090, TIGR01091, TIGR01134, TIGR01161, TIGR01162, TIGR01203, TIGR01302, TIGR01354, TIGR01368, TIGR01369, TIGR01694, TIGR01704, TIGR01736, TIGR01737, TIGR01744, TIGR02075, TIGR02127, TIGR02170, TIGR02274, TIGR02487, TIGR02491, TIGR02504, TIGR02506, TIGR02961, TIGR03263, TIGR03284

Model Classification	Count	Model Names
Regulatory functions	9	TIGR00242, TIGR00244, TIGR00331, TIGR00498, TIGR01394, TIGR01529, TIGR01693, TIGR01950, TIGR03968
Transcription	19	TIGR00082, TIGR00188, TIGR00690, TIGR00766, TIGR00767, TIGR00922, TIGR01388, TIGR01951, TIGR01953, TIGR01966, TIGR02013, TIGR02027, TIGR02191, TIGR02258, TIGR02273, TIGR02386, TIGR02393, TIGR02949, TIGR03988
Transport and binding proteins	20	TIGR00383, TIGR00400, TIGR00750, TIGR00754, TIGR00773, TIGR00972, TIGR00974, TIGR00975, TIGR01104, TIGR01256, TIGR02135, TIGR02138, TIGR02275, TIGR03409, TIGR03410, TIGR03770, TIGR03771, TIGR03772, TIGR03851, TIGR04520
Unclassified	64	TIGR00177, TIGR00180, TIGR00369, TIGR00759, TIGR01350, TIGR01412, TIGR01417, TIGR01428, TIGR01430, TIGR01581, TIGR01698, TIGR01701, TIGR01751, TIGR01788, TIGR01885, TIGR01919, TIGR02133, TIGR02188, TIGR02200, TIGR02234, TIGR02278, TIGR02288, TIGR02349, TIGR02412, TIGR02631, TIGR02753, TIGR02857, TIGR02927, TIGR02947, TIGR02952, TIGR02960, TIGR03003, TIGR03005, TIGR03081, TIGR03083, TIGR03086, TIGR03089, TIGR03178, TIGR03180, TIGR03356, TIGR03449, TIGR03450, TIGR03452, TIGR03459, TIGR03464, TIGR03465, TIGR03467, TIGR03535, TIGR03539, TIGR03625, TIGR03664, TIGR03815, TIGR03817, TIGR03819, TIGR03843, Cpn10, GrpE, Methyltransf_5, PGK, SHMT, TIGR03869, TIGR03873, TIGR03936, TIGR03997
Unknown function	31	TIGR00044, TIGR00092, TIGR00103, TIGR00149, TIGR00164, TIGR00196, TIGR00250, TIGR00257, TIGR00368, TIGR00481, TIGR00486, TIGR00494, TIGR00724, TIGR00726, TIGR00762, TIGR00977, TIGR01303, TIGR01304, TIGR01393, TIGR01448, TIGR01490, TIGR01764, TIGR01967, TIGR01970, TIGR01976, TIGR03156, TIGR03816, TIGR03941, TIGR03954, TIGR03960, TIGR04047

Figure S3
Core gene comparison to the Database of Essential Genes (DEG) sorted by function. Red columns indicate hits not found in the database, compared to all core gene classifications shown in blue

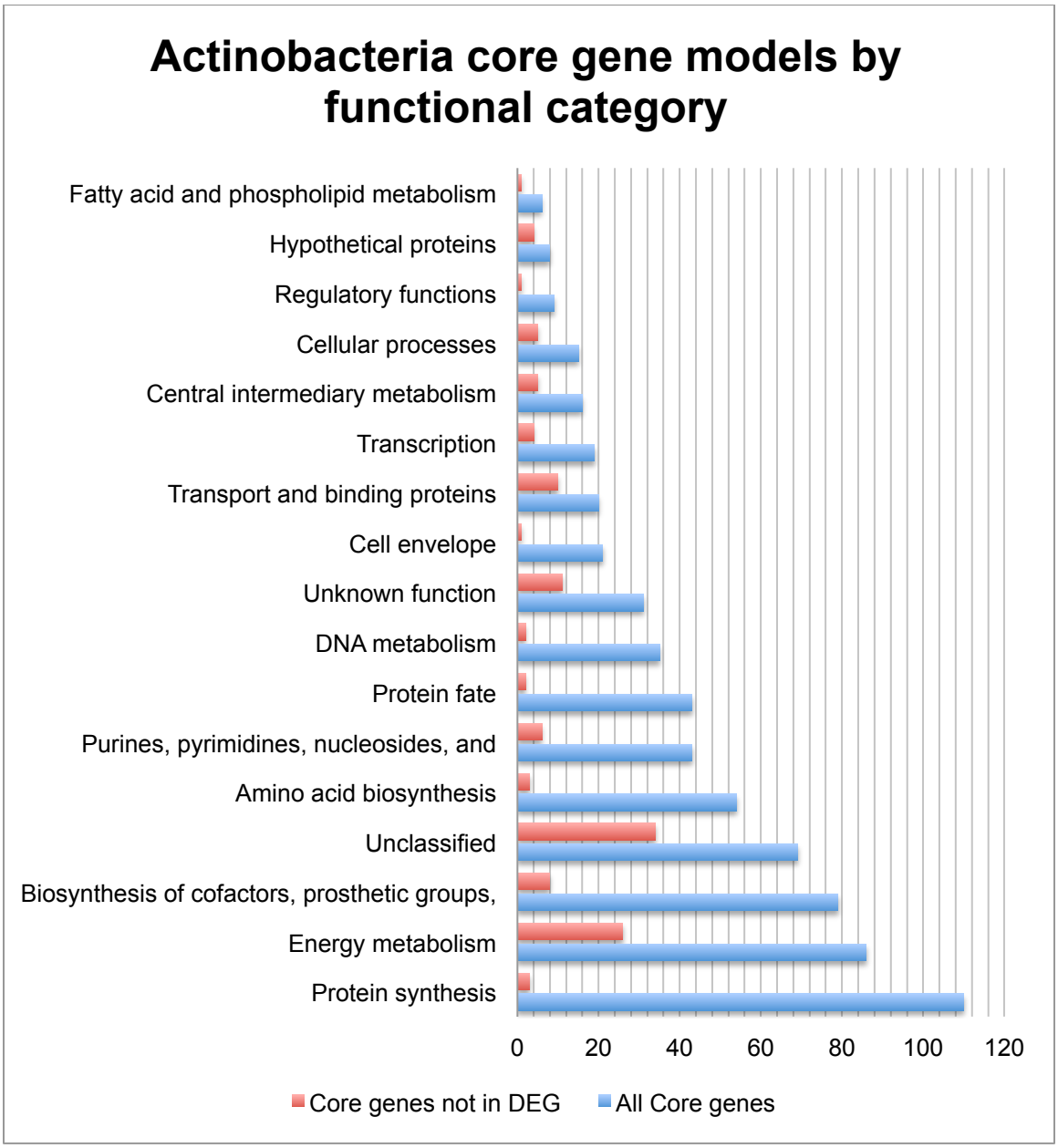


Figure S4
Comparison of counts for all proteins in the reference set sorted by function. Red columns indicate distribution of all core gene hits and blue indicate distribution of all TIGRFAM model hits

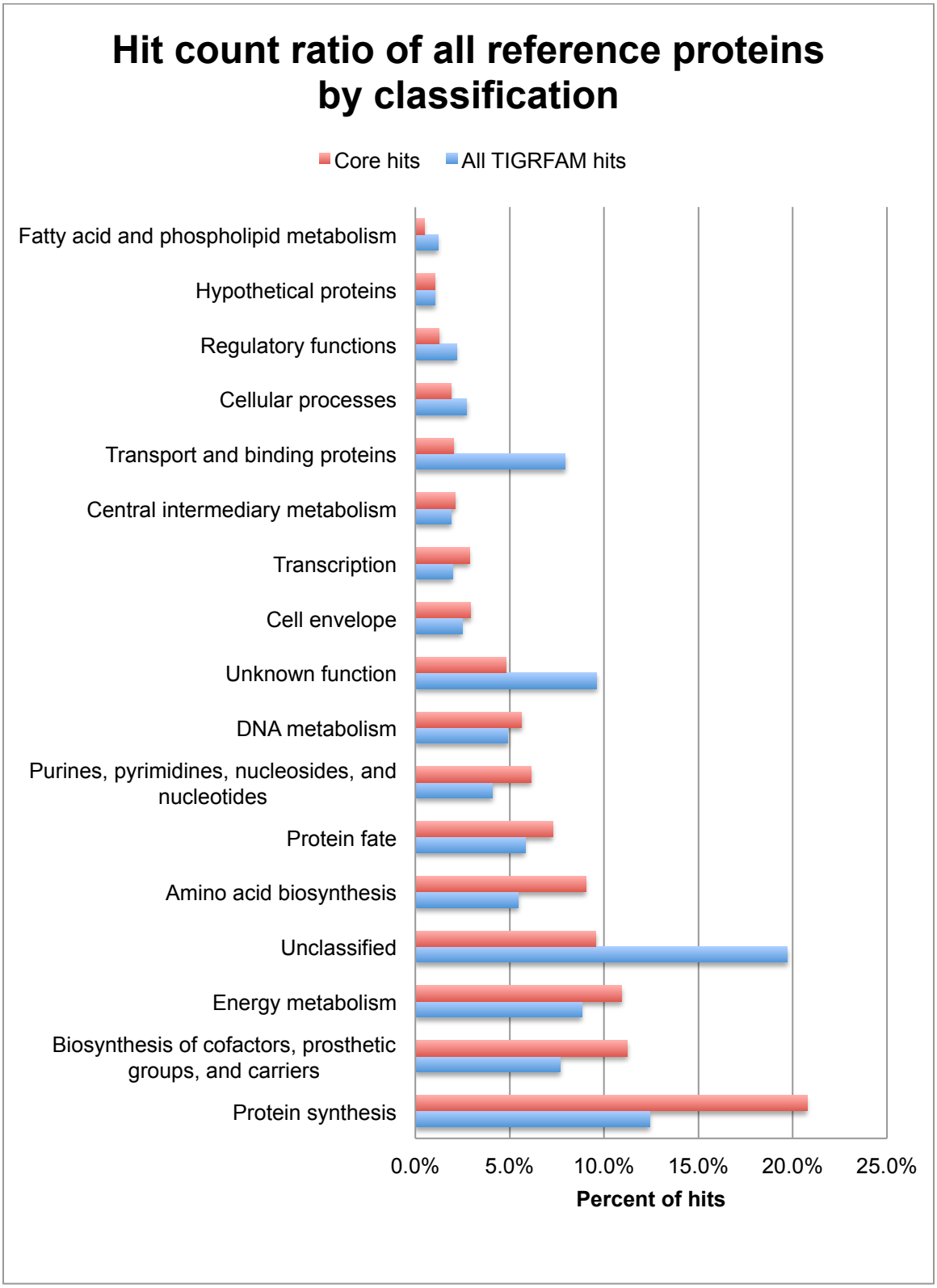


Figure S5
Histograms of essentiality measures. (a) Median values of pairwise dN/dS calculations for all alignments. (b) Reference ubiquity by species represents presence in all genomes. If the majority of organisms in a genus have a gene it is counted and these counts are shown relative to total genera in the set. (c) Single copy ratio shows number of occurrences a gene is found only once divided by total genomes that have one or more copies

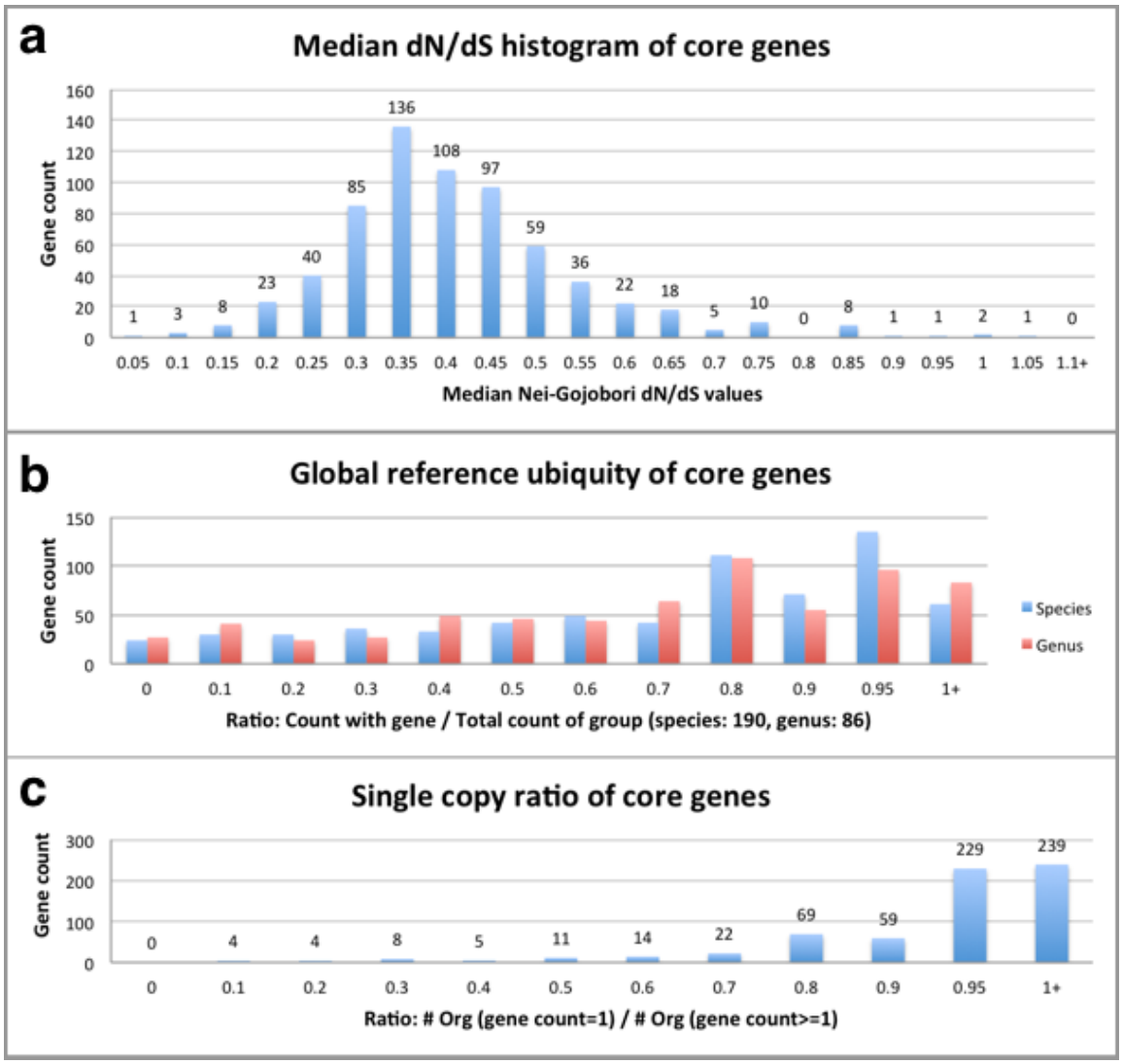


Figure S6
Median and mode distributions of all bootstrap support values of reference trees using all nodes and terminal nodes respectively

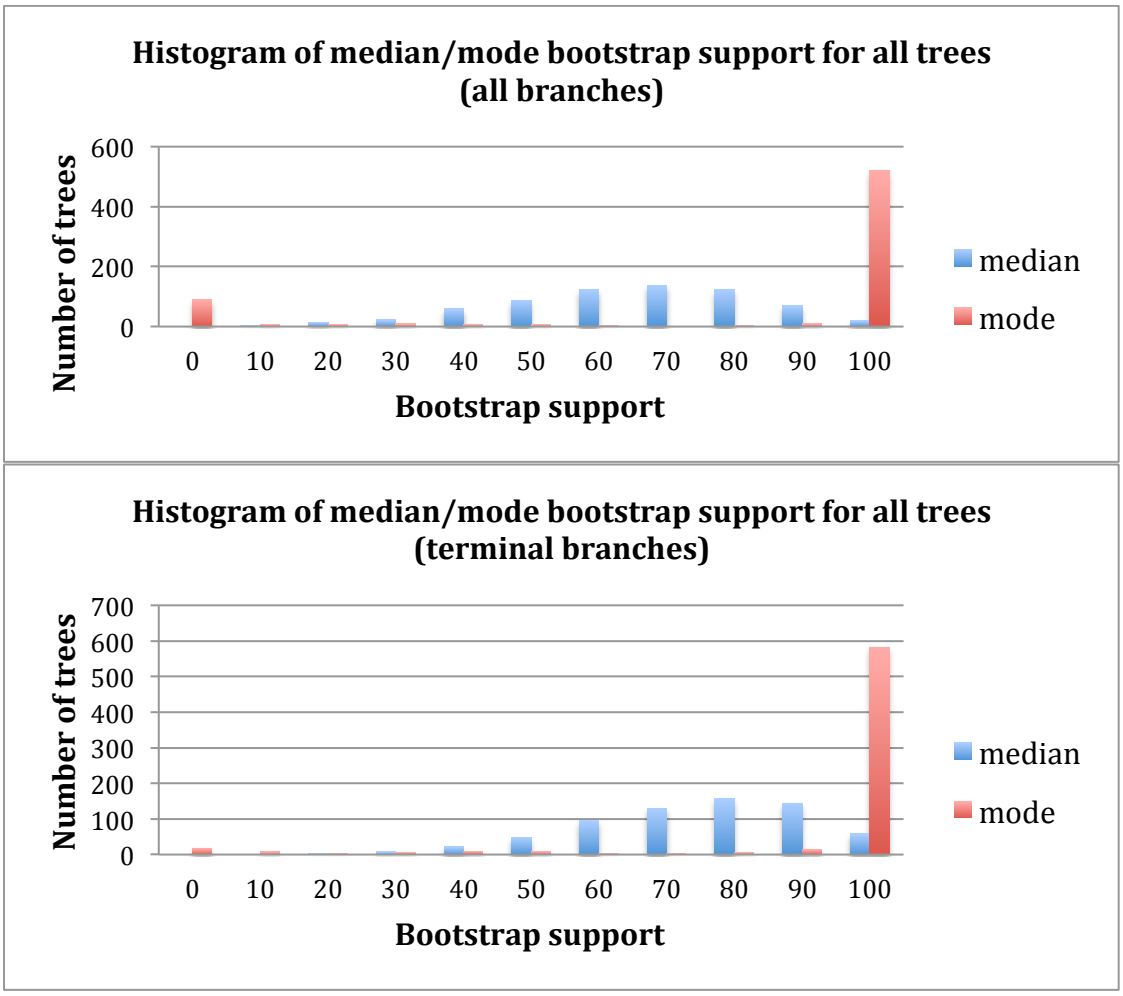


Figure S7
(a) Functional classification of all MIBiG cluster core gene hits. Inner ring shows hits with exploration mode enabled while outer ring is default filtered core hits. Actual counts are presented in parenthesis (b) Distribution of hit counts for clusters where blue, red and purple columns represent exploration core, resistance, and default core model hits respectively

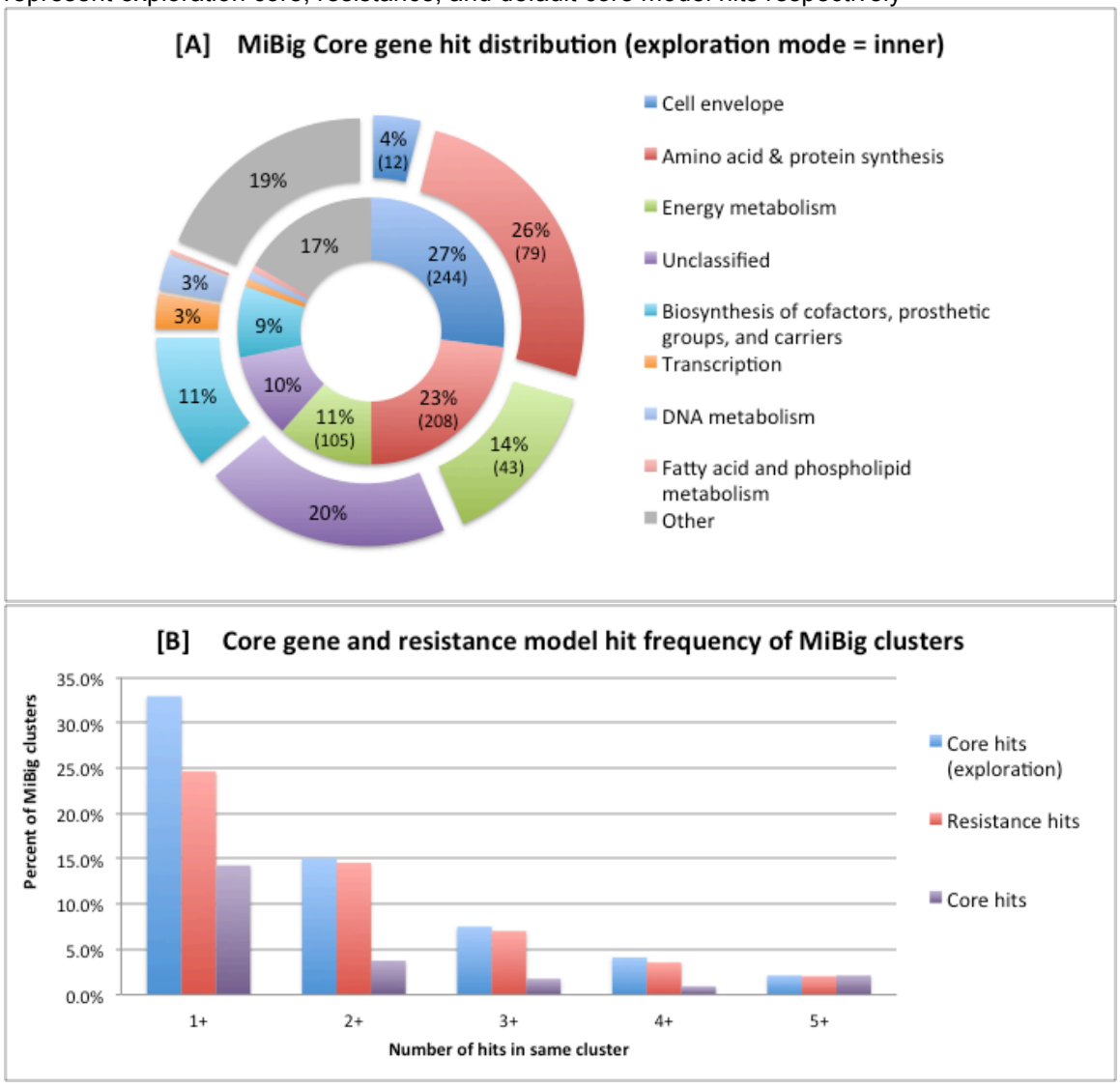


Table S8
Detection frequency statistics from complete actinobacterial genomes using combined 189 reference and 11 positive controls run through the ARTS pipeline.

Hit type	Avg	Std	min	max
Core	489	79	271	653
Duplicate	27	23	0	96
BGC prox.	27	24	0	140
Phylogeny	125	88	7	422
Two +	16	16	0	83
Three +	2	4	0	22

Percent of core genes showing criteria				
Hit type	**Avg**	**Std**	**min**	**max**
Duplicate	**5.07%**	**3.85%**	**0.00%**	**16.03%**
BGC prox.	**5.03%**	**4.34%**	**0.00%**	**25.78%**
Phylogeny	**25.74%**	**17.53%**	**1.68%**	**70.69%**
Two +	**3.32%**	**3.00%**	**0.00%**	**13.74%**
Three +	**0.35%**	**0.65%**	**0.00%**	**3.69%**

Table S9

Positive examples of genomes with known self-resistance mechanisms analyzed with ARTS exploration mode. Hits to ARTS criteria are shown as; D: Duplication, B: BGC proximity, P: Phylogeny, R: Resistance model. Rows in grey indicate non-actinobacterial genomes where phylogeny criteria do not apply. Tan rows indicate resistance that is not within a BGC. Notes marked with stars are explained in the bottom row

Product	Resistance gene	Organism	Gene Accession (ref)	ARTS hits	Criteria hits (>2, >3)	BGCs (total, core hit, res hit)	Genes (core, total)
Novobiocin	duplicated *gyrB*	*Streptomyces niveus* NCIMB 11891	WP_03123 2360 (1)	**D , B , R, *P**	46 , 13	30 , 24 , 8	577 , 7815
Clorobiocin	duplicated *gyrB*	*Streptomyces roseochromogenes* DS 12.976	AAN65247 (2)	**D , B , R, *P**	71 , 20	43 , 24 , 13	590 , 9055
Albicidin	pentapeptide repeat protein for GyrB	*Xanthomonas albilineans* GPE PC73	CBA16025 (3)	**B , R**	2, 0	8 , 7 , 3	434 , 3208
Streptolydigin	mutated *rpoB*	*Streptomyces lydicus* NRRL2433	AAQ19729 (4)	**R**	42 , 3	35 , 17 , 13	576 , 8518
Rifamycin	mutated *rpoB*	*Amycolatopsis mediterranei* S699	AAS07760 (5)	**R**	49 , 5	30 , 16 , 8	564 , 9575
Rifampicin	duplicated *rpoB*	*Nocardia farcinica* IFM 10152	BAD59497. 1 (6)	**D**	21 , 3	17, 13, 7	550 , 5946
Thiocillin	duplicated ribosomal L11	*Bacillus cereus* ATCC 14579	AAP11944, AAP11947 (7)	**D , B**	6 , 0	10 , 7 , 2	449 , 5255
Erythromycin	duplicated 23S rRNA methyltransferase	*Saccharopolyspora erythraea* NRRL23338	WP_00995 0391 (8)	****D , B , P , R**	83 , 17	36 , 16 , 13	653 , 7198
Agrocin 84	duplicated Leu-tRNA synthase	*Agrobacterium radiobacter* K84	ACM31456 (9)	**D**	4 , 0	10 , 7 , 0	475 , 6684
Thiolactomycin	duplicated FabB/F	*Salinispora pacifica* DSM 45543	ALJ49913 (10)	****D , B , P**	61 , 20	25 , 20 , 9	641 , 4784
Salinospora-mide A	duplicated beta-proteasome subunit	*Salinispora tropica* CNB-440	ABP53490 (11)	**D , B , P , R**	30 , 7	19 , 15 , 7	531 , 4536
Vancomycin	Peptidoglycan remodeling	*Amycolatopsis orientalis* DSM 40040	CCD33128, CCD33129, CCD33130 (12)	**B, R *****	50 , 8	39 , 22 , 17	563 , 8194
Cephamycin	duplicated beta-lactamase	*Streptomyces clavuligerus* ATCC 27064	AAF86620 (13)	**B, D, R**	26 , 3	45 , 20 , 15	546 , 7730
GE2270	duplicated elongation factor	*Planobispora rosea* ATCC 53733	AGY49599, AGY49600 (14)	**D , B , P**	34 , 6	26 , 15 , 9	549 , 8176

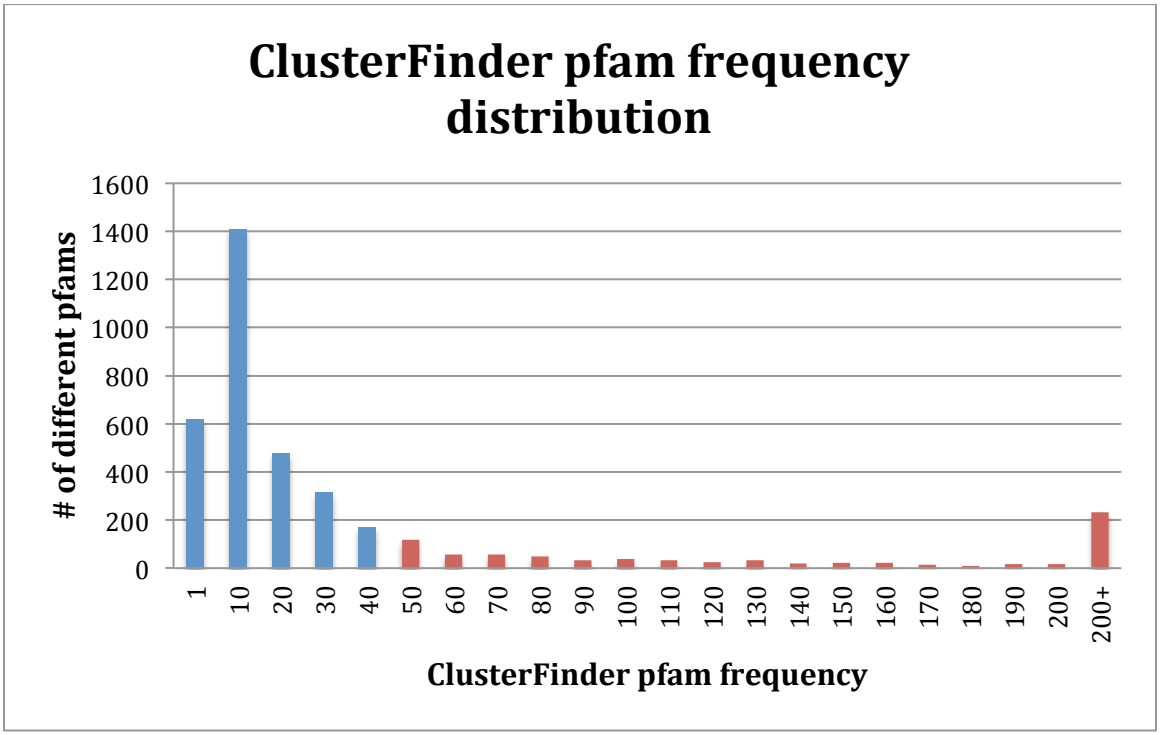
Notes: * Intra-genus phylogeny hits only seen; ** Other hits found using the advanced noise cutoff "E1" (90% model noise cutoff); *** Only *vanX* gene is detected

Table S10
ARTS results of example BGCs from the MIBiG database that contain known self-resistance mechanisms in their sequence. “Core” and “Res.” Indicate a match to a gene for core and resistance model respectively. Where both models match the same protein, a “+” is indicated. Notes indicated with a star are listed in the row below

Product	Resistance gene	Organism	Accession (ref)	BGC ID	ARTS Hit
Griselimycin	Copy of *dnaN*	*Streptomyces sp. DSM 40835*	AKC91855 (15)	BGC0001414	**Core + Res.**
Coumermycin	Copy of *gyrB*	*Streptomyces rishiriensis DSM 40489*	AAO47226 (16)	BGC0000833	**Core + Res.**
Novobiocin	Copy of *gyrB*	*Streptomyces niveus NCIMB 9219*	AFI47646 (1)	BGC0000834	**Core + Res.**
Albicidin	pentapeptide repeat protein for GyrB	*Xanthomonas albilineans GPE PC73*	CBA16025 (3)	BGC0001088	**Res.**
Cystobactamide	pentapeptide repeat protein for GyrB	*Cystobacter sp. Cbv34*	AKP45389 (17)	BGC0001413	**Res.**
Rifamycin	mutated *rpoB*	*Amycolatopsis mediterranei*	AAS07760 (5)	BGC0000136	***Core + Res**
Yatakemycin	Copy of DNA glycosylase	*Streptomyces sp. TP-A2060*	ADZ13541 (18)	BGC0000466	**No hit**
Azinomycin	Copy of DNA glycosylase	*Streptomyces sahachiroi*	ABY83174 (19)	BGC0000960	**No hit**
Rubradirin	two copies of Initiation factor ****	*Streptomyces achromogenes subsp. rubradiris*	CAI94679, CAI94684 (20)	BGC0000141	**Core,Core**
Thiocillin	two copies of Ribosomal protein L11 ****	*Bacillus cereus ATCC 14579*	AAP11944, AAP11947 (7)	BGC0000612	**Core,Core**
GE2270	two copies of elongation factor	*Planobispora rosea ATCC 53733*	AGY49599, AGY49600 (14)	BGC0001155	**Core,Core**
Erythromycin	duplicated 23S rRNA methyltransferase	*Saccharopolyspora erythraea NRRL2338*	WP_009950391 (8)	BGC0000055	**Res.**
Pikromycin	duplicated 23S rRNA methyltransferase	*Streptomyces venezuelae ATCC 15439*	AAC69328, AAC69327 (21)	BGC0000094	**Res.**
Avilamycin	duplicated 23S rRNA methyltransferase	*Streptomyces viridochromogenes Tue57*	AAG32067, AAG32066 (22)	BGC0000026	**No hit**
Mupirocin	duplicated Ile-tRNA synthetase	*Pseudomonas fluorescens NCIMB 10586*	AAM12927 (23)	BGC0000182	**Core**
Borrelidin	duplicated Thr-tRNA synthetase	*Streptomyces parvulus*	CAE45679 (24)	BGC0000031	**Core**
Indolmycin	duplicated Trp-tRNA synthase	*Streptomyces griseus ATCC12648*	AJT38681 (25)	BGC0001206	**Res.**
Salinosporamide A	duplicated beta-proteasome subunit	*Salinospora tropica CNB-440*	ABP53490 (11)	BGC0001041	**Core + Res.**
Eponemycin	duplicated beta-proteasome subuint	*Streptomyces hygroscopicus ATCC 53709*	AHB38505 (26)	BGC0000345	**Core + Res.**
Vancomycin	Peptidoglycan remodeling	*Amycolatopsis orientalis* HCCB 10007	CCD33128, CCD33129, CCD33130 (12)	BGC0000455	**Res.**
Cephamycin	duplicated beta-lactamase	*Streptomyces clavuligerus ATCC 27064*	AAF86620 (13)	BGC0000319	**Res.**
Platencin	duplicated FabB/F	*Streptomyces platensis MA7339*	ACS13710 (27)	BGC0001156	***Core** ****No hit**

Thiolactomycin	duplicated FabB/F	*Salinispora pacifica DSM 45543*	ALJ49913 (28)	BGC0001237	*Core **No hit
Thiotetroamide	two copies of FabB/F	*Streptomyces afghaniensis NRRL5621*	ALJ49924, ALJ49919 (28)	BGC0001236	*Core,*Core **No hit
Kalimantacin	duplicated FabI	*Pseudomonas fluorescens BCCM_ID9359*	ADD82948 (29)	BGC0001099	No hit
Andrimid	One copy of acetyl-CoA carboxyltransferase	*Pantoea agglomerans*	AAO39114 (30)	BGC0000956	Res.
* Search used exploration mode with relaxed cutoff: 90% Noise cutoff (E1) ** No hits when using default search mode *** Using the provided MIBiG cluster annotation by running though "Existing Antismash Job" section					
**** Putative resistance, in vitro experiments not shown					

Figure S11
Distribution of pfam frequencies taken from the biosynthetic pfam frequency file in ClusterFinder. These were frequencies defined from a manually curated set of 732 known BGCs. Here we have highlighted those that are most consistently found to remove higher confidence biosynthetic functions from the core gene list



SUPPLEMENTAL METHODS

Core gene and known resistance searches
Initial steps of the analyses consist of searching for BGCs, essential genes, and known resistance models. Submitted sequences that do not have BGC annotations are first processed with antiSMASH v3 using default settings to identify BGCs. The annotated Genbank is then parsed using Biopython (31) to identify all protein coding sequences, rRNAs, and cluster annotations. DNA sequences and protein translations are written to FASTA files for Hidden Markov Model (HMM) searches using HMMER (32) v3. These models include those for known

resistance factors and core genes, as determined from previous analysis detailed in the reference set section. HMM domain results are parsed and the best model hit for a gene is extracted if it passes 50% coverage length thresholds of both model and gene. This value was chosen to allow for missing domains and incomplete sequences while reducing fragmented hits. Resistance models are based on Resfam (33) which include known resistance factors from several databases (34–36). Domain of Unknown Function (DUF) models are from the Pfam (37) database and are used to highlight potential novel chemistry in a cluster. Custom submitted models are appended to corresponding core gene and known resistance models then searches are performed using model specific trusted cutoffs by default.

Reference set and core gene detection

The current *Actinobacteria* reference set is comprised of complete genomes from 189 species representing 22 different families that are available through NCBI's RefSeq (38) database (Supplemental S1). Essential genes are inferred by a comparative genomics approach where ubiquitous "core genes" are those consistently found in reference organisms as detected using HMMER and Hidden Markov Models (HMMs) from the TIGRFAM (39) protein family database; In addition, predefined core genes from the TIGRFAM v15 "bacterial core gene set" (GenProp0799) are included. All TIGRFAM homologous proteins with emphasis on conserved function, "equivologs", and their hypothetical and domain variants are used for essential gene analysis. After HMM detection, counts for genes are recorded in a gene matrix consisting of all reference genomes. Family specific core genes are then defined as genes present in greater than 95% of genomes relative to each family based on the count matrix. Families with less than 10 genomes were combined and a lowered ubiquity threshold of 90% is used instead to account for variations due to more distant relations.

For accelerated gene tree creation, all core gene sequences are extracted into corresponding multi-record FASTA files and, where applicable, out-group sequences added using model matches from various sequences of Proteobacteria, which harbored many non-actinobacterial genomes with core gene matches. Each core gene protein FASTA file is then aligned with MAFFT (40) followed by a codon alignment with Pal2nal (41). Trimmed copies of each codon alignment are made using TrimAl (42) with the maximum likelihood optimized "automated1" setting. RaxML (43) is used to build each tree with 100 bootstrap replicates using GTRGAMMAI model selection. Pairwise selection (dN/dS) values were calculated for each alignment using the yn00 tool from the PAML(44) package and the median of all Nei-Gojobori dN/dS values were logged to the model metadata. Metadata for functional classification are taken from model descriptions and associated main categories in TIGRFAM Roles. Additional statistics such as global ubiquity and how often a gene appears as a single copy are also recorded using the matrix of gene counts in the reference.

Core gene filtering and exploration mode

To reduce false positives while retaining the ability to search for more potential targets, a second search mode is provided as an option. The default search filters core genes for unlikely targets whereas "exploration mode" omits this step. Core genes are filtered for transport and regulatory functions by terms found in the descriptions and additional biosynthetic genes are removed if the protein sequences from the corresponding HMM seed alignments yield positive hits for high frequency BGC Pfams. High frequency biosynthetic Pfams are determined using ClusterFinder (45) Pfam frequency data where those above a frequency of 50 were used. This threshold was conservatively chosen based on the histogram of different Pfams at each frequency (Supplemental S11).

Duplication screening
Duplication is determined comparatively using the sum of median and standard deviation of corresponding sequence counts in the reference set. Genes with counts greater than this baseline are recorded along with the bit-score, scaffold location, and reference count statistics provided for manual review.

BGC proximity screening
BGC proximity is calculated by finding all core gene locations that intersect with BGC boundaries on the same scaffold. Visualizations are appended to the antiSMASH generated graphics and colored by criteria to quickly identify the type of proximity hits. Results from DUF and resistance model hits are appended to cluster annotations in a similar manner where hits for both resistance and core models are marked as “CoreRes” indicating a potential known target (a core model that is also in the known resistance set).

Phylogenetic screening
Queries of genomes, which are part of the reference phyla, will be screened for genes subject to HGT. Input sequences that do not have enough core genes, or sequences not part of the reference phyla will fail this screen or produce inaccurate results so it is advised to switch this screen off when using these inputs. The screening for HGT involves making sequence alignments to build many gene trees from which tree reconciliation with species tree is used to infer HGT for each gene. Alignments are created by adding the extracted nucleotide core sequences to the pre-trimmed reference codon alignments using the add method in MAFFT (mafft --add). Appended alignments are stored for user export and trimmed copies are made using the automated1 method in TrimAl. Trimmed alignments and existing reference trees are used with the Evolutionary Placement Algorithm (46)(EPA) option available in RAxML to produce gene trees with the query sequences added. The species tree is then inferred from a coalescent of multi-locus gene trees using ASTRAL(47). The set of gene trees used are all single copy genes present in every reference and query organism with dN/dS values < 1; 16S rRNA sequences are also included if present. Each gene tree is then compared to the species tree to delineate incongruences due to duplications, transfers, and loss; this is determined using the ranger-dtl-U(48) tool with default cost values. All transfers involving the query organism are then parsed and sorted and an additional filtering of intra-genus transfers is applied if genus names match with reference.

References

1. Steffensky,M., Mühlenweg,A., Wang,Z.X., Li,S.M. and Heide,L. (2000) Identification of the novobiocin biosynthetic gene cluster of Streptomyces spheroides NCIB 11891. *Antimicrob. Agents Chemother.*, **44**, 1214–1222.
2. Schmutz,E., Mühlenweg,A., Li,S.M. and Heide,L. (2003) Resistance genes of aminocoumarin producers: Two type II topoisomerase genes confer resistance against coumermycin A1 and clorobiocin. *Antimicrob. Agents Chemother.*, **47**, 869–877.
3. Hashimi,S.M., Huang,G., Maxwell,A. and Birch,R.G. (2008) DNA gyrase from the albicidin producer Xanthomonas albilineans has multiple-antibiotic-resistance and unusual enzymatic properties. *Antimicrob. Agents Chemother.*, **52**, 1382–1390.
4. Sánchez-Hidalgo,M., Núñez,L.E., Méndez,C. and Salas,J.A. (2010) Involvement of the beta subunit of RNA polymerase in resistance to streptolydigin and streptovaricin in the producer organisms Streptomyces lydicus and streptomyces spectabilis. *Antimicrob. Agents Chemother.*, **54**, 1684–1692.
5. Floss,H.G. and Yu,T.-W. (2005) RifamycinMode of Action, Resistance, and Biosynthesis. *Chem. Rev.*, **105**, 621–632.

6. Ishikawa,J., Yamashita,A., Mikami,Y., Hoshino,Y., Kurita,H., Hotta,K., Shiba,T. and Hattori,M. (2004) The complete genomic sequence of Nocardia farcinica IFM 10152. *Proc. Natl. Acad. Sci. U. S. A.*, **101**, 14925–30.
7. Wieland Brown,L.C., Acker,M.G., Clardy,J., Walsh,C.T. and Fischbach,M. a (2009) Thirteen posttranslational modifications convert a 14-residue peptide into the antibiotic thiocillin. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 2549–2553.
8. Bibb,M.J., White,J., Ward,J.M. and Janssen,G.R. (1994) The mRNA for the 23S rRNA methylase encoded by the ermE gene of Saccharopolyspora erythraea is translated in the absence of a conventional ribosome-binding site. *Mol. Microbiol.*, **14**, 533–545.
9. Ryder,M.H., Slota,J.E., Scarim,A. and Farrand,S.K. (1987) Genetic analysis of agrocin 84 production and immunity in Agrobacterium spp. *J. Bacteriol.*, **169**, 4184–4189.
10. Tang,X., Li,J., Millán-Aguiñaga,N., Zhang,J.J., O'Neill,E.C., Ugalde,J.A., Jensen,P.R., Mantovani,S.M. and Moore,B.S. (2015) Identification of Thiotetronic Acid Antibiotic Biosynthetic Pathways by Target-directed Genome Mining. *ACS Chem. Biol.*, 10.1021/acschembio.5b00658.
11. Kale,A.J., McGlinchey,R.P., Lechner,A. and Moore,B.S. (2011) Bacterial self-resistance to the natural proteasome inhibitor salinosporamide A. *ACS Chem. Biol.*, **6**, 1257–1264.
12. Marshall,C.G., Lessard,I.A.D., Park,I.S. and Wright,G.D. (1998) Glycopeptide antibiotic resistance genes in glycopeptide-producing organisms. *Antimicrob. Agents Chemother.*, **42**, 2215–2220.
13. Liras,P. (1999) Biosynthesis and molecular genetics of cephamycins. Cephamycins produced by actinomycetes. *Antonie Van Leeuwenhoek*, **75**, 109–124.
14. Sosio,M., Amati,G., Cappellano,C., Sarubbi,E., Monti,F. and Donadio,S. (1996) An elongation factor Tu (EF-Tu) resistant to the EF-Tu inhibitor GE2270 in the producing organism Planobispora rosea. *Mol. Microbiol.*, **22**, 43–51.
15. Kling,A., Lukat,P., Almeida,D. V, Bauer,A., Fontaine,E., Sordello,S., Zaburannyi,N., Herrmann,J., Wenzel,S.C., Konig,C., *et al.* (2015) Antibiotics. Targeting DnaN for tuberculosis therapy using novel griselimycins. *Science*, **348**, 1106–1112.
16. Wang,Z.-X., Li,S.-M. and Heide,L. (2000) Identification of the Coumermycin A1Biosynthetic Gene Cluster of Streptomyces rishiriensisDSM 40489. *Antimicrob. Agents Chemother.*, **44**, 3040–3048.
17. Baumann,S., Herrmann,J., Raju,R., Steinmetz,H., Mohr,K.I., Huttel,S., Harmrolfs,K., Stadler,M. and Muller,R. (2014) Cystobactamids: myxobacterial topoisomerase inhibitors exhibiting potent antibacterial activity. *Angew. Chem. Int. Ed. Engl.*, **53**, 14605–14609.
18. Xu,H., Huang,W., He,Q.-L., Zhao,Z.-X., Zhang,F., Wang,R., Kang,J. and Tang,G.-L. (2012) Self-resistance to an antitumor antibiotic: a DNA glycosylase triggers the base-excision repair system in yatakemycin biosynthesis. *Angew. Chem. Int. Ed. Engl.*, **51**, 10532–10536.
19. Wang,S., Liu,K., Xiao,L., Yang,L., Li,H., Zhang,F., Lei,L., Li,S., Feng,X., Li,A., *et al.* (2016) Characterization of a novel DNA glycosylase from *S. sahachiroi* involved in the reduction and repair of azinomycin B induced DNA damage. *Nucleic Acids Res.*, **44**, 187–197.
20. Kim,C.-G., Lamichhane,J., Song,K.-I., Nguyen,V.D., Kim,D.-H., Jeong,T.-S., Kang,S.-H., Kim,K.-W., Maharjan,J., Hong,Y.-S., *et al.* (2008) Biosynthesis of rubradirin as an ansamycin antibiotic from Streptomyces achromogenes var. rubradiris NRRL3061. *Arch. Microbiol.*, **189**, 463–473.
21. Almutairi,M.M., Park,S.R., Rose,S., Hansen,D.A., Vázquez-Laslop,N., Douthwaite,S., Sherman,D.H. and Mankin,A.S. (2015) Resistance to ketolide antibiotics by coordinated expression of rRNA methyltransferases in a bacterial producer of natural ketolides. *Proc. Natl. Acad. Sci.*, 10.1073/pnas.1512090112.
22. Mosbacher,T.G., Bechthold,A. and Schulz,G.E. (2005) Structure and function of the antibiotic resistance-mediating methyltransferase AviRb from Streptomyces viridochromogenes. *J. Mol. Biol.*, **345**, 535–545.
23. Thomas,C.M., Hothersall,J., Willis,C.L. and Simpson,T.J. (2010) Resistance to and synthesis of the antibiotic mupirocin. *Nat. Rev. Microbiol.*, **8**, 281–289.
24. Olano,C., Wilkinson,B., Sanchez,C., Moss,S.J., Sheridan,R., Math,V., Weston,A.J., Brana,A.F., Martin,C.J., Oliynyk,M., *et al.* (2004) Biosynthesis of the angiogenesis inhibitor borrelidin by Streptomyces parvulus Tu4055: cluster analysis and assignment of functions.

Chem. Biol., **11**, 87–97.
25. Vecchione,J.J. and Sello,J.K. (2009) A novel tryptophanyl-tRNA synthetase gene confers high-level resistance to indolmycin. *Antimicrob. Agents Chemother.*, **53**, 3972–3980.
26. Schorn,M., Zettler,J., Noel,J.P., Dorrestein,P.C., Moore,B.S. and Kaysser,L. (2014) Genetic basis for the biosynthesis of the pharmaceutically important class of epoxyketone proteasome inhibitors. *ACS Chem. Biol.*, **9**, 301–309.
27. Peterson,R.M., Huang,T., Rudolf,J.D., Smanski,M.J. and Shen,B. (2014) Mechanisms of self-resistance in the Platensimycin- and platencin-producing streptomyces platensis MA7327 and MA7339 strains. *Chem. Biol.*, **21**, 389–397.
28. Tang,X., Li,J., Millán-Aguiñaga,N., Zhang,J.J., O'Neill,E.C., Ugalde,J.A., Jensen,P.R., Mantovani,S.M. and Moore,B.S. (2015) Identification of Thiotetronic Acid Antibiotic Biosynthetic Pathways by Target-directed Genome Mining. *ACS Chem. Biol.*, **10**, 2841–2849.
29. Mattheus,W., Masschelein,J., Gao,L.-J., Herdewijn,P., Landuyt,B., Volckaert,G. and Lavigne,R. (2010) The kalimantacin/batumin biosynthesis operon encodes a self-resistance isoform of the FabI bacterial target. *Chem. Biol.*, **17**, 1067–1071.
30. Liu,X., Fortin,P.D. and Walsh,C.T. (2008) Andrimid producers encode an acetyl-CoA carboxyltransferase subunit resistant to the action of the antibiotic. *Proc. Natl Acad. Sci. USA*, **105**, 13321–13326.
31. Cock,P.J.A., Antao,T., Chang,J.T., Chapman,B.A., Cox,C.J., Dalke,A., Friedberg,I., Hamelryck,T., Kauff,F., Wilczynski,B., *et al.* (2009) Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, **25**, 1422–1423.
32. Finn,R.D., Clements,J. and Eddy,S.R. (2011) HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Res.*, **39**, 29–37.
33. Gibson,M.K., Forsberg,K.J. and Dantas,G. (2014) Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *ISME J.*, **9**, 1–10.
34. McArthur,A.G., Waglechner,N., Nizam,F., Yan,A., Azad,M.A., Baylay,A.J., Bhullar,K., Canova,M.J., De Pascale,G., Ejim,L., *et al.* (2013) The comprehensive antibiotic resistance database. *Antimicrob. Agents Chemother.*, **57**, 3348–3357.
35. Thai,Q.K., Bös,F. and Pleiss,J. (2009) The Lactamase Engineering Database: a critical survey of TEM sequences in public databases. *BMC Genomics*, **10**, 390.
36. Bush,K. and Jacoby,G.A. (2010) Updated functional classification of ??-lactamases. *Antimicrob. Agents Chemother.*, **54**, 969–976.
37. Finn,R.D., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Mistry,J., Mitchell,A.L., Potter,S.C., Punta,M., Qureshi,M., Sangrador-Vegas,A., *et al.* (2016) The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Res.*, **44**, D279–D285.
38. O'Leary,N.A., Wright,M.W., Brister,J.R., Ciufo,S., Haddad,D., McVeigh,R., Rajput,B., Robbertse,B., Smith-White,B., Ako-Adjei,D., *et al.* (2016) Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.*, **44**, D733–D745.
39. Haft,D.H., Selengut,J.D. and White,O. (2003) The TIGRFAMs database of protein families. *Nucleic Acids Res.*, **31**, 371–373.
40. Katoh,K. and Standley,D.M. (2013) MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.*, **30**, 772–780.
41. Suyama,M., Torrents,D. and Bork,P. (2006) PAL2NAL: Robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.*, **34**, 609–612.
42. Capella-Gutiérrez,S., Silla-Martínez,J.M. and Gabaldón,T. (2009) trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, **25**, 1972–1973.
43. Stamatakis,A. (2006) RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, **22**, 2688–2690.
44. Yang,Z. (2007) PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.*, **24**, 1586–1591.
45. Cimermancic,P., Medema,M.H., Claesen,J., Kurita,K., Wieland Brown,L.C., Mavrommatis,K., Pati,A., Godfrey,P.A., Koehrsen,M., Clardy,J., *et al.* (2014) Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell*, **158**, 412–

421.
46. Berger,S.A., Krompass,D. and Stamatakis,A. (2011) Performance, accuracy, and web server for evolutionary placement of short sequence reads under maximum likelihood. *Syst. Biol.*, **60**, 291–302.
47. Mirarab,S., Reaz,R., Bayzid,M.S., Zimmermann,T., S. Swenson,M. and Warnow,T. (2014) ASTRAL: Genome-scale coalescent-based species tree estimation. *Bioinformatics*, **30**, 541–548.
48. Bansal,M.S., Alm,E.J. and Kellis,M. (2012) Efficient algorithms for the reconciliation problem with gene duplication, horizontal transfer and loss. *Bioinformatics*, **28**, 283–291.