

Supplemental Materials

	Search filter	pHMM	Graph	Post process	Output DNA	Output Protein
Xander	N	Assembly	DBG	Y	Y	Y
MegaGTA	N	Assembly	DBG	Y	Y	Y
SAT	Y	Filter	Overlap	N	Y	N
HMMGRASPx	Y	Filter, assembly	Overlap	N	Y	Y
Genseed-HMM	Y	Filter	NA	N	Y	N
MEGAN	N	NA	Overlap	N	Y	N

Table S1. Overview of gene-targeted tools. Finder: function to filter reads that are fragments of target genes. pHMM: where pHMM is used in algorithms. Post assembly process: steps after assembly to enable diversity analyses and comparison among samples. Y - function is provided. N – function is not provided. Genseed-HMM’s assembly graph is “NA” since it depends on the third party tool that was chosen in its pipeline. MEGAN’s pHMM is “NA” since it is not pHMM-based tool.

ID	Description	Phylum
NZ_CP009257.1	<i>Acinetobacter baumannii</i> strain AB030	<i>Gammaproteobacter</i>
NZ_CP007793.1	<i>Azospirillum brasilense</i> strain Az39	<i>Alphaproteobacteria</i>
NZ_CP007794.1	<i>Azospirillum brasilense</i> strain Az39 plasmid AbAZ39_p1	<i>Alphaproteobacteria</i>
NC_012483.1	<i>Acidobacterium capsulatum</i> ATCC 51196	<i>Acidobacteria</i>
NZ_CP010415.1	<i>Azotobacter chroococcum</i> NCIMB 8003	<i>Gammaproteobacteria</i>
NZ_CP010421.1	<i>Azotobacter chroococcum</i> NCIMB 8003 plasmid pAcX50f	<i>Gammaproteobacteria</i>
NC_012560.1	<i>Azotobacter vinelandii</i> DJ	<i>Gammaproteobacteria</i>
NC_014638.1	<i>Bifidobacterium bifidum</i> PRL2010 chromosome	<i>Actinobacterium</i>
NC_004722.1	<i>Bacillus cereus</i> ATCC 14579 chromosome	<i>Firmicutes</i>
NC_008255.1	<i>Cytophaga hutchinsonii</i> ATCC 33406	<i>Bacterioidetes</i>
NC_012489.1	<i>Gemmatimonas aurantiaca</i> T-27 DNA	<i>Gemmatimonadetes</i>
NZ_CP016094.1	<i>Lacunisphaera limnophila</i> strain IG16b chromosome	<i>Verrucomicrobia</i>
NC_002516.2	<i>Pseudomonas aeruginosa</i> PAO1 chromosome	<i>Gammaproteobacteria</i>
NZ_CP014784.1	<i>Pseudomonas alcaligenes</i> strain NEB 585	<i>Gammaproteobacteria</i>

NC_016830.1	<i>Pseudomonas fluorescens</i> F113	<i>Gammaproteobacteria</i>
NC_009439.1	<i>Pseudomonas mendocina</i> ymp	<i>Gammaproteobacteria</i>
NC_002947.4	<i>Pseudomonas putida</i> KT2440 chromosome	<i>Gammaproteobacteria</i>
NZ_CP027543.1	<i>Pseudomonas stutzeri</i> strain DW2-1 chromosome	<i>Gammaproteobacteria</i>
NC_007005.1	<i>Pseudomonas syringae</i> pv. <i>syringae</i> B728a chromosome	<i>Gammaproteobacteria</i>

Table S2. Genomes used to create the synthetic data set.

Method	Contigs	Length	Non-target	<97%	97%	98%	99%	100%
Xander	11	585-876	0	3	6/8	5/5	3/3	3/3
MegaGTA	11	585-876	0	3	6/8	5/5	3/3	3/3
HMM-GRASP ^{x1}	43	66-354	0	14	6/29	6/29	6/27	6/27
HMM-GRASP ^{x2}	0	>=450	-	-	-	-	-	-
MEGAN ³	20	204-1,464	11	0	6/9	6/9	6/9	6/7
MEGAN ⁴	11	474-1,464	6	0	5/5	5/5	5/5	3/3
SAT-assembler	NF	-	-	-	-	-	-	-
Genseed-HMM ⁵	127	33-1360	71	0	6/56	6/56	6/56	6/56
Genseed-HMM ⁶	48	453-1360	46	0	2/2	2/2	2/2	2/2

Table S3. BLAST summary for *nifH* contigs assembled from the synthetic data. There were 6 *nifH* sequences present in the synthetic data. HMM-GRASP^{x1} - all contigs; HMM-GRASP^{x2} – filtered to at least 450 bp in length; MEGAN³ – all contigs; MEGAN⁴ – filtered to at least 450 bp in length. Genseed-HMM⁵: all contigs assembled; Genseed-HMM⁶: contigs were filtered to a minimum length of 450 bp.

Method	Contigs	Length	Non- target	<97%	97%	98%	99%	100%
Xander	2	456-465	0	0	2/2	2/2	2/2	2/2
MegaGTA	2	456-465	0	0	2/2	2/2	2/2	2/2
HMM-GRASPx ¹	21	78-195	0	11	4/10	4/10	4/10	4/10
HMM-GRASPx ²	0	>=450	-	-	-	-	-	-
MEGAN ³	8	423-1,497	3	0	4/5	4/5	4/5	4/5
MEGAN ⁴	7	567-1497	3	0	4/4	4/4	4/4	4/4
SAT-assembler ⁵	53	155-947	46	0	4/7	4/7	4/7	4/7
SAT-assembler ⁶	5	460-947	0	0	4/5	4/5	4/5	4/5
Genseed-HMM ⁷	29	94-1376	25	0	4/4	4/4	4/4	4/4
Genseed-HMM ⁸	18	492-1376	14	0	4/4	4/4	4/4	4/4

Table S4. BLAST summary for *nirK* contigs assembled from the synthetic data. There were 4 *nirK* sequences in the synthetic data. HMM-GRASPx¹ - all contigs; HMM-GRASPx² - filtered to at least 450 bp in length; MEGAN³ - all contigs; MEGAN⁴ - filtered to at least 450 bp in length; SAT-Assembler⁵ - all contigs; SAT-Assembler⁶ - contigs were de-replicated, and filtered to at least 450 bp in length. Genseed-HMM⁷: all contigs assembled; Genseed-HMM⁸: contigs were de-replicated and filtered to a minimum length of 450 bp.

Method	Contigs	Length	Non-target	<97%	97%	98%	99%	100%
Xander	27	537-879	0	5	16/22	16/22	16/22	14/15
MegaGTA	28	537-879	0	5	16/23	16/23	16/23	15/16
MEGAN ¹	68	204-3702	45	1	17/22	17/21	17/21	16/20
MEGAN ²	39	450-3702	25	1	12/13	12/12	12/12	11/11
SAT-assembler	NF							
Genseed-HMM ³	439	31-1614	290	6/13	18/136	18/132	18/100	18/100
Genseed-HMM ⁴	198	452-1614	184	2/2	11/12	11/12	10/11	10/11

Table S5. BLAST summary for *nifH* contigs assembled from the mock data. MEGAN¹: all contigs assembled. MEGAN²: contigs filtered to a minimum length of 450 bp. NF – not found. Mock data by HMM-GRASP_x failed to complete. Genseed-HMM³: all contigs assembled; Genseed-HMM⁴: contigs were filtered to a minimum length of 450 bp. There were 18 *nifH* sequences in the strains included in the mock data.

Method	Contigs	Length	Non-target	<97%	97%	98%	99%	100%
Xander	2	621-702	0	0	2/2	2/2	2/2	2/2
MegaGTA	3	507-702	0	0	2/3	2/3	2/3	2/2
MEGAN ¹	30	210-2238	26	0	4/4	4/4	4/4	4/4
MEGAN ²	13	480-2,238	9	0	4/4	4/4	4/4	4/4
SAT-assemblber ³	256	60-533	156	0	5/100	5/100	5/98	5/88
SAT-assembler ⁴	2	533-533	0	0	1/2	1/2	1/2	0
Genseed-HMM ⁵	43	61-1343	35	0	5/8	5/8	5/8	5/8
Genseed-HMM ⁶	27	509-1343	21	0	5/6	5/6	5/6	4/5

Table S6. BLAST summary for *nirK* contigs assembled from the mock data. MEGAN¹: all contigs assembled. MEGAN²: contigs filtered to a minimum length of 450 bp. SAT-Assembler³: all contigs assembled. SAT-Assembler⁴: contigs were de-replicated, duplicates removed, and filtered to a minimum length of 450 bp. There were 5 *nirK* sequences in the strains included in the mock data.

Assembler	Contigs	Chimera	% Chimera	Same Genus
Xander	28	12	42.9	11
MegaGTA	28	12	42.9	11
SAT-assembler	176	79	44.9	57
SAT-assembler-450	106	74	69.8	54
MEGAN	55	0	0.0	NA
MEGAN-450	20	0	0.0	NA
Genseed-HMM	97	0	0.0	NA
Genseed-HMM-450	9	0	0.0	NA

45 Table S7. UCHIME summary for *rplB* contigs assembled from the synthetic data.

46

Assembler	Contigs	Chimera	% Chimera	Same Genus
Xander	95	5	5.3	5
MegaGTA	94	5	5.3	5
SAT-assembler	2,765	104	3.8	60
SAT-450	61	18	29.5	13
MEGAN	93	0	0	NA
MEGAN-450	50	0	0	NA
Genseed-HMM	408	1	0.2	1
Genseed-HMM-450	44	0	0	NA

47 Table S8. UCHIME summary for *rplB* contigs assembled from the mock data.

Assembler	Contigs	Chimera	% Chimera	Same Genus
Xander	11	5	45.5	5
MegaGTA	11	5	45.5	5
MEGAN	20	2	10.0	2
MEGAN-450	11	2	18.2	2
Genseed- HMM	127	0	0	NA
Genseed- HMM -450	48	0	0	NA

49 Table S9. UCHIME summary for *nifH* contigs assembled from the synthetic data.