This file contains the following:

- 1. Supplementary Figures, page 2-11
- 2. Supplementary Tables, page 12-21
- 3. Supplementary Methods & Materials, page 22-29
 - 4. Supplementary Notes, Page 30-36
 - 5. Supplementary Files 1-4 Page 37-42



Supplementary Figures

Figure S1. Relationship between FastANI and ProbMASH-ANI (a) and FastANI and MASH-ANI (b) for closely related genomes. Spearman and Pearson correlations were significant at p<0.001 in both (a) and (b). ProbMASH-ANI is transformed from ProbMinHash distance according to the MASH distance equation (See Materials and Methods, ProbMinHash section). Mean absolute error (MAE) were 0.023 and 0.021 respectively for (a) and (b) for data > 80% ANI, slightly larger than FastANI (MAE=0.014) for the same comparisons; see Figure S13.



Figure S2. Relationship between AAI (entire proteome) and ProbMASH distance for (a) all AAI values and (b) AAI values between 0.52 and 0.95 among 2000 genomes randomly selected from the GTDB v207 database. Both Spearman and Pearson correlation significance test showed p<0.01.



Figure S3. Relationship between AAI (universal gene set, or uAAI) and probMASH distances for AAI values between 0.5 and 0.9 among 2000 genomes randomly selected from the GTDB v207 database. Both Spearman and Pearson correlations were significant at p<0.01.



Figure S4. Performance of GSearch search step. (a) Search time (blue) and speed (orange) for an increasing number of query genomes against a fixed database of 67503 genomes and, (b) search time (blue) and speed (orange) for an increasing number of database genomes and a fixed number of 1059 query genomes. Search speed is defined as the number of database genomes searched, on average, per query genome per second. The orange line with squared data points represents Mash, whose speed was constant for any database size for 1059 query genomes, for comparison.



Figure S5. GSearch request speed comparison with BinDash software for (a) all bacterial genomes (~318K, 2TB) and (b) all viral genomes (~3 million) as database. A total of 8,466 and 10,000 query genomes were used, respectively, for this analysis. Note that GSearch and BinDash are plotted on different axes, primary y and secondary y axis, respectively. A total of 24 threads were used. Sketch size m=10^5 was used to achieve similar accuracy with ProbMinHash.



Figure S6. Database build time for the IMGVR phage species database (935,122) at the amino acid level (a), and total search (request) time for 10,000 query phages (b) against this database. Database size was about 15.8 GB, and thus loading the database takes a substantial fraction of the total searching time.



Figure S7. Rooted Mean Squared Error (RMSE) of SetSektch LSH algorithm for estimation of Jaccard index with respect to true Jaccard index for different values of parameter b. Note that changing b requires changing m in SetSektch. Three b values were used, 0.001, 0.0005, 0.0001 corresponding to m=4096, 6144 and 8192 respectively. The red dashed line indicates J=0.015 with RMSE 0.81%, corresponding to ANI 77.99%, with variation of ANI from 77.94% to 78.04% for the case that b=0.0005 and m=6144.



Figure S8. Parallel and concurrent computing model of GSearch tohnsw (a) and request or add (b) modules. Standard dispatcher and worker model was used with Reader and Collector to handle file input, collecting results from sketching workers, and initializing HNSW graph building/searching workers. Tohnsw/add module will dump sketching results and graph to disk while searching module will load prebuilt sketching and graph files and then dump answers for the query genomes to disk. Note that the number of Workers for HNSW building/searching/adding is to use all available computing cores while the number of Workers for sketching can be controlled to balance memory consumption and sketch speed. The concurrent model was based on Rayon library, and communication among threads was based on Crossbeam library.



Figure S9. Parallel efficiency and memory consumption of GSearch tohnsw and request modules (ProbMinHash) at the nucleotide level. (a) tohnsw build module real-time CPU usage (left axis) and memory consumption (right axis) for the GTDB v207 (65,703 genomes) database using 24-thread node. (b) The same with (a) but on a 64-thread node. (c) Real-time CPU usage and memory consumption for searching 8,466 genomes against the GTDB prebuilt database in (a) on a 24-thread node. (d) Real-time CPU usage and memory consumption for searching 8,466 genomes against entire NCBI/Ref_Seq genomes (~318K). The two phases in all figures are sketching and graph building/searching, respectively.



Figure S10. Parallel efficiency and memory consumption of GSearch request modules (SetSketch) at nucleotide level. (a) 8,466 query genomes against the GTDB v207 database (65,703 genomes) using 24 threads. (b) 8,466 query genomes against NCBI's RefSeq database (~318K genomes) using 64 threads.



Figure S11. Correlation between ANI estimates from our implementation of Densified MinHash (optimal densification) and Mash bottom-m implementation. A sketch size m=12,000 was used for improved accuracy in densified MinHash while 15,000 was

used in Mash. Our implementation, as also reported in the BinDash paper, is at least 10 times faster than Mash.



Figure S12. Root Mean Squared Error (RMSE) of Densified MinHash and Classic MinHash theoretical RMSE ($RMSE = \frac{\sqrt{J^*(1-J)}}{\sqrt{m}}$) for Jaccard similarity, where m is the sketch size. We use set size (number of elements in the set) n=300,000 for all cases, m is 10^5 for classic MinHash.



Figure S13. FastANI (y-axis) versus orthoANI (x-axis). Results are based on a testing dataset of 3000 genomes randomly subsampled from the large collection pf genome used in Figure S1. MAE is 1.4%.

Supplementary Tables

Table S1. Comparisons of MinHash-like and HyperLogLog algorithms. In bold face are results obtained as part of this study (via gsearch --algo prob/super/hll option); remaining results shown are from the literature. J_p is closer to real Jaccard index than J_w despite both algorithms considering k-mer weights.

	Weighted element	Set Size bias	Speed (More is better)	Space (Less is better)	Mergeability ¹⁰	Approximated Jaccard-like index
Classic MinHash ¹	Х	\checkmark	****	***	\checkmark	J
Densifed MinHash (with OD ^{2,11})	Х	\checkmark	*****	☆☆☆	Х	J
SuperMinHash ³	Х	\checkmark	***	***	\checkmark	J
ProbMinHash (default) ⁴	\checkmark	Х	*****	***	\checkmark	Jp
BagMinhash ⁵	\checkmark	\checkmark	**	***	\checkmark	J_{w}
DartMinHash ⁶	\checkmark	\checkmark	****	***	\checkmark	J_{w}
HyperLogLog ⁷	Х	Х	***	**	\checkmark	J
HyperLogLog ⁸	Х	Х	**	*	\checkmark	J
SetSketch ⁹	X (√) ¹²	X	***	**	\checkmark	J
UltraLogLog	Х	Х	**	*	\checkmark	J

¹Broder (1997), ²Shrivastava (2017), ³Ertl (2017), ⁴Ertl (2020), ⁵Ertl (2018), ⁶Christiani (2020), ⁷Flajolet et al. (2007), ⁸Karppa and Pagh (2022), ⁹Ertl (2021) ¹⁰ which means that adding an element or taking the union of multiple subsets can be performed in sketch space, important feature for large scale application when new data need to be added to existing sketches or distributed environment.

¹¹OD indicates optimal densification. Since it is not mergeable, it is difficult to be applied in distributed environment.

¹²SetSketch can be applied in a weighted fashion but not in this paper

Table S2. Theoretical comparison of estimator variance (relative mean standard error) of MinHash-like algorithms for Jaccard index, where m is the number of registers (sketch size). The same m was used for all tools expect for SetSketch, which requires a smaller m to have similar accuracy. J or J_P represents Jaccard index or J_P , respectively.

	RMSE
Classic MinHash	Sqrt(J*(1-J)/m)
ProbMinHash	$sqrt(J_p^*(1-J_p)/m)$
SuperMinHash ¹	$sqrt(\frac{J(1-J)}{m}(1-\frac{\sum_{l=1}^{m-1}l^{u}((l+1)^{u}+(l-1)^{u})-2l^{u}}{(m-1)^{u}m^{u}(u-1)}))$
SetSketch ²	[sqrt(1/m), sqrt(1.08/m)]
Dashing (original&Improved)3	<pre>sqrt(1.079/m) + union/intersection error > sqrt(1.079/m)</pre>
Dashing (MLE) ⁴	sqrt(1.074/m)
Densifed MinHash ⁵	sqrt(1/m)
τ-GRA (τ=0.8898) ⁶	sqrt(1.075/m)

¹The second long term in the variance is always smaller than 1, thus SuperMinHash RMSE is always smaller than classic MinHash, the smallest MinHash RMSE known until today; ²As b is close to 1, RMSE is close to lower bound; ³Both the original and improved estimator need correction at small cardinality. Dashing2 has the same variance with SetSketch because it reimplemented SetSketch; ⁴MLE is Maximum Likelihood Estimator, theoretically optimal for HyperLogLog sketches but much slower; ⁵One Permutation MinHash with Optimal Densification, variance goes to 0 much faster than classic MinHash/SuperMinHash/ProbMinHash as m increases, see Shrivastava (2017). Note that the variance is larger than classic MinHash or ProbMinHash for any J because J(1-J) is always smaller than 0.25; ⁶Generalized Remaining Area with τ=0.8898 has the lowest variance. Detailed analysis of RSME for Tau-GRA can be found in Pettie and Wang (2022)

Table S3. GSearch's performance on major CPU platforms for searching 8,466 query genomes against the RefSeq reference database (~318K genomes). ProbMinHash was used as the genomic distance algorithm for this benchmark, with 50 neighbors requested.

CPU	Number of threads	Clock speed (GHz)	Request time for nt (min)	Gene Prediction- FGSrs (min)⁰	Request time for proteome (min)	hmmsear ch time (min) ^d	Request time for USCG (min)ª
Intel (R) Xeon (R) Gold 6226ª	24	2.70	9.329	3.348	6.234	3.524	0.445
Intel (R) Core i9-8950HK⁵	16	2.30	13.654	6.764	7.91	6.214	0.613
AMD EPYC 7513aª	32(24 used)	2.60	10.245	4.452	6.021	3.345	0.402
Apple M1 Pro ^b	10	3.22	9.669	4.31	6.166	4.603	0.568

^aRHEL v7.9, Linux v3.10.0-1160, all threads used. 512G RAM available.

^bMacOS v12.3, Darwin 21.4.0 on MacBook Pro laptops, all threads used. Database was split into two pieces to have smaller memory requirement on those 16GB RAM laptops since database size was ~15GB for entire NCBI/RefSeq.

•Parallel package was used to run multiprocess at the same time. FGSrs stands for FragGeneScanRs (comparisons with Prodigal can be found in Table S11). Note that in practice only those genomes failing in the request step at the nucleotide (nt) level (best match found was less than 78% ANI) will be used in this amino-acid step.

^dOnly 1000 genomes were used for testing hmmsearch and USCG request in GSearch because this step was for assessing a few genomes that were novel at the order level or above. Parallel Packages was used to run multiple processes of hmmsearch, one thread per process for hmmsearch.

Table S4. GSearch 3-step classification pipeline accuracy using GTDB-tk as reference for 1,000 test genomes of various degrees of novelty relative to the database genomes.

	GSearch	GTDB
Species level (>95% ANI)	100% (699/699)	100% (reference)
Genus to family level or above (<62% AAI)	87.1%(266/301)	100% (reference)

Table S5. GSearch search (request) performance on major CPU platforms using the GTDB v207 database for graph building and 1000 genome queries. Default ProbMinHash option was used.

CPU	Number of threads	Clock speed (GHz)	Request time for nt (min)	Gene Prediction- FGSrs (min)°	Request time for proteome (min)	hmmsear ch time (min) ^d	Request time for USCG (min)
Intel (R) Xeon (R) Gold 6226ª	24	2.70	2.329	1.348	1.334	0.524	0.117
Intel (R) Core i7-7770HQ♭	8	2.80	8.654	6.764	2.041	1.534	0.510
AMD EPYC 7513aª	32(24 used)	2.60	1.937	1.120	1.021	0.345	0.102
Apple M1 Prob	10	3.22	2.369	2.12	0.866	0.498	0.168

^aRHEL v7.9, Linux v3.10.0-1160, all threads used.

^bMacOS v12.3, Darwin 21.4.0, all threads used.

^cParallel package was used to run multiprocess at the same time. FGSrs stands for FragGeneScanRs. Note that in practice only those genomes failing at the Request step for nucleotide-level search (best match found is <78% ANI) will be used at this step. ^dOnly 100 genomes were used for testing hmmsearch because this step is for very novel (deep-branching) genomes at order level or above, which are not very common in real-world dataset. The Parallel Package was used to run multiple processes of hmmsearch, one thread per process for hmmsearch.

Table S6. GSearch performance and accuracy for the viral database using the BLASTbased AAI as the reference standard. MASH dist command was run using 24 threads for this analysis.

	GSearch (aa)	Recall (top 5)	MASH (aa)
1000 phage	4 min 24 s	98.32%	2.12 h
10,000 Phage	39 min 58 s	96.04%	17.4 h

Table S7. Benchmarking of hnswlib-rs library using the MNIST fashion (70,000) dataset.

	Mª	ef_construct	recall	Build time (s)	Query time⁵ (s)	Speed ^c (queries/s)
Run #1	32	200	0.9849	4.88	0.27	36569.63
Run #2	64	400	0.9959	9.09	0.46	21699.18
Run #3	128	400	0.9981	9.66	0.78	12740.01
Run #4	128	800	0.9984	18.00	0.86	11574.46
Run #5	128	1600	0.9986	34.53	0.93	10747.78
Run #6	256	1200	0.9995	30.00	1.61	6209.15
Run #7	256	1600	0.9996	39.18	1.67	5972.75
Run #8	256	2400	0.99999	56.33	1.72	5804.61

^a ef_search=M, the width of search for querying equals to M

^b Large M and ef_construct take more time for both build and query search. In the MNIST fashion dataset, Euclidean distance is precomputed and not considered in this test. The distance used is the same for Table S8, Table S9 and table S10.

° Tests were run on a 24 threads Intel (R) Xeon (R) Gold 6226 CPU@2.70 Ghz. Speed is tested using 10,000 queries.

Table S8. Benchmarking of hnswlib-rs library using the SIFT1M (1,000,000) dataset.

	М	ef_construct	recall	Build time (min)	Query time (s)	Speed ª(queries/s)
Run #1	128	1600	0.9980	21.18	1.81	5528.93
Run #2	256	3200	0.9986	40.83	3.82	2618.20

^a Tests were run on a 24 threads Intel (R) Xeon (R) Gold 6226 CPU@2.70 Ghz. Speed is tested using 10,000 queries.

Table S9. Benchmarking	of hnswlib-rs librar	y for scalabilit	y using the	SIFT1M dataset.
------------------------	----------------------	------------------	-------------	-----------------

	Number of threads	М	ef_construct	Build time (min)	Query time (s)	Speedª (queries/s)
Run #1	8	256	3200	87.97	7.90	1265.27
Run #2	12	256	3200	67.02	6.26	1596.35
Run #3	24	256	3200	40.83	3.82	2618.20

^a Tests were run on an Intel (R) Xeon (R) Gold 6226 CPU@2.70 Ghz. For each run, the number of threads shown in column #2 were requested. Speed is tested using 10,000 queries.

Table S10. Comparison of hnswlib-rs with C++ hnswlib using the MNIST fashion (70,000) dataset. For this test, 60,000 genomes were used for building as database and 10,000 as query, requesting 50 best neighbors.

	Number of threads	Μ	ef_construc t	Build time (s)	Query time (s)	Speedª (queries/s)
Run #1 (hnswlib-rs)	24	128	1600	34.53	0.93	11225
Run #2 (hnswlib-c++1)	24	128	1600	31.02	1.05	10204
Run #3 (hnswlib-rs)	24	256	3200	92.83	2.32	4250
Run #4 (hnswlib-c++)	24	256	3200	51.83	1.85	5235

¹https://github.com/nmslib/hnswlib

Table S11. The table shows benchmark results of FGSrs against truth (full version v0.0.1) and comparison with other tools. True positives were denoted as the number of base pair in a prediction on the correct strand, false positives the number of base pair in a prediction outside gene annotations or on the wrong strand, true negatives the number of bp outside

gene annotations that weren't in any prediction, and false negatives the number of bp inside gene annotations that weren't in any prediction. Prec denotes precision, Sens denotes sensitivity, Spec denotes specificity. Negative Predictive Value and Matthew's Correlation Coefficient are provided as percentages.

tool	ТР	FP	TN	FN	prec	sens	F1 score	spec	NPV	мсс
FGS	2494195	231257	656879	283010	91.51	89.81	0.90	73.96	69.89	62.58
prodigal	2488137	117535	745673	313996	95.49	88.79	0.92	86.38	70.37	70.36
FGSrs	2494195	231257	656879	283010	91.51	89.81	0.90	73.96	69.89	62.58

Table S12. Effect of genome completeness on GSearch recall. Genomes of different degrees of completeness were obtained by randomly sampling the gene sequences from the predicted gene collection of the complete genome. *Hydrogenimonas urashimensis*, an H₂ dependent chemolithoautotrophic bacteria isolated from deep sea hydrothermal vent (Mino et al., 2021), was used as query. There is no identical genome to *H. urashimensis* in the GTDB database.

Completeness	Top Hits found/10	Top 5 hits found/5	Recall (top 10)	Recall (top 10)
0.95	10/10	5/5	100%	100%
0.90	10/10	5/5	100%	100%
0.85	9/10	5/5	90%	100%
0.80	8/10	5/5	80%	100%
0.75	8/10	5/5	80%	100%
0.70	8/10	5/5	80%	100%
0.65	8/10	5/5	80%	100%
0.60	8/10	5/5	80%	100%
0.55	8/10	5/5	80%	100%
0.50	7/10	5/5	70%	100%
0.45	5/10	5/5	50%	100%
0.40	4/10	4/5	40%	80%
0.35	4/10	4/5	40%	80%
0.30	4/10	4/5	40%	80%

0.20	2/10	2/5	20%	40%
0.10	1/10	2/5	10%	40%

Table S13. GSearch, Sourmash, Dashing and BinDash benchmark against blastn-ANI at nucleotide level. Query genomes is "OceanDNA-b42278.fa" from (Nishimura and Yoshizawa, 2022). Database genomes were all NCBI/RefSeq genomes. Mash, FastANI and Blastn-based ANI give the same top 10 for this query genomes. For each column, matches are ranked by the respective distance, taken directly from the software output. Top 10 found by blastn-ANI were about 80% to 97% ANI.

Blastn-ANI (top10), ground truth	GSearch (top10, ProbMinHash), recall 100%, 2s	Sourmash (top10), recall 90%, 13min	Dashing (top10), recall 90%, 26s	BinDash (top 10), recall 100%, 3.7s
GCA_902591925.1_genomic.fna	GCA_902591925.1_genomic.fna	GCA_902591925.1_genomic.fna	GCA_902591925.1_genomic.fna	GCA_902591925.1_genomic.fna
GCA_902617045.1_genomic.fna	GCA_902617045.1_genomic.fna	GCA_902617045.1_genomic.fna	GCA_902617045.1_genomic.fna	GCA_902617045.1_genomic.fna
GCA_902541175.1_genomic.fna	GCA_902547295.1_genomic.fna	GCA_902541175.1_genomic.fna	GCA_902612915.1_genomic.fna	GCA_902541175.1_genomic.fna
GCA_902612915.1_genomic.fna	GCA_902612915.1_genomic.fna	GCA_902612915.1_genomic.fna	GCA_902541175.1_genomic.fna	GCA_902612915.1_genomic.fna
GCA_004212975.1_genomic.fna	GCA_004212975.1_genomic.fna	GCA_902547295.1_genomic.fna	GCA_902547295.1_genomic.fna	GCA_902547295.1_genomic.fna
GCA_902547295.1_genomic.fna	GCA_902541175.1_genomic.fna	GCA_004212975.1_genomic.fna	GCA_004212975.1_genomic.fna	GCA_004212975.1_genomic.fna
GCA_902630885.1_genomic.fna	GCA_902586925.1_genomic.fna	GCA_902586925.1_genomic.fna	GCA_000252525.1_genomic.fna	GCA_902630885.1_genomic.fna
GCA_902586925.1_genomic.fna	GCA_902630885.1_genomic.fna	GCA_902631785.1_genomic.fna	GCA_902630885.1_genomic.fna	GCA_902586925.1_genomic.fna
GCA_902631785.1_genomic.fna	GCA_902631785.1_genomic.fna	GCA_000252525.1_genomic.fna	GCA_902631785.1_genomic.fna	GCA_902631785.1_genomic.fna
GCA_000252525.1_genomic.fna	GCA_000252525.1_genomic.fna	GCA_902582355.1_genomic.fna	GCA_902582355.1_genomic.fna	GCA_000252525.1_genomic.fna

Table S14. GSearch (ProbMinHash), Sourmash, Dashing and Sourmash benchmark against Blastn-ANI at nucleotide level. Query genome is "OceanDNA-b42278.fa" from (Nishimura and Yoshizawa, 2022). Database genomes are all NCBI/RefSeq genomes (318k) after removing the top 10 ground truth genomes found in **Table S13** above. For each column, matches are ranked by the respective distance provided in the output of each tool. Top 10 matches found by blastn-ANI showed between 75% to 80% to the query genome, corresponding to Jaccard index 0.009 to 0.015 respectively. Boldface denotes the genomes found by each method compared to the ground truth in column 1.

Blastn-ANI (top10), ground truth	GSearch (Prob) (top10), recall 50%, 2s	Sourmash (top10), recall 60%, 14min	Dashing (default Ertl-MLE estimator) (top10), recall 10%, 26s	BinDash recall (90%). 3.7s
GCA 902582355.1 genomic.fna (79.24%)	GCA 002690725.1 genomic.fna	GCA 002690725.1 genomic.fna	GCA 902582355.1 genomic.fna	GCA 902582355.1 genomic.fna
GCA_002690725.1_genomic.fna (78.79%)	GCA_902560315.1_genomic.fna	GCA_902560315.1_genomic.fna	GCA_003282945.1_genomic.fna	GCA_902560315.1_genomic.fna
GCA_902560315.1_genomic.fna (77.92%)	GCA_902556105.1_genomic.fna	GCA_902556105.1_genomic.fna	GCA_012735275.1_genomic.fna	GCA_002690725.1_genomic.fna
GCA_902517505.1_genomic.fna (77.57%)	GCA_902558095.1_genomic.fna	GCA_902582355.1_genomic.fna	GCF_000165465.1_genomic.fna	GCA_902517505.1_genomic.fna
GCA_002169625.2_genomic.fna (76.41%)	GCA_902517505.1_genomic.fna	GCA_902558095.1_genomic.fna	GCF_003111605.1_genomic.fna	GCA_902583575.1_genomic.fna
GCA_902556105.1_genomic.fna (76.37%)	GCF_017873235.1_genomic.fna	GCA_003213495.1_genomic.fna	GCA_017532025.1_genomic.fna	GCA_902579825.1_genomic.fna
GCA_902583575.1_genomic.fna (75.30%)	GCA_002169625.2_genomic.fna	GCA_002721465.1_genomic.fna	GCA_902592715.1_genomic.fna	GCA_003213495.1_genomic.fna
GCA_003213495.1_genomic.fna (75.24%)	GCA_002704625.1_genomic.fna	GCA_902517505.1_genomic.fna	GCA_014653355.1_genomic.fna	GCA_902556105.1_genomic.fna
GCA_902579825.1_genomic.fna (75.15%)	GCA_902559345.1_genomic.fna	GCA_902557965.1_genomic.fna	GCA_902626385.1_genomic.fna	GCA_902558095.1_genomic.fna
GCA_902563835.1_genomic.fna (75.00%)	GCA_902557965.1_genomic.fna	GCA_902559345.1_genomic.fna	GCA_016288795.1_genomic.fna	GCA_902563835.1_genomic.fna

Table S15. GSearch (SetSketch) and Dashing (Ertl's Joint MLE) benchmarking against blastn-ANI. The table is similar to **Table S14** above but using the SetSketch algorithm.

Blastn-ANI (top10), ground truth	GSearch with SetSketch HLL (top10), recall 60%, 2s	GSearch with SuperMinHash (top10), recall 60%, 2s	Dashing (JMLE methods, ~10x slower) (top 10), recall 10%, 17.4 min
GCA_902582355.1_genomic.fna (79.24%)	GCA_002690725.1_genomic.fna	GCA_902560315.1_genomic.fna	GCA_902582355.1_genomic.fna
GCA_002690725.1_genomic.fna (78.79%)	GCA_902560315.1_genomic.fna	GCA_002690725.1_genomic.fna	GCA_902586925.1_genomic.fna
GCA_902560315.1_genomic.fna (77.92%)	GCA_902517505.1_genomic.fna	GCA_902583575.1_genomic.fna	GCA_902626385.1_genomic.fna
GCA_902517505.1_genomic.fna (77.57%)	GCA_902583575.1_genomic.fna	GCA_902517505.1_genomic.fna	GCA_003282945.1_genomic.fna
GCA_002169625.2_genomic.fna (76.41%)	GCA_902579825.1_genomic.fna	GCA_902559345.1_genomic.fna	GCA_012735275.1_genomic.fna
GCA_902556105.1_genomic.fna (76.37%)	GCA_902559345.1_genomic.fna	GCA_902579825.1_genomic.fna	GCF_000165465.1_genomic.fna

GCA_902583575.1_genomic.fna	GCA_902558095.1_genomic.fna	GCA_902556105.1_genomic.fna	GCF_003111605.1_genomic.fna
GCA_003213495.1_genomic.fna	GCA_902556105.1_genomic.fna	GCA_902592715.1_genomic.fna	GCA_017532025.1_genomic.fna
(75.24%) GCA_902579825.1_genomic.fna	GCA_902528875.1_genomic.fna	GCA_014653355.1_genomic.fna	GCA_902592715.1_genomic.fna
GCA_902563835.1_genomic.fna	GCA_902551185.1_genomic.fna	GCF_000165465.1_genomic.fna	GCA_014653355.1_genomic.fna

Table S16. Recall (top 10) of ProbMinHash, Densified MinHash, SetSketch and SuperMinHash against the BLAST-ANI results. Same query genome as **Table S13** was used.

Blastn-ANI (top10), ground truth	GSearch ProbMinHash (top10), recall 100%, 2s	GSearch Densified Minhash recall 100%, 2s	GSearch setsketch recall 100%, 5s	GSearch SuperMinHash recall 100%, 8s
GCA_902591925.1_genomic.fna	GCA_902591925.1_genomic.f na	GCA_9025919 25.1_genomic. fna	GCA_902591925.1_genomic.fna	GCA_902591925.1_genomic.fna
GCA_902617045.1_genomic.fna	GCA_902617045.1_genomic.f na	GCA_9026170 45.1_genomic. fna	GCA_902617045.1_genomic.fna	GCA_902617045.1_genomic.fna
GCA_902541175.1_genomic.fna	GCA_902547295.1_genomic.f na	GCA_9025411 75.1_genomic. fna	GCA_902612915.1_genomic.fna	GCA_902541175.1_genomic.fna
GCA_902612915.1_genomic.fna	GCA_902612915.1_genomic.f na	GCA_9026129 15.1_genomic. fna	GCA_902541175.1_genomic.fna	GCA_902612915.1_genomic.fna
GCA_004212975.1_genomic.fna	GCA_004212975.1_genomic.f na	GCA_9025472 95.1_genomic.	GCA_004212975.1_genomic.fna	GCA_004212975.1_genomic.fna
GCA_902547295.1_genomic.fna	GCA_902541175.1_genomic.f na	GCA_0042129 75.1_genomic.	GCA_902547295.1_genomic.fna	GCA_902547295.1_genomic.fna
GCA_902630885.1_genomic.fna	GCA_902586925.1_genomic.f na	GCA_9026308 85.1_genomic. fna	GCA_902630885.1_genomic.fna	GCA_902630885.1_genomic.fna
GCA_902586925.1_genomic.fna	GCA_902630885.1_genomic.f na	GCA_9025869 25.1_genomic.	GCA_902586925.1_genomic.fna	GCA_902586925.1_genomic.fna
GCA_902631785.1_genomic.fna	GCA_902631785.1_genomic.f na	GCA_9026317 85.1_genomic.	GCA_902631785.1_genomic.fna	GCA_902631785.1_genomic.fna
GCA_000252525.1_genomic.fna	GCA_000252525.1_genomic.f na	GCA_0002525 25.1_genomic. fna	GCA_902582355.1_genomic.fna	GCA_000252525.1_genomic.fna

Table S17. Recall (top 10) of GSearch (ProbMinHash option), Mash and Sourmash against the BLAST-AAI results. The same query genome as shown in **Table S14** was used but at the proteome level after gene prediction. Dashing and BinDash were not included because these tools do not provide an amino acid level search option. Same query genome as Table S13.

Blastp-AAI (top10), ground truth, AAI	GSearch ProbMinHash (top10), recall 90%, 2s	Mash, recall 90%, 13 min	Sourmash, recall 80%, 3 min
GCA_902591925.1_genomic.fna (83.2%)	GCA_902591925.1_genomic.fna	GCA_902591925.1_genomic.fna	GCA_902591925.1_genomic.fna
GCA_902617045.1_genomic.fna (83.16%)	GCA_902617045.1_genomic.fna	GCA_902617045.1_genomic.fna	GCA_902617045.1_genomic.fna
GCA_902541175.1_genomic.fna (82.57%)	GCA_004212975.1_genomic.fna	GCA_902612915.1_genomic.fna	GCA_902541175.1_genomic.fna
GCA_902612915.1_genomic.fna (82.15%)	GCA_902612915.1_genomic.fna	GCA_902541175.1_genomic.fna	GCA_902612915.1_genomic.fna
GCA_004212975.1_genomic.fna (81.86%)	GCA_902547295.1_genomic.fna	GCA_902547295.1_genomic.fna	GCA_902558095.1_genomic.fna
GCA_902630885.1_genomic.fna (80.99%)	GCA_902541175.1_genomic.fna	GCA_004212975.1_genomic.fna	GCA_902547295.1_genomic.fna
GCA_902547295.1_genomic.fna (80.73%)	GCA_902586925.1_genomic.fna	GCA_902559345.1_genomic.fna	GCA_902630885.1_genomic.fna
GCA_902586925.1_genomic.fna (80.50%)	GCA_902558095.1_genomic.fna	GCA_902586925.1_genomic.fna	GCA_902586925.1_genomic.fna
GCA_000252525.1_genomic.fna (79.96%)	GCA_902630885.1_genomic.fna	GCA_902631785.1_genomic.fna	GCA_902631785.1_genomic.fna
GCA_902631785.1_genomic.fna (79.48%)	GCA_000252525.1_genomic.fna	GCA_902582355.1_genomic.fna	GCA_902559345.1_genomic.fna

Table S18. Database build time, maximum memory and database size for GSearch (ProbMinHash option), Densified MinHash, SuperMinHash and SetSketch. GTDB v207 was used (65,703 genomes). 24 threads are used.

	Build time	Max Memory	Database size
ProbMinHash (default)	1.3h	15.3G	3.0G
Densified MinHash	0.74h	9.4G	1.9G
SuperMinHash	4.7h	13.2G	3.0G
SetSketch	2.4h	3.7G	0.6G

Table S19. Database build time, maximum memory and database size for GSearch (ProbMinHash option), Densified MinHash, SuperMinHash and SetSketch. NCBI/RefSeq genomes (~318K) were used. 24 threads are used.

	Build time	Max Memory	Database size
ProbMinHash (default)	4.1h	21G	15G
Densified MinHash	1.4h	14G	11G
SuperMinHash	27h	16G	15G
SetSketch	13h	3.7G	3.0G

Table S20. GSearch accuracy benchmark using orthoANI as an additional ground truth. Query genomes is "OceanDNA-b42278.fa" from (Nishimura and Yoshizawa, 2022). Database genomes were all NCBI/RefSeq genomes. Showing top 10 nearest genomes, ANI from ~94% to ~81% for the top 10.

Ortho-ANI (top10), ground truth	Blastn-ANI (ANI calculator), top 10	GSearch (top10, ProbMinHash), recall 100%, 2s
GCA_902591925.1_genomic.fna (95.33%)	GCA_902591925.1_genomic.fna (95.27%)	GCA_902591925.1_genomic.fna (95.38%)
GCA_902617045.1_genomic.fna (94.12%)	GCA_902617045.1_genomic.fna (94.18%)	GCA_902617045.1_genomic.fna (94.29%)
GCA_902541175.1_genomic.fna (93.35%)	GCA_902541175.1_genomic.fna (93.22%)	GCA_902547295.1_genomic.fna (93.31%)
GCA_902612915.1_genomic.fna (92.43%)	GCA_902612915.1_genomic.fna (92.46%)	GCA_902612915.1_genomic.fna (92.83%)
GCA_004212975.1_genomic.fna (92.06%)	GCA_004212975.1_genomic.fna (92.10%)	GCA_004212975.1_genomic.fna (91.91%)
GCA_902547295.1_genomic.fna (90.15%)	GCA_902547295.1_genomic.fna (90.09%)	GCA_902541175.1_genomic.fna (90.35%)
GCA_902630885.1_genomic.fna (88.56%)	GCA_902630885.1_genomic.fna (88.51%)	GCA_902586925.1_genomic.fna (88.12%)
GCA_902586925.1_genomic.fna (85.23%)	GCA_902586925.1_genomic.fna (85.49%)	GCA_902630885.1_genomic.fna (85.37%)
GCA_902631785.1_genomic.fna (83.91%)	GCA_902631785.1_genomic.fna (83.88%)	GCA_902631785.1_genomic.fna (83.48%)
GCA_000252525.1_genomic.fna (81.08%)	GCA_000252525.1_genomic.fna (81.32%)	GCA_000252525.1_genomic.fna (81.51%)

Supplementary Table References

Broder, A.Z. (1997) On the resemblance and containment of documents. In *Proceedings Compression and Complexity of SEQUENCES 1997 (Cat No 97TB100171)*: IEEE, pp. 21-29. Christiani, T. (2020) DartMinHash: Fast Sketching for Weighted Sets. *arXiv preprint arXiv:200511547*.

Ertl, O. (2017) Superminhash-A new minwise hashing algorithm for jaccard similarity estimation. *arXiv preprint arXiv:170605698*.

Ertl, O. (2018) BagMinHash - Minwise Hashing Algorithm for Weighted Sets. *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & amp; Data Mining*: 1368–1377.

Ertl, O. (2020) ProbMinHash – A Class of Locality-Sensitive Hash Algorithms for the (Probability) Jaccard Similarity. *IEEE Transactions on Knowledge and Data Engineering*: 1-1. Ertl, O. (2021) SetSketch: filling the gap between MinHash and HyperLogLog. *Proc VLDB Endow* **14**: 2244–2257.

Flajolet, P., Fusy, É., Gandouet, O., and Meunier, F. (2007) Hyperloglog: the analysis of a nearoptimal cardinality estimation algorithm. *Discrete Mathematics and Theoretical Computer Science*: 137-156.

Karppa, M., and Pagh, R. (2022) HyperLogLogLog: Cardinality Estimation With One Log More. *arXiv preprint arXiv:220511327*.

Mino, S., Shiotani, T., Nakagawa, S., Takai, K., and Sawabe, T. (2021) Hydrogenimonas urashimensis sp. nov., a hydrogen-oxidizing chemolithoautotroph isolated from a deep-sea hydrothermal vent in the Southern Mariana Trough. *Systematic and Applied Microbiology* **44**: 126170.

Nishimura, Y., and Yoshizawa, S. (2022) The OceanDNA MAG catalog contains over 50,000 prokaryotic genomes originated from various marine environments. *Scientific Data* **9**: 305. Pettie, S., and Wang, D. (2022) Simpler and Better Cardinality Estimators for HyperLogLog and PCSA. *arXiv preprint arXiv:220810578*.

Shrivastava, A. (2017) Optimal densification for fast and accurate minwise hashing. *International Conference on Machine Learning*: 3154-3163.

Supplementary Methods & Materials

2 **ProbMinHash vs. traditional MinHash**

1

3 Mash is a hashing-based algorithm based on MinHash¹, which is very efficient for comparing 4 genome/metagenome overall similarity². Mash distances represent a k-mer-based overall 5 overlap between sequences according to a minimal evolutionary model. Essentially, Mash 6 distance is the Jaccard similarity value of kmer shared between k-mer sets A and B extracted 7 from genome A and B. However, Mash, and similar MinHash-based tools, have several 8 limitations; most notably, the loss of k-mer frequency information (only presence/absence of kmer 9 is counted) and the impact of relative set size (e.g., completeness level of a genome or genome 10 size) on the Jaccard similarity estimates (for example, Mash distance, where MinHash based 11 estimation of Jaccard index is biased by different set size, or total k-mer count of the genome due 12 to bottom-k sketches; HyperLogLog is not affected by different set size)^{2,3} Although some recent 13 MinHash implementations address the relative set size limitation, e.g., the over-sketching and 14 track-abundance methods of the MinHash-based tools 'finch, 'sourmash' or FracMinHash 15 (sketching a subset of k-mers according to size of set) and HyperLogLog 4-7, they do not utilize 16 the frequencies of all observed k-mers in generating the k-mer-profile (sketch) for a given 17 sequence set. More recently, in the HULK software, consistent weighted sampling (D²histosketch 18 which is the same with P-MinHash algorithm and it was proposed for J_p, we called it P-MinHash 19 because it was invented before D²histosketch and is equivalent to P-MinHash⁸) ⁹ was utilized to 20 incorporate k-mer frequency information when estimating weighted and standard Jaccard 21 similarity, which effectively addresses these limitations mentioned above ¹⁰. Notably, the hash 22 algorithm (P-MinHash) used in D²histosketch could be further optimized to achieve a time 23 complexity below O(nm) (where m denotes the signature size and n is the number of elements 24 with nonzero weight in two sequence sets), further improving the performance of applications

22

25 such as HULK. Motivated by the SuperMinHash for conventional Jaccard similarity estimation ¹¹ 26 and BagMinHash algorithm for weighted Jaccard similarity estimation ¹², ProbMinHash 27 (probminhash 3(a) and 4 algorithm) is orders of magnitude faster than the original algorithm P-28 MinHash proposed in D²histosketch ¹⁴. Probminhash estimates the Jaccard probability J_{p} 29 similarity, and 1- Jp is indeed a metric on the probability distributions and is Pareto optimal 30 (Supplementary Note 1)^{8, 14}. Densified MinHash, or One Permutation MinHash with Optimal/faster 31 Densification, is the fastest MinHash algorithm due to theoretical breakthrough (average case 32 O(n+m)) despite large variance than classic MinHash (Table S2). It uses only one hash function 33 but copy values from empty bins to non-empty bins ("densified") in the sketch vector either 34 mapping forward or backward or both ^{13, 15, 16}. We choose Sourmash, Mash, Dashing 1/2 and 35 BinDash for benchmark because all provide estimation of Jaccard index, which correlates very 36 well with ANI after transformation, and had been all previously benchmarked against BLAST-37 based ANI and fastANI^{2, 19}.

38

39 Comparison with other genome/sequence search algorithm

40 There are many other data structures designed for general purposes sequence search problems 41 such as Sequence Bloom Tree (SBT) and its variants ²⁰⁻²², COBS/BIGSI ^{23, 24}, Layered LSH 42 (approximating ungapped alignment, not applicable to ANI like distance, which is gapped alignment) ²⁵ and RAMBO (Repeated and Merged Bloom Filter) ²⁶ that can achieve linear or 43 44 sometimes even sub-linear genomic database search performance alone. However, those tools 45 have never been benchmarked against BLAST-based ANI/AAI in the context of microbial genomic 46 search (that is, whether the best hits/genomes found by the mentioned tools are the same with 47 ANI/AAI comparison best hits/genomes), which is a popular reference standard for measuring 48 microbial genomic distance/identity, and thus infer microbial taxonomy ^{27, 28}. MinHash-based tool 49 such as Mash and FastANI have been benchmarked against BLAST-based ANI, and were shown

to be the most accurate k-mer-based sub-linear algorithms for ANI comparisons and/or searching
 ^{2, 19}.

52

53 HNSW in Rust benchmark against testing dataset

54 To benchmark our reimplementation of hnswlib, we followed standard ANN benchmark 55 procedures using two popular testing datasets (MINST and SIFT1M) based on their Euclidean 56 distance²⁹. Our results showed that, for the MINST fashion dataset (784 dimensions, 60,000 57 vectors), recall for top 100 neighbors of 10,000 query vectors is greater than 98% for a smaller 58 number of M and ef construct, and even higher recall rate (99.86%) for a medium M and 59 ef construct while query speed is not compromised (Supplemental Table S7). For the SIFT1M 60 dataset (128 dimensions, 1,000,000 vectors), recall for top 100 neighbors of 10,000 guery vectors 61 was 99.77% for a medium M and ef construct (Supplemental Table S7 and S8). In terms of 62 speed, we compare it with hnswlib using the MINST-fashion dataset and it is as fast as hnswlib: 63 it took 18.06s and 0.89s for database building and searching for hnswlib-rs while it took 18.47s 64 and 1.07s for database building and searching for the C++ hnswlib) (Supplementary Table S9). 65 The Rust package hnswlib-rs can be found at: https://github.com/jean-pierreBoth/hnswlib-rs. For 66 each genomic database, we chose M and ef construct experimentally, by gradually increasing M 67 and ef_construct while monitoring query speed and recall, similar to what is shown in 68 Supplementary Table S2 for MNIST dataset. We stopped the assessment when there was only a 69 marginal increase in accuracy but decent decrease in speed. To leverage between recall and 70 speed, we use M=128 and ef search=1600 for graph building for GTDB database fungal 71 database while M=128, ef search=3200 for phage database.

72

73 Details of program implementation in Rust

24

74 Tohnsw starts by reading database genomes and generating k-mer profile and sketches using 75 the ProbMinHash3a algorithm (or SuperMinHash, SetSketch, Densified MinHash) for distance 76 calculation (Figure 1a and c). Next, tohnsw selects, at random, the first batch of genomes to insert 77 into the graph (Figure 1a (1)), following HNSW constructing rules mentioned above and taking 78 into account the computed ProbMinHash distance or SetSketch approximated Jaccard distances 79 (1-J) between genomes to connect genomes based on their relatedness (Figure 1b) until all 80 genomes in database have been inserted (Figure 1a (2), (3) and (4)). Finding nearest genomes 81 for the genome to be inserted is essentially a search process but search in a partially built graph, 82 which is similar to the request/search module: whenever a genome is going to be searched 83 against the existing graph, each genome in the graph is associated with a list that stores the M 84 closest neighbors/genomes to the genome and the distance to these neighbors. Then, the 85 distances of this genome with the nearest neighbors (M) of entry genome in each layer will be 86 computed (ef construct times) using the Probminhash3a algorithm or the SetSketch, and the 87 smallest distance of the neighbor genomes will be the new entry genome (Figure 1d and e). This 88 process will be repeated until the nearest genomes ($\leq M$) in the layer are found and subsequently, 89 the program will go to the layer below, using the genome that was represented by the nearest 90 genome in the above layer as new entry genome in the new layer. The search layer algorithm is 91 repeated until the bottom layer is reached/analyzed (Figure 1c). In contrast to the default settings 92 in the original hnswlib, we allow the two parameters of neighbor selecting heuristics, 93 extendCandidates to be true and keepPrunedConnections to be false because our genomic data 94 is extremely clustered and there is no need to fix the number of connections per element 95 considering the maximum connection allowed. The "Add" module is to add new genomes to pre-96 built HNSW database using default parameters loaded in the pre-built database files. Request 97 module will load the graph database and then search guery genomes against it to return the best

25

neighbors of each query, following exactly the same procedure with building step but withoutupdating the database.

100

101 Details of the SetSketch implementation

102 We implemented SetSketch algorithm 1 locality sensitivity section (LSH) according to lower and

upper bound of Jaccard $(J_{low} = max(0, \frac{b\frac{D_0/m+1}{2}}{b-1} - 1), J_{up} = \frac{b^{D_0/m}-1}{b-1})$ where D₀ is the number of 103 104 registers in the sketch of genome A that are equal to those in the sketch of genome B. In practice, 105 J_{low} is closer to true Jaccard for small J thus we use it. We use parameter m = 6144, b = 1.0005, 106 a = 20, and g = 65534 instead of the default ones to have smaller RMSE (<0.8%) around J=0.015 107 (corresponding to ANI 77.99% according to Mash equation) and acceptable running time (Figure 108 S7), which is equivalent to HyperLogLog for estimating Jaccard index in terms of space but with 109 smaller variance (with b close to 2, SetSketch will then be equivalent to SuperMinHash). A 110 SetSketch using this configuration is suitable to represent any set with up to 10¹⁹ distinct 111 elements/k-mers (much larger than total-number of k-mers from microbial genomes). The expected error of cardinality estimates or LSH is very small $(\sqrt{\frac{1}{m}}(\frac{b+1}{b-1}log(b)-1))$ as b 112 approximates $1(\frac{1}{\sqrt{m}})^{30}$, close to that of MinHash for large m, like 10⁴ or above but use much 113 114 smaller space. We also implemented the Joint Maximum Likelihood Estimator (JMLE): the ML 115 estimate for Jaccard was found by standard univariate optimization algorithm called Brent's 116 method, based on the argmin Rust package since the ML function is strictly concave for the 117 parameter mentioned above ³⁰. The RMSE of JMLE based on its Fisher information can be found 118 in Supplementary Note 7. Since RMSE of JMLE is smaller than LSH, it is much slower in practice, 119 but we only use it for filtering false positives after top k best neighbors were found for each query 120 by LSH and HNSW, thus it is not a problem for overall speed.

121 Details of the Densified MinHash implementation

122 We reimplemented both the One Permutation MinHash with Optimal Densification ¹⁵ and also the 123 Faster Densification ¹⁶ in Rust in the probminhash package. Specifically, one predefined hash 124 function is applied to all elements/k-mers in the set. Then, one-permutation MinHash 125 deterministically partitions the hash values into a predefined maximum number B of buckets, 126 extracts the smallest hash value in each bucket, and extracts the b lowest bits of each smallest 127 hash value. These B*b bits are used as the signature of the set. Usually, a hash value v is 128 assigned to the $\lceil v/(M/B) \rceil$ bucket, where M is the maximum possible value for v. Although fast at 129 both constructing and comparing sketches, one-permutation MinHash may produce a bucket that 130 contains hash values for one set but no hash values for another set. All buckets are totally ordered 131 in some way, each empty bucket uses the smallest hash value in the next non-empty bucket as 132 its own hash value (densified), and an additional bucket containing a special value is ordered after 133 all other buckets. We either map non-empty bins to empty bins (faster densification) or the other 134 way around to copy values (optimal densification). The densified sketch vectors for 2 sets were 135 then used for calculating collision probability/Jaccard index (See densminhash.rs in prominhash 136 package). We can use 2 times larger sketch size m for densified MinHash to achieve a smaller 137 variance compared to original MinHash without increasing running time.

138

139 Supplementary Methods & Materials References

- Broder, A.Z. in Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171) 21-29 (IEEE, 1997).
 Ondov, B.D. et al. Mash: fast genome and metagenome distance estimation using
- 1422.Ondov, B.D. et al. Mash: fast genome and metagenome distance estimation using143MinHash. Genome Biology 17, 132 (2016).

1443.Koslicki, D. & Zabeti, H. Improving MinHash via the containment index with applications145to metagenomic analysis. Applied Mathematics and Computation **354**, 206-215 (2019).

- 1464.Brown, C.T. & Irber, L. sourmash: a library for MinHash sketching of DNA. Journal of147Open Source Software 1, 27 (2016).
- 1485.Bovee, R. & Greenfield, N. Finch: a tool adding dynamic abundance filtering to genomic149MinHashing. Journal of Open Source Software 3, 505 (2018).

150	6.	Irber, L. et al. Lightweight compositional analysis of metagenomes with FracMinHash
151		and minimum metagenome covers. <i>bioRxiv</i> , 2022.2001.2011.475838 (2022).
152	7.	Baker, D.N. & Langmead, B. Dashing: fast and accurate genomic distances with
153		HyperLogLog. <i>Genome Biology</i> 20 , 265 (2019).
154	8.	Moulton, R. & Jiang, Y. Maximally Consistent Sampling and the Jaccard Index of
155		Probability Distributions. 2018 IEEE International Conference on Data Mining (ICDM),
156		347-356 (2018).
157	9.	Yang, D., Li, B., Rettig, L. & Cudré-Mauroux, P. D ² histoSketch: Discriminative and
158		Dynamic Similarity-Preserving Sketching of Streaming Histograms. IEEE Transactions
159		on Knowledge and Data Engineering 31, 1898-1911 (2019).
160	10.	Rowe, W.P.M. et al. Streaming histogram sketching for rapid microbiome analytics.
161		<i>Microbiome</i> 7 , 40 (2019).
162	11.	Ertl. O. Superminhash-A new minwise hashing algorithm for jaccard similarity estimation.
163		arXiv preprint arXiv:1706.05698 (2017).
164	12.	Ertl. O. BagMinHash - Minwise Hashing Algorithm for Weighted Sets. <i>Proceedings of the</i>
165		24th ACM SIGKDD International Conference on Knowledge Discovery & amp: Data
166		Mining 1368–1377 (2018)
167	13	Jia P et al in Proceedings of the 2021 International Conference on Management of
168	10.	Data 830-842 (2021)
169	14	ErtLO ProbMinHash – A Class of Locality-Sensitive Hash Algorithms for the
170	• ••	(Probability) Jaccard Similarity IEEE Transactions on Knowledge and Data Engineering
171		
172	15.	Shrivastava, A. Optimal densification for fast and accurate minwise hashing.
173		International Conference on Machine Learning, 3154-3163 (2017).
174	16.	Mai, T. et al. in Uncertainty in Artificial Intelligence 831-840 (PMLR, 2020).
175	17.	Baker, D.N. & Langmead, B. Dashing 2: genomic sketching with multiplicities and
176		locality-sensitive hashing. <i>bioRxiv</i> (2022).
177	18.	Agret, C., Cazaux, B. & Limasset, A. Toward optimal fingerprint indexing for large scale
178		genomics. <i>bioRxiv</i> . 2021.2011.2004.467355 (2022).
179	19.	Jain, C., Rodriguez-R. L.M., Phillippy, A.M., Konstantinidis, K.T. & Aluru, S. High
180		throughput ANI analysis of 90K prokarvotic genomes reveals clear species boundaries.
181		Nature Communications 9. 5114 (2018).
182	20.	Solomon, B. & Kingsford, C. Fast search of thousands of short-read sequencing
183		experiments. Nature biotechnology 34, 300-302 (2016).
184	21.	Solomon, B. & Kingsford, C. Improved search of large transcriptomic sequencing
185		databases using split sequence bloom trees. Journal of Computational Biology 25, 755-
186		765 (2018).
187	22.	Sun, C., Harris, R.S., Chikhi, R. & Medvedev, P. Allsome sequence bloom trees. Journal
188		of Computational Biology 25, 467-479 (2018).
189	23.	Bingmann, T., Bradley, P., Gauger, F. & Igbal, Z. in String Processing and Information
190		Retrieval: 26th International Symposium, SPIRE 2019, Segovia, Spain, October 7–9,
191		2019, Proceedings 26 285-303 (Springer, 2019).
192	24.	Bradley, P., Den Bakker, H.C., Rocha, E.P., McVean, G. & Igbal, Z. Ultrafast search of
193		all deposited bacterial and viral genomic data. <i>Nature biotechnology</i> 37 , 152-159 (2019).
194	25.	Chakraborty, A. & Bandyopadhyay, S. in 2018 Fifth International Conference on
195		Emerging Applications of Information Technology (EAIT) 1-4 (IEEE, 2018).
196	26.	Gupta, G. et al. in Proceedings of the 2021 International Conference on Management of
197		Data 2226-2234 (2021).

198 199	27.	Konstantinidis, K.T. & Tiedje, J.M. Towards a Genome-Based Taxonomy for Prokarvotes. <i>Journal of Bacteriology</i> 187 , 6258-6264 (2005).
200	28.	Goris, J. et al. DNA–DNA hybridization values and their relationship to whole-genome
201		sequence similarities. International Journal of Systematic and Evolutionary Microbiology
202		57 , 81-91 (2007).
203	29.	Aumüller, M., Bernhardsson, E. & Faithfull, A. ANN-Benchmarks: A benchmarking tool
204		for approximate nearest neighbor algorithms. Information Systems 87, 101374 (2020).
205	30.	Ertl, O. SetSketch: filling the gap between MinHash and HyperLogLog. Proc. VLDB
206		Endow. 14, 2244–2257 (2021).
207		
208		
209		
210		
211		
212		
213		
214		
215		
216		
217		
218		
219		
220		
221		
222		
223		
224		
225		
226		
227		
228		
229		
230		
231		
232		
233		
234		
235		
236		
237		
238		
239		
240		
241		
242		
243		

Supplementary Notes

Note 1. Calculation of metric (Weighted) Jaccard similarity indices. $J_W = \frac{\sum_{d \in D} min(\omega_A(d), \omega_B(d))}{\sum_{d \in D} max(\omega_A(d), \omega_B(d))}$ (where ω is the weight function of each element d) and total k-mer count in genomes in biasing the Jaccard index estimation (or set size) ¹. K-mer-weighted hashing approaches (taking into account the abundance of k-mers, not just presence/absence) are more space-expensive (e.g. higher memory requirement) for large dataset ², but are advantageous for genomes with frequent repeats. They have not been widely adopted yet ³. ⁴. To consider multiplicity of k-mers in the kmer set of genomes, traditional MinHash algorithms will not be a good choice since they assume unique set element (k-mer). New MinHash algorithms such as ICWS, BagMinHash and DartMinHash were designed for weighted set to address this limitation, with DartMinHash being the fastest ⁵⁻⁸. Still, those weighted MinHash algorithms do not solve the problem of different genome size (set size) in biasing estimation of weighted Jaccard index ⁹. A possible solution is to normalize the abundance of k-mer by the total k-mer count of each dataset, thereby providing a probability distribution of each k-mer. This then leads to the normalized weighted Jaccard index

 $J_{N} = \frac{\sum_{d \in D} min(\frac{\omega_{A}(d)}{\sum_{d' \in D} \omega_{A}(d')}, \frac{\omega_{B}(d)}{\sum_{d' \in D} \omega_{B}(d')})}{\sum_{d \in D} max(\frac{\omega_{A}(d)}{\sum_{d' \in D} \omega_{A}(d')}, \frac{\omega_{B}(d)}{\sum_{d' \in D} \omega_{B}(d')})}, \text{ where } \omega_{A} \text{ and } \omega_{B} \text{ describe the weight of each k-mer. Also,}$

those weighted MinHash algorithm could be further optimized computationally to be orders of magnitude faster, like it was done in BagMinHash and DartMinHash ¹⁰, similar to the computational optimizations implemented in MinHash in SuperMinHash ¹¹. Recently, ProbMinHash was proposed to take into account both weighted set (k-mer multiplicity) and total set size (total k-mer count, or genome size) ¹². Accordingly, new Locality Sensitive Hashing algorithms (P-MinHash) considering weighted set and different set size was proposed to estimate weighted and normalized (to account for set size difference) Jaccard-like index J_P =

 $\sum_{d \in D} \frac{1}{\sum_{d' \in D} max(\frac{\omega_A(d')}{\omega_A(d)}, \frac{\omega_B(d')}{\omega_B(d)})}$. P-MinHash is considered a more general case for Jaccard Index

estimation and is more close to true Jaccard than J_W and J_N ^{13, 14}. More importantly, J_p is Pareto optimal and 1-J_p is a proper metric ¹³. It has been shown that ANI estimated from J_p, which was computed by ProbMinHash (ProbMinHash2 algorithm) in Dashing 2, is slight better for bacterial genomes than traditional MinHash like Mash ¹⁵. ProbMinHash was built upon the new MinHash algorithm with further computational optimization for speed ¹². Hence, ProbMinHash is the currently the default option in GSearch. HyperLogLog can also be used to estimate Jaccard index: distinct element count of k-mers in set/genome A and B via HyperLogLog sketch (memory efficiency) and then use the inclusion-exclusion rule to have Jaccard index (*Jaccard*(*A*, *B*) = $\frac{|A|+|B|-|AUB|}{|AUB|}$) as implemented in Dashing, despite being overall less accurate for small cardinalities (e.g. virus genomes) ^{16, 17}. The Maximum Likelihood estimator (MLE) ¹⁶, in addition to the original estimator in HyperLogLog and improved estimator in ¹⁶ for distinct element counting, was thought to have the smallest variance, which met the Cramér-Rao lower bound, despite slower and difficult to compute and update ¹⁸. This is the idea behind Dashing, which we benchmark against GSearch.

Note 2. Kmer selecting rationale for nucleotide (nt) and amino-acid (aa) level searches for MinHash/probminhash tools. Assume amino acid/nucleotide sequences evolve at a constant rate and alphabets $|\Sigma|$ for AA is 20 and 4 for nucleotide. Consider two sequences x , y \in A^N drawn randomly over the alphabet of size #A = 20, N is the genome length, and let v_x,v_y \in R^{20^k} denote the k-mer frequency profile (multiplicity of each k-mer divided by total number of k-mers) for each sequence. For long sequences with N \gg 20^k, k-mer frequencies will converge to their mean, that

is 20^{-k} for all their components, which implies that $IIv_x - v_y II_1 \rightarrow 0$. Therefore, any k-mer profilebased method will severely underestimate distance between these two random sequences. In order to avoid this, k has to be restricted to large values $k' \ge log_{|\Sigma|}(N)$. In practice, since genomes sequences are not completely random, k can be slightly smaller than $\log_4(N)$ without underestimating genomic distance. Ondov et al., in the MASH paper, came up with a probability term (1-q)/q to take into account the probability of having a random kmer: $k' \ge log_{|\Sigma|}(N(1-q)/q)$ ¹⁹. Assume a probability q=0.01, for a typical bacterial proteome, $k \ge \log_{20}(1000000^*0.99/0.01)$ =6.146, so at least a kmer of size 6 or above should be used. For a typical bacterial genome at nucleotide level, N is about $4*10^{6}$, k >= $\log_4(4*10^{6}*0.99/0.01) = 14.280$, so at least a kmer of size 14 or above should be used. For universal genes only, N is about 350 AA * 120=42000, so $k \ge \log_{20}(42000*0.99/0.01) = 5.087$, so at least a kmer of size 5 or above should be used. Since universal gene alphabets are not randomly evolving (e.g., some regions in genes that encode key metabolic functions such as RNA processing units rarely mutate), k=5 should be appropriate. To optimize between sensitivity and specificity for kmer, we followed the practice suggested by Ondov et al., and Jain et al. 19, 20; that is, use k=16 for bacterial nucleotide genome sequences and a k=7 for bacterial proteome sequences. For fungal genomes, genome size is 10 to 20 times larger than tht of bacteria but the same formula applies and so, we have $k \ge 1$ $\log_4(20^{4}10^{6}0.99/0.01) = 16.441$, and accordingly we use kmer=21. For fungal proteome, coding density is much smaller than that of bacteria (we use an empirical 0.5 coding density) thus, we have $k \ge \log_{20}(10^{*}1^{*}10^{6} \cdot 0.99/0.01) = 6.914$. We use k=11 for fungal proteome to increase the specificity of the searches. For virus/bacteriophage genomes, we use k=11 for nt and k=7 for aa, considering also the jumbo/giant phage genomes that were found recently.

Note 3. Assessment of the time complexity of ProbMinHash and HNSW. The search time complexity is estimated based on the equation (or big O annotation) O(vdlog(n)), where n is the dataset size while v and d are maximum out-degree of the graph to be built and the number of dimensions (kmer features in a genome) of the dataset, respectively ^{21, 22}. In nearest neighbor search studies, L2 distance (Euclidean distance) is often used and is related to dataset dimension, which is very large for k-mer features of genomes ($\sim 10^{6}$). In the case of ProbMinHash distance, which is a MinHash-based method to sample the k-mer space of genomes and serve as a dimensionality reduction in approximating genomic distance, the sampled number of k-mers (or meanwise hashes) is always a constant and much smaller number (sketch size, normally around 10,000) than the total number of k-mers of a genome (number of dimensions of dataset). Therefore, d can be ignored (MinHash samples only a small fraction of k-mer space/dimension). Also, v is a small constant considering the graph structure to be built (normally smaller than 100). Therefore, the time complexity, as it was also explained in the main text, can be safely written as O(log(n)) for low dimension datasets ²². Similar rules applied for the time complexity of the graph build step, which is O(dnlog(n)) and thus, O(nlog(n)) in our case, where d is dataset dimension. Time complexity O(n + mlog(m)) of the Probminhash3a algorithm, where n is the set size (total sampled k-mer number) while m is the number of weighted elements (k-mers), is also a constant in the context of high-quality genomes given also that m is very small for prokaryotic and bacteriophage genomes (normally about 5% genomic sequences are multi-copy). Therefore, ProbMinHash time complexity is very close to O(n) in practice. For a given genome, for example, bacterial genome or viral genome, n is the sampled k-mer number of the genome, and thus also a constant number.

Note 4. Kmer-based genomic distance estimation is less accurate for distantly related genomes. For a random mutation rate $r \in (0,1)$, the probability that a k-mer (k ls length of the k-

mer) belonging to sequence X and Y is not mutated is $1 - (1 - r)^k \approx 1 - exp(-\frac{r}{k})$, indicating that $k \leq r^{-1}$, so k must be small enough to capture the mutation and also big enough to avoid underestimation (**Note 2**). This means that for close related bacterial genomes (r<0.01 for example), k could easily meet both conditions mentioned above while for distantly related genomes -for example- r=0.125 (corresponding to an ANI value of 87.5%) k <= 8, which is contradictory with k>log₄(N)=14.28 (Note 2). Thus, k-mer based method will lose accuracy for distantly related genomes. In practice, we observed that this r threshold is around 0.215 (ANI=78.5%) because mutation is not completely random as assumed above.

Note 5. Theoretical guarantee of graph based NNS search algorithms. Until recently, a theoretical analysis based on a dataset evenly distributed on an d-dimension Euclidean sphere (d≪log(n), n is the dataset size) showed that under certain conditions, there is a guarantee that the best neighbors could be found compare to brute-force distance metric comparisons ²³. However, a theoretical analysis under more general conditions, e.g., other metric space, or d not being much smaller than log(n) is still not available, to the best of our knowledge. Despite the lack of theoretical analysis under more general conditions, graph-based algorithms work well in practice, as also reflected by the large number of graph-based NNS libraries available ^{21, 24}.

Note 6. Parallelism of probminhash and HNSW. Due to ownership mechanisms of Rust (so called memory and thread safety), fearless concurrency (e.g., no data competition/race, memory and thread safety) is made possible. The crossbeam crate package was used in GSearch for communication of data among threads for task level parallelism while Rayon crate was used to do data level parallelism ²⁵. Accordingly, by default, GSearch uses all available processors/threads to make full use of multi-processor CPUs.

34

Note 7. Details of Joint Maximum Likelihood estimator in SetSketch algorithm. For cardinalities n_U and n_v of set U and V estimated by SetSketch cardinality estimation section according to equation 12 in ²⁶, the maximum likelihood method can be used to estimate J. Specifically, the log-likelihood function as a function of J is: $log \mathcal{L}(J) = D_+ log(p_b(u - vJ)) + D_- log(p_b(v - uJ)) + D_0 log(1 - p_b(u - vJ) - p_b(v - uJ)),$

where $D_+ = |\{i: K_{Ui} > K_{Vi}\}|, D_- = |\{i: K_{Ui} < K_{Vi}\}|, D_0 = |\{i: K_{Ui} = K_{Vi}\}|$ are number of registers in the sketch of U that are greater than, less than, or equal to those in the sketch of V, respectively; K_{Ui} and K_{Vi} are registers of set U and V, respectively while v and u is relative cardinalities u =

$$\frac{n_U}{n_U+n_V}$$
 and $v = \frac{n_v}{n_U+n_V}$, respectively. $p_b(x)$ is defined as

 $p_b(x) = -log_b(1 - x \frac{b-1}{b})$. The RMSE of the ML estimate is expected to be $I^{-1/2}(J)$, where I

denotes the Fisher information with respect to J for n_U and n_v :

$$I(J) = \frac{m(b-1)^2}{b^2 \log^2(b)} \left(\frac{(vb^{p_b(u-vJ)})^2}{p_b(u-vJ)} + \frac{(ub^{p_b(v-uJ)})^2}{p_b(v-uJ)} + \frac{(vb^{p_b(u-vJ)} + ub^{p_b(v-uJ)})^2}{1 - p_b(u-vJ) - p_b(v-uJ)}\right)$$

The SetSketch paper showed that the estimation error (RMSE) for J will be almost the same for

MinHash with the same number of registers m.

Supplementary Notes References

- 1. Rowe, W.P.M. et al. Streaming histogram sketching for rapid microbiome analytics. *Microbiome* **7**, 40 (2019).
- 2. Marchet, C., Iqbal, Z., Gautheret, D., Salson, M. & Chikhi, R. REINDEER: efficient indexing of k-mer presence and abundance in sequencing datasets. *Bioinformatics* **36**, i177-i185 (2020).
- 3. Brown, C.T. & Irber, L. sourmash: a library for MinHash sketching of DNA. *Journal of Open Source Software* **1**, 27 (2016).
- 4. Bovee, R. & Greenfield, N. Finch: a tool adding dynamic abundance filtering to genomic MinHashing. *Journal of Open Source Software* **3**, 505 (2018).
- 5. loffe, S. in 2010 IEEE International Conference on Data Mining 246-255 (2010).

- 6. Shrivastava, A. Simple and efficient weighted minwise hashing. *Advances in Neural Information Processing Systems* **29** (2016).
- 7. Christiani, T. DartMinHash: Fast Sketching for Weighted Sets. *arXiv preprint arXiv:2005.11547* (2020).
- 8. Wu, W., Li, B., Chen, L., Gao, J. & Zhang, C. A review for weighted minhash algorithms. *IEEE Transactions on Knowledge and Data Engineering* **34**, 2553-2573 (2020).
- 9. Koslicki, D. & Zabeti, H. Improving MinHash via the containment index with applications to metagenomic analysis. *Applied Mathematics and Computation* **354**, 206-215 (2019).
- 10. Ertl, O. in Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & amp; Data Mining 1368–1377 (Association for Computing Machinery, London, United Kingdom; 2018).
- 11. Ertl, O. Superminhash-A new minwise hashing algorithm for jaccard similarity estimation. *arXiv preprint arXiv:1706.05698* (2017).
- 12. Ertl, O. ProbMinHash A Class of Locality-Sensitive Hash Algorithms for the (Probability) Jaccard Similarity. *IEEE Transactions on Knowledge and Data Engineering*, 1-1 (2020).
- 13. Moulton, R. & Jiang, Y. in 2018 IEEE International Conference on Data Mining (ICDM) 347-356 (2018).
- 14. Yang, D., Li, B., Rettig, L. & Cudré-Mauroux, P. D²histoSketch: Discriminative and Dynamic Similarity-Preserving Sketching of Streaming Histograms. *IEEE Transactions on Knowledge and Data Engineering* **31**, 1898-1911 (2019).
- 15. Baker, D.N. & Langmead, B. Dashing 2: genomic sketching with multiplicities and localitysensitive hashing. *bioRxiv* (2022).
- 16. Ertl, O. New cardinality estimation algorithms for HyperLogLog sketches. *arXiv preprint arXiv:1702.01284* (2017).
- 17. Baker, D.N. & Langmead, B. Dashing: fast and accurate genomic distances with HyperLogLog. *Genome Biology* **20**, 265 (2019).
- 18. Pettie, S. & Wang, D. Simpler and Better Cardinality Estimators for HyperLogLog and PCSA. *arXiv preprint arXiv:2208.10578* (2022).
- 19. Ondov, B.D. et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology* **17**, 132 (2016).
- 20. Jain, C., Rodriguez-R, L.M., Phillippy, A.M., Konstantinidis, K.T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications* **9**, 5114 (2018).
- 21. Lu, K., Kudo, M., Xiao, C. & Ishikawa, Y. HVS: hierarchical graph structure based on voronoi diagrams for solving approximate nearest neighbor search. *Proceedings of the VLDB Endowment* **15**, 246-258 (2021).
- 22. Malkov, Y.A. & Yashunin, D.A. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **42**, 824-836 (2020).
- 23. Prokhorenkova, L. & Shekhovtsov, A. Graph-based Nearest Neighbor Search: From Practice to Theory. *Proceedings of the 37th International Conference on Machine Learning* **119**, 7803--7813 (2020).
- 24. Fu, C., Xiang, C., Wang, C. & Cai, D. Fast approximate nearest neighbor search with the navigating spreading-out graph. *arXiv preprint arXiv:1707.00143* (2017).
- 25. Moraza, I.E. Rust High Performance. (Packt Publishing, Birmingham, UK; 2018).
- 26. Ertl, O. SetSketch: filling the gap between MinHash and HyperLogLog. *Proc. VLDB Endow.* **14**, 2244–2257 (2021).

Supplementary File 1

user genome FastANI	/GTDB-Tk bes	st hit	GSearch	best hit	,
GWMC127 GWMC127 002 fa	asta oz	GCF	000738435 1	GCF 000	0738435 1
GWMC127_GWMC127_023 fa	asta oz	GCA	002685195 1	GCA_00	2685195 1
GWMC127_GWMC127_045_fa	asta oz	GCA	002389265 1	GCA_00	2389265 1
GWMC127_GWMC127.14 fac	sta.gz		002711735 1		2711725 1
	sta.yz		_002711733.1		1640045 1
	sta.yz		_001642945.1		1042945.1
GWMC127_GWMC127.37.fas	sta.gz	GCF_	900197625.1	GCF_900	J197625.1
GWMC127_GWMC127.47_st	ib.fasta.gz	GCA_	_003045825.1	GCA_00	3045825.1
GWMC127_GWMC127.51.fas	sta.gz	GCA_	_003483155.1	GCA_00	3483155.1
GWMC127_GWMC127.64.fas	sta.gz (GCA_	_002387615.1	GCA_00	2387615.1
GWMC131_GWMC131.002.fa	asta.gz	GCF_	_000738435.1	GCF_000)738435.1
GWMC131_GWMC131.024.fa	asta.gz (GCA_	_003045825.1	GCA_00	3045825.1
GWMC131_GWMC131.067.fa	asta.gz	GCA_	003482475.1	GCA_00	3482475.1
GWMC131 GWMC131.18.fas	sta.oz (GCA	002711735.1	GCA 00	2711735.1
GWMC132 GWMC132.003.fa	asta.oz (GCF	000738435.1	GCF 00	0738435.1
GWMC132 GWMC132 019 fa	asta oz	GCA	002711735 1	GCA 00	2711735 1
GWMC132 GWMC132 049 fa	asta oz	GCA	003483155 1	GCA_00	3483155 1
GWMC132 GWMC132 3 fast	a 0.7	GCA	002684605 1		268/605 1
GWMC132 GWMC132 53 fac	a.yz		002/83505 1		2483505 1
GWWC132_GWWC132.55.1as	sta.yz		_003463303.1	GCA_00	2463303.1
	sta.yz		_003045625.1	GCA_00	3043625.1
GWMC132_GWMC132.80.fas	sta.gz	GCF_	_000315525.1	GCF_000	J315525.1
GWMC244_GWMC244.021.fa	asta.gz (GCA_	_003486095.1	GCA_00	3486095.1
GWMC244_GWMC244.22.fas	sta.gz	GCA_	_003023665.1	GCA_00	3023665.1
GWMC244_GWMC244.28.fas	sta.gz	GCA_	_008081045.1	GCA_00	3081045.1
GWMC244_GWMC244.31.fas	sta.gz (GCA_	_013911165.1	GCA_01	3911165.1
GWMC245_GWMC245.104.fa	asta.gz	GCA_	_009921445.1	GCA_00	9921445.1
GWMC245_GWMC245.24.fas	sta.gz (GCF	003335675.1	GCF_003	3335675.1
GWMC245 GWMC245.34.fas	sta.oz (GCA	011525065.1	GCA 01	1525065.1
GWMC245_GWMC245_36 fas	sta oz (GCA	003023665 1	GCA 00	3023665 1
GWMC245_GWMC245.8 fast	a 07 (GCA	008081045 1	GCA 00	80810451
GWMC247 GWMC247 32 fas	a.gz	GCA	003023665 1		3023665 1
GWMC247_GWMC247.28 for	to gz		0000200000.1		2700205 1
	sta.yz		_002700365.1	GCA_00	2700365.1
GWWC247_GWWC247.42.185	sta.yz		_006061045.1	GCA_00	5061045.1
GWMC247_GWMC247.55_st	ib.tasta.gz	GCA_	_009919335.1	GCA_00	9919335.1
GWMC264_GWMC264.0007.	fasta.gz	GCA_	_003482475.1	GCA_00	3482475.1
GWMC264_GWMC264.10.fas	sta.gz (GCF_	_900197625.1	GCF_900)197625.1
GWMC264_GWMC264.35.fas	sta.gz (GCA_	_009919335.1	GCA_00	9919335.1
GWMC264_GWMC264.45.fas	sta.gz (GCF_	_000738435.1	GCF_000)738435.1
GWMC264_GWMC264.5.fast	a.gz (GCA_	_002390525.1	GCA_00	2390525.1
GWMC265_GWMC265.0031_	_sub.fasta.gz	GCF_	_001642945.1	GCF_00	1642945.1
GWMC265_GWMC265.26.fas	sta.gz	GCF_	900197625.1	GCF_900	0197625.1
GWMC265 GWMC265.47.fas	sta.az (GCF	000738435.1	GCF 00	0738435.1
GWMC265 GWMC265.51 fas	sta.gz	GCA	002390525.1	GCA 00	2390525.1
GWMC266_GWMC266.021_s	sub fasta oz	GCA	003045825 1	GCA 00	3045825 1
GWMC266 GWMC266 12 fas	sta oz	GCA	009919335 1	GCA 00	9919335 1
GWMC266_GWMC266_18 fac	sta gz	GOR	000107625 1		107625 1
GWMC266_GWMC266.24 for	to gz		000720425.1		720425 1
	sta.yz		_000736435.1		J736435.1
	sta.gz		_003482475.1	GCA_00	3482475.1
GWMC266_GWMC266.3.fast	a.gz	GCA_	_002390525.1	GCA_00	2390525.1
GWMC281_GWMC281.019.ta	asta.gz	GCA_	_003486095.1	GCA_00	3486095.1
GWMC281_GWMC281.021.fa	asta.gz (GCA_	_009919335.1	GCA_00	9919335.1
GWMC281_GWMC281.049.fa	asta.gz	GCA_	_003045935.1	GCA_00	3045935.1
GWMC281_GWMC281.066_s	sub.fasta.gz	GCA_	_003482475.1	GCA_00	3482475.1
GWMC281_GWMC281.22_su	ub.fasta.gz	GCA_	_003045825.1	GCA_00	3045825.1
GWMC281_GWMC281.48_st	ub.fasta.gz	GCF_	000738435.1	GCF_000	0738435.1
GWMC282 GWMC282.004.fa	asta.gz	GCA	003486095.1	GCA 00	3486095.1
GWMC282 GWMC282.006.fa	asta.oz (GCA	002691565.1	GCA 00	2691565.1
GWMC282 GWMC282 020 s	sub fasta oz	GCA	010023085 1	GCA 01	0023085 1
GWMC282 GWMC282 20 fas	ta nz (GCA	011525015 1	GCA_01	1525015 1
GWMC282 GWMC282 22 fas	sta.gz	GCE	000738/35 1	GCE 00	1738/35 1
GW/MC282 GW/MC282 27 for	sta oz		002/57055 1	GCA 00	2457055 1
	sta.yz		_002437033.1	GCA_00	2437033.1
	sta.yz		-90019/025.1	GOF_900	19/020.1
	sta.gz	GUA_		GCA_00	2000015.1
GVVMC282_GVVMC282.44.fas	sta.gz (_003045935.1	GCA_00	3045935.1
GWMC282_GWMC282.53.fas	sta.gz	GCF_	_003335675.1	GCF_00	3335675.1
GWMC282_GWMC282.56.fas	sta.gz	GCA_	_003023665.1	GCA_00	3023665.1
GWMC282_GWMC282.57.fas	sta.gz	GCA_	_002480045.1	GCA_00	2480045.1
GWMC284_GWMC284.010_s	sub.fasta.gz	GCA_	_003486095.1	GCA_00	3486095.1
GWMC284_GWMC284.039_s	sub.fasta.gz	GCA_	_009919335.1	GCA_00	9919335.1
GWMC284_GWMC284.28.fas	sta.gz (GCA_	_003023665.1	GCA_00	3023665.1

GWMC284_GWMC284.3.fasta.gz	GCF_900197605.1	GCF_900197605.1
GWMC284_GWMC284.84.fasta.gz	GCA_009923785.1	GCA_009923785.1
GWMC304_GWMC304.038.fasta.gz	GCA_009921445.1	GCA_009921445.1
GWMC304_GWMC304.24.fasta.gz	GCA_008081045.1	GCA_008081045.1
GWMC497_GWMC497.11.fasta.gz	GCA_002390485.1	GCA_002390485.1
GWMC497_GWMC497.24.fasta.gz	GCA_002457245.1	GCA_002457245.1
GWMC497_GWMC497.29.fasta.gz	GCA_009919335.1	GCA_009919335.1
GWMC497_GWMC497.34.fasta.gz	GCA_002389265.1	GCA_002389265.1
GWMC497_GWMC497.41.fasta.gz	GCF_001735715.1	GCF_001735715.1
GWMC497_GWMC497.45.fasta.gz	GCA_009936895.1	GCA_009936895.1
GWMC497_GWMC497.58.fasta.gz	GCF_900197625.1	GCF_900197625.1
GWMC497_GWMC497.6.fasta.gz	GCF_000738435.1	GCF_000738435.1
GWMC497_GWMC497.62.fasta.gz	GCA_003485335.1	GCA_003485335.1
GWMC539_GWMC539.009.fasta.gz	GCA_002691565.1	GCA_002691565.1
GWMC539_GWMC539.16.tasta.gz	GCF_000384415.1	GCF_000384415.1
GWMC539_GWMC539.23.fasta.gz	GCA_003023665.1	GCA_003023665.1
GWMC539_GWMC539.2_sub.fasta.gz	GCF_000738435.1	GCF_000738435.1
GWMC539_GWMC539.35.fasta.gz	GCA_003486095.1	GCA_003486095.1
GWMC539_GWMC539.8.1asta.gz	GCA_003045935.1	GCA_003045935.1
GWMC540_GWMC540.003.fasta.gz	GCA_003486095.1	GCA_003486095.1
GWMC540_GWMC540.012.1asta.gz GWMC540_GWMC540.18.fasta.gz GWMC540_GWMC540.28 fasta.gz	GCA_002091505.1 GCA_011525015.1	GCA_002031505.1
GWMC540_GWMC540.28.1asta.gz GWMC540_GWMC540.38.fasta.gz GWMC550_GWMC550_037_fasta.gz	GCA_003045935.1 GCE_000179255.1	GCA_003045935.1 GCE_000179255.1
GWMC550_GWMC550.037.iasta.gz GWMC550_GWMC550.14.fasta.gz GWMC550_GWMC550.28 fasta.gz	GCF_003335675.1	GCF_003335675.1
GWMC550_GWMC550.45.fasta.gz	GCA_003023665.1	GCA_003023665.1
GWMC551_GWMC551.14.fasta.gz	GCA_008081045.1	GCA_008081045.1
GWMC551_GWMC551_4 fasta.gz	GCA_008081045.1	GCA_008081045.1
GWMC552_GWMC552.021.fasta.gz	GCA_008081045.1	GCA_008081045.1

Supplementary File 2

The same query genome BRRP3A.13 for the following

Top 10 from split database 1

0 path .././GWMC 1000 HQ faa aai 635/BRRP3A.13.faa, fasta id total sequence, len : 800987

distance : 9.156E-1 answer fasta id ../database/database_1/GB_GCA_016699615.1_protein.faa answer fasta id total sequence, seq len : 906076 distance : 9.452E-1 answer fasta id ./database/database_1//3B_GCA_017984135.1_protein.faa answer fasta id total sequence, seq len : 797872 distance : 9.822E-1 answer fasta id ./database/database_1//3B_GCA_001897535.1_protein.faa Distalfoe : 9.6262-1 airswer fasta id tota sequence, seq len : 957913 distance : 9.835E-1 answer fasta id total sequence, seq len : 957913 distance : 9.835E-1 answer fasta id ../database/database_1//GB_GCA_018268055.1_protein.faa answer fasta id total sequence, seq len : 974493 distance : 9.839E-1 answer fasta id ../database/database_1//GB_GCA_016787045.1_protein.faa distance : 9:03L-1 answer fasta id total sequence, seq len : 1127633 distance : 9:839E-1 answer fasta id .../database/database_1/GB_GCA_018268135.1_protein.faa answer fasta id total sequence, seq len : 960055 answer fasta io total sequence, seq ien : 950055 distance : 9.438E-1 answer fasta id./database/database/atabase/atabase/atabase/database/database/database/database/database/database/database/database/database/database/atabase/da distance : 9.843E-1 answer fasta id ../database/database_1/0B_GCA_017303895.1_protein.faa answer fasta id total sequence, seq len : 1121215 distance : 9.844E-1 answer fasta id ../database/database_1/0B_GCA_002281875.1_protein.faa answer fasta id total sequence, seq len : 962695

Top 10 from split database 2

0 path .././GWMC 1000 HQ faa aai 635/BRRP3A.13.faa. fasta id total sequence. len : 800987

distance : 9.551E-1 answer fasta id/database/database_2/GB_GCA_016711605.1_protein.faa
answer fasta id total sequence, seq len : 1152198
distance : 9.832E-1 answer fasta id/database/database_2/GB_GCA_018267475.1_protein.faa
answer fasta id total sequence, seq len : 858052
distance : 9.846E-1 answer fasta id/database/database_2/GB_GCA_903920455.1_protein.faa
answer fasta id total sequence, seq len : 710803
distance : 9.850E-1 answer fasta id/database/database_2/RS_GCF_003384935.1_protein.faa
answer fasta id total sequence, seq len : 1036633
distance : 9.850E-1 answer fasta id/database/database_2/RS_GCF_004342685.1_protein.faa
answer fasta id total sequence, seq len : 1066355
distance : 9.851E-1 answer fasta id/database/database_2/GB_GCA_002307135.1_protein.faa
answer fasta id total sequence, seq len : 1018135
distance : 9.852E-1 answer fasta id/database/database_2/GB_GCA_018267625.1_protein.faa
answer fasta id total sequence, seq len : 901729
distance : 9.852E-1 answer fasta id/database/database_2/RS_GCF_000348685.1_protein.faa
answer fasta id total sequence, seq len : 920194
distance : 9.852E-1 answer fasta id/database/database_2/GB_GCA_018267875.1_protein.faa
annual fasta iditatel annuale ann lan (000014

answer fasta id total sequence, seq len : 822214 distance : 9.853E-1 answer fasta id ../database/database_2/GB_GCA_018267925.1_protein.faa answer fasta id total sequence, seq len : 1037058

Top 10 from database 3

0 path .././GWMC_1000_HQ_faa_aai_635/BRRP3A.13.faa, fasta_id total sequence, len : 800987

- distance : 9.711E-1 answer fasta id ./database/database_3/GB_GCA_016705125.1_protein.faa answer fasta id total sequence, seq len : 973431 distance : 9.843E-1 answer fasta id ./database/database_3/GB_GCA_903840505.1_protein.faa distance : 9.647E-1 answer fasta id total sequence, seq len : 884513 distance : 9.847E-1 answer fasta id total sequence, seq len : 884513 distance : 9.847E-1 answer fasta id ../database/database_3/GB_GCA_001567275.1_protein.faa distance : 9.647E-1 answer fasta id ../database/database_3/GB_GCA_903857255.1_protein.faa distance : 9.647E-1 answer fasta id total sequence, seq len : 903318 distance : 9.847E-1 answer fasta id .../database/database_3/GB_GCA_019136655.1_protein.faa answer fasta id total sequence, seq len : 909617 distance : 9.849E-1 answer fasta id ../database/database 3/RS GCF 900167075.1 protein.faa distance : 9.4912+1 aitsefi fasta di Judatabase/otatabase_JGB_GCA_016786925.1_protein.faa answer fasta id total sequence, seq len : 1236990 distance : 9.851E-1 answer fasta id .../database/database_JGB_GCA_016786925.1_protein.faa answer fasta id total sequence, seq len : 1087491 distance : 9.851E-1 answer fasta id../database/database./database.3/GB_GCA_016715625.1_protein.faa answer fasta id total sequence, seq len : 1063770 distance : 9.851E-1 answer fasta id ../database/database_3/GB_GCA_903828245.1_protein.faa
- answer fasta id total sequence, seq len : 615471 distance : 9.853E-1 answer fasta id ../database/database_3/GB_GCA_017303835.1_protein.faa answer fasta id total sequence, seq len : 1008010

Top 10 from database 4

0 path ../../GWMC_1000_HQ_faa_aai_635/BRRP3A.13.faa, fasta_id total sequence, len : 800987

distance : 9.540E-1 answer fasta id ../database/database_4/GB_GCA_016721245.1_protein.faa

- distance : 9.8402-1 answer fasta id total sequence, seq len : 1111343 distance : 9.829E-1 answer fasta id .../database/database_4/GB_GCA_018262795.1_protein.faa answer fasta id total sequence, seq len : 1041576 distance : 9.838E-1 answer fasta id ../database/database 4/GB GCA 002352045.1 protein.faa
- distance : 9.600L-1 answer fasta id total sequence, seq len : 1030255 distance : 9.843E-1 answer fasta id .../database/database_4/GB_GCA_018268195.1_protein.faa answer fasta id total sequence, seq len : 923024
- distance : 9.843E-1 answer fasta id ../database/database/database.4/GB_GCA_903861055.1_protein.faa answer fasta id total sequence, seq len : 1018870 distance : 9.846E-1 answer fasta id ../database/database_4/GB_GCA_016183765.1_protein.faa

answer fasta id total sequence, seq len : 1094501 distance : 9.847E-1 answer fasta id ../database/database_4//BB_GCA_004295015.1_protein.faa answer fasta id total sequence, seq len : 1160623 distance : 9.847E-1 answer fasta id ../database/database_4/RS_GCF_009363055.1_protein.faa distance : 9.847E-1 answer fasta id Jdatabase/database_4/GB_GCA_016786165.1_protein.faa answer fasta id total sequence, seq len : 662789 distance : 9.847E-1 answer fasta id Jdatabase/database_4/GB_GCA_016267725.1_protein.faa distance : 9.849E-1 answer fasta id Jdatabase/database_4/GB_GCA_016788165.1_protein.faa

answer fasta id total sequence, seq len : 1025440

Top 10 from split database 5

0 path .././GWMC_1000_HQ_faa_aai_635/BRRP3A.13.faa, fasta_id total sequence, len : 800987

distance : 9.539E-1 answer fasta id ../database/database_5/GB_GCA_016719635.1_protein.faa answer fasta id .:/database/database_5/GB_GCA_001567165.1_protein.faa distance : 9.841E-1 answer fasta id .:/database/database_5/GB_GCA_001567165.1_protein.faa answer fasta id total sequence, seg len : 638956 distance : 9.842E-1 answer fasta id ../database/database_5/RS_GCF_900100625.1_protein.faa answer fasta id total sequence, seq len : 872517 distance : 9.842E-1 answer fasta id ../database/database/database_JGME_GCA_016722375.1_protein.faa distance : 9.844E-1 answer fasta id Jdatabase/database_5/GB_GCA_016787165.1_protein.faa distance : 9.844E-1 answer fasta id Jdatabase/database_5/GB_GCA_016787165.1_protein.faa answer fasta id total sequence, seq len : 1849407 distance : 9.844E-1 answer fasta id Jdatabase/database_5/GB_GCA_002455375.1_protein.faa distance : 9.845E-1 answer fasta id ./database/database_5/GB_GCA_03883945.1_protein.faa answer fasta id total sequence, seq len : 948379 distance : 9.845E-1 answer fasta id ./database/database_5/GB_GCA_03883945.1_protein.faa distance : 9.845E-1 answer fasta id ./database/database_5/GB_GCA_015655125.1_protein.faa distance : 9.845E-1 answer fasta id ../database/database_b/GB_dCA_U1bob1/zb.1_protein.taa answer fasta id total sequence, seq len : 1100101 distance : 9.847E-1 answer fasta id ../database/database_5/GB_GCA_017495115.1_protein.faa answer fasta id total sequence, seq len : 763790 distance : 9.847E-1 answer fasta id ../database/database_5/RS_GCF_002217405.1_protein.faa

answer fasta id total sequence, seq len : 770090

Top 10 from one large database

0 path ../GWMC 1000 HQ faa aai/BRRP3A.13.faa.gz, fasta id total sequence, len : 800987

distance : 9.156E-1 answer fasta id ./protein_faa_reps_r207/all/GB_GCA_016699615.1_protein.faa answer fasta id total sequence, seq len : 906076 distance : 9.452E-1 answer fasta id ./protein_faa_reps_r207/all/GB_GCA_017984135.1_protein.faa answer fasta id total sequence, seq len : 797872 distance : 9.530E-1 answer fasta id ./protein_faa_reps_r207/all/GB_GCA_016719635.1_protein.faa answer fasta id total sequence, seq len : 104205 distance : 9.540E-1 answer fasta id ./protein_faa_reps_r207/all/GB_GCA_016719635.1_protein.faa answer fasta id total sequence, seq len : 111343 distance : 9.551E-1 answer fasta id ./protein_faa_reps_r207/all/GB_GCA_016711605.1_protein.faa answer fasta id total sequence, seq len : 111343 distance : 9.551E-1 answer fasta id ./protein_faa_reps_r207/all/GB_GCA_016705125.1_protein.faa answer fasta id total sequence, seq len : 1152189 distance : 9.422E-1 answer fasta id ./protein_faa_reps_r207/all/GB_GCA_016705125.1_protein.faa answer fasta id total sequence, seq len : 1152189 distance : 9.822E-1 answer fasta id ./protein_faa_reps_r207/all/GB_GCA_001897535.1_protein.faa answer fasta id total sequence, seq len : 1152189 distance : 9.829E-1 answer fasta id ./protein_faa_reps_r207/all/GB_GCA_001897535.1_protein.faa answer fasta id total sequence, seq len : 157813 distance : 9.829E-1 answer fasta id ./protein_faa_reps_r207/all/GB_GCA_018262795.1_protein.faa answer fasta id total sequence, seq len : 1041576 distance : 9.832E-1 answer fasta id ./protein_faa_reps_r207/all/GB_GCA_018267475.1_protein.faa distance : 9.156E-1 answer fasta id ../protein_faa_reps_r207/all/GB_GCA_016699615.1_protein.faa

distance : 9.832E-1 answer fasta id ../protein_faa_reps_r207/all/GB_GCA_018267475.1_protein.faa answer fasta id total sequence, seq len : 858052 distance : 9.835E-1 answer fasta id ../protein_faa_reps_r207/all/GB_GCA_018268055.1_protein.faa answer fasta id total sequence, seq len : 974493

Supplementary File 3

Commands used for the pieces of software compared to GSearch

Sourmash (v4.2.3): (1) create sketches: find ./gtdb v207/ -name *.fna.gz' | parallel -j 24 "sourmash sketch dna -p k=16,noabund {} -o {.}.sourmash"

sourmash sketch dna -p k=16,noabund OceanDNA-b42278.fa -o OceanDNA-b42278.fa.sourmash (2) Create SBT index from the database sketches in (1): sourmash index -k 16 gtdb_sourmash_index --from-file gtdb_v207_sourmash_sketch.sig.txt -q

Note: here gtdb_v207_sourmash_sketch.sig.txt is a list of *.sourmash file from step (1) (3) Search sketch against (2) SBT index: sourmash search --threshold 0.01 --num-results 50 OceanDNA-b42278.fa.sourmash gtdb sourmash index.sbt.zip -o OceanDNAb42278.sourmash.dist.txt

Note: here OceanDNA-b42278.fa.sourmash is obtained by similar commands in (1) but only for this one query genome file. This step takes about 15 minutes for just one query genome (only one thread can be used) in sourmash. We use GNU parallel to run 24 queries at a time.

find ./Tara_query/ -name '*.fna.sourmash' | parallel -j 24 "sourmash search --threshold 0.01 --num-results 50 {} qtdb sourmash index.sbt.zip -o {.}.sourmash.dist.txt"

Mash (v2.3):

(1) Create sketches: For nucleotide: mash sketch -p 24 -s 12000 -k 16 -o ./query_genome -l query_name.txt mash sketch -p 24 -s 12000 -k 16 -o ./reference_genome -l reference_name.txt

For Amino Acid: mash sketch -p 24 -s 12000 -k 7 -a -o ./query_genome_aa -l query_name_aa.txt mash sketch -p 24 -s 12000 -k 7 -a -o ./reference_genome_aa -l reference_name_aa.txt (2) Compute distance mash dist -p 24 query_genome.msh reference_genome.msh > dist.txt

mash dist -p 24 query_genome_aa.msh reference_genome_aa.msh > dist aa.txt Dashing (v1.0.2-4-g0635): default HyperLogLog option with Ertl's Joint MLE estimator (--ertl-imle) or default estimator (1) create sketches: dashing sketch -k 16 --nthreads 24 -S 14 --suffix dashing_hll -F ./GTDB_nt_name.txt

dashing sketch -k 16 --nthreads 24 -S 14 --suffix dashing_hll -F ./Query_name.txt Note: here GPD_test10000_nt_name.txt is a list of all bacteria database genomes, -S 14 (14*64) is similar to (or better) -s 12000 in Mash above (2) Compute distance:

dashing dist -F ./GTDB_nt_dashing_sketch.txt -Q Query_nt_name.txt --nthreads 24 --presketched -O dist.txt Note: here GTDB_nt_dashing_sketch.txt is a list of .hll sketches from step (1) while Query_nt_name.txt is the list of sketches of query genomes generated in similar way to (1).

Dashing 2 (version v2.1.11):

dashing2 dashing2 sketch -k 16 -S 12000 --threads 24 --pminhash -Q Tara_Ocean_MAGs_name.txt -F name.txt --cmpout Tara_Ocean_MAGs_Dashing2_pminhash_NCBI.txt

To make fair comparison with GSearch (default probminhash), we use the probminhash (--pminhash) option. BinDash (v0.2.1): bindash sketch --kmerlen=16 --outfname=Tara_bindash --sketchsize64=250 --listfname=Tara_Ocean_MAGs_name.txt --dens=1 -bbits=32 --minhashtype=2 --nthreads=24

bindash dist --nthreads=24 --outfname=Tara_refseq_dist.txt Tara_bindash refseq_ncbi_all_bindash_optdens

FastANI (v1.33):

fastANI --ql query_name.txt --rl reference_name.txt -t 24 -o ANI.txt

Note: guery_name.txt and reference_name.txt are files storing genome path.

Supplementary File 4

query genome name	GSearch taxon	omy guery genome name Sourmash ica taxonomy		-
AUBR2B.34.fasta.gz	d_Bacteria; p_	_Bacteroidetes; c_unassigned; o_unassigned; f_unassigned; g_unassigned; s_Bacteroidetes bacterium OLB8 AU	UBR2B.34.fasta.gz	"found, Bacteria, Bacteroidetes, unassigned, unassigned, unassigned, Bacteroidetes bacterium OLB8,"
AUBR2C.63.fasta.gz	d_Bacteria; p_	_Bacteroidetes; c_unassigned; o_unassigned; f_unassigned; g_unassigned; s_Bacteroidetes bacterium OLB8 AU	UBR2C.63.fasta.gz	"found, Bacteria, Bacteroidetes, unassigned, unassigned, unassigned, Bacteroidetes bacterium OLB8,"
AUEP2BC.56.fasta.gz	d_Bacteria; p_		OCH26.2.tasta.gz A245	Touno, bacteria, Proteobacieria, Gammaproteobacteria, iniorichales, iniorichales, iniorichales, inioritik, exelicionali, AUEP28C, 55, fasta, az. "found Bacteria, Bacteria, Bacterio dets, unassioned unassioned unassioned successioned society and the society of t
AUXP4AB.87.fasta.gz	d_Bacteria; p_	Proteobacteria; c_Deltaproteobacteria; o_Myxococcales; f_Anaeromyxobacteraceae; g_Anaeromyxobacter; s_unas	assigned	AUXP4AB.87.fasta.gz "found, Bacteria, Proteobacteria, Deltaproteobacteria, Myxococcales, Anaeromyxobacteraceae, Anaeromyxobacter, unassigned,"
AUXP4AB.163.fasta.gz	d_Bacteria; p_	_Proteobacteria; c_Alphaproteobacteria; o_Rhizobiales; f_Rhizobiaceae; g_Kaistia; s_Kaistia sp. SCN 65-12 AU	UXP4AB.163.fasta.gz	"lound, Bacteria, Proteobacteria, Alphaproteobacteria, Rhizobiales, Rhizobiaceae, Kaistia, Kaistia sp. SCN 65-12,"
BRBH3C 63 fasta oz	d_Bacteria; p_ d_Bacteria; p_	_actinobacteria; cactinobacteria; ocoryneoacteriales; icorooniaceae; gskermania; sskermania piniformis Au Racteroidetes; c_unassigned; o_unassigned; f_unassigned; g_unassigned; s_Racteroidetes hacterium ADurb RinA;	0AP4Ab.12.tasta.gz A245	"round, bacteria, Acunobacteria, Acunobacteria, Convectariaes, opronacteria, Skermania, Skermania, Skermania prinormis," BBRH3C K3 fasta oz. "forund Bacteria Bacteria Bacteria Bacteria del unassigned unassigned unassigned Bacteria del subacteria del su
BRBH1C.68.fasta.gz	d_Bacteria; p_	Bacteroidetes; c_unassigned; o_unassigned; f_unassigned; g_unassigned; s_Bacteroidetes bacterium ADurb.BinA	A245	BRBH1C.68.fasta.gz "Tound, Bacteria, Bacteroidetes, unassigned, unassigned, unassigned, unassigned, Bacteroidetes bacterium ADurb. BinA245,"
BRBH4C.97.fasta.gz	d_Bacteria; p_	_Bacteroidetes; c_unassigned; o_unassigned; f_unassigned; g_unassigned; s_Bacteroidetes bacterium ADurb BinA	A245	BRBH4C.97.fasta.gz "found, Bacteria, Bacterioidetes, unassigned, unassigned, unassigned, unassigned, Bacterioidetes bacterium ADurb. BinA245,"
BHOS2C 21.tasta.gz BBOS2C 42 fasta oz	d_Bacteria; p_ d_Bacteria; p_	_Spirochaetes; cSpirochaetia; oSpirochaetales; !Spirochaetaceae; gSpirochaeta; sSpirochaeta sp. EBM-43BH Proteobacteria; cDeltanoteobacteria; oDesulfobacterales; !Desulfobulbaceae; oDesulfobulbu; sDesulfobulb	HUS2C:21.tasta.gz bus pronionicus	"Tound, Bacteria, Spirochaetia, Spirochaetia, Spirochaetiaes, Spirochaetia, Spirochaet
BRRP2B.69.fasta.gz	d_Bacteria; p_		RRP2B.69.fasta.gz	¹ Ourd, Bacteria, Bacteria/Belevidete, unassigned, unassigned, unassigned, Bacterio/detes bacterium OLBB,"
BRRP3A.53.fasta.gz	d_Bacteria; p_	_Bacteroidetes; c_unassigned; o_unassigned; f_unassigned; g_unassigned; s_Bacteroidetes bacterium OLB8 BR	RRP3A.53.fasta.gz	"found, Bacteria, Bacterioidetes, unassigned, unassigned, unassigned, Bacterioidetes bacterium OLB8,"
BRRP3A.37.fasta.gz	d_Bacteria; p_	_Chlorobi; c_unassigned; o_unassigned; f_unassigned; g_unassigned; s_Chlorobi bacterium OLB7 BF	RRP3A.37.fasta.gz	"found, Bacteria, Chlorobi, unassigned, unassigned, unassigned, Chlorobi bacterium OLB7."
CLVPM1.1.fasta.gz	d_Bacteria; p_	Proteobacteria; cGammaproteobacteria; oKannomonabales; tunassigned; gunassigned; sunassigned Proteobacteria: cGammaproteobacteria: oThiotrichales: fThiotrichaceae: gThiothrix: sThiothrix eikelboomiiCL	EVPM1.1.fasta.gz	"ound, Bacteria, Proteobacteria, Gammaproteobacteria, Antinomonapaes, unassigneo, unassigneo, unassigneo, inassigneo, inassigneo, "" "ound, Bacteria, Proteobacteria, Gammaproteobacteria, Thiotrichales, Thiotrichasea, Thiotrichastea, Thiotricha
CLVPM1.2.fasta.gz	d_Bacteria; p_	Proteobacteria; c_Gammaproteobacteria; o_Thiotrichales; f_Thiotrichaceae; g_Thiothrix; s_Thiothrix eikelboomiiCL	LVPM1.2.fasta.gz	"found, Bacteria, Proteobacteria, Gammaproteobacteria, Thiotrichales, Thiotrichaceae, Thiothrix, Thiothrix eikelboomii,"
CLVPM2.2.fasta.gz	d_Bacteria; p_	_Proteobacteria; cGammaproteobacteria; oThiotrichales; fThiotrichaceae; gThiothrix; sThiothrix eikelboorniiCL	LVPM2.2.fasta.gz	"found, Bacteria, Proteobacteria, Gammaproteobacteria, Thiotrichales, Thiotrichaceae, Thiothrix, Thiothrix eikelboornii,"
CLVPM3.2.tasta.gz	d_Bacteria; p_	_Proteobacteria; cGammaproteobacteria; oIniotnchales; tIniotnchaceae; gIniotnnx; sIniotnrx eikelboomii GL Proteobacteria; cGhanardeobacteria; ounaeeinnad; tunaeeinnad; ounaeeinnad; eGhanardeobacteria; bacteri	LVPM3.2.tasta.gz rium ADurb BinA280	"Tound, Bacteria, Proteobacteria, Gammaproteobacteria, Throthchales, Thiothchales, Iniothnix, Thiothnix elikelboomi," (2) VDM3 54 fasta oz. "Tound Bactaria: Dirutachastaria a Abhanztatoria unsestimad unsestimad labhanztanbactaria bactaria m
ADurb.BinA280,"	ubabiena, p_	_rioleobaciena, cAphaproleobaciena, obhassigned, 1bhassigned, gbhassigned, sAphaproleobaciena bacien	num Aburb.biriA280	CCYFMSJ04.iassgreu, hundbautena, Proteobautena, Aphiaproteobautena, unassgreu, unassgreu, unassgreu, Aphiaproteobautena bautenum
CLVPT1.2.fasta.gz	d_Bacteria; p_	_Proteobacteria; cGammaproteobacteria; oThiotrichales; fThiotrichaceae; gThiothrix; sThiothrix eikelboomii CL	LVPT1.2.fasta.gz	"found, Bacteria, Proteobacteria, Gammaproteobacteria, Thiotrichales, Thiotrichaceae, Thiothrix, Thiothrix eikelboomii,"
CLVPT1.1.fasta.gz	d_Bacteria; p_	_Proteobacteria; c_Gammaproteobacteria; o_Thiotrichales; f_Thiotrichaceae; g_Thiothrix; s_Thiothrix eikelboomiiCL	LVPT1.1.fasta.gz	"found, Bacteria, Proteobacteria, Gammaproteobacteria, Thiotrichales, Thiotrichaese, Thiothrix, elikelboomi,"
CLVP12.2.Iasta.gz	d Bacteria: p	Proteobacteria: cGammaproteobacteria: oThiotrichales; fThiotrichaeea; gThiotrix: s_Thiotrix ekkelboomiCL	LVP12.2.Idsta.gz	Tourio, sacenta, rotebacenta, sammaprovebacenta, monicinaes, monicinaese, monicina, monicinae electronim, "
CNBJ6BC.12.fasta.gz	d_Bacteria; p_	Bacteroidetes; c_unassigned; o_unassigned; 1_unassigned; g_unassigned; s_Bacteroidetes bacterium ADurb.BinA:	A245	CNBJ6BC.12.fasta.gz "found, Bacteria, Bacterioidetes, unassigned, unassigned, unassigned, unassigned, Bacterioidetes bacterium ADurb. BinA245,"
CNCD4C.3.fasta.gz	d_Bacteria; p_	_Bacteroidetes; c_unassigned; o_unassigned; f_unassigned; g_unassigned; s_Bacteroidetes bacterium ADurb.Bin1	141	CNCD4C.3 fasta.gz Tound Bacteria, Bacteroidetes, unassigned, unassigned, unassigned, unassigned, Bacteroidetes bacterium ADurb. Bin 141,"
CNDU4C.33.tasta.gz	d_Bacteria; p_	_bacteroidetes; c_unassigned; o_unassigned; !_unassigned; g_unassigned; s_bacteroidetes bacterium Aburb.bin1- Actionhactaria: c_Actionhactaria: o_Connehactarialae: {_Microhactariaceae: o_Mucrohactarium: e_unaesionad_CN	NDI 1AC 13 feets or	CNCJ4C33.1asta.gz Tound, Bacteria, Bacterioidetes, inassigned, ina
CNDL1BC.25.fasta.gz	d Bacteria: p	Actinobacteria: c Actinobacteria: o Corvnebacteriales; Mycobacteriaceae: g Mycobacterium; s unassigned Ch	NDL1BC.25.fasta.gz	"Cund. Bacteria Actinobacteria. Convebacteriales. Woobacteriaceae. Woobacterium.unassigned."
CNDL2AC.49.fasta.gz	d_Bacteria; p_	Bacteroidetes; c_Flavobacteriia; o_Flavobacteriales; t_Flavobacteriaceae; g_Flavobacterium; s_Flavobacterium can	auense	CNDL2AC.49.fasta.gz "found, Bacteria, Bacterioidetes, Flavobacteria, Flavobacteriales, Flavobacteriaceae, Flavobacterium, Flavobacterium cauense,"
CNDL2AC.4.fasta.gz	d_Bacteria; p_	_Bacteroidetes; cSphingobacteria; oSphingobacteriales; funassigned; gunassigned; sSphingobacteriales bact	sterium 39-19	CNDL2AC.4.fasta.gz "found Bacteria Bacteriodetes, Sphingobacteriales, unassigned, unassigned, Sphingobacteriales bacterium 39-19," CNUNIC 201 Itanto gg. Towned Bacteria, Bacteriodetes, Sphingobacteriales, unassigned, unassigned, Sphingobacteriales bacterium 39-19,"
CNJN1C 22 fasta oz	d Bacteria: p	Proteobacteria; c_Populmicrobinado, c_Populmicrobinado, i_microbinado estas (icrostitix parviceta	CNUNIC 22 fasta az "inund Bachela Proteihacheda Alhannehechachera Shihonomonadaneae unassimed inassimed "
CNJN2C.1.fasta.gz	d_Bacteria; p	Bacteroidetes; c unassigned; o unassigned; f unassigned; g unassigned; s Bacteroidetes bacterium ADurb.Bin1	141	CNJN2C.1.fasta.gz Tound, Bacteria, Bacterioidetes, unassigned, unassigned, unassigned, unassigned, Bacterioidetes bacterium ADurb. Bin141,"
CNJN4C.36.fasta.gz	d_Bacteria; p_	_Bacteroidetes; c_unassigned; o_unassigned; f_unassigned; g_unassigned; s_Bacteroidetes bacterium ADurb.Bin1+	141	CNJN4C.36.fasta.gz "found, Bacteria, Bacterioidetes, unassigned, unassigned, unassigned, unassigned, Bacterioidetes bacterium ADurb. Bin141,"
CNXA3C 36 fasta.gz	d_Bacteria; p_ d_Bacteria: p_	Proteopactena; c_Alphaproteopactena; o_Hnodobacterales; f_Rhodobacteraceae; g_unassigned; s_unassigned; s_unassigned; f_unassigned; a_unassigned; s_Bactervirtetee bacterium ADurb BinA	www.X2C.18.tasta.gz A245	Tourno, bacteria, introecosceria, Aphaproteco acteria, Hhodobacteriales, Hhodobacteriales, enassigned, unassigned, " "DNXABC 36 fasta oz." "Fruind Racteria Racteriordetas unassinned unassinned unassigned unassigned,"
CNXM4C.17.fasta.gz	d_Bacteria: p		NXM4C.17.fasta.gz	"Jound, Bacteria, Bacterioidetes, unassigned, unassigned, unassigned, Bacterioidetes bacterium OLB8,"
CNXM3C.28.fasta.gz	d_Bacteria; p_	_Bacteroidetes; c_unassigned; o_unassigned; f_unassigned; g_unassigned; s_Bacteroidetes bacterium OLB8 CN	NXM3C.28.fasta.gz	"found, Bacteria, Bacteroidetes, unassigned, unassigned, unassigned, Bacteroidetes bacterium OLB8,"
COCL1B.006_sub.fasta.ga	zd_Bacteria; p_	_Proteobacteria; cGammaproteobacteria; ounassigned; funassigned; gunassigned; sunassigned CC	OCL1B.006_sub.fasta.gz	"Tound, Bacteria, Proteobacteria, Gammaproteobacteria, unassigned, unassigned, unassigned, unassigned, "
COCLIB.11.tasta.gz	d_Bacteria; p_	_rroteobacteria; cGammaproteobacteria; oKammomoriadales; funassigned; gunassigned; sunassigned; C Proteobacteria; cBetannoteobacteria; o_Buikholderiales; funassigned; sunassigned; sunassigned; G	OCLIB.II.tasta.gz	"round sacteria, Proteosacteria, cammaproteosacteria, Xannonmonadaes, unassigned, unassigned, unassigned, " "fuund Bacteria Proteosacteria Reamplechacteria Reurkholderiales unassigned unassigned unassigned"
COCL2B.15.fasta.gz	d_Bacteria; p	Bacteroidetes; c_unassigned; o_unassigned; 1_unassigned; g_unassigned; s_Bacteroidetes bacterium ADurb.BinA	A395	COCL2B.15.fasta.gz "Yound, Bacteria, Bacterioidetes, unassigned, unassigned, unassigned, unassigned, Bacterioidetes bacterium ADurb. BinA395,"
COCL2B.60.fasta.gz	d_Bacteria; p_	_Proteobacteria; cDeltaproteobacteria; oDesulfobacterales; fDesulfobulbaceae; gDesulfobulbus; sDesulfobulb	bus propionicus	COCL2B.60.fasta.gz "found, Bacteria, Proteobacteria, Detaproteobacteria, Desulfobaltos, Desulfobulbus, Desulfob
COCL3C.005_sub.tasta.g	zd_Bacteria; p_	Proteobacteria; c_unassigned; o_unassigned; t_unassigned; g_unassigned; s_unassigned C.	OCL3C.005_sub.tasta.gz	"Tound, Bacteria, Proteobacteria, unassigned, unassigned, unassigned, unassigned, "
COCL3C.3.fasta.gz	d Bacteria: p	Proteobacteria: C Betaproteobacteria: o Burkholderiales: f unassigned; g unassigned; s unassigned CC	OCL3C.3.fasta.gz	Tourin backeria, Protecoacteria, Bearrotecoacteria, Alarin kontakaares, anassined, unassigned, unassigned, masteria, Protecoacteria, Bearrotecoacteria, Berkholderiales, unassigned, unassigned, unassigned, anassined, "
-			-	
DEHE1C.2.fasta.gz	d_Bacteria; p_	_Proteobacteria; cGammaproteobacteria; oThiotrichales; fThiotrichaceae; gThiothrix; sThiothrix eikelboomiiDE	EHE1C.2.fasta.gz	"found, Bacteria, Proteobacteria, Gammaproteobacteria, Thiotrichales, Thiotrichaceae, Thiothrix, elikelboornii,"
DERA1C 41 feets or	d_Bacteria; p_	_Proteobacteria; c_betaproteobacteria; o_burkholdenales; r_unassigned; g_unassigned; s_unassigned CC Proteobacteria; c_Gammanytachacteria; o_unassigned; f_Compatibilitacteriassic; a_Candidatus Contacthibacter; s	Candidatus Contandoh	Touno, bacteria, rhoteoacceria, betaproteoacceria, burknooenaies, unassigneo, unassigneo, inassigneo, inas
Contendobacter, Candidate	us Contendobact	er odensis,"		
ITLF2C.7.fasta.gz	d_Bacteria; p_	_Proteobacteria; c_Betaproteobacteria; o_unassigned; f_unassigned; g_unassigned; s_unassigned TL	LF2C.7.fasta.gz	"found, Bacteria, Proteobacteria, Betaproteobacteria, unassigned, unassigned, unassigned, unassigned, "
NZBL2.101.fasta.gz	d_Bacteria; p_	Bacteroidetes; c Bacteroida; o unassigned; 1 unassigned; g unassigned; s Bacteroida bacterum GaUU//bSNZ	ZBL2.101.fasta.gz	"Tound, Bacteria, Bacteriodetes, Bacteriode, unassigned, unassigned, unassigned, Jacteria de Bacterium (BaUU/1558,"
ITLP26.6.1asta.gz	d_Bacteria; p_	Proteobacteria; cbetaproteobacteria; ounassigned; funassigned; gunassigned; sunassigned TIL Proteobacteria; cBetaproteobacteria; ounassigned; funassigned; ounassigned; sunassigned TIL	MC3C 16 fasta oz	"round Bacteria Proteobacteria betaproteobacteria unassigned, unassigned, unassigned, unassigned, "
	d Destades			
ITRB1C.37.fasta.gz	d_bacteria; p_	_Proteobacteria; cbetaproteobacteria; oHnodocyclales; tHnodocyclaceae; gInauera; sInauera prenylacetica	1	ITRB1C.37.fasta.gz "found,Bacteria,Proteobacteria,Betaproteobacteria,Rhodocyclales,Rhodocyclaceae,Thauera,Thauera phenylacetica,"
ITRB1C.37.fasta.gz ITRB1C.24.fasta.gz	d_Bacteria; p_ d_Bacteria; p_	_Proteodacteria; cbetaproteodacteria; ohnodocyclales; thnodocyclalesae; ginauera; sinauera pnenylacerica _Bacteroidetes; cunassigned; ounassigned; tunassigned; gunassigned; sBacteroidetes bacterium ADurb.Bin1+	a 141	ITFB1C.37.fasta.gz "found, Bacteria, Proteobacteria, Betaproteobacteria, Rhodocyclales, Rhodocyclaesae, Thauera, Thauera
ITRB1C.37.fasta.gz ITRB1C.24.fasta.gz NZDM2.40.fasta.gz	d_Bacteria; p_ d_Bacteria; p_ d_Bacteria; p_	Proteobacteria; C_Detaproteobacteria; O_Indoco/plates; L_Indoco/platese; G_Inauera; S_Inauera; S	a 141 ZDM2.40.fasta.gz ZDM2.121 feete en	ITTBE1C3 faista ground Bacteria Protosobacteria Betepotebacteria Photosobacteria Photosobacter
ITRB1C.37.fasta.gz ITRB1C.24.fasta.gz NZDM2.40.fasta.gz NZDM2.131.fasta.gz SADP3.8.fasta.gz	d_Bacteria; p d_Bacteria; p d_Bacteria; p d_Bacteria; p d Bacteria; p	Proteodadenia ;energiprocedicacenta; 0m/docuycases;minococycaizea;minococycaizea;minococycaizea;minococycaizea;minococycaizea;minococycaizea;minococycaizea;minococycaizea;minococycaizea;m	s 141 IZDM2.40.fasta.gz IZDM2.131.fasta.gz Candidatus Competibact	ITTRB 10.27 January Tood Taberina Protectatorian Baltgrontedizatinia (Pricodocyclates, Pricodoc, Taumar, Tausara, Panylaces, Carl TTRB 10.24 Anara, S. Tood Sackerta Approximative Academic Approximation (Constraints, Protectatorian Approximation), Constraints, Panylaces Approximation, Constraints,
ITRB1C.37.fasta.gz ITRB1C.24.fasta.gz NZDM2.40.fasta.gz NZDM2.131.fasta.gz SADP3.8.fasta.gz Competibacter,Candidatus	d_Bacteria; p d_Bacteria; p d_Bacteria; p d_Bacteria; p d_Bacteria; p s Competibacter	Interessations (a 141 ZDM2.40.fasta.gz ZDM2.131.fasta.gz _Candidatus Competibact	ITTRB 10.27 Jarlan g. "Tound Bacteria Rotechaderia Belaporteofizateria (Rhodocyclases, Rhouese, Thauera, Thauras (Phenylaeotea, * TIRB 10.27 Jatala g. "Tound Bacteria Rotechadeus, aussigned, aussigned, Janasigned, Janasign
ITRB1C.37.fasta.gz ITRB1C.24.fasta.gz NZDM2.40.fasta.gz NZDM2.131.fasta.gz SADP3.8.fasta.gz Competibacter,Candidatus SAKB5.19.fasta.gz	d_Bacteria; p d_Bacteria; p d_Bacteria; p d_Bacteria; p d_Bacteria; p s Competibacter d_Bacteria; p	Protectuaries C., Designotectuaries C., Inductorybask, L., Inductybasker, E., Takutybasker, B., Takut, B., Takut, B., Takut, B., Takut, S.,	a 141 ZDM2.40.fasta.gz ZDM2.131.fasta.gz _Candidatus Competibact acterium UTPRO2 Accurum/literature.gr BA C	ITTRE IG. 21 fals g "ound, Elacteria, Protectacteria, Elisophie Casteria, Prico Casteria, Price Casteria, Pric
ITHB1C.37.fasta.gz ITHB1C.24.fasta.gz NZDM2.40.fasta.gz NZDM2.131.fasta.gz SADP3.8.fasta.gz Competibacter,Candidatus SAKB5.19.fasta.gz SAKR2.109.fasta.gz Accumulibacter.Candidatus	d_Bacteria; p d_Bacteria; p d_Bacteria; p d_Bacteria; p d_Bacteria; p s Competibacter d_Bacteria; p d_Bacteria; p is Accumulibacte	Detectoblets - Lansangerid - L	a 141 IZDM2.40.fasta.gz IZDM2.131.fasta.gz Candidatus Competibact acterium UTPRO2 Accumulibacter sp. BA-9	TIRBIG 22.1 States g "Nound, Bacterina Revelocitations (Rivologycieles, Phoseber, Thusers, Thusers
IFHB1C.37.fasta.gz IFHB1C.24.fasta.gz NZDM2.40.fasta.gz NZDM2.40.fasta.gz SADP3.8.fasta.gz Competibacter,Candidatus SAKB5.19.fasta.gz SAKB5.20.fasta.gz Accumulibacter,Candidatus SAZK3.32.fasta.gz	d_Bacteria; p d_Bacteria; p d_Bacteria; p d_Bacteria; p d_Bacteria; p s Competibacter d_Bacteria; p d_Bacteria; p s Accumulibacter d_Bacteria; p	Podcebadenia C. podkytołoście z umasignej C. prospisate C. procesje C. procesj	a 141 IZDM2.40.fasta.gz IZDM2.131.fasta.gz .Candidatus Competibact acterium UTPRO2 .Accumulibacter sp. BA-5 ium Ga0077550	TIRBI C3. Tatala g "tourd, Elactrica Enderdise Landardise Calculations (Phodocyclases, Thusane, Th
IFIBITC.37.fasta.gz IFIBITC.24.fasta.gz NZDM2.40.fasta.gz NZDM2.131.fasta.gz SAVB3.19.fasta.gz SAVB5.19.fasta.gz SAVB5.19.fasta.gz SAVR2.109.fasta.gz Accumulibacter,Candidatu SAVR2.209.fasta.gz SEGL2C.1.fasta.gz	d_Bacteria; p d_Bacteria; p	networkset _ Designeed can be unasigned _ unasigned _ unasigned _ unasigned _ unasigned & unasigned _	a 141 ZDM2.40.fasta.gz Candidatus Competibact acterium UTPRO2 Accumulibacter sp. BA-5 ium Ga0077550 ficrothrix parvicella	TIRBIT C2 Tarling and the set of
ITHBI C.37. fasta.gz ITHBI C.24. (kasta.gz NZDM2.40. fasta.gz NZDM2.10. fasta.gz SADP3.8. fasta.gz SAVR5.19. fasta.gz SAVR5.19. fasta.gz SAVR5.2109. fasta.gz SAVR5.2109. fasta.gz SAVR5.2109. fasta.gz SESDIC.50. fasta.gz SESDIC.50. fasta.gz	d_Bacteria; p d_Bacteria; p	networkstein - Destagioperio - Land Landon - Joseph - Toronalization - Tor	a Id1 ZDM2.131.fasta.gz Candidatus Competibact acterium UTPRO2 Accumulibacter sp. BA-5 ium Ga0077550 ficrofthrix parvicella Ficrofthrix parvicella FSIPQC 72 fasta.gz	TIRBI C3.2 Tatalag Tound Exterine Antechnickeus Engingender Statistics Riverbodeckeuses. Thusers, Thusers and Panelysockeux, " TIRBI C3.2 Tatalag Tound Exterine Antechnickeus Langarourge La
ITHBI C.37.fasta.gz ITHBI C.24.fasta.gz NZDM2.40.fasta.gz NZDM2.40.fasta.gz SADP3.8.fasta.gz Competibacter, Candidatu SAKB5.19.fasta.gz SAKR2.109.fasta.gz SAKR2.109.fasta.gz SEGL2C.1.fasta.gz SESDIC.50.fasta.gz SESDIC.50.fasta.gz SESDIC.50.fasta.gz	d_Bacteria; p. d_Bacteria; p.	Telephonetesis C., Desagnored care de l'accordinates en la construction de la construction	a ZDM2.431.fasta.gz ZDM2.131.fasta.gz Candidatus Competibaci acterium UTPRO2 Accumulitbacter sp. BA-5 ium Ga0077550 ium Ga0077550 ium Ga0077550 ium Ga0077550 ium Ga0077550 ium Ga077550 ium Ga077550 Galla Stata.gz	TIRBI C3.2 fata ga "Nourid Externis Detectioateris Edisposed Exteris in Produces teams (Paulos consequent, Tausara phenylasetes." TIRBI C3.2 fata ga "Nourid Externis Approximation", assigned Externisory and Ex
ITHBIC37 lastia gg ITHBIC34 lastia gg NZDM240 fastia gg NZDM240 fastia gg NZDM240 fastia gg Competitizater, Candidatu SAK051 91 lasta gg SAK072 109 lasta gg SAK072 109 lasta gg SEGL263 fasta gg SEGL263 fasta gg SEGL263 fasta gg SEGL273 fasta gg SEGL273 fasta gg	d_Bacteria; p. d_Bacteria; p.	Telescoberts C_beaptorecounts C_instances C_instances C_instances C_instances Telescoberts C_beaptorecounts C_instances C_instances C_instances Telescoberts C_instances C_instances Telescoberts Telescobert	A ZDM2.40.fasta.gz ZDM2.131.fasta.gz ,Candidatus Competibact acterium UTPRO2 Accumulibacter sp. BA-5 lium Ga0077550 ficrofhrix parvicella ESDC 7.2 fasta.gz GEA1C.73.fasta.gz GEA1C.73.fasta.gz	TIRBI C3. Tatalag Tourd, Elacteria Enderdideu, traasigned, caracity endersideu, traasigned, traasi
ITFBIG 1.37 Assta ge ITFBIG 2.37 Assta ge NZDM2-40 Insta ge NZDM2-40 Insta ge NADPA 2.40 Insta ge Competibacter, Candidata: SAKB5.19 Insta ge SAKB2.19 Insta ge SAKB2.19 Insta ge SEGL2.57.1 Insta ge SEGL2.57.1 Insta ge SESDIC 50 Insta ge SESDIC 50 Insta ge SESDIC 50 Insta ge SEGL2.72 Insta ge SEGL2.72 Insta ge SEGL2.72 Insta ge SEGL2.73 Insta ge	d_Bacteria; p. d_Bacteria; p. d_Bacteria; p. d_Bacteria; p. s Competibacteria; p. d_Bacteria; p.	Telescolaris C., Delaptorebactinis G., Jencory Janes, L., Procupscieller, S., Tataler, L., Tataler, Janoberg, J., Tataler, J., Peterbacheris, C., Analynovacharis, C., Janasignet, L., Lanasignet, J., Karasignet, J., Belaptorebacheris, and Karasignet, J., Lanasignet, J., Lanasignet, J., Karasignet, J.,	4 20142. 40 fasta gz ZDM2.131.tasta gz Candidatus Competibad acterium UTPRO2 Accumulibacter sp. BA-6 tium Ga0077550 tium Ga0077550 tikmoterius parvicella ferothrix parvicella ferothrix parvicella GENTC.731.tasta gz GWE 10.14.tasta gz	TITTBI C12 Altarga Tourish Electrical Released Electrical Productions (Production Research Tauters) TTBI C12 Altarga Tourish Electrical Released
ITTRB1C24 Areata ge NZDM2 40 Insta ge NZDM2 40 Insta ge NZDM2 131 Insta ge Competibilities of the Insta Competibilities of the Insta ge SAKR2 109 Insta ge SAKR2 109 Insta ge SAKR2 109 Insta ge SEGU2 50 Insta ge SESUC 201 I	d_Bacteria; p. d_Bacteria; p. d_Bacteria; p. d_Bacteria; p. d_Bacteria; p. s Competibacter d_Bacteria; p. d_Bacteria; p.	Bederoldesis c. unsaligned c. unsaligned j. unsaligned j. unsaligned s. Backeroldesis background ba	A 1 2DM2. 40. fasta.gz ZDM2. 131. fasta.gz Candidatus Competibaci acterium UTPR02 Accumulibacter sp. BA-5 ium Ga0077550 ficrothrix parvicella ESDC2.72. fasta.gz GEA1C.73. fasta.gz GWE1C1.4. fasta.gz GWE1C3.4. fasta.gz GWE1C3.4. fasta.gz GUR1C.8. fasta.gz	TIRBI C3. Tatalag Tourd, Elacteria Enderdideut, analogo et al. Competitional et al. Competitional. Competitional et
ITTBB1C24 Assta gg ITTBB1C24 Assta gg NZDM240 Insta gg NZDM240 Insta gg NZDM240 Insta gg NZDM2518 Insta gg SAREP 38 Insta gg SAREP 38 Insta gg SAREP 38 Insta gg SERIC 50 Insta gg SERIC 50 Insta gg SERIC 50 Insta gg SERIC 73 Insta gg SGWE1C 73 Insta gg SGWE1C 73 Insta gg SGWE1C 74 Insta gg	d_Bacteria; p. d_Bacteria; p.	relexaciones C. Designoceccianis E. Jeccoryclassis, L. Jeccoryclassis, L. Testaro, S. Testari, L. Testari, Jenny, Jeccory, J. K.	2012 20182 40 fasta gg ZDM2 40 fasta gg ZDM2 131 fasta gg ZDM2 40 fasta gg Learning UTPR02 Accumulibacter sp. BA-6 furm Garoy7550 furmftir parvicella furmftir parvicella ESDC 7.2 fasta gg GEA1C.73 fasta gg GWE1C3.7 fasta gg GWE1C3.7 fasta gg end WKS2C.51 fasta gg	TIRBI C3.2 failing and the set of the set o
ITTRB1C24 Asta gg ITTRB1C24 Asta gg NGDNg-0 Insta gg NGDNg-0 Insta gg Competibacter, Candidatu SADP3 & Batta gg SAKF2 109 Insta gg SAKF2 109 Insta gg SAKF2 109 Insta gg SAKF2 109 Insta gg SGWE CC 14 Insta gg SGWE CC 14 Insta gg SGWE CC 37 Insta gg	d_Bacteria; p. d_Bacteria; p.	Decisioalistes c. unassigned c. unassigned f. unassigned g. unassigned s. Unassigned a. Unassigned a	211 2010/2.40 fasta gz ZDM2.40 fasta gz ZDM2.131 fasta gz ZDM2.131 fasta gz ZDM2.40 fasta gz Accumulibacter sp. BA-5 fandhrung parkoella terrothic parkoella terrothic parkoella ESD2C.72 fasta gz GWE1C1.41 kasta gz GWE1C1.41 kasta gz GWE1C3.71 kasta gz MKSSC.71 fasta gz	TIRBI C3. Strates "Nourd, Bacteria Asteriolete, anteriorgen Lanagerout, massigned, Lanagerout, massigned, Lanagerout, massigned, Lanagerout, massigned, Lanagerout, massigned, Lanagerout, anteriorgen, Lanagerout, anterister, Managerout, anteriorgen, Lanagerout, anteriorgen, Manager
ITHBIC37 Asita ge ITTBIC224 Asita ge NZTMC131 Iseta ge NZTMC131 Iseta ge SADP38. Iseta ge SADP38. Iseta ge SAM521. 109 Iseta ge SAM52. 109 Iseta ge SAM52. 109 Iseta ge SEGL20. 11 Iseta ge SEGL20. 11 Iseta ge SEGL20. 11 Iseta ge SEGL20. 11 Iseta ge SEGL20. 12 Iseta ge TWKS20. 11 Iseta ge	d_Bacteria; p. d_Bacteria; p.	Telescobers C, Delaptorebactinis G, Jecobo Jassi, L., Paccopciales, E., Totale, L., Tataler, Jerny, Jecobo J, Pacebacheris C, Anghorebacheris A, Jenseignet C, Jenseignet G, Jenseignet K, Jenseignet J, Pacebacheris C, Jenseignet C, Jenseignet C, Jenseignet K, Jenseignet J, Pacebacheris C, Jenseignet C, Jenseignet C, Jenseignet K, Jenseignet J, Pacebacheris C, Bacterobekers L, Beckerobekers L, Jenseignet J, Jenseign	2006 40 Jasta gz ZDM6 40 Jasta gz ZDM6 40 Jasta gz ZDM6 40 Jasta gz ZDM6 40 Jasta gz Accumulibacter sp. BA-5 um Ga0077550 Grothirx parvicella Grothirx parvicella Grothirx parvicella Grothirx parvicella Grothir La Jasta gz GMFC 17 Jasta gz GMFC 14 Jasta gz MKSC 12 Jasta gz MKSC 12 Jasta gz MKSC 12 Jasta gz MKSC 12 Jasta gz	TIRBI CL 2 failing "Novid Rackeria Protectadentia Biologonade Rackeria Produces Thanking Provided Exact Transport Classing Cl
ITTRIC 527 Jasting: TTRIC 122 Jasting: TTRIC 123 Jasting: TTRIC 123 Jasting: TTRIC 123 Jasting: SAPRE 16 Intel of 27 SAPRE 16 Intel of 27 SAPRE 16 Jasting: SAPRE 16 Jasting: SAPRE 16 Jasting: SAPRE 16 Jasting: SAPRE 123 Jasting: SESSEC 123 Jasting: THRESC 244 Jastin	d_Bacteria; p d_Bacteria; p	neerootates -, Desegreto La subscripted L junces Jung - L junces jone - L tocknown - L managing - Monthaline	2004 Al Isata gr ZDM2 Al Isata gr ZDM2 Al Isata gr ZDM2 Al Isata gr Accumulation UTPRO2 Accumulationer sp. BA-6 um Ga0077550 Kordhiriy parwicella ESDBC 72 Isata gr Genthiri parwicella ESDBC 72 Isata gr GWTE C1 Al Isata gr GWTE C3 Al Isata gr WESC 2.5 Isata gr WESC 2.5 Isata gr WESC 2.5 Isata gr MESC 2.5 Isata gr MESC 2.5 Isata gr MESC 2.5 Isata gr MESC 2.5 Isata gr	TIRBE 12.5.1 failing "Nourid Educinia Enderoideux Lessinger Langelingen La
ITTRIC 22 A statu gr TTRIC 12 A statu gr TTRIC 12 A statu gr NCDM 2 13 I statu gr NCDM 2 13 I statu gr Competitisater, Canrolisater, SAMES, 10 Sastar, gr SAMES, 10 Sastar, gr SAMES, 10 Sastar, gr SAMES, 10 Sastar, gr SAMES, 10 Sastar, gr SEGU 2 1, I statu gr TWKS2 2 5, I statu gr SEGU 2 1,	d Bacteria p d	Telescobers C, Delaptorebactinis G, Jeccobyclasse, L, Jeccobyclasse, S, Totale L, Tataler Jernycheck Proteobacteris C, Jenaptorebactinis C, Jencopyclasse, L, Jencopyclasse, S, Tatale L, Jenaptorebactinis C, Jencopyclasse J,	2012/074-01 fastia.gz ZDM2-01 fastia.gz ZDM2-01 fastia.gz Zanddatus.Competibadi acterium UTPRO2 Accumulibacier sp. BA-4 Um Ga0077550 Erochtiv: parvicelia ESD202: 72 heata.gz GWFEI-02 fastia.gz GWFEI-02 fastia.gz GWFEI-02 fastia.gz GWFEI-02 fastia.gz met GWFEI-02 fastia.gz met MWFEI-02 fastia.gz met MWFEI-02 fastia.gz met MWFEI-02 fastia.gz MWFEI-02 fast	TIRBI C3.2 failing with the second
ITTRICC27 Janisage ITTRICC27 Janisage ITTRICC27 Janisage NCDMc2 131 Institu ge Competitionation (Competitional Competition Competitionation (Competitional Competition Competitional Competitional Competition Competitional Competitional Competition Competitional Competitional Competition SCHC2 109 Institute Competition SCHC2 100 Institute Competition Institute Competition I	a. Sacieria; p. d. Bacteria; y. d. Bacteria	neekoolastes - Desagoreado canada de la consegue de la consegue - La con	2012/02/2014/2014/2014/2014/2014/2014/20	TIRBI C3.2 Marg. "Novid Radenia Redecideus a Bisligneted Staferia (Producing Lasses) The Bisligneted Staferin (Producing Lasses) The Bisligneted Staferia (Pro
11781-1C327 Judia ga WC2Ne 40 Judia ga WC2Ne 10	d. Bacteria: p. d	Protectosettes C., Delaprotectostes C., Janacijovel C., Janasijovel C., Janasijovel S., Janasijovel J., Janasijovel J., Janasijovel S., Jan	2012/074-20 fasta gz ZDM2-40 fasta gz ZDM2-40 fasta gz ZDM2-40 fasta gz Candidatus Competibadi acterium UTPRO2 Accumulizacter sp. BA-4 fasta gz Microthrx parvicella Sicrothrx parvicella GEA1C-7.2 hasta gz GWHE C3-7 lasta gz MWHE C3 fasta gz MWHE C3 fast	TIRB 10:2 Taking Yound Edicine Released Bestgerede Casteria (Produce) steems (Packet on Section (Packet on
ITTEN IC 327 Junits ge ICCDU2 42 Junits ge NCDU2 43 Junits ge NCDU2 43 Junits ge NCDU2 43 Junits ge NCDU2 43 Junits ge SAVES 15 Junits ge SAVES 15 Junits ge SAVES 15 Junits ge SESDIC 53 Junits ge SESDIC 54 Junits ge SESDIC 55 Jun	d. Bacteria; p. d. Bacteria; p. d. Bacteria; p. d. Bacteria; p. d. Bacteria; p. d. Bacteria; p. d. Bacteria; p	neekootakis L. Desagoreado canadigued L. Janasigned L. Janasigned S. Lansaking L. Janasigned M. Janasigned J. Jana	2012/074-20 fasta gr ZDM2-10 fasta gr ZDM2-10 fasta gr Candidatus Competibadi actarium IUTPRO2 Accumulibater sp. BA-6 ium Ga007750 bin facothris parvicella Ecothris parvicella ESDC272 lasta gr UME/CG 27 lasta gr UME/CG 27 lasta gr UME/CG 27 lasta gr UME/CG 26	THRE 10.2 Taking The Constanting of th
11781-1C327 Judia ga WC2NL42 Al Judia ga SCWTC 10 A Judia ga SCWTC	 a societaria p. d. a societaria p. d. Bacteria p. d. Bacteria	necessories	2012/02/2014/2014/2014/2014/2014/2014/20	TIRB 10:2 Altariag Yound Educinal Enderrich Protectacients (Processpreame) Provide Startinal Enderrich Protectacients (Processpreame) TIRB 10:2 Altariag Yound Educinal Enderriche Protectacients (Protectacients) Provide Startinal Enderriche Protectacients) Yound Educinal Enderrich Protectacients (Protectacients) Provide Educinal Enderrich Protectacients) Provide Educinal Enderrich Protectacients SAUSS 10:1 Educinal Protectacients Yound Educinal Enderrich Protectacients Provide Educinal Enderrich Protectacients SAUSS 10:1 Educinal Protectacients Yound Educinal Enderrich Protectacients Provide Educinal Enderrich Protectacients SAUSS 10:1 Educinal Protectacients Yound Educinal Enderrich Protectacients Provide Educinal Enderrich Protectacients SSDD 20:1 Educinal Protectacients Yound Educinal Protectacients Protectacients Protectacients SSDD 20:1 Educinal Protectacients Yound Educinal Protectacients Protectacients Protectacients SSDD 20:1 Educinal Protectacients Yound Educinal Protectacients Protectacients Protectacients SSDD 20:1 Educinal Protectacients Protectacients Protectacients Protectacients Protectacients SSDD 20:1 Educinal Protectacients Protectacienta Protectacienta <t< td=""></t<>
ITTEN ICS27 Junits.gz ICS2042 40 Junits.gz NCDMC 41 Junits.gz NCDMC 41 Junits.gz NCDMC 41 Junits.gz NCDMC 41 Junits.gz SAVR51 10 Junits.gz SAVR51 10 Junits.gz SAVR52 10 Junits.gz SAVR52 10 Junits.gz SESDIC 12 Junits.gz USARIC 16 Junits.gz USARIC 16 Junits.gz USARIC 16 Junits.gz USARIC 10 Junits.gz USARIC 10 Junits.gz USARIC 10 Junits.gz USARIC 10 Junits.gz USARIC 10 Junits.gz USARIC 10 Junits.gz	A sacrieria p. A sacrie	neesoneste L. Beagnotecker in St. L. Hondon Josep L. Janasgord E. Janasgord E. Janasgord J. Baradi persystemic Neurophysics and the second state of the second state o	2014 2014 2014 2014 2014 2014 2014 2014	 TITBE 102.1 Altarga Void Backeria Proteoclastica Biological Science (Protocycleses Thusers, Thusers Phenylastics, * 1990). TITBE 102.1 Altarga Void Backeria Biological Biological Science (Protocycleses Thusers), Thusers Phenylastics, * 1990. TITBE 102.1 Altarga Void Backeria Biological Biol
ITTEN IC 327 Junits ge NCDM2 40 Junits ge NCDM2 40 Junits ge NCDM2 41 Junits ge SAV05 31 Setting ge SAV05 32 Junits ge SAV05 32 Junits ge SAV05 32 Junits ge SSV05 12 Junits ge S	La carteria p. La carte	neekoostasis L. Desagonado cana su managonet (unasagonet G. Jonasagonet S. Locational J. Marchanich (anasagonet J. Protechaberistics J. Aphaprochebacteris o unasagonet (unasagonet G. Jonasagonet S. Jonasagonet S. Protechaberistics J. Aphaprochebacteris o unasagonet (unasagonet G. Jonasagonet J. Protechaberistics J. Aphaprochebacteris o unasagonet (unasagonet G. Jonasagonet J. Protechaberistics J. Defaprochebacteris o protechaberistics J. Protechaberistics J. Defaprochebacteris o protechaberistics J. Protechaberistics J. Defaprochebacteris o protechaberistics J. Protechaberistics J. Defaprochebacteris o prasagonet (unasagonet J. Defaprochebacteris J. Defaprochebateristics J. Defaprochebacteris o prasagonet (unasagonet J. Jonasagonet J. Defaprochebacteris J. Protechaberistics J. Defaprochebacteris o prasagonet (unasagonet J. Jonasagonet J. Scholler J. Protechaberistics J. Defaprochebacteris o unasagonet (Jonasagonet J. Jonasagonet J. Jonasagonet J. Protechaberistics J. Defaprochebacteris o unasagonet J. Jonasagonet J. Jonasagonet J. Defabrochebacteris J. Defaprochebacteris J. Protechaberistics J. Defaprochebacteris J. Defabrochebacteris J. Jonasagonet J. Jonasagonet J. Jonasagonet J. Bederocheberis J. Jonasagonet J. Jonasagonet J. Jonasagonet J. Jonasagonet J. Bederocheberis J. Jonasagonet J. Jonasagonet J. Jonasagonet J. Bederocheberis betterin (J.B. S. Defabrocheberis J. Jonasagonet J. Jonasagonet J. Jonasagonet J. Bederocheberis betterin (J.B. S. Defabrocheberis J. Jonasagonet J. Jonasagonet J. Jonasagonet J. Jonasagonet J. Jonasagonet J. Bederocheberis J. Jonasagonet J. Jonasagonet J. Jonasagonet J. Bederocheberis betterin (J.B. S. Defabrocheberis J. Jonasagonet J. Jonasagonet J. Jonasagonet J. Bederocheberis betterin (J.B. S. Defabrocheberis J. Jonasagonet J. Jonasagonet J. Jonasagonet J. Bederocheberis betterin (J.B. S. Defabrocheberis J. Jonasagonet J. Jonasagonet J. Jonasagonet J. Jonasagonet J. Bederocheberis J. J	2012/02/2014/2014/2014/2014/2014/2014/20	TITTBI C.S. Tatang "Novid Allicentia Protectications Englightenetications (Photopoliticsen, Thuesen, Thuese
ITTRICE27 Junits ge ITTRICE27 Junits ge INCDNA 20 Junits ge INCDNA 20 Junits ge INCDNA 21 Istains ge INCDNA 20 Junits 20 JUNIT	a Discretisa, p.	neesocaesta C. Delaptorebacteria C. Janasignet C. Janasignet S. Janasignet S. Janasignet S. Janasignet J. Janasign	2012 COMA: 40 Anata, gr 2014 2014 Anata, gr 2014 2014 31 Anata, gr 2014 2014 31 Anata, gr 2014 2014 31 Anata, gr Accumulabated sep, BA-4 Accumulabated sep, BA-4 BA-4 BA-4 Comparison of Accumulabated sep, BA-4 Accumulabated sep, BA-4 Ac	 TITBE 102. Taking Varian Strand States (Comparison of Comparison of Compa
ITTEN ICS 21 Main get NCDMC 40 Institute 11 Main get NCDMC 41 Institute 12 Main get NCDMC 42 Main get NCD	A societia, p. A societia, p. Bacteria, p. Competibility of the society o	neessoonset, C. Jesalpored Carlos K. J. Moor J. Josef J. Lanser, J. L. Conserver, S. L. Tabart, S. L. Tabart, J. M. Santar, J. K. S. Santar, J. K. S. Santar, J. K. S. Santar, J. K. Santar, J. K. Santar, J. K. Santar, J. S. Santar, J. Santa	2014 2014 2014 2014 2014 2014 2014 2014	TITBE IG. 21 Allang "North Glackins Enderside Linestry Performance Statistics IPRoduces Tables The Statistics Allands and Statistics Allands an
ITTRICC27 Junits ge ITTRICC27 Junits ge INCDNC 40	A Badretia, p. P. Badretia, p.	neesocaesta, C. Josephorecontenis, E. Josephorecontenis, C. Joseph	2012 COMA-01 Alasta, gz CZDMA-20 Alasta, gz CZDMA-20 Hasta, gz CZDMA-20 Hasta, gz CZDMA-20 Hasta, gz Accumulibateter sp. BA-4 kum GadOY7550 Gradbar sp. Alasta, gz GZMA-27 Alasta, gz GZMA-27 Alasta, gz GZMA-27 Alasta, gz GZMA-27 Alasta, gz HOM-10 C-37 Alasta, gz HOM-10 C-37 Alasta, gz HOM-10 Alasta, gz SHO-10 Alasta, gz	 TITBE 10:2 Altarga Void Backeria Protectacienta Ribertato Exteriori Riboto Costenea, Thauera Thauera Promytostet." ("Toron Science Ribertato Ribertato Activity Ribertato Rib
ITTEN C32 A status get NCDMC 43 O status get SAVR53 S O status get SAVR53 S O status get SAVR53 S O status get SESTOC 72 I status get SESTOC 73 I status get SESTOC 73 I status get SESTOC 73 I status get SESTOC 74 I stat	de Bacteria p. de Bacte	Telescolares, C., Delaporebacteris, C., Janasgord, C., Janasgord, S., Lanasgord, S., Lanasgord, S., Maran, J., Maran, Maran, J.,	2014 2004 2014 and a gr 2014 2014 2014 and a gr 2014 2014 2014 2014 2014 2014 2014 2014	 TITBE 10.2.1 A tang model of the second secon
ITTRICC2 21 Main ge NCDM2 40 Main ge NCDM2 40 Main ge NCDM2 40 Main ge NCDM2 41 have general taxes general taxes g	A Bardenia P A Barde	Decembers C, Desegnered C, Janasginet L, Janasginet J,	2012 CDM2-01 fastia ag 2014 2014 2014 fastia ag 2014 2014 2014 fastia ag 2014 2014 2014 2014 2014 2014 2014 Leaderium (DTPRO2 Accumulationate or ga Accumulationate or ga Accumulationate or ga Accumulationate or ga 2014 2014 2014 2014 2014 2014 ESIZEC 2014 2014 2014 2014 Control to particular based particular based or ga 2014	 ITTER 12.5.1 Altarg. Novid Scherick Redevideu, Largeropkelscher, Scherpotekscher, Floudoryckesse, Thusers, Thusers phenylosette, " Thord Scheric Redevideu, Largeropkelscher, Largeropkelscher, Largeropkelscher, Scherpotekscher, Thusers, Thusers phenylosette, " Thord Scherick, Proteckacheris, Bergeropkelscher, Largeropkelscher, Scherpotekscher, Largeropkelscher, Thusers, Thusers phenylosette, " Thord Scherick, Proteckacheris, Bergeropkelscher, Scherpotekscher, Scherpotekscher, Largeropkelscher, Thusers, T
ITTEN IC 227 Junits ge ICCD/02 40 Junits ge NCD/02 40 Junits ge SCOPE 10 JUNIT	de Bacteria p. de Bacte	neessoenset, L. Salaporesko tank t neesson y L neesson y L.	2014 2014 2014 2014 2014 2014 2014 2014	 THRI G.2. Starling Workshow Constraints (Networkshow Construction) THRI G.2. Starling Workshow Construction (Networkshow Construction) SADRE St
ITTEN C32 A balance of the CSD	A Bardenia P A Barde	ndeerootains - L. Desagned and S. L. Lansagned - L. Lansagned - L. Lansagned - L. Lansagned - L. Marasigned - Lansagned - Lans	2012 CDMA 501 fastia, gr 2014 CDMA 501 fast	 TITBE 102. A fairing ' Tool Radientin Proteodoctinal Biologoted caterial, Prodocyclase, Thousen, Thauera, Thauera phenylosote, '' Tool Radientia Anteriode Jackstrain, Devologoted and Staterial Anteriode Staterial Antonia Stateria Antonia Stateria Antonia Stateria Antonia Stateria Antonia Stateria Proteodoctina (''''''''''''''''''''''''''''''''''''
ITTEN IC 327 Junits ge ICT/2014 20 Junits ge NCDMC 31 Intelling 20 NCDMC 30 Intelling 20 NCDMC 30 Intelling 20 NCDMC 30 Intelling 20 NCDMC 30 Intelling 20 NCMC 30		Telescolaris, C., Diagonobactinis, E., Jeccolaris, E., Jeccolaris, E., Jeccolaris, E., Tatalin, P., Portechardiet, C., Maraliportecharis, C., Jenssignet, L., Jenssignet, J., Jenssignet, J., Lenssignet, J	2014 2014 2014 2014 2014 2014 2014 2014	 THRI G.2. Mang, Yoo K. Shara, Sharakarakar, Yang Yuo, Yuo Kakaraka, Sharakarakarakarakarakarakarakarakarakara
ITTEN C32 A latest ge NCDM2 40 latest ge NCDM2 40 latest ge NCDM2 41 latest ge NCM2 41 late	d d	needecoders C, Josephoredo C, Joseph	2014 2014 2014 2014 2014 2014 2014 2014 2014	TITBE IG. 21 Allang "Nord Radenia Robectaderia Arterological Englisher Englisher and State State III Arterological Englisher Arterological Englisherologiste Englisherol Arterological Englisherol Arterologic
ITTEN ICS2 A latest ge ICS2 A latest ge NCDMC 43 latest ge SAVES 16 latest ge SAVES 16 latest ge SAVES 16 latest ge SAVES 10 latest ge USCR10 LC 10 latest	d d	Telescopers C. Designorescience L. Decomposition L. Jacobio Construction L. Telescopic Construction L. Longologic Cons	2014 2014 2014 2014 2014 2014 2014 2014	 TITBE 102.7.1.8 Integ Tourd, Backeria, Proteckaderia, Jassinger, Langinger, Langinge
ITTEN IC 237 Junits ge NCDMC 40 Junits ge NCDMC 41 Junits ge NCDMC 41 Junits ge NCDMC 41 Junits ge NCDMC 41 Junits ge NCDMC 42 Junits ge NCMC 42 Juni		Technologies C, Designedo C, Janagiged C, Janagiged C, Janagiged S,	2014 2014 2014 2014 2014 2014 2014 2014	 THBI 10:3 Taking Novel Allocation Proteoclastical Biologoted Statistics (Produce) states (Packooper Statistics) (Produce) statistics (Packooper Statistics) (Packooper Sta
THEN C32 A latest get THEN C32 A latest get NCDMC 40 latest get NCDMC 41 latest NCDMC 41 latest SMR55 h latest get SMR55	d d	Telescolaris C., Delaprotectarias L., Decorporation, L., Telescolaris C., Telescolaris L., Telescolaris C., Telescolaris L., Delaprotecharis L., Cardidata M., Africactaris L., Delaprotecharis L., Activitationa L., Telescolaris L., Cardidata M., Africactaris L., Addinicolaris L., Machina L., Telescolaris L., Cardidata M., Africactaris L., Addinicolaris L., Addinicolaris L., Cardidata M., Adricactaris L., Addinicolaris L., Addinicolaris L., Cardidata M., Adricactaris L., Addinicolaris L., Cardidata M., Cardidata M., Cardidata M., Africactaris L., Addinicolaris L., Cardidata M., Cardidata M., Cardidata M., Cardidata M., Cardidata M., Advince L., Social Addinicolaris L., Lanesignet J., Lanesig	2014 2004 2014 and a gr 2004 201	 TITBE 102. Taking Novel Research Protectacients Responses Team Providence "Team Providence" ("Team Providence Team Providence Tea
ITTEN IC 237 Junits ge ICCDIM2 40 Junits ge NCDIM2 40 Junits ge NCDIM2 41 Junits ge NCDIM	d d	Technologies C, Designored C, Lanagarde L, Lanagarde J, Lanagarde J	2014 2004 2014 and a gr 2014 2014 2014 and a gr 2014 2014 2014 2014 2014 2014 2014 2014	 THRE 1G.2.1 Attaing Yound Backetina Andersider, Kanada Sharkine Andersider, Kanada Yuung Yuung
ITTEN C23 Junits ge NCDM2 40 Junits ge NCDM2 40 Junits ge NCDM2 41 Junits ge NCDM2	d d	ndeexostes - Lossingeró a constructure d'invession - Lossingeró - Loss	2012 CDMA: 510 fastia, gr 2014 CDMA: 510 fastia	 TITBE 102. Taking No. 2007. Bioinform Proteocolarism Engineering Statistics Produces Tourses, Tour
THEN C32 A status get THEN C32 A status get NCDMC 43 basis get NCDMC 44 basis get	d Bacteria p. Bacteria p.	Telescobers C, Delaporebacteris L, Decompany L, Janasgord S, Janasgord J, Janasgord S, Janasgord J, Janasgord S, Janasgord J, Janasg	2014 2004 2014 and a gr 2014 2014 2014 and a gr 2014 2014 2014 2014 2014 2014 2014 2014	 THRE 10.2.1 Allang Water And States Proteocolstratic Biological States (Produce) Cases (Pauce Constraints, Theorem Phenylastics, " The Case Allang Water And States (Pauce Constraints, Pauce Constrai
ITTEN IC 23 / Junits ge NCDM2 40 Junits ge NCDM2 40 Junits ge NCDM2 41 Junits ge N	d Bacteria: p. p. Bacteria: p. Bacter	Decembers C., Desegnered C., anasignet J., anasignet J., anasignet S., anasignet S.	2012 CDMA: 510 fastia, gr 2014 CDMA: 510 fastia	TITBE IG. 21 Altang "North Glacenia Proteocidents Belgenoted States in Proceedings 1. TITBE IG. 21 Altang "North Glacenia Proteocidents and States (North Glacenia States) Proteocidents and States) Proteocidents and States (North Glacenia States) Proteocidents and States) Proteocidents a
THEN C32 A latest get THEN C32 A latest get NCDMC 40 latest get NCDMC 41 latest get SAV655 16 latest get SAV655 16 latest get SAV655 16 latest get SAV655 10 latest get SAV6	d d	Telescolaris, C., Belgorebockeris, C., Janagiped, L., Janagiped, S., Janasiped, B., Janasiped, S., Janasiped, J., Janasiped	2014 2014 2014 2014 2014 2014 2014 2014	 TITBE 102.7.1.1.tang Markan Protectacients Religious Carteria (Fractory classes) Religious Religious Religious Carteria (Fractory Classes) Religious Carteria (Fractory Classes) Religious Religi
THEN C32 A latest get THEN C32 A latest get NCDMC 40 latest get NCDMC 41 latest get NCDMC 41 latest get NCDMC 41 latest get NCDMC 41 latest get NCDMC 42 latest get NCMC 4	d d	needenoomie L. Deslopende canadigue de l'analignet L'analignet à l'analignet à l'analignet de l'	2014 2004 201 January 2014 2014 2014 2014 2014 2014 2014 2014	 THER 10.2 Taking Value of the second s
ITTEN IC32 A latella ge ITTEN IC32 A latella ge IC3204 20 latell	d Bacteria p. d Bacte	Telescobers C, Delaporebactina L, Decompany L, Janasgord J, Janasgo	2014 2007 2014 and a gr 2014 2014 2014 and a gr 2014 2014 2014 2014 2014 2014 2014 2014	 THRI GL 21 Alargia Yound Backeria Alexelective Langebrock and the sequence of the se
THEN C32 A balance of the control of		 Indexposite C., Designed C., Lansagned L., Lansagned L., Lansagned S., Lansagned M. K., Marchall J., Marker J., Marker	2014 2004 2014 and a gr 2014 2014 2014 and a gr 2014 2014 2014 2014 2014 2014 2014 2014	 THER 10.5.7 Manuary Yound Backeria Andersideux Langsing Calculation (Productionses) Thus and productions (Productionses) Thus and Productionses) Thus and Productionses (Productionses) Thus and Productionses (Productionses) Thus and Productionses) Thus and Productionses (Productionses) (Productionses)
THEN C32 A latest get THEN C32 A latest get NCDMC 41 latest get SAVR5 15 latest get SAVR5	d d	 Protechanies C., Designored C., analignet C., Lancasgord S., Lancasgord S., Lancasgord N., Santan J., Peterbaharis C., Analignet C., Lancasgord J., Lancasgord	2014 2004 2014 and a gr 2014 2014 2014 and a gr 2014 2014 2014 2014 2014 2014 2014 2014	 TITBE 102.7.1.8. Integ Tourd, Backeria, Protectoaleria, Floredocaleria, Floredocaleria, Floredocaleria, Floredocaleria, Floredocaleria, Protectoaleria, Teurna province analysis, Electronic Autors, Carolina (1), ************************************