J. Indian Inst. Sci.
A Multidisciplinary Reviews Journal
ISSN: 0970-4140 Coden-JIISAD

REVIEW ARTICLE

Comparative Genomics Using the Integrated Microbial Genomes and Microbiomes (IMG/M) System: A *Deinococcus* Use Case

Rekha Seshadri[*], *Nikos C. Kyrpides and Natalia N. Ivanova*

Abstract | The Integrated Microbial Genomes and Microbiomes (IMG/M) system is a web-based platform that provides access to the wealth of public sequence data arising from diverse environments and enables the user to answer biological questions. In this review, we explore IMG's tools and features using genome data for genus *Deinococcus* isolates as well as metagenome-assembled genomes (MAGs). We use various comparative genomic and visualization tools to investigate this genus and address specific research questions.

1 Background

Extreme environments on Earth include hypersaline lakes, arid regions, deep sea, acidic sites, cold and dry polar regions, permafrost, and extremophiles native to these environs are conjectured to survive the harsh conditions of extraterrestrial settings—and possibly serve as model organisms to understand the fate of biological systems in such environments. The survival of organisms inside rocks (endolithic communities) or their survival on Mars has also been an area of focus for understanding the likelihood of life on other planets.[1,2] Metagenome samples and available isolates from such types of environments can be accessed via the IMG/M data portal.

One such model organism is *Deinococcus radiodurans*, a polyextremophile, famously resistant to radiation, desiccation, and many toxic chemicals.[3] This resistance is linked to its ability to recover following exposure to diverse kinds of damage, which is lethal to most organisms, and is mediated by hundreds of proteins involved in DNA repair, oxidative stress defense, proteome protection, regulation, and various unknown functions. Bacteria belonging to the family *Deinococcaceae* are some of the most radiation-resistant organisms discovered and there is large diversity in the molecular mechanisms involved in this resistance within the *Deinococcus* genus.[4,5] Members of this group are not only model organisms for the study of DNA damage and repair, but candidates for practical applications such as cleanup of radioactive waste sites (e.g., *D. radiodurans* engineered to express enzymes for metal detoxification or degradation of organic pollutants).[6]

Here, we use *Deinococcus* as the biological case study to explore and highlight useful data resources and tools for comparative genomics within the IMG/M system. IMG/M allows the benchtop biologist to formulate and answer biological questions quickly using an easy-to-use web interface. We identify and compare cultured isolate and uncultivated genomes of the genus *Deinococcus* to address the following objectives: (1) explore the general genome properties and diversity of available genomes with particular attention to the *D. radiodurans* clade, (2) identify unique functional gene content in this clade that may be associated with any relative resistance prowess compared to other *Deinococcus* species, (3) assess the occurrence of Deinococci in environmental samples and the relative diversity of uncultivated genomes (MAGs) recovered from these samples, and (4) examine the distribution of a previously characterized mutagenesis cassette

[1] *U.S. DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA.*
**rseshadri@lbl.gov*

to highlight strains that may possibly serve as a chassis for strain engineering efforts.

2 Exploring Genomes and Their Statistics

Using the advanced genome search function that allows the user to query and retrieve datasets based on elaborate metadata criteria (sourced from the GOLD database mentioned in the accompanying article in this issue[7]), we retrieve 47 isolate genomes (finished and draft assemblies) by specifying a total genome scaffold count of ≤ 600. The scaffold count serves as a measure of the assembly quality, helping to eliminate highly fragmented draft genomes from the analysis. Metadata filters and other utilities of the advanced search builder are presented in this YouTube tutorial. *Deinococcus* spp. genome sizes range from 2.75 to 6.65 Mbp (Fig. 1). *D. radiodurans* R1 was the first isolated (from irradiated canned meat) and its finished genome consists of two chromosomes and two plasmids (177 and 45 kb) totaling 3.34 Mbp.[8]

From the genome cart, additional metadata can be explored for these genomes, such as genome statistics like G + C%, sequencing or assembly methodology, environmental classification, and much more. As reported previously for this genus,[9] there is little correlation between genome size and G + C % (Fig. 2).

Recently added and highly useful fields include contamination and completeness estimates from checkM[10] and GTDB-Tk taxonomy.[11] CheckM scores are provided to help with assessment of the quality of individual genomes, while GTDB-Tk provides a genome-based taxonomy, which is based on a set of universally conserved marker genes. It complements the NCBI lineage, which is based on multiple classification criteria ranging from average nucleotide identity (ANI) to 16S rRNA-based classification. Data tables are easily exportable into a spreadsheet for graphing or other analysis or visualization.

Environmental metadata show that *Deinococcus* spp. were isolated from a wide range of environments as summarized in Fig. 2b. Reported sites of isolation include lake sediment, weathered granite, irradiated medical instruments, and air purification systems among others.[12]

The system provides the results of searches of proteomes performed against various public functional annotation sources (as described in the accompanying article[7])—offering maximal opportunity to make biological inferences based on these various annotations (COG, Tigrfam, Pfam, Kegg Orthology, Superfam, etc.), which are partially overlapping, but have distinct foci and strengths. The genome table can be reconfigured to show summaries of all these search results and more custom analyses like CRISPR arrays or biosynthetic gene clusters (BGCs, based on AntiSMASH v5[13]). For example, the four *D. radiodurans* (*D.r.*) strains encode 4 BGCs—2 terpene clusters and 2 T3PKS. Secondary metabolites are specialized compounds produced by these BGCs enabling organisms to respond to environmental stresses or mediate interactions with each other—and could have applications in many fields. A genome details page is available for each genome which displays all these genome statistics and also provides useful tools for exploration such as chromosome maps (Fig. 3), Kmer frequency analysis, phylogenetic distribution of BLAST hits, etc.

3 Comparing Genomes

The relationship of the *Deinococcus* isolates can be examined using the ANI tool or visualized on a phylogram employing 16S or other marker gene alignments. Here again, and advanced gene builder can be used to retrieve appropriate markers (e.g., using Pfams for RpoB) or 16S rRNA (using minimum sequence length constraints), and sequence alignment and phylogram can be generated from the gene cart. Narrowing in on *D.r.*, four strains (isolated from irradiated meat) form a distinct clade (as expected) with *D. xibeiensis* and *D. wulumuqiensis* (both isolated from radiation-contaminated soils) as their closest available relatives in this cohort (Fig. 4). This result is corroborated by the top BLAST hits of *D.r.* 16S rRNA genes that can be accessed from the gene details page.

Pairwise ANI[14] (under "Compare Genomes") of *D.r.* R1 strain against all other "species" echoes this observation (Fig. 5). Highest ANI and AF values are shared between the near-identical strains of *D.r.* (99.9% ANI/96% AF), followed by *D. xibeiensis* and *D. wulumuqiensis*. Precomputed ANI clusters can also be browsed from the ANI tool and filtering "Contributing species" for *Deinococcus*—over 40 distinct "species" of *Deinococcus* may be discerned using ANI for species classification[14].

Functional content (based on CDS protein sequence assignments to COG, Pfam, KO, and Tigrfam) may be compared across all genomes

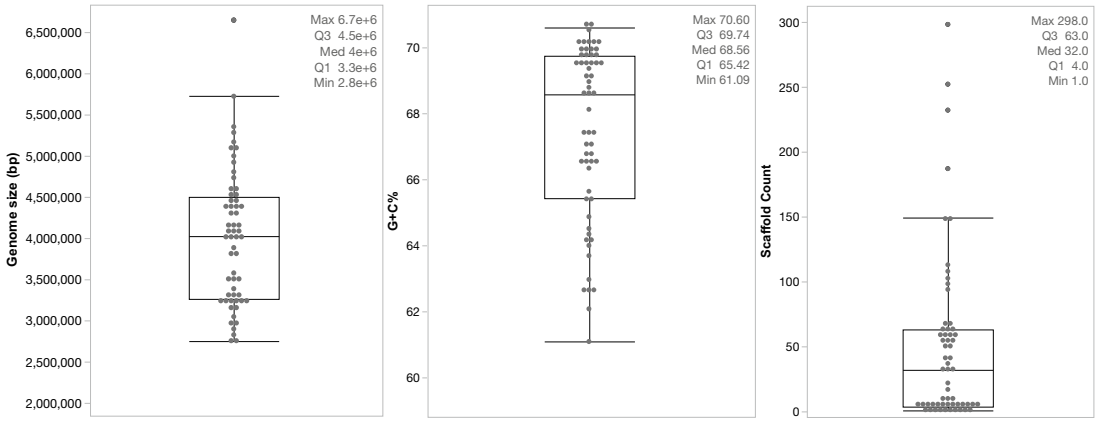


Figure 1: Genome size (panel 1), G + C% (panel 2) and total scaffold counts distribution (panel 3) for 60 *Deinococcus* spp isolates is shown. Five quantile summary is included in each plot.

using a variety of tools under "Compare genomes", such as genome clustering, abundance profiles, and more. To directly compare proteomes based on pairwise sequence similarities rather than protein family assignments, tools under Compare genomes > Phylogenetic Profilers may be appropriate. For example, the "Single Genes" tool may be employed to find genes that are unique to the *D.r.* clade by specifying all other genomes in the "without homologs" box (Fig. 6)—only 306 genes are retrieved from the reference strain (R1) that meet the specified criteria (Supplementary Table 1). Of these, over half are conserved hypothetical proteins and the remainder notably comprise many regulatory proteins, transposases, biotin uptake, osmoprotectant transporter, bacterial sensor and chemotaxis proteins, erythromycin esterase, triacylglycerol lipase, and cytochrome c-type biogenesis proteins. Some of these have been implicated in their hallmark radiation resistance phenotype, although many mechanisms of exist[15,16].

4 Discovering Genomes "in the Wild"

The distribution or occurrence of *D.r.* and related strains in metagenome samples can be assessed preliminarily using the "top IMG metagenomes 16S rRNA hits" option from RNA homologs in the 16S gene detail page. This can be repeated individually for 16S genes from each genome. From the BLAST results (filtering ≥ 97% identity hits, which roughly corresponds to a species level

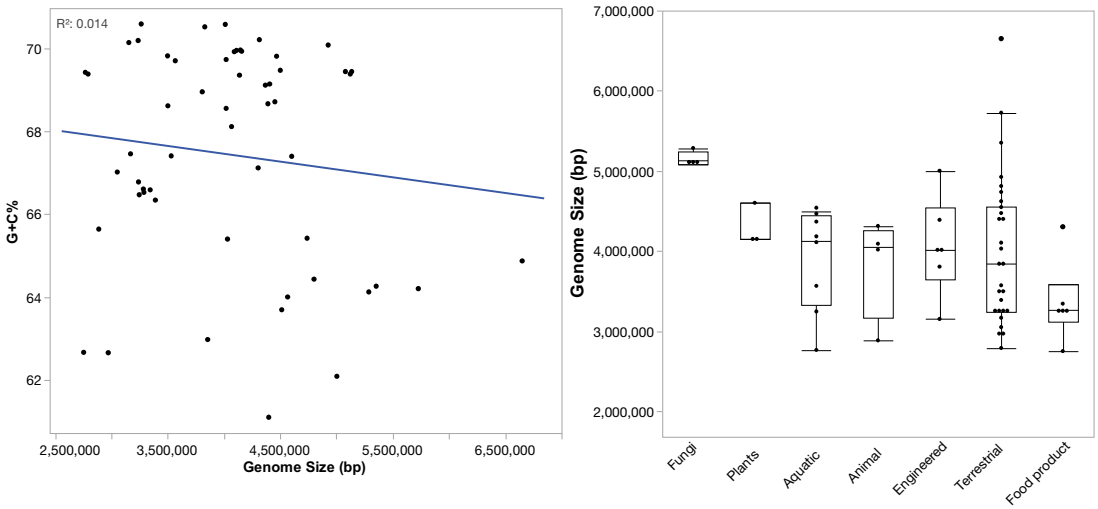


Figure 2: Plot of genome size versus G + C for each of the isolate genomes (left). Unlike reported trends in other taxa, there is little correlation between these two metrics in this ancient phylum. Genome size distribution of *Deinococcus* spp. isolated from distinct environmental sources (right). Points indicate individual genome sizes.

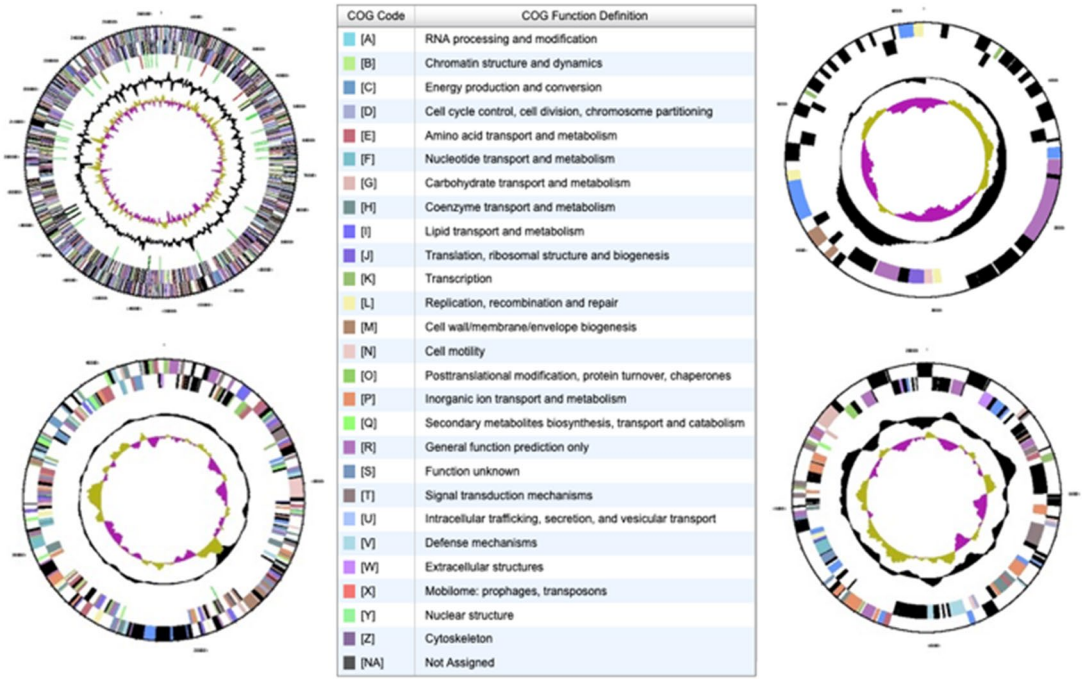


Figure 3: Circular representation of the *Deinococcus radiodurans* R1 overall genome structure comprised of four individual replicons. The outer scale designates coordinates in base pairs. The outer circles show predicted CDS on the plus strand color-coded by function role categories.

16S rRNA sequence similarity), it is apparent that these species are not present in the thousands of metagenome samples available in IMG. The *D.r.* clade members are only found in a mock synthetic community (example—BLAST hits from R1). While the distribution of Deinococci has not been explored systematically, a preliminary assessment of 16S matches suggests that some other *Deinococcus* species like *D. grandis* ATCC 43672 or *D. soli* N5, or *D. sp* UR-1, are detected in freshwater samples from certain lakes, rivers, or aquifers around the world (Fig. 7—BLAST hits from UR1).

Another approach pursuing this objective is to assess whether any ***Deinococcus*** metagenome-assembled genomes (**MAGs**) have been recovered from metagenome samples. Over 200,000 auto-computed MAGs referred to as "metagenome scaffold bins" can be explored using the "metagenome bin" search tools under "Find genomes". Again, an advanced search builder can be employed to constrain the search based on various MAG statistics, taxonomy, and more. Here, we retrieve 9 *Deinococcus* MAGs with ≥ 95% checkM completeness from primarily endolithic communities (Fig. 8). Additional 18 MAGs are available with lower completeness (as low as 54%) recovered from a variety of host-associated, soil and freshwater samples. In addition to checkM measures, the quality of

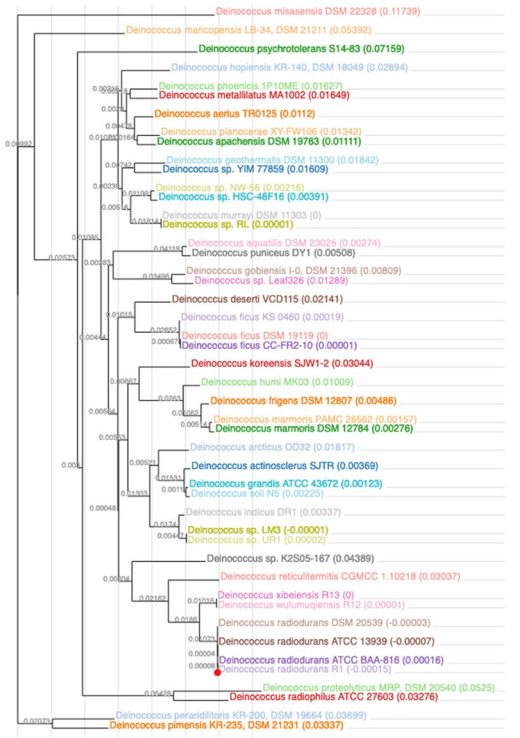
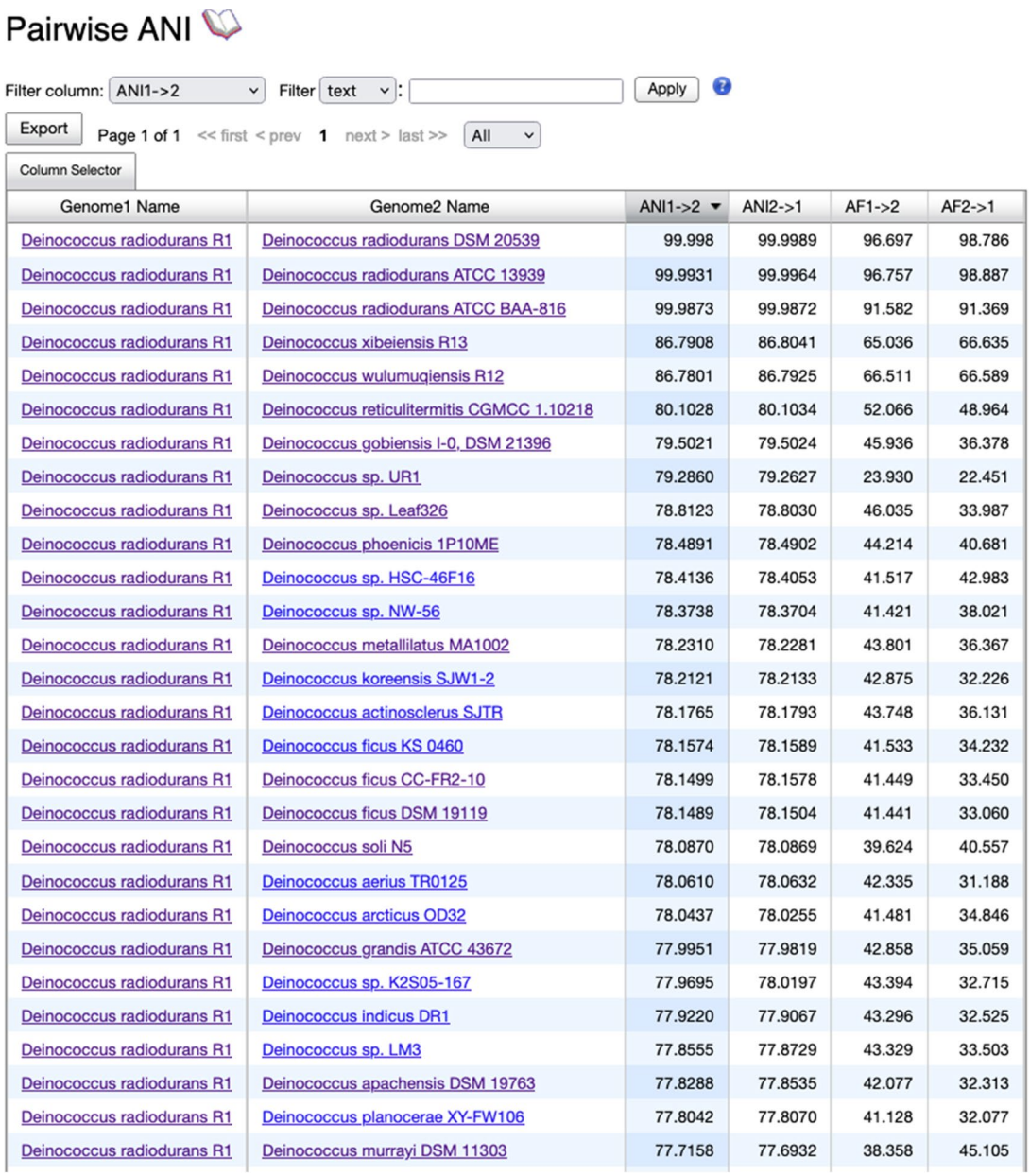


Figure 4: Rectangular phylogram inferred from RpoB protein sequence alignments (by ClustalW) using the "sequence alignment" tool accessible from the gene cart. *D.r.* str. R1 is highlighted toward the bottom of the tree. This and neighboring genomes form the 6-genome "D.r clade" referenced below.



Pairwise ANI

Filter column: ANI1->2 Filter text : Apply

Export Page 1 of 1 << first < prev 1 next > last >> All

Column Selector

Genome1 Name	Genome2 Name	ANI1->2	ANI2->1	AF1->2	AF2->1
Deinococcus radiodurans R1	Deinococcus radiodurans DSM 20539	99.998	99.9989	96.697	98.786
Deinococcus radiodurans R1	Deinococcus radiodurans ATCC 13939	99.9931	99.9964	96.757	98.887
Deinococcus radiodurans R1	Deinococcus radiodurans ATCC BAA-816	99.9873	99.9872	91.582	91.369
Deinococcus radiodurans R1	Deinococcus xibeiensis R13	86.7908	86.8041	65.036	66.635
Deinococcus radiodurans R1	Deinococcus wulumuqiensis R12	86.7801	86.7925	66.511	66.589
Deinococcus radiodurans R1	Deinococcus reticulitermitis CGMCC 1.10218	80.1028	80.1034	52.066	48.964
Deinococcus radiodurans R1	Deinococcus gobiensis I-0, DSM 21396	79.5021	79.5024	45.936	36.378
Deinococcus radiodurans R1	Deinococcus sp. UR1	79.2860	79.2627	23.930	22.451
Deinococcus radiodurans R1	Deinococcus sp. Leaf326	78.8123	78.8030	46.035	33.987
Deinococcus radiodurans R1	Deinococcus phoenicis 1P10ME	78.4891	78.4902	44.214	40.681
Deinococcus radiodurans R1	Deinococcus sp. HSC-46F16	78.4136	78.4053	41.517	42.983
Deinococcus radiodurans R1	Deinococcus sp. NW-56	78.3738	78.3704	41.421	38.021
Deinococcus radiodurans R1	Deinococcus metallilatus MA1002	78.2310	78.2281	43.801	36.367
Deinococcus radiodurans R1	Deinococcus koreensis SJW1-2	78.2121	78.2133	42.875	32.226
Deinococcus radiodurans R1	Deinococcus actinosclerus SJTR	78.1765	78.1793	43.748	36.131
Deinococcus radiodurans R1	Deinococcus ficus KS 0460	78.1574	78.1589	41.533	34.232
Deinococcus radiodurans R1	Deinococcus ficus CC-FR2-10	78.1499	78.1578	41.449	33.450
Deinococcus radiodurans R1	Deinococcus ficus DSM 19119	78.1489	78.1504	41.441	33.060
Deinococcus radiodurans R1	Deinococcus soli N5	78.0870	78.0869	39.624	40.557
Deinococcus radiodurans R1	Deinococcus aerius TR0125	78.0610	78.0632	42.335	31.188
Deinococcus radiodurans R1	Deinococcus arcticus OD32	78.0437	78.0255	41.481	34.846
Deinococcus radiodurans R1	Deinococcus grandis ATCC 43672	77.9951	77.9819	42.858	35.059
Deinococcus radiodurans R1	Deinococcus sp. K2S05-167	77.9695	78.0197	43.394	32.715
Deinococcus radiodurans R1	Deinococcus indicus DR1	77.9220	77.9067	43.296	32.525
Deinococcus radiodurans R1	Deinococcus sp. LM3	77.8555	77.8729	43.329	33.503
Deinococcus radiodurans R1	Deinococcus apachensis DSM 19763	77.8288	77.8535	42.077	32.313
Deinococcus radiodurans R1	Deinococcus planocerae XY-FW106	77.8042	77.8070	41.128	32.077
Deinococcus radiodurans R1	Deinococcus murrayi DSM 11303	77.7158	77.6932	38.358	45.105

Figure 5: Results of pairwise ANI query comparing *D.r.* R1 against other strains and species (top 25 rows are shown). ANI is average nucleotide identity and AF is alignment fraction.

individual MAGs can be assessed from within the scaffold workspace or the scaffold cart using the Kmer frequency analysis plot [Fig. 9—result for HQ MAG from a sandstone endolithic community (3300039401_2) with 67 scaffolds]. Potentially incorrectly binned scaffolds could be assessed and removed from the metagenome bin as desired.

The phylogenetic relationship of these MAGs can again be explored by recovering RpoB sequences from the 9 HQ MAGs and recomputing the RpoB phylogram as before including isolates (Fig. 10). The near-identical endolithic MAGs may represent a new *Deinococcus* species compared to available isolate genomes.

5 Gene Cassette Search

D. radiodurans is renowned for efficient DNA repair pathways including excision repair, mismatch repair, and recombination repair. It is not mutable by UV radiation due to this very accurate DNA repair and the absence of error-prone

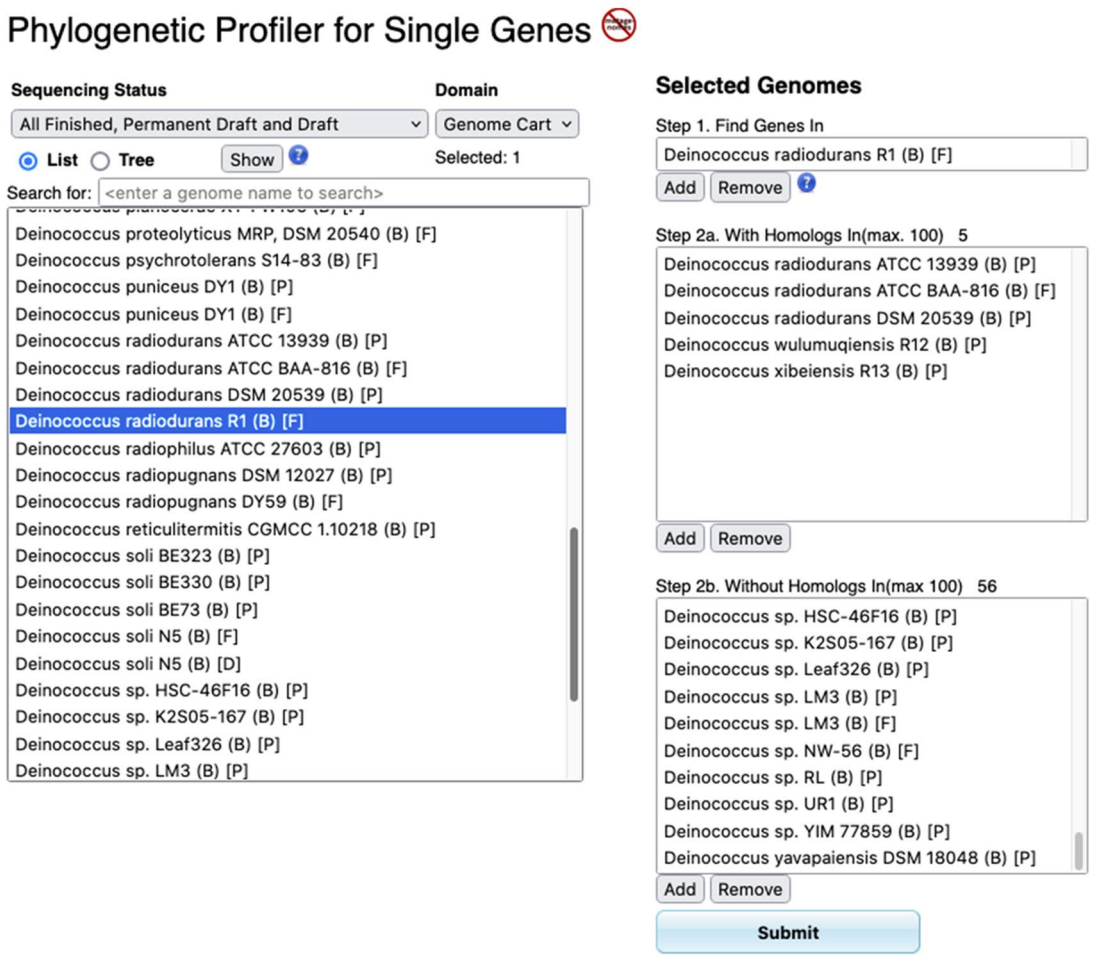


Figure 6: Screenshot of data input page for pairwise sequence-based (USEARCH) comparisons of specified genomes. Tool can be found under Compare Genomes > Phylogentic Profilers > Single Genes.

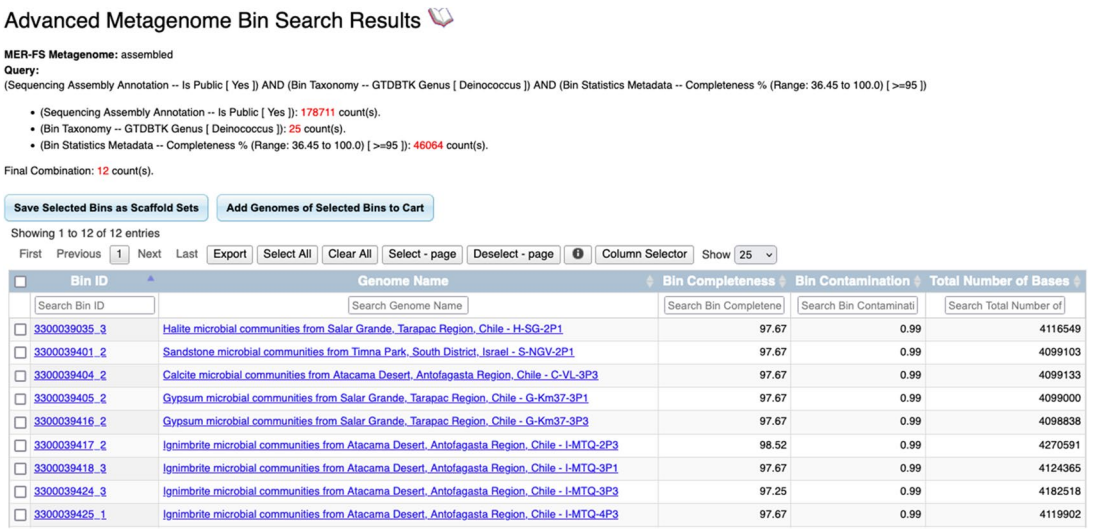
translesion (TLS) DNA polymerase. In contrast, other radiation-resistant species (*D. deserti and D. ficus*) do possess TLS polymerase genes, and it is possible to obtain UV-radiation-induced mutants in these strains. The ability to generate mutants using UV stress is mediated by a mutagenesis cassette described previously[17]. We will use the Find Genes > Cassette Search tool to survey the presence of colocalized genes encoding this mutagenesis cassette (*lexA–imuB–dnaE—PLASMID origin*) across all *Deinococccus* isolate genomes. Using the Pfams and Tigrfam representing these functions (Fig. 11a), we can retrieve 11 instances of colocalized error-prone DNA repair genes (mutagenesis cassettes) from nine distinct *Deinococcus* spp. (including previously known *D. deserti* and *D. ficus*) (Fig. 11b).

Various tools can be employed within the Cassette Workspace to assess the similarity of the 11 mutagenesis cassettes such as a function heatmap or similarity network plot (Fig. 12). These 11 cassettes organize roughly into three clusters and two singletons.

The relative distribution of this cassette in the context of species phylogeny suggests potential acquisition via horizontal acquisition in some clades (Fig. 13). In *D. deserti*, the gene cassette is encoded on a plasmid, and plasmid-mediated acquisition in other species is possible. While currently in beta-testing, IMG implemented a new "GeNomad" pipeline (https://zenodo.org/record/7015982) that can predict plasmid scaffolds in draft genomes and metagenomes, and at least seven cassettes are suggested to be plasmid-borne.

Homolog	Percent Identity	E-value	Bit Score	Genome Name	Contig Length	Contig GC	Contig Read Depth
Ga0194112_101864911	99.58%	0.0e+00	2577	Freshwater microbial communities from Lake Tanganyika, Tanzania - TA2015016 Mahale Deep Cast 400m	1700	0.55	67.00
Ga0194111_100753282	99.50%	0.0e+00	2571	Freshwater microbial communities from Lake Tanganyika, Tanzania - TA2015033 Kigoma Deep Cast 300m	2812	0.58	20.00
Ga0194113_100942543	99.50%	0.0e+00	2571	Freshwater microbial communities from Lake Tanganyika, Tanzania - TA2015017 Mahale Deep Cast 200m	2651	0.60	18.00
Ga0194110_100847092	99.15%	0.0e+00	2542	Freshwater microbial communities from Lake Tanganyika, Tanzania - TA2015032 Kigoma Deep Cast 1200m	2678	0.60	40.00
Ga0209023_101508251	99.08%	0.0e+00	2538	Freshwater and sediment microbial communities from Lake Erie, Canada (SPAdes)	1588	0.55	19.00
Ga0352974_1025421	96.47%	0.0e+00	2331	Aquifer fluids microbial communities from Pleistocene sands, Araihazar, Bangladesh - B3 Site	1917	0.57	1.00
Ga0209617_100604031	96.40%	0.0e+00	2326	Freshwater microbial communities from dead zone in Lake Erie, Canada - CCB epilimnion July 2011 (SPAdes)	1576	0.56	17.00
Ga0233424_100508132	96.40%	0.0e+00	2326	Freshwater microbial communities from Lake Towuti, South Sulawesi, Indonesia - Watercolumn_Towuti2014_125_MG	1919	0.57	171.00

Figure 7: Screenshot showing top 8 BLAST results of *Deinococcus* sp. UR1 16S rRNA against a custom reference database of 16S rRNA genes arising from diverse metagenome samples available in IMG.



Advanced Metagenome Bin Search Results

MER-FS Metagenome: assembled
Query:
(Sequencing Assembly Annotation -- Is Public [Yes]) AND (Bin Taxonomy -- GTDBTK Genus [Deinococcus]) AND (Bin Statistics Metadata -- Completeness % (Range: 36.45 to 100.0) [>=95])

- (Sequencing Assembly Annotation -- Is Public [Yes]): 178711 count(s).
- (Bin Taxonomy -- GTDBTK Genus [Deinococcus]): 25 count(s).
- (Bin Statistics Metadata -- Completeness % (Range: 36.45 to 100.0) [>=95]): 46064 count(s).

Final Combination: 12 count(s).

Save Selected Bins as Scaffold Sets | Add Genomes of Selected Bins to Cart

Showing 1 to 12 of 12 entries

First Previous 1 Next Last Export Select All Clear All Select - page Deselect - page Column Selector Show 25

Bin ID	Genome Name	Bin Completeness	Bin Contamination	Total Number of Bases
Search Bin ID	Search Genome Name	Search Bin Completene	Search Bin Contaminati	Search Total Number of
3300039035_3	Halite microbial communities from Salar Grande, Tarapac Region, Chile - H-SG-2P1	97.67	0.99	4116549
3300039401_2	Sandstone microbial communities from Timna Park, South District, Israel - S-NGV-2P1	97.67	0.99	4099103
3300039404_2	Calcite microbial communities from Atacama Desert, Antofagasta Region, Chile - C-VL-3P3	97.67	0.99	4099133
3300039405_2	Gypsum microbial communities from Salar Grande, Tarapac Region, Chile - G-Km37-3P1	97.67	0.99	4099000
3300039416_2	Gypsum microbial communities from Salar Grande, Tarapac Region, Chile - G-Km37-3P3	97.67	0.99	4098838
3300039417_2	Ignimbrite microbial communities from Atacama Desert, Antofagasta Region, Chile - I-MTQ-2P3	98.52	0.99	4270591
3300039418_3	Ignimbrite microbial communities from Atacama Desert, Antofagasta Region, Chile - I-MTQ-3P1	97.67	0.99	4124365
3300039424_3	Ignimbrite microbial communities from Atacama Desert, Antofagasta Region, Chile - I-MTQ-3P3	97.25	0.99	4182518
3300039425_1	Ignimbrite microbial communities from Atacama Desert, Antofagasta Region, Chile - I-MTQ-4P3	97.67	0.99	4119902

Figure 8: High-quality *Deinococcus* metagenome bins (or MAGs). High quality is delineated based on bin completeness of ≥ 95% (CheckM).

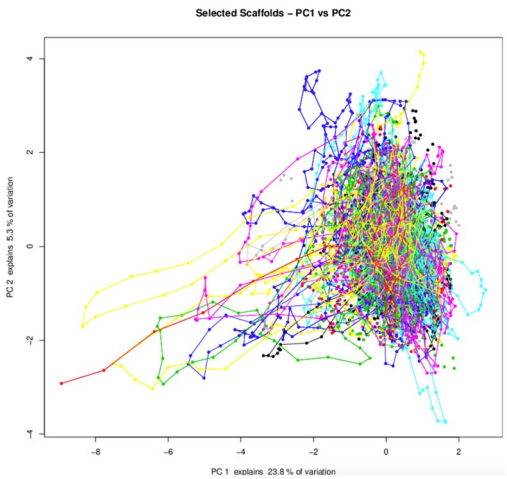


Figure 9: Kmer frequency plot of sandstone endolithic community metagenome bin (3300039401_2). Individual scaffolds are colored. Tool is accessed either through the scaffold cart or scaffold workspace. For isolate genomes, tool is also accessible through the genome details pages.

6 Statistical Analysis Tool

To see whether the presence of mutagenesis cassette is correlated with other differences in functional complement of *Deinococcus spp*. we can perform further genome comparisons by delineating two "genotype groups"—genomes with a TLS polymerase mutagenesis cassette and those without. Functional differences between these two groups of *Deinococcus* strains can be explored using a statistical analysis tool accessible through the Genome Workspace. Two workspace genome sets are first created—9 genomes in the TLS + group versus 40 genomes (dereplicated to remove near-identical strains or low-quality genomes) in the TLS group. A Mann–Whitney test with multiple testing correction is performed based on counts of genes assigned to functions (e.g., Pfam) between the two groups. Other statistical methods and other input choices are also available (Fig. 14).

Comparing the presence versus absence using "Absolute" gene counts of individual Pfams, only 45 Pfams are found to be significantly enriched (FDR adjusted P-value ≤ 0.05) in the TLS + group. As expected, these include the Pfams corresponding to the 3-gene cassette that was used to delineate the TLS + group (Fig. 15). Interestingly, among other Pfams, there is a domain of unknown function (DUF6504) encoded by a small intervening conserved hypothetical protein found in many TLS cassettes. Annotation of DUF6504 can be examined more closely through the gene details page (example). The gene is only about 80 amino acids with no other available annotations. Compare this to four Pfam domains and many other

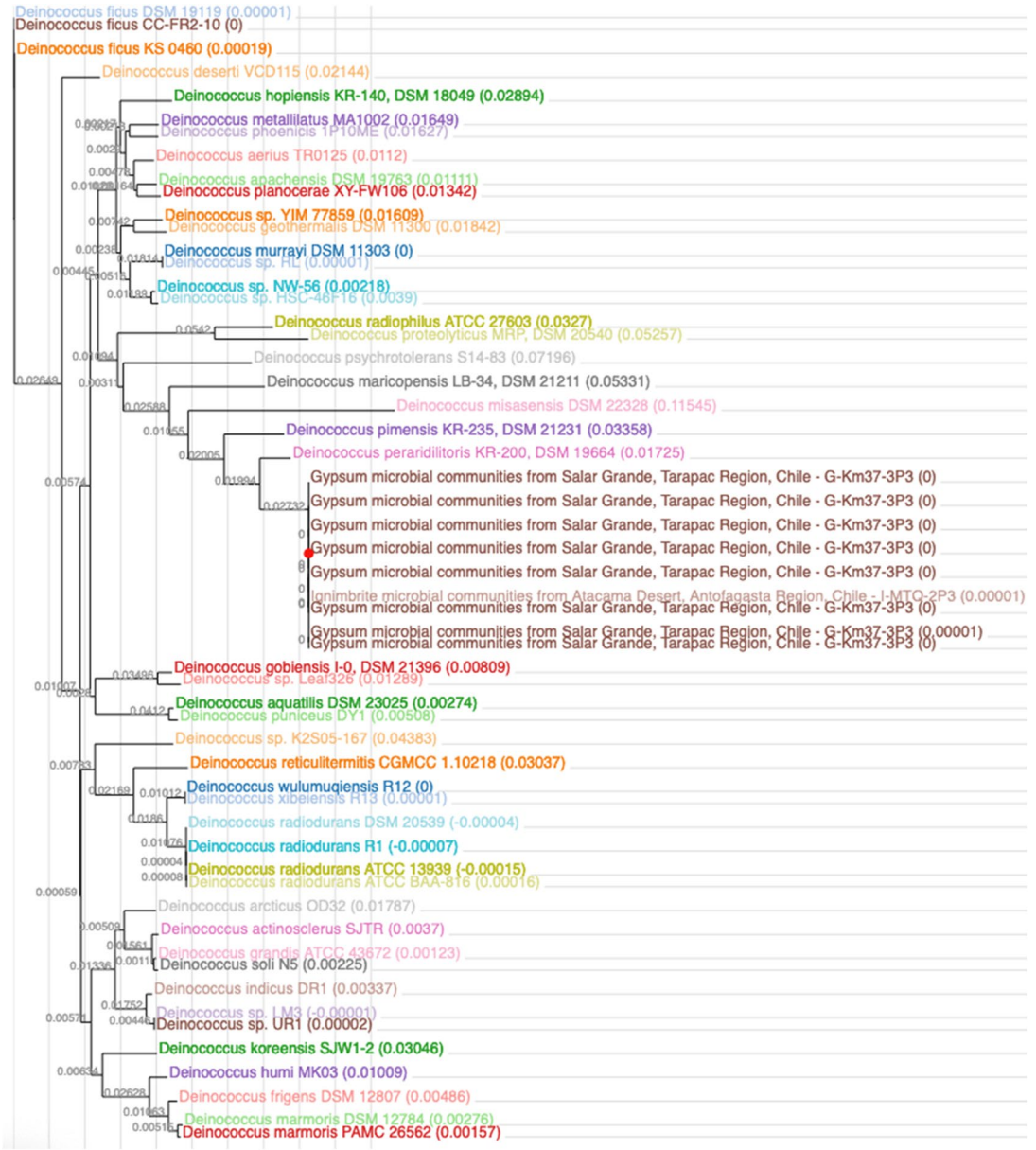


Figure 10: RpoB sequences from metagenome bins or MAG are added to isolate reference tree.

annotation details available for DnaE protein (example).

Results and a full matrix of all Pfam gene counts per genome can be downloaded and explored further. This comparative tool can be particularly valuable when discrete groups can be discerned based on distinct genotype or phenotype traits. A YouTube tutorial and documentation provide more details about this tool. Comparisons can also be extended to combine MAGs with isolates if desired using the analysis data groups feature.

It is undeniable that there are many limitations innate to genomic analysis starting with potential sequencing errors, misassembly, and annotation inaccuracies arising from underlying assumptions in the gene finding process. However, the IMG system attempts to mitigate some of these issues by implementing state-of-the-art data processing and computational analysis tools as detailed in the accompanying paper in this issue[7] Many genome assembly and annotation quality metrics, and other contextual data are also available to aid the

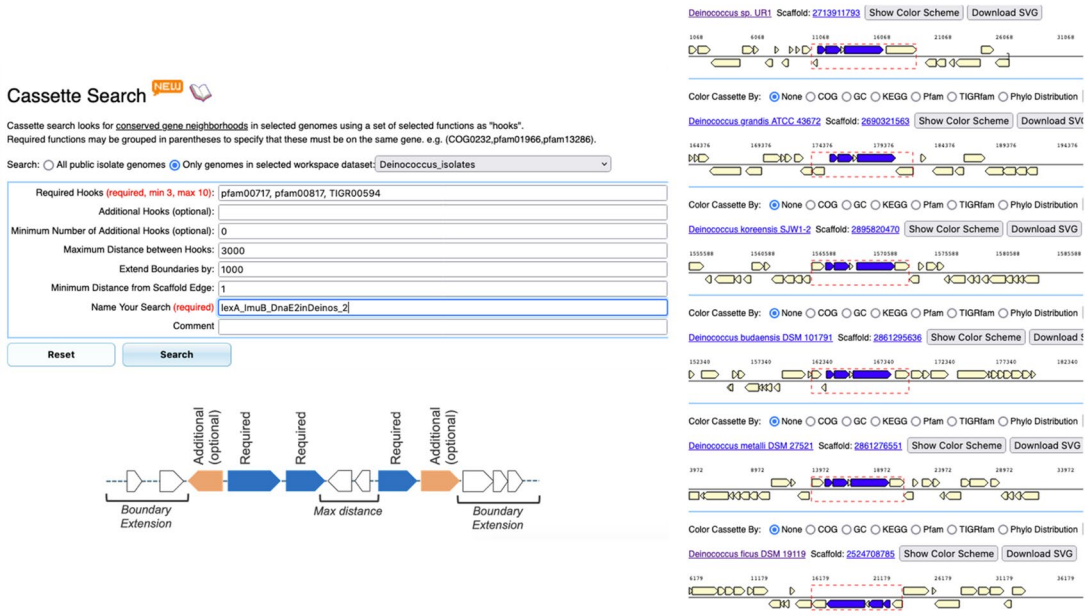


Figure 11: **a** Cassette search input page and **b** results of cassette search can be saved as "Cassette Sets" in Workspace. Gene neighborhoods of cassettes can be viewed alongside from the cassette workspace. Genes corresponding to "required hooks" specified in the search input as colored blue. The boundaries of the requested cassettes as specified are outlined by the red box. Only a subset of cassette results are shown for ease of display.

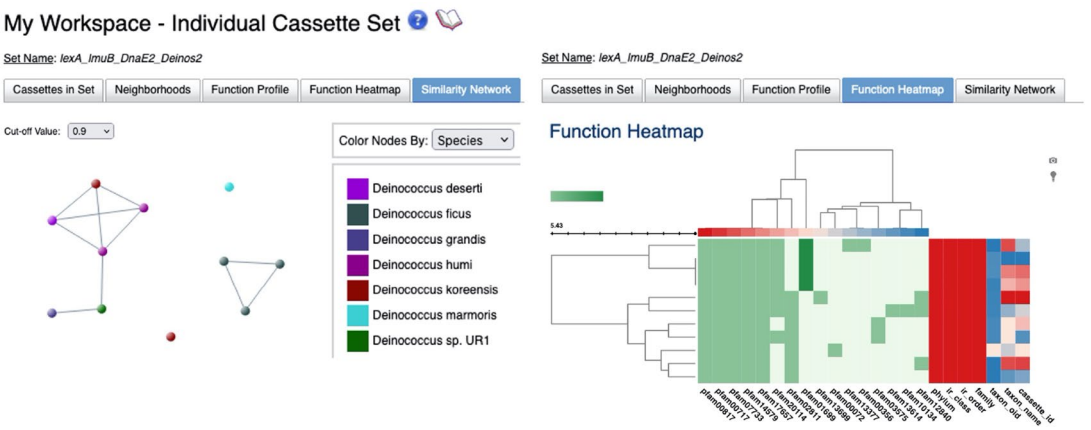


Figure 12: Additional tools in the Cassette Workspace. Similarity network graph (left) to summarize the data showing 3 *D. ficus* cassettes forming a distinct group (dark green) and separated from other partially linked clusters and singletons. Nodes are colored by species. Function heatmap (right) can be used to visualize the Pfam content of cassettes. Cells are colored with hues of green based on the number of copies of the selected Pfam in the cassette (darker signifies a higher copy number). Rows are individual cassettes, while columns are Pfams that occur in all cassettes and define the core functions of the cassette.

researcher. Tools like Kmer frequency plots and precomputed ANI clusters are also valuable for quality assessment. Isolate genome proteins are uniformly annotated and updated allowing the user to reliably compare proteomes. Furthermore, annotations resulting from multiple

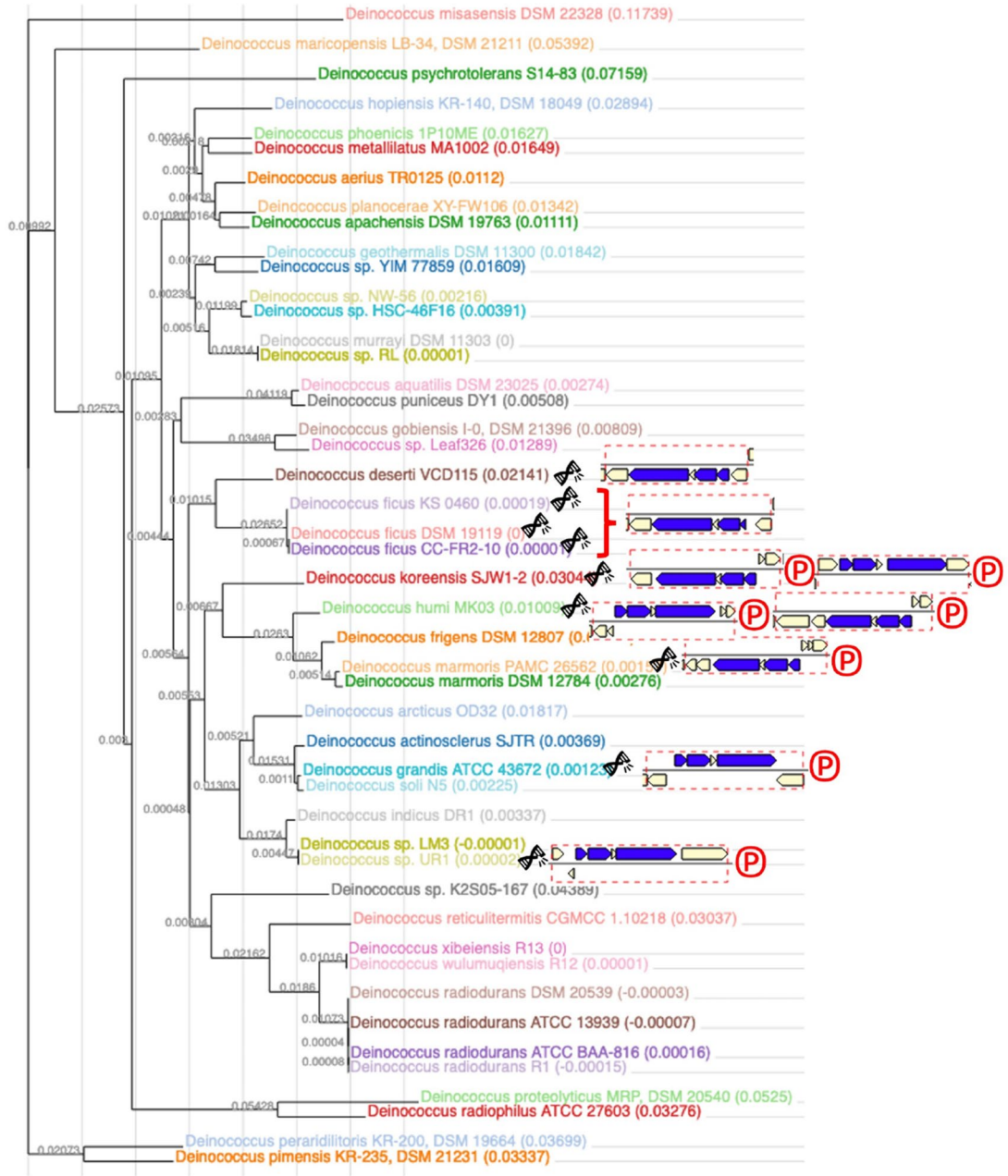


Figure 13: Isolate reference RpoB tree showing genomes containing TLS gene cassettes. For genomes with more than one cassette, both gene clusters are shown alongside. Putative plasmid-borne prediction is indicated by the "P".

databases (e.g., Pfam, KEGG, Tigrfam, COG, and SuperFam) provide an internal consistency check while maximizing the opportunity to make biological inferences.

While an exhaustive description of all tools and features is beyond the scope of this article, a suite of features and tools supporting metagenome analysis is also available. Video tutorials and other help documents are offered through

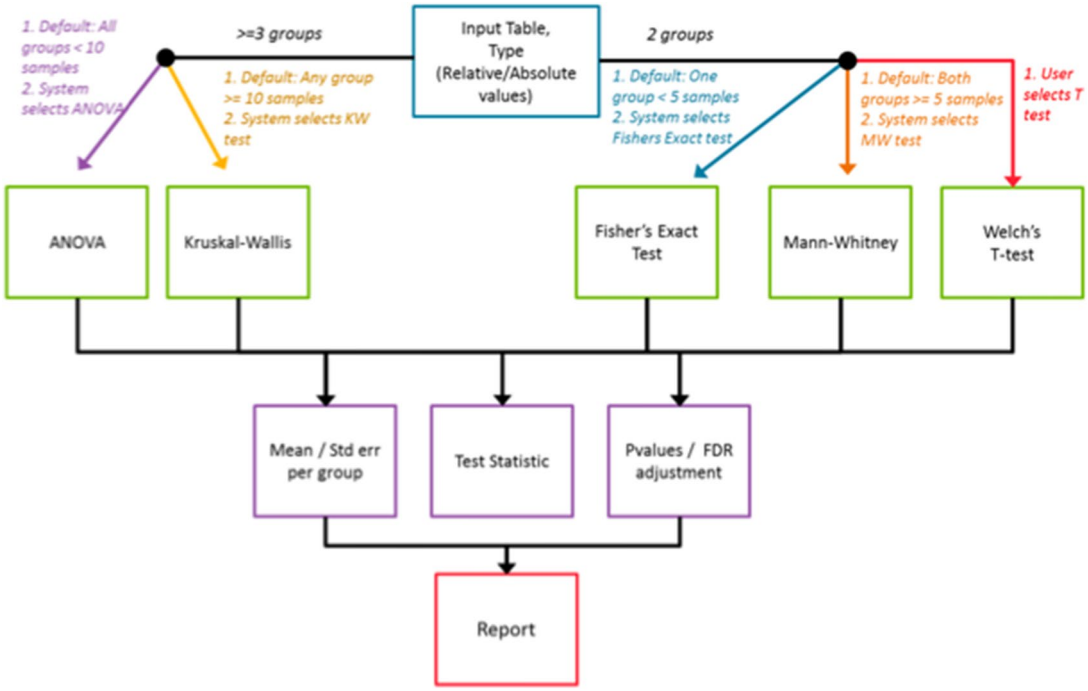
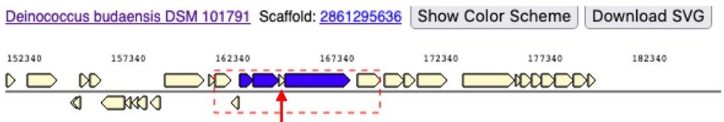


Figure 14: Decision tree for selection of default statistical test method. One of five statistical methods may be applied depending on size and number of input datasets (FDR—false discovery rate).



Feature	Description	Mean TLS_negative_Deino(n=40)	Mean TLS_polymerase_Deino(n=13)	MWTest UStat	MWTest adjPval
pfam07733	Bacterial DNA polymerase III alpha NTPase domain DnaE	1.15	2.76923	14	2.010e-06
pfam20114	Family of unknown function (DUF6504)	0.025	1.30769	43.5	2.010e-06
pfam14579	Helix-hairpin-helix motif DnaE	1.175	2.76923	19	2.276e-06
pfam17657	Bacterial DNA polymerase III alpha subunit finger domain DnaE	1.175	2.76923	19	2.276e-06
pfam00817	impB/mucB/samB family ImuB	0.825	2.38462	51.5	0.00252462
pfam01558	Pyruvate ferredoxin/flavodoxin oxidoreductase	0	0.461538	140	0.00252462
pfam01855	Pyruvate flavodoxin/ferredoxin oxidoreductase, thiamine diP-bdg	0	0.461538	140	0.00252462
pfam02811	PHP domain DnaE	3.125	4.30769	59.5	0.00252462
pfam02574	Homocysteine S-methyltransferase	1.05	1.61538	113	0.0026126
pfam00717	Peptidase S24-like LexA	1.6	3.61538	58	0.00591072

Figure 15: Top ten Pfam results of Mann–Whitney statistical comparison of genomes in TLS+ versus TLS− groups. Pfams corresponding to known TLS genes are indicated with red text. Location of DUF6504 is highlighted in the context of an example TLS gene cassette.

the user interface, and in-person or virtual workshops are offered worldwide upon request.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Funding

The work was conducted by the Joint Genome Institute (https://ror.org/04xm1d337), supported by the Office of Science of the U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231. We also provide proposal DOIs associated with the JGI-generated datasets analyzed in this review (Supplementary Table 2).

Declarations

Conflict of interest

The authors declare no competing financial interests.

Supplementary Information

Below is the link to the electronic supplementary material.Supplementary file1 (XLSX 36 KB)
Supplementary file2 (XLSX 10 KB)

Received: 15 December 2022 Accepted: 4 March 2023
Published online: 21 June 2023

References

1. Horne WH et al (2022) Effects of desiccation and freezing on microbial ionizing radiation survivability: considerations for mars sample return. Astrobiology 22:1337–1350
2. Olsson-Francis K, Cockell CS (2010) Experimental methods for studying microbial survival in extraterrestrial environments. J Microbiol Methods 80:1–13
3. Lown JW, Sim SK, Chen HH (1978) Hydroxyl radical production by free and DNA-bound aminoquinone antibiotics and its role in DNA degradation. Electron spin resonance detection of hydroxyl radicals by spin trapping. Can J Biochem 56:1042–1047
4. Krisko A, Radman M (2013) Biology of extreme radiation resistance: the way of *Deinococcus radiodurans*. Cold Spring Harb Perspect Biol 5(7):a012765
5. Lim S, Jung JH, Blanchard L, de Groot A (2019) Conservation and diversity of radiation and oxidative stress resistance mechanisms in Deinococcus species. FEMS Microbiol Rev 43:19–52
6. Makarova KS et al (2001) Genome of the extremely radiation-resistant bacterium *Deinococcus radiodurans* viewed from the perspective of comparative genomics. Microbiol Mol Biol Rev 65:44–79
7. Mukherjee S et al (2023) Bioinformatics analysis tools for studying microbiomes at the DOE joint genome institute. J Indian Inst Sci
8. White O et al (1999) Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. Science 286:1571–1577
9. Li XQ, Du D (2014) Variation, evolution, and correlation analysis of C+G content and genome or chromosome size in different kingdoms and phyla. PLoS ONE 9:e88339
10. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res 25:1043–1055
11. Chaumeil PA et al (2019) GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics 36(6):1925–1927
12. Andersson AM, Weiss N, Rainey F, Salkinoja-Salonen MS (1999) Dust-borne bacteria in animal sheds, schools and children's day care centres. J Appl Microbiol 86:622–634
13. Blin K et al (2019) antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. Nucleic Acids Res 47:W81–W87
14. Varghese NJ et al (2015) Microbial species delineation using whole genome sequences. Nucleic Acids Res 43:6761–6771
15. Dulermo R et al (2015) Identification of new genes contributing to the extreme radioresistance of *Deinococcus radiodurans* using a Tn5-based transposon mutant library. PLoS ONE 10:e0124358
16. Jin M et al (2019) The diversity and commonalities of the radiation-resistance mechanisms of Deinococcus and its up-to-date applications. AMB Express 9:138
17. Zeng YH, Shen FT, Tan CC, Huang CC, Young CC (2011) The flexibility of UV-inducible mutation in *Deinococcus ficus* as evidenced by the existence of the imuB-dnaE2 gene cassette and generation of superior feather degrading bacteria. Microbiol Res 167:40–47

Rekha Seshadri is a highly experienced computational biologist affiliated with the US Department of Energy (DoE) Joint Genome Institute (JGI), a leading research institution dedicated to advancing the field of genomics. With a strong focus on studying the diversity, ecology, and functional capabilities of microbes in various environments, Dr. Seshadri's research aids the development of biotechnological solutions for sustainable agriculture, environmental remediation, carbon sequestration, bioenergy production, and much more. She is also a co-developer of the IMG/M system for comparative microbial genomics and metagenomics which enables scientists to bridge the gap from sequence to biology.

Nikos C. Kyrpides has over 20 years of research and leadership experience in genomics, bioinformatics and microbiome data science. At the US DoE Joint Genome Institute (JGI), he led the Genome Biology Program managing large interdisciplinary research teams and built state-of-the-art data management systems for the analysis of genomics and metagenomics data and metadata. He has received many awards and recognitions for his contributions to microbiology and genomic research. His innovative research has led to the development of new methods and techniques for studying microbial diversity and function, pushing the boundaries of genomics research and opening new possibilities for understanding the microbial world. His current research focus is on Microbiome Data Science and the analysis of Big Data. His group is developing novel methods for enabling large-scale comparative analysis, as well as mining and visualization of big data.

Natalia N. Ivanova is a highly accomplished research scientist at the US DoE Joint Genome Institute (JGI) with over 20 years' experience in microbial genomics and bioinformatics. She has been at the forefront of genomic research leading the development of high-throughput annotation and analysis pipelines and integration and analysis of various types of omics data. She is also the architect and co-developer of the IMG/M system for comparative microbial genomics and metagenomics. Her research has expanded our understanding of microbial diversity and the potential applications of genomics in environmental and biotechnological research.