Supplementary Table 1. Summary of the synthetic metagenomes used in this study.

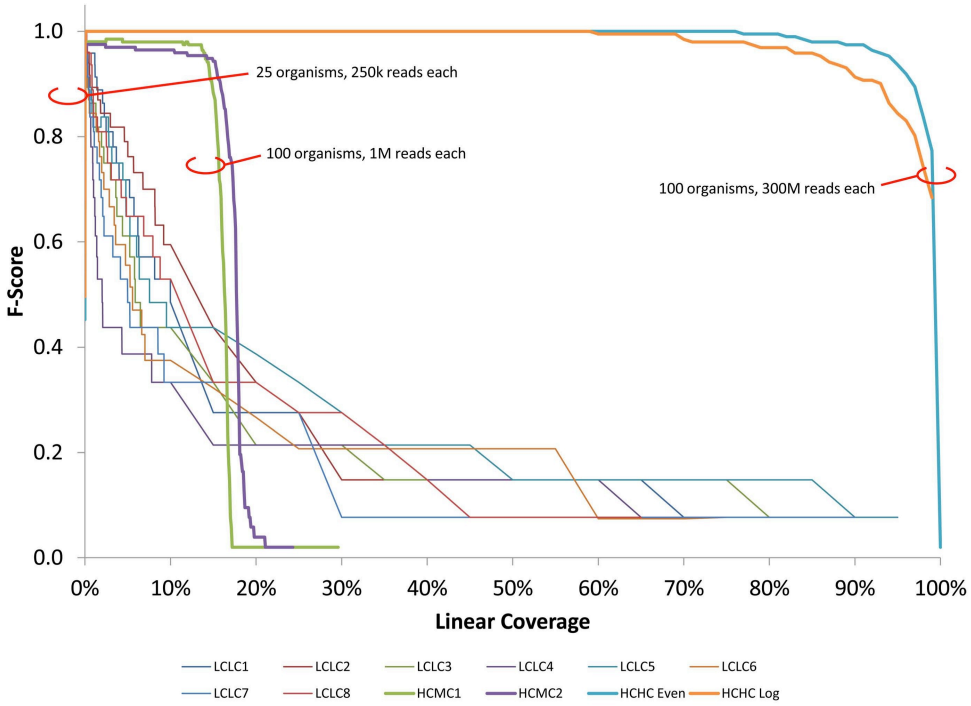
Dataset	Designation	Complexity	Coverage	# of species	Error Model Basis	Read Distribution	Total # of Reads	Mean Read Length (bp)	Total Input Bases (bp)	Accuracy
MG1	**HCHC**	High	High	100	Illumina	Even	300M	100	29,714,584,400	0.998225
MG2						Log-normal		100	29,714,381,400	0.999113
MG3	**HCHC**	High	High	251	Illumina	Even	300M	100	29,714,802,400	0.999113
MG4				250		Log-normal		100	29,714,322,000	0.999113
MG5	**HCHC**	High	High	322	Illumina	Even	300M	100	29,714,761,200	0.999113
MG6				403		Log-normal		100	29,714,454,200	0.999113
MG7	**HCMC**	High	Medium	100	Illumina	Even	1M	89	88,642,935	0.996451
MG8								89	88,628,116	0.996451
MG9	**LCLC**	Low	Low	25	Illumina	Log-normal	250k	89	22,151,627	0.998225
MG10								89	22,157,912	0.999113
MG11								89	22,159,621	0.995563
MG12								89	22,149,068	0.995563
MG13								89	22,169,311	0.995563
MG14								89	22,151,916	0.994676
MG15								89	22,159,780	0.994676
MG16								89	22,166,184	0.998225
MG17	**LCLC**	Low	Low	20	Illumina	Even	6,562,065	75	492,154,875	0.996451
MG18						Stag	7,932,819	75	594,961,425	0.997338
MG19	**LCLC**	Low	Low	20	454	Even	1,386,198	534	740,531,674	0.999113
MG20						Stag	1,225,169	533	653,060,044	0.999113

Supplementary Table 2. Community profiling tools used in this study

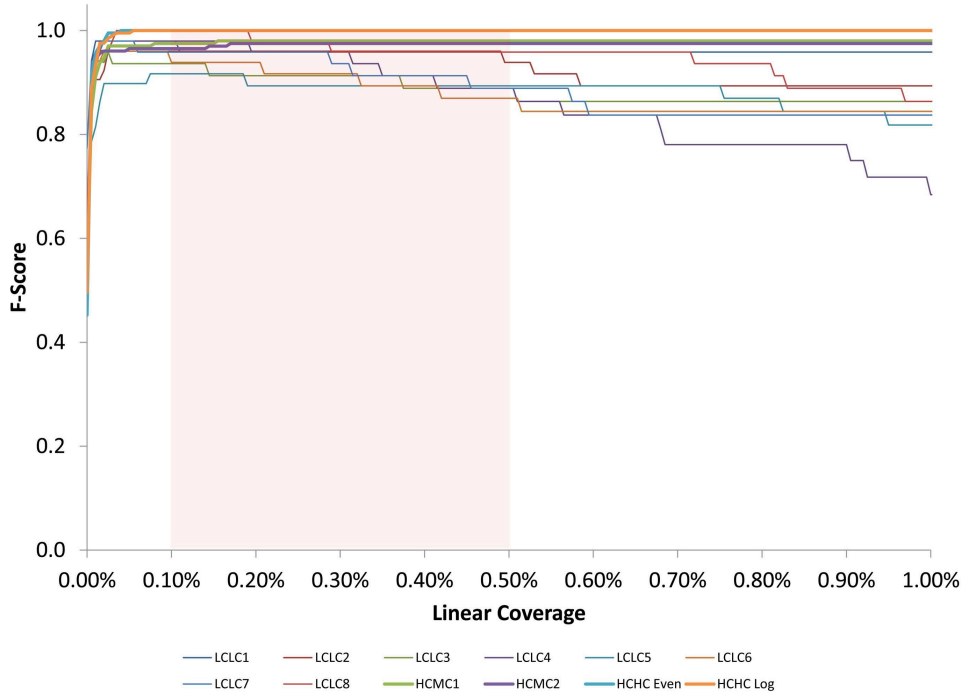
Tool	Version	Threads Used	Input Data	Input size (reads)	Options	Native Cutoff	Post-processing	Taxonomy Level Used	Capable
GOTTCHA	1	12	all	all	BWA mem	>= 10 reads, >= 100 bp, >= 0.5% signatures	none	species	strain
BLASTn	2.2.28+	12	Fecal RNA	1M read subsample	default	> 10 E-value	LCA with >= 10 reads	species	strain
			all others	all				species	strain
BWA	0.7.4-r385	12	all	all	aln/samse	<= 1 gap openning, <=2 mismatches in seed, < 4% missing alignments	>= 10 reads	species	strain
MetaPhlAn	1.7.7	12	all	all	default (bowtie2)	> 10Kbp marker genes, >1 quantile	none	species	species
mOTUs	1	12	all	all	default	< 97% identity, < 45bp reads	none	species	species
Kraken-mini	0.10.2-beta	12	all	all	default with preloaded database	keep most weighted path	none	species	strain

Supplementary Figure 1. Filtering parameter selection using F-score as a function of linear coverage across multiple synthetic metagenomes of varying coverage. An optimal value of linear coverage was selected by overlapping the F-score profiles of the two HCHC metagenomes (MG1, MG2), the two HCMC metagenomes (MG7, MG8), and the eight LCLC metagenomes (MG9-MG16), and selecting a region where the F-scores were maximal. (*A*) F-score across the full range of linear coverage: 0 – 100%. (*B*) Close-up of linear coverage in the range of 0.0 – 1.0%. The shaded region reflects a range of linear coverages (0.1 – 0.5%) that yield an optimal F-score across the datasets tested.

A

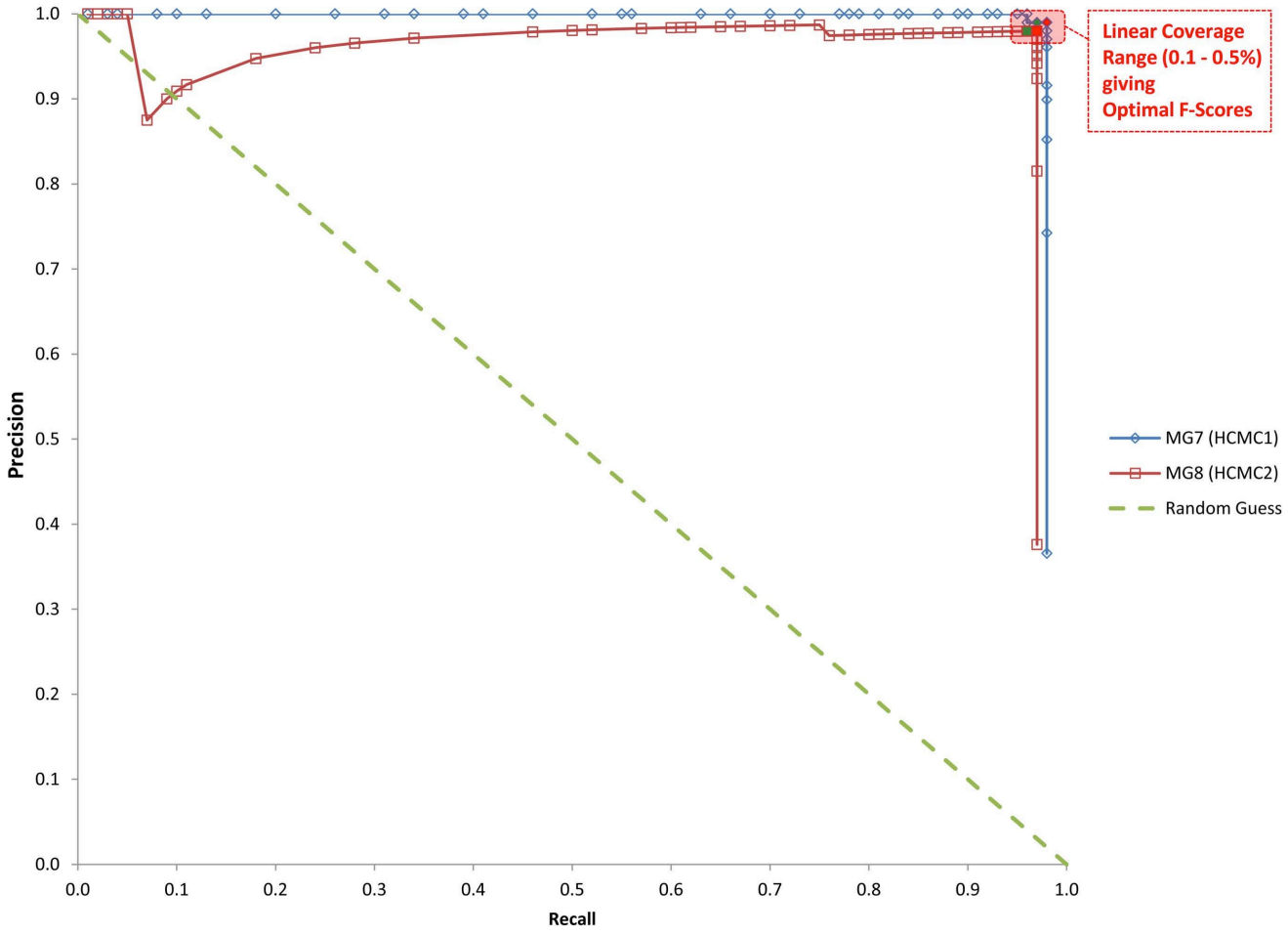


B

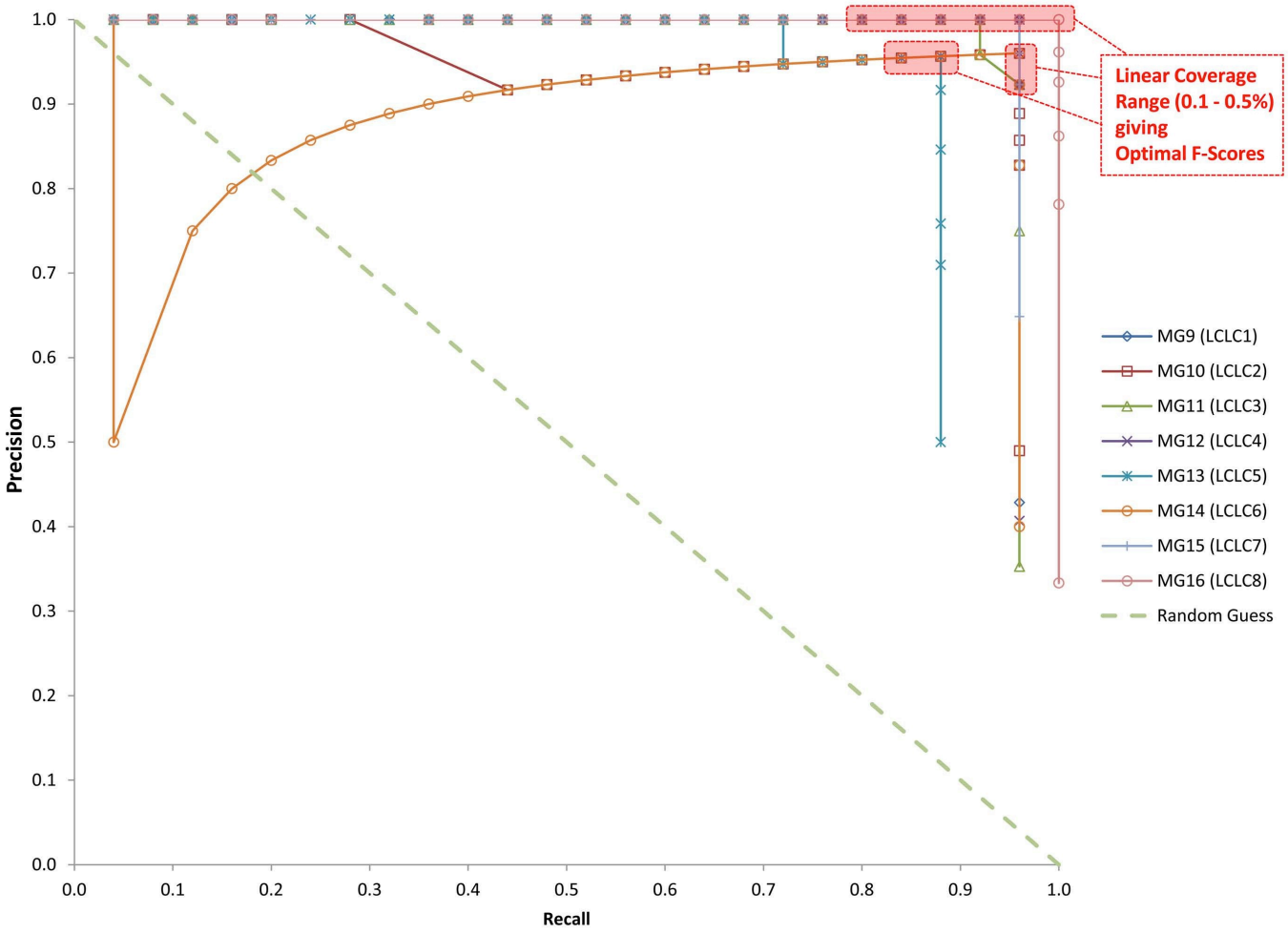


Supplementary Figure 2. GOTTCHA precision-recall curves for two high-complexity (MG7-MG8) and eight low-complexity (MG9-MG16) synthetic metagenomes and their associated F-scores. Precision and recall values were determined and plotted for each synthetic metagenome (solid lines) throughout the entire range of linear coverages (0 – 100%). For comparison, the precision and recall values for random guessing are shown (dashed line). The range of linear coverages yielding optimal F-scores across all HCHC, HCMC, and LCLC datasets (shaded pink regions) is shown for (*A*) the high complexity, medium coverage (HCMC: MG7-MG8) and (*B*) low complexity, low coverage (LCLC: MG9-MG16) synthetic metagenomes: 0.1% (green-filled marker) to 0.5% (red-filled marker). (*C*) Each tool's classification performance in terms of the F-score.

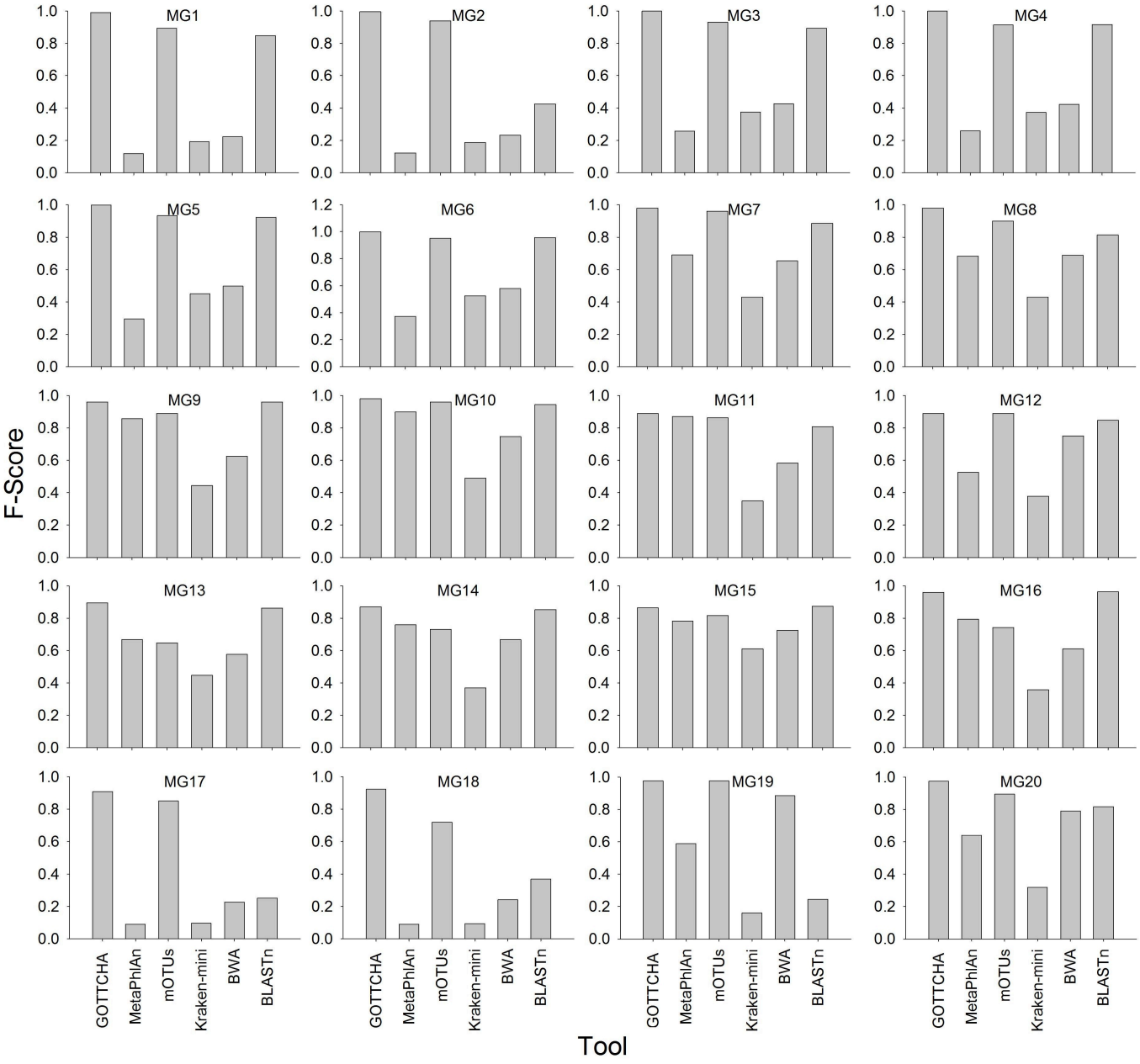
2A



2B

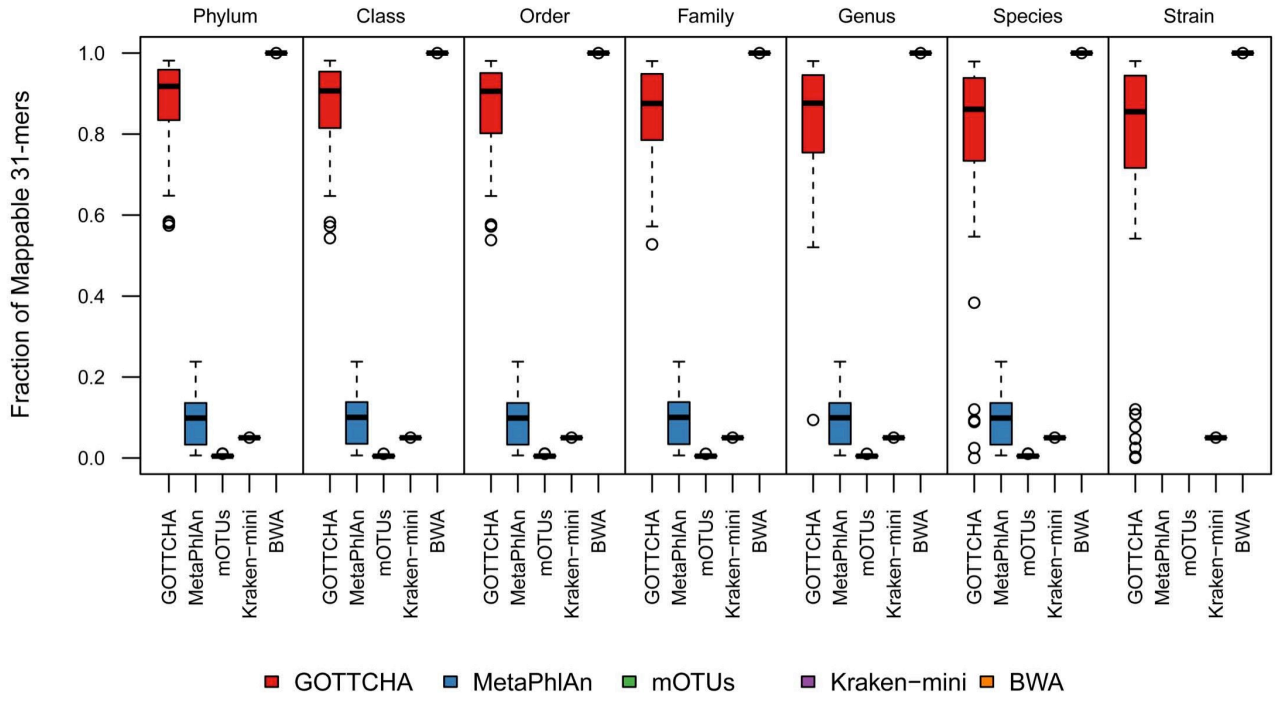

Linear Coverage Range (0.1 - 0.5%) giving Optimal F-Scores
MG9 (LCLC1)
MG10 (LCLC2)
MG11 (LCLC3)
MG12 (LCLC4)
MG13 (LCLC5)
MG14 (LCLC6)
MG15 (LCLC7)
MG16 (LCLC8)
Random Guess
Precision
Recall
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0

2C

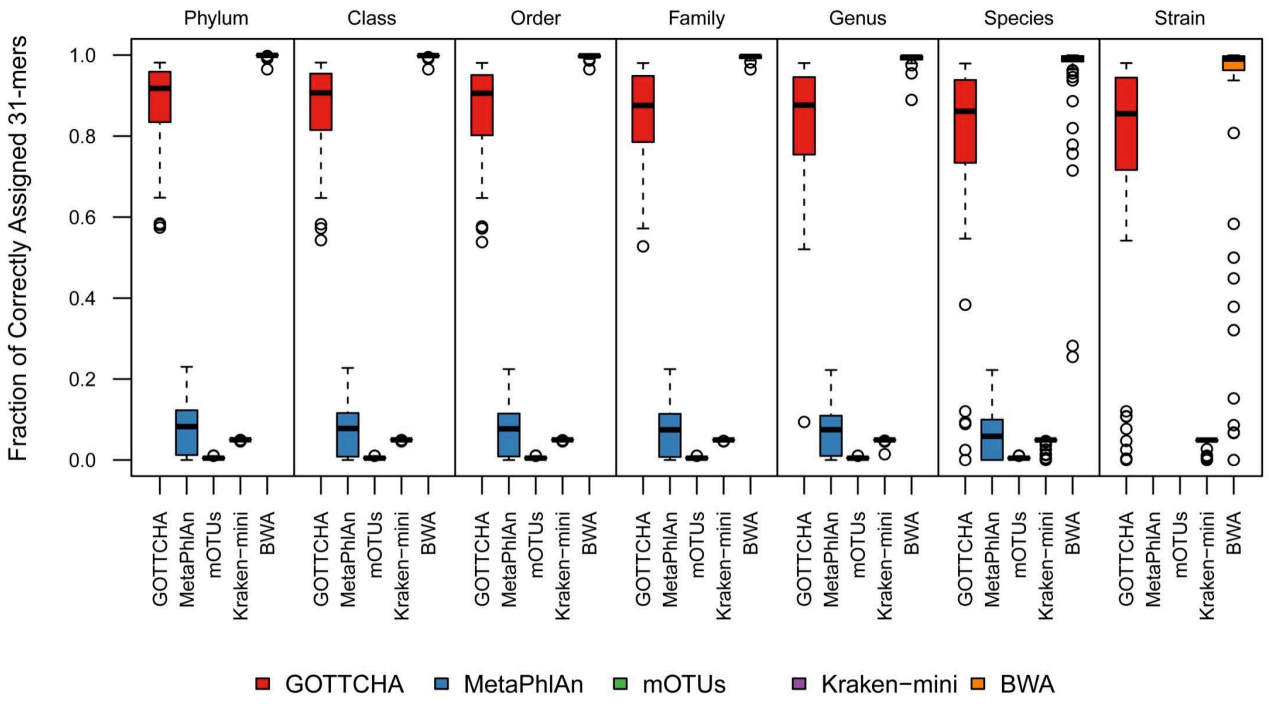

MG1
MG2
MG3
MG4
MG5
MG6
MG7
MG8
MG9
MG10
MG11
MG12
MG13
MG14
MG15
MG16
MG17
MG18
MG19
MG20
F-Score
GOTTCHA
MetaPhlAn
mOTUs
Kraken-mini
BWA
BLASTn
Tool

Supplementary Figure 3. Hierarchical capture of genomic content from 100 randomly selected genomes. (*A*) Fraction of all reads extractable from the 100 randomly selected genomes that could be mapped back to the tool-specific references, regardless of correctness, and organized by taxonomic level. Each data point represents the reported fraction of a single genome. (*B*) Fraction of the classified reads that were correctly assigned. (*C*) Fraction of the classified reads that was incorrectly assigned. The median (solid bar), interquartile range (boxes: 25%-75%), range without outliers (whiskers), and outliers (open circles) are shown. The full list of 100 selected genomes is provided in **Supplementary List 1**.

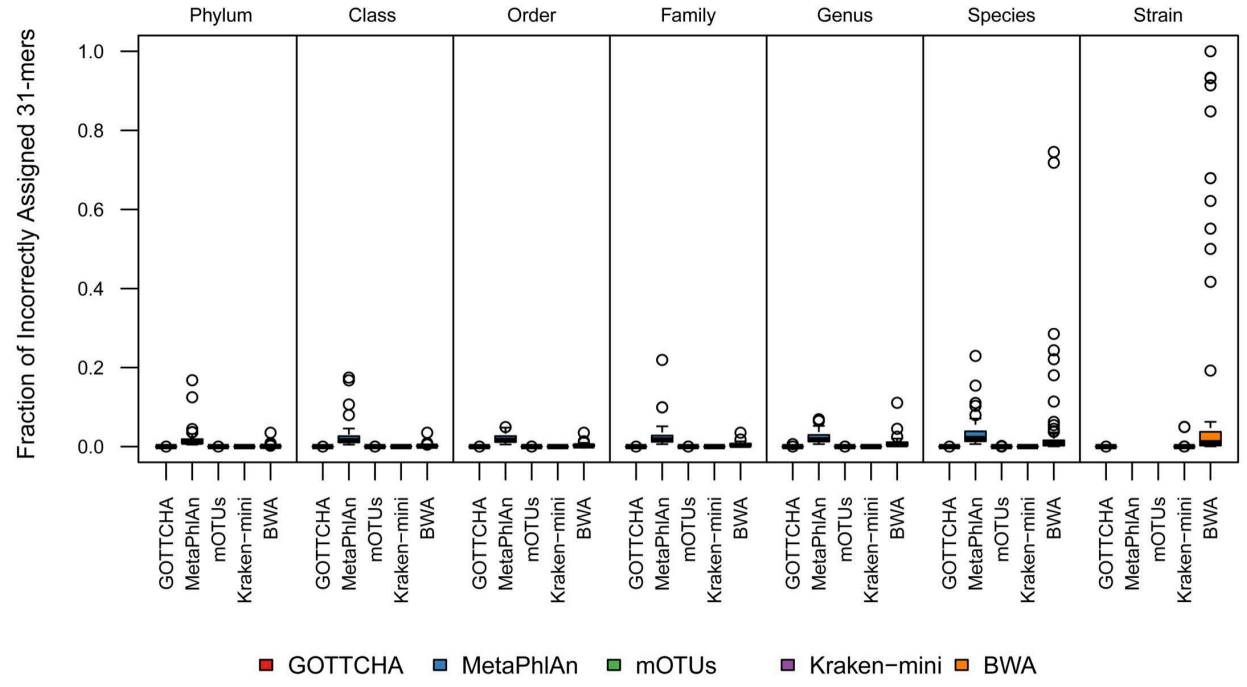
3A



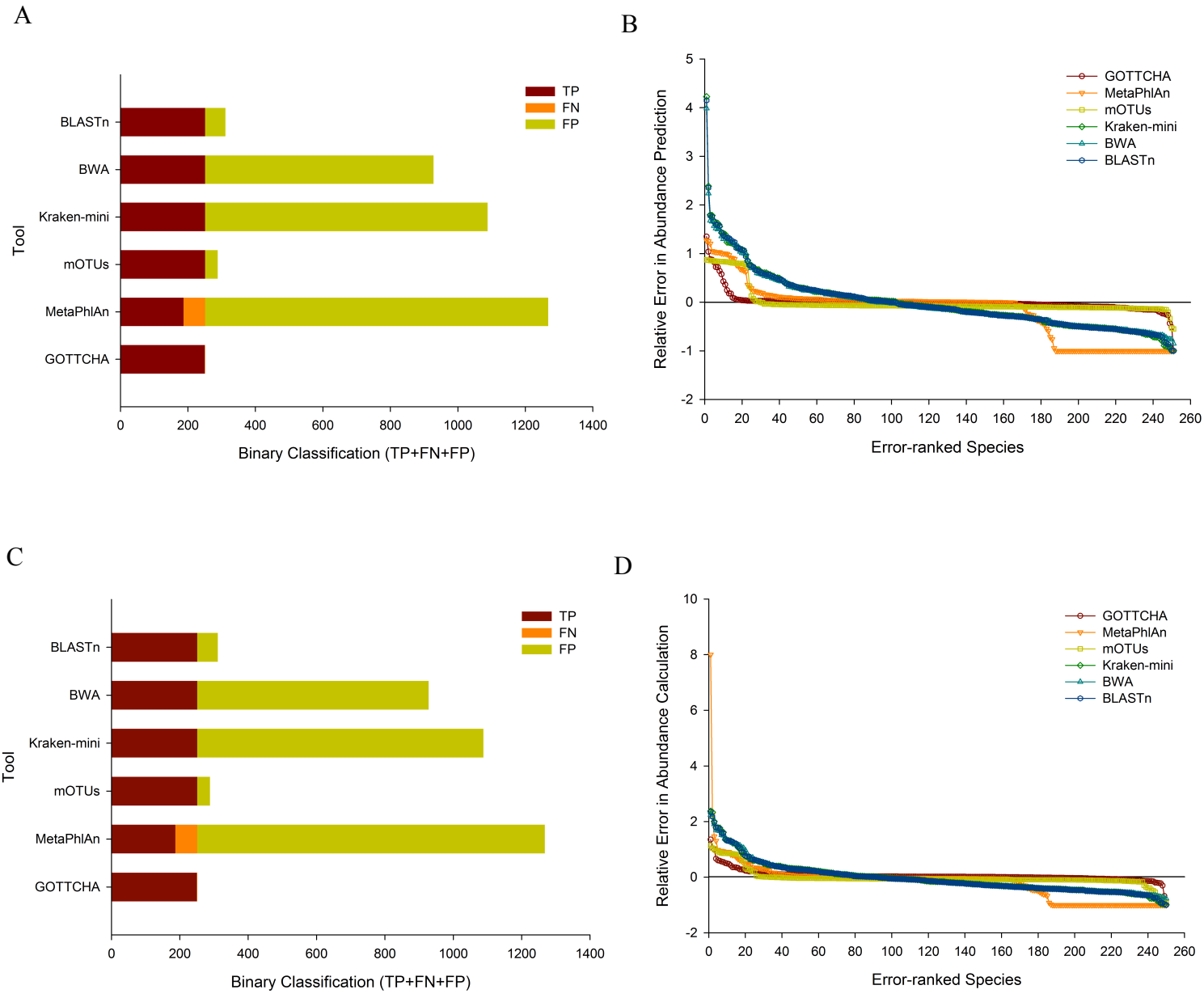
3B


Phylum
Class
Order
Family
Genus
Species
Strain
Fraction of Correctly Assigned 31-mers
1.0
0.8
0.6
0.4
0.2
0.0
GOTTCHA
MetaPhlAn
mOTUs
Kraken−mini
BWA
GOTTCHA
MetaPhlAn
mOTUs
Kraken−mini
BWA

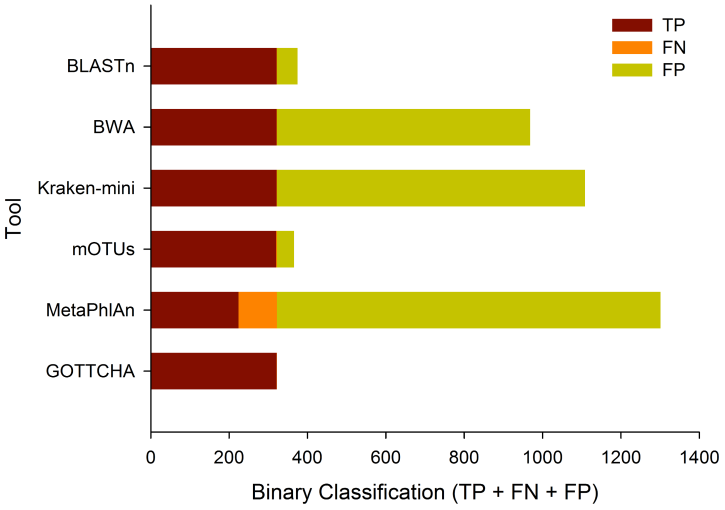
3C


Phylum
Class
Order
Family
Genus
Species
Strain
Fraction of Incorrectly Assigned 31-mers
1.0
0.8
0.6
0.4
0.2
0.0
GOTTCHA
MetaPhlAn
mOTUs
Kraken−mini
BWA
GOTTCHA
MetaPhlAn
mOTUs
Kraken−mini
BWA

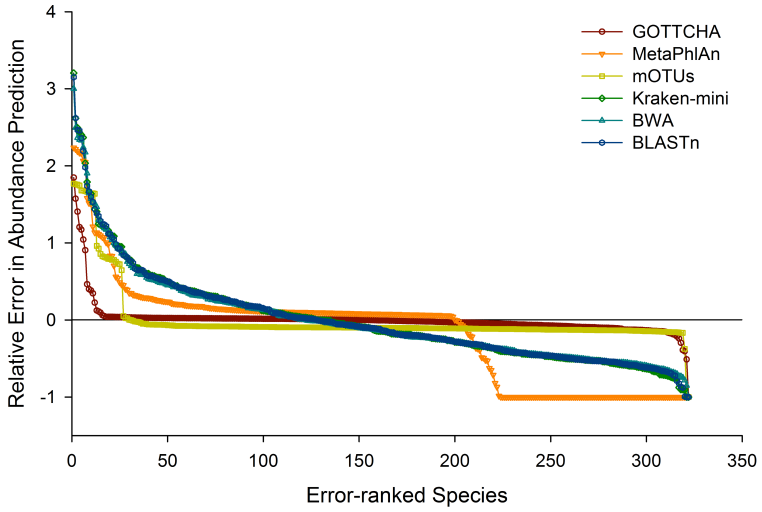
Supplementary Figure 4. Classification and relative abundance predictions for four additional high complexity, high coverage (HCHC) synthetic metagenomes. Two communities with >200 organisms: The EVEN community (*A*) classification and (*B*) relative abundances; and the LOG-NORMAL community (*C*) classification and (*D*) relative abundances. Two communities with >300 organisms: The EVEN community (*E*) classification and (*F*) relative abundances; and the LOG-NORMAL community (*G*) classification and (*H*) relative abundances.



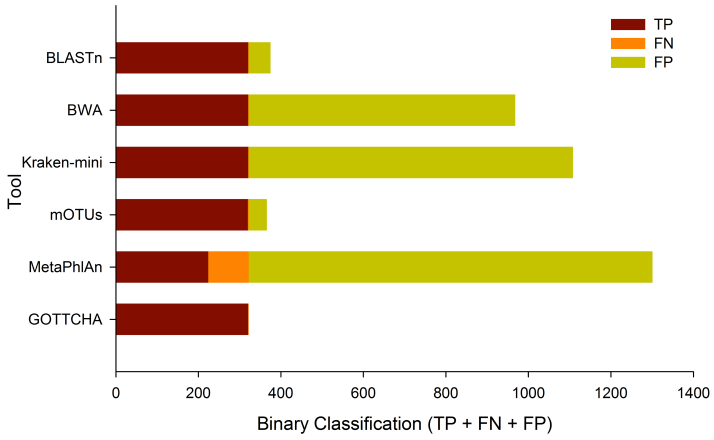
E


TP
FN
FP
BLASTn
BWA
Kraken-mini
mOTUs
MetaPhlAn
GOTTCHA
Tool
0
200
400
600
800
1000
1200
1400
Binary Classification (TP + FN + FP)

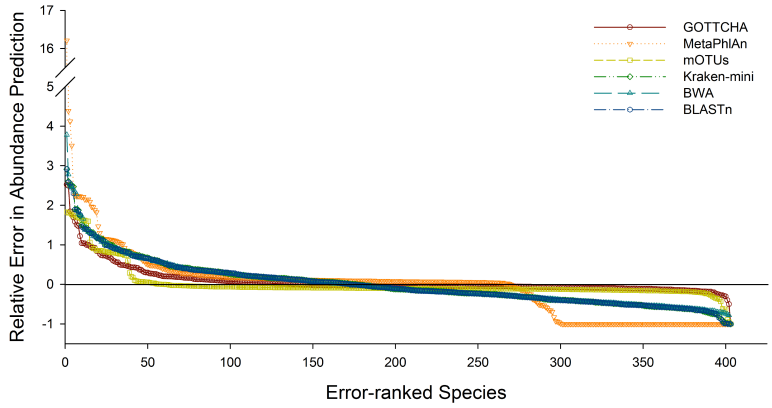
F


GOTTCHA
MetaPhlAn
mOTUs
Kraken-mini
BWA
BLASTn
Relative Error in Abundance Prediction
4
3
2
1
0
-1
0
50
100
150
200
250
300
350
Error-ranked Species

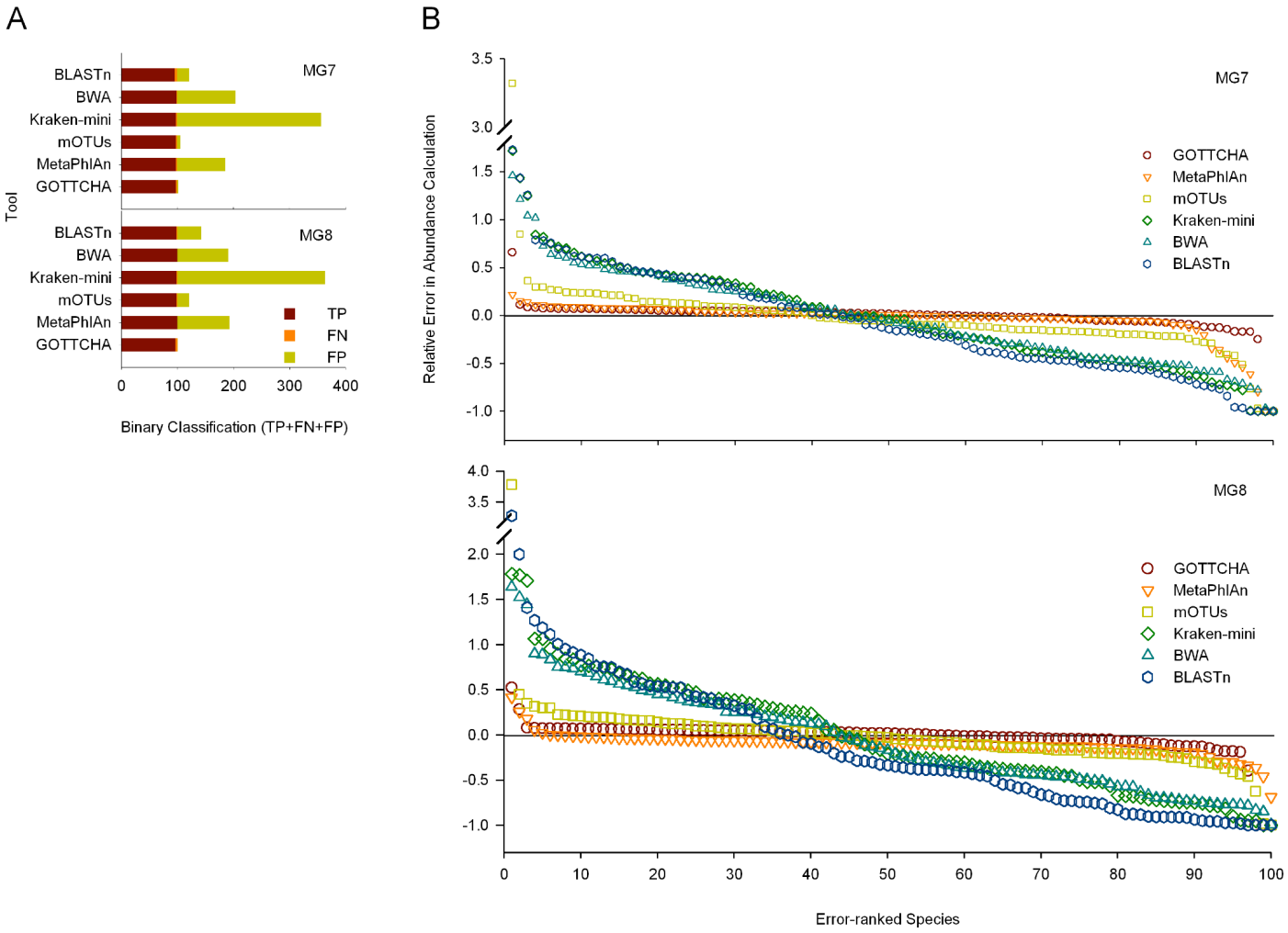
G


TP
FN
FP
BLASTn
BWA
Kraken-mini
mOTUs
MetaPhlAn
GOTTCHA
Tool
0
200
400
600
800
1000
1200
1400
Binary Classification (TP + FN + FP)

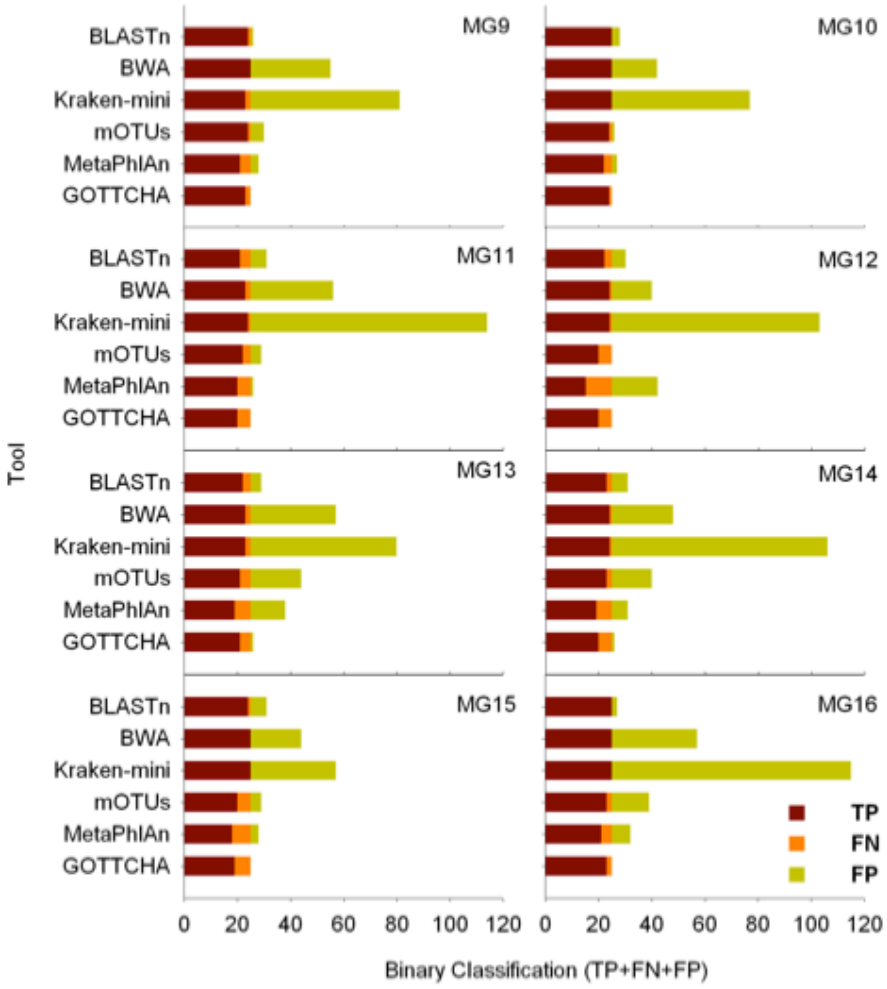
H


GOTTCHA
MetaPhlAn
mOTUs
Kraken-mini
BWA
BLASTn
Relative Error in Abundance Prediction
17
16
5
4
3
2
1
0
-1
0
50
100
150
200
250
300
350
400
Error-ranked Species

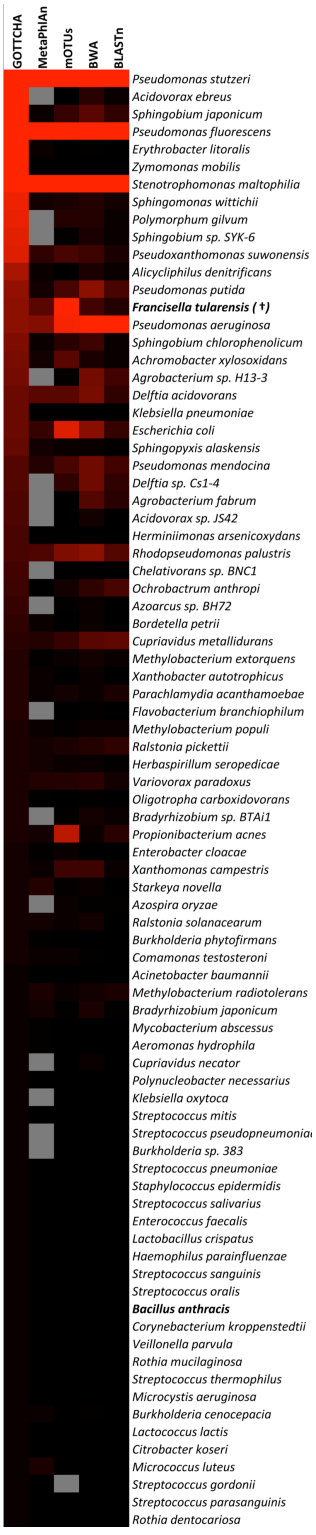
Supplementary Figure 5. Classification and relative abundance predictions for two high complexity, medium coverage (HCMC) synthetic metagenomes (MG7-MG8). Species classification (*A*) and relative abundance comparisons (*B*) of MG7 and MG8. Both metagenomes consist of 1M reads distributed evenly among 100 organisms.



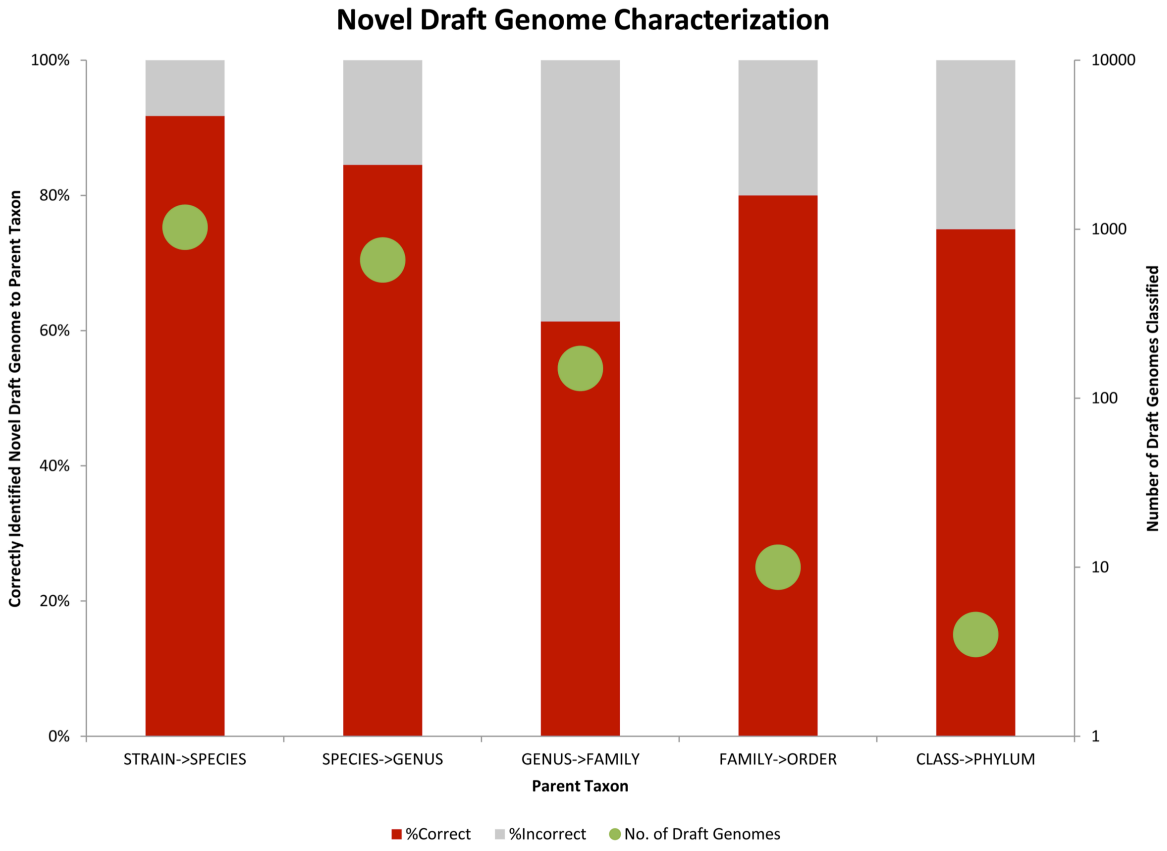
Supplementary Figure 6. Classification of eight low complexity, low coverage (LCLC) synthetic metagenomes (MG9-MG16). Species classification comparisons of MG9 – MG16. Each metagenome consists of 250k reads of mean length 89-bp log-normally distributed among 25 organisms.



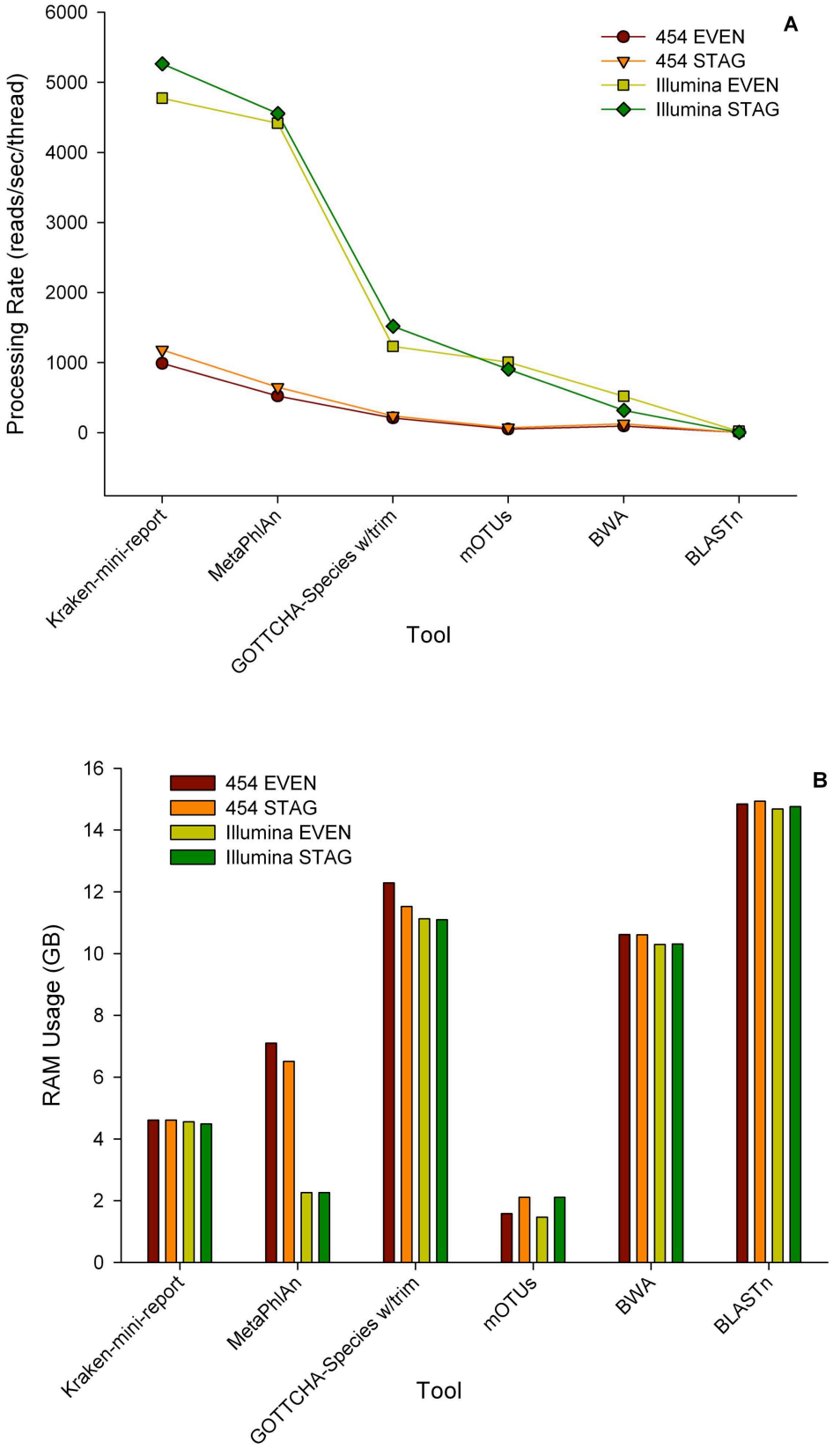
Supplementary Figure 7. Identification of the pathogen, *Francisella tularensis*, spiked in an environmental (air filter) microbiome sample. The environmental metagenome of an air filter spiked with *F. tularensis* (†) was profiled with GOTTCHA, MetaPhlAn, mOTUs, Kraken, BWA, and BLASTn. The range of organisms identified in the heat map was ordered and truncated down to the 83 bacteria identified by GOTTCHA. Relative abundances range from 3.1x10-5 (black) to 0.031 (red), while grey cells indicate absence.



Supplementary Figure 8. Classification of 1849 novel draft genomes at various taxonomic hierarchies. Draft genomes without representation at the *Strain*, *Species*, *Genus*, *Family*, or *Class* levels were assigned to their parent taxa using the GOTTCHA workflow.



Supplementary Figure 9. Community profiling computational performance of the six tools on the four HMP mock datasets. (*A*) Processing rate and (*B*) RAM usage.



Supplementary List 1. List of the 100 genomes used in **Supplementary Figure 3**

Acaryochloris_marina_MBIC11017_uid58167
Acetobacterium_woodii_DSM_1030_uid88073
Acetohalobium_arabaticum_DSM_5501_uid51423
Acinetobacter_baumannii_ATCC_17978_uid58731
Actinosynnema_mirum_DSM_43827_uid58951
Aeromonas_salmonicida_A449_uid58631
Agrobacterium_H13_3_uid63403
Alteromonas_macleodii__Deep_ecotype__uid58251
Aminobacterium_colombiense_DSM_12261_uid47083
Anoxybacillus_flavithermus_WK1_uid59135
Azorhizobium_caulinodans_ORS_571_uid58905
Bacteriovorax_marinus_SJ_uid82341
Bartonella_clarridgeiae_73_uid62131
Beijerinckia_indica_ATCC_9039_uid59057
Brevundimonas_subvibrioides_ATCC_15264_uid42117
Brucella_suis_ATCC_23445_uid59015
Burkholderia_pseudomallei_K96243_uid57733
Caulobacter_segnis_ATCC_21756_uid41709
Cellulophaga_lytica_DSM_7489_uid63401
Chitinophaga_pinensis_DSM_2588_uid59113
Collimonas_fungivorans_Ter331_uid70793
cyanobacterium_UCYN_A_uid43697
Cyclobacterium_marinum_DSM_745_uid71485
Cytophaga_hutchinsonii_ATCC_33406_uid57651
Deinococcus_maricopensis_DSM_21211_uid62225
Hydrogenobaculum_Y04AAS1_uid58857
Hyphomonas_neptunium_ATCC_15444_uid58433
Jannaschia_CCS1_uid58147
Laribacter_hongkongensis_HLHK9_uid59265
Leadbetterella_byssophila_DSM_17132_uid60161
Mahella_australiensis_50_1_BON_uid66917
Marinithermus_hydrothermalis_DSM_14884_uid65783
Methanobrevibacter_smithii_ATCC_35061_uid58827
Methanocella_paludicola_SANAE_uid42887
Methanohalophilus_mahii_DSM_5219_uid47313
Methylocella_silvestris_BL2_uid59433
Methylovorus_MP688_uid60723
Moorella_thermoacetica_ATCC_39073_uid58051
Nakamurella_multipartita_DSM_44233_uid59221
Nitrobacter_hamburgensis_X14_uid58293
Nocardioides_JS614_uid58149
Orientia_tsutsugamushi_Boryong_uid61621
Pantoea_ananatis_uid86861
Paracoccus_denitrificans_PD1222_uid58187
Pelobacter_propionicus_DSM_2379_uid58255
Polynucleobacter_necessarius_asymbioticus_QLW_P1DMWA_1_uid58611
Prochlorococcus_marinus_MIT_9515_uid58313
Pyrococcus_yayanosii_CH1_uid68281
Pyrolobus_fumarii_1A_uid73415
Rubrobacter_xylanophilus_DSM_9941_uid58057
Desulfarculus_baarsii_DSM_2075_uid51371
Desulfatibacillum_alkenivorans_AK_01_uid58913
Desulfurobacterium_thermolithotrophum_DSM_11699_uid63405
Dichelobacter_nodosus_VCS1703A_uid57643
Dickeya_dadantii_Ech586_uid42519
Elusimicrobium_minutum_Pei191_uid58949
Erythrobacter_litoralis_HTCC2594_uid58299
Escherichia_coli_O157_H7_Sakai_uid57781
Ethanoligenens_harbinense_YUAN_3_uid46255
Eubacterium_limosum_KIST612_uid59777
Ferrimonas_balearica_DSM_9799_uid53371
Fluviicola_taffensis_DSM_16823_uid65271
Gallibacterium_anatis_UMN179_uid66567
Gardnerella_vaginalis_409_05_uid43211
Geobacillus_Y412MC61_uid41171
Glaciecola_nitratireducens_FR1064_uid73759
Gordonia_bronchialis_DSM_43247_uid41403
Granulicella_mallensis_MP5ACTX8_uid49957
Halogeometricum_borinquense_DSM_11551_uid54919
Halomicrobium_mukohataei_DSM_12286_uid59107
Halopiger_xanaduensis_SH_6_uid68105
Halorhodospira_halophila_SL1_uid58473
Halorubrum_lacusprofundi_ATCC_49239_uid58807
Halothermothrix_orenii_H_168_uid58585
Halothiobacillus_neapolitanus_c2_uid41317
Segniliparus_rotundus_DSM_44985_uid49049
Simkania_negevensis_Z_uid68451
Sphingobium_chlorophenolicum_L_1_uid52597
Sphingopyxis_alaskensis_RB2256_uid58351
Staphylococcus_pseudintermedius_HKU10_03_uid62125
Starkeya_novella_DSM_506_uid48815
Sulfobacillus_acidophilus_TPY_uid68841
Sulfurospirillum_deleyianum_DSM_6946_uid41861
Synechococcus_WH_7803_uid61607
Tannerella_forsythia_ATCC_43037_uid83157
Terriglobus_saanensis_SP1PR4_uid53251
Tetragenococcus_halophilus_uid74441
Thermaerobacter_marianensis_DSM_12885_uid61727
Thermobifida_fusca_YX_uid57703
Thermodesulfobacterium_OPB45_uid68283
Thermodesulfobium_narugense_DSM_14796_uid66601
Thermodesulfovibrio_yellowstonii_DSM_11347_uid59257
Thermoproteus_tenax_Kra_1_uid74443
Thermovirga_lienii_DSM_17291_uid77129
Thermus_scotoductus_SA_01_uid62273
Thiobacillus_denitrificans_ATCC_25259_uid58189
Veillonella_parvula_DSM_2008_uid41927
Waddlia_chondrophila_WSU_86_1044_uid49531
Yersinia_enterocolitica_palearctica_105_5R_r__uid63663
Yersinia_pestis_CO92_uid57621