

# Lawrence Berkeley National Laboratory

## LBL Publications

### Title

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

### Permalink

<https://escholarship.org/uc/item/2w77t5fd>

### Journal

Journal of the Indian Institute of Science, 103(3)

### ISSN

0970-4140

### Authors

Seshadri, Rekha  
Kyrpides, Nikos C  
Ivanova, Natalia N

### Publication Date

2023-07-01

### DOI

10.1007/s41745-023-00368-7

### Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

## **Comparative Genomics using the Integrated Microbial Genomes & Microbiomes (IMG/M) System - a *Deinococcus* Use Case**

**Rekha Seshadri\*, Nikos Kyrpides, Natalia N. Ivanova**

DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720

### **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.

\*To whom correspondence should be addressed: [rseshadri@lbl.gov](mailto:rseshadri@lbl.gov)

## 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.

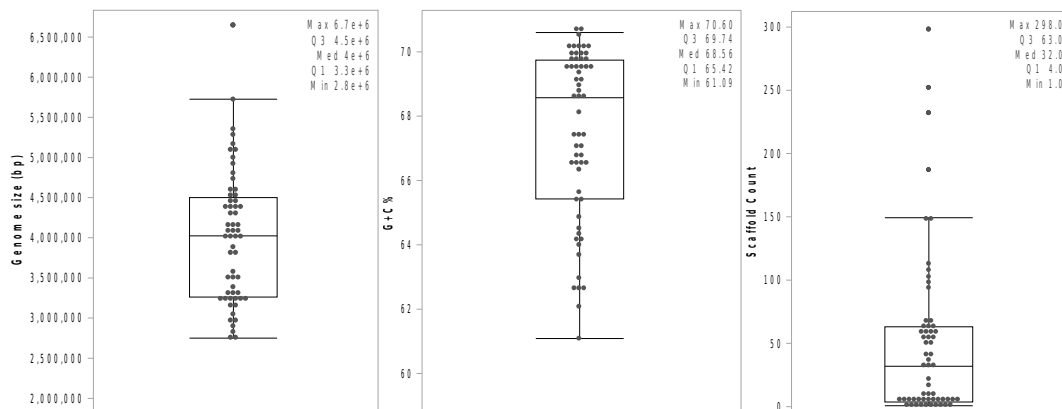
One such model organism is *Deinococcus radiodurans*, a polyextremophile, famously resistant to radiation, desiccation, and many toxic chemicals (3). This resistance is linked to its ability to recover following exposure to diverse kinds of damage, which is lethal to most organisms, and is mediated by 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* in order 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 (4) examine the distribution of a previously characterized mutagenesis cassette to highlight strains that may possibly serve as a chassis for strain engineering efforts.

## Exploring Genomes & 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 by Reddy et. al.), we retrieve 47 isolate genomes (finished and draft assemblies) by specifying a total genome scaffold count of  $\leq 600$ . 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 [You tube tutorial](#). *Deinococcus* spp. genome sizes range from 2.75 - 6.65 Mbp (Fig. 1). *Deinococcus 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 (7).

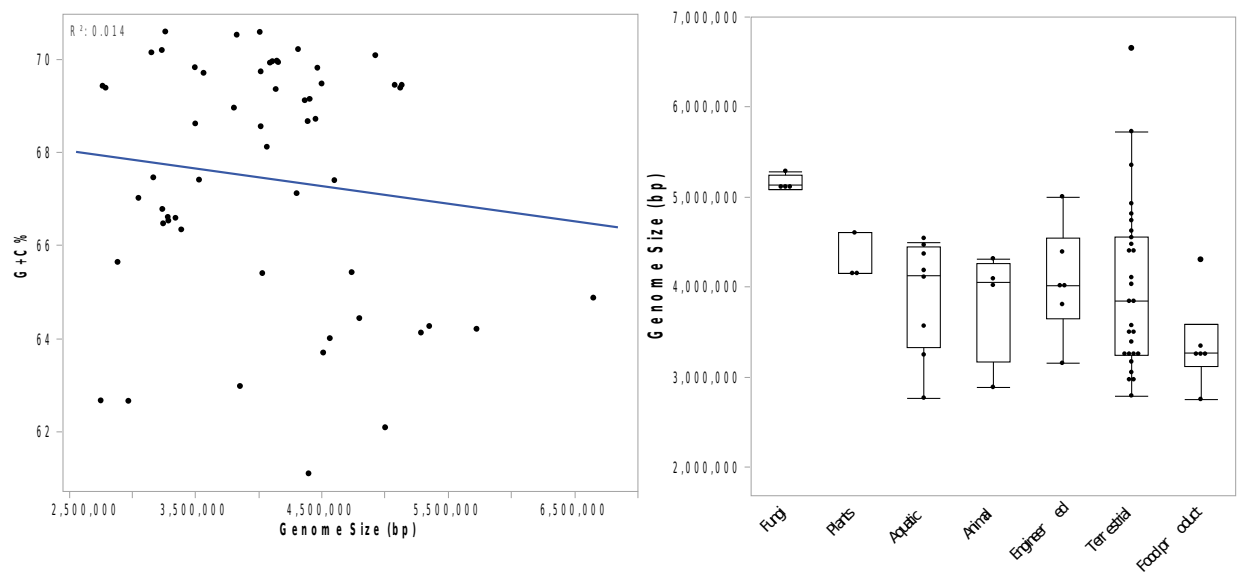


**Figure 1.** Genome Size (panel 1), G+C% (panel 2) and total scaffold counts distribution for 60 *Deinococcus* spp isolates is shown. Five quantile summary is included in each plot.

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 (8), there is little correlation between genome size and G+C % (Fig. 2a).

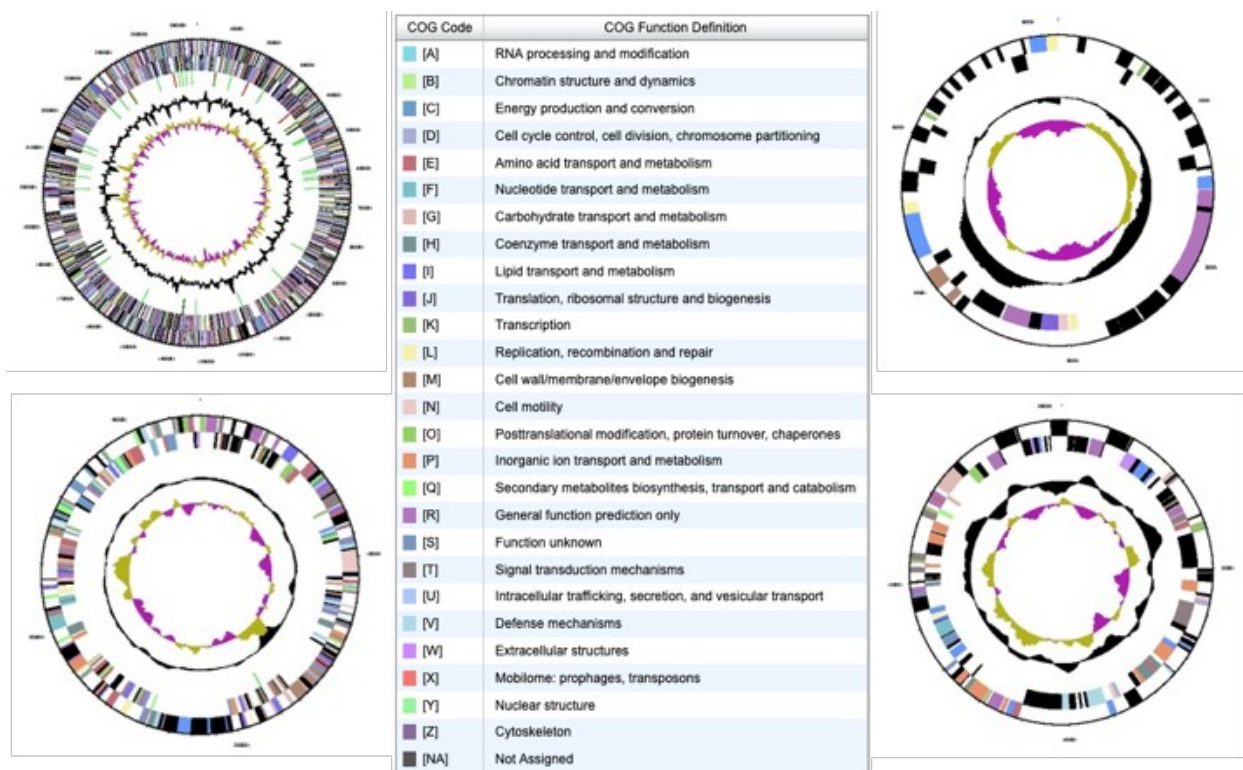
Recently added and highly useful fields include contamination and completeness estimates from checkM (9) and GTDB-Tk taxonomy (RE). CheckM scores are provided to help with assessment of the quality of individual genomes, while GTDB-Tk provides a genome-based taxonomy, which is based on a set of universally 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 (10).



**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.

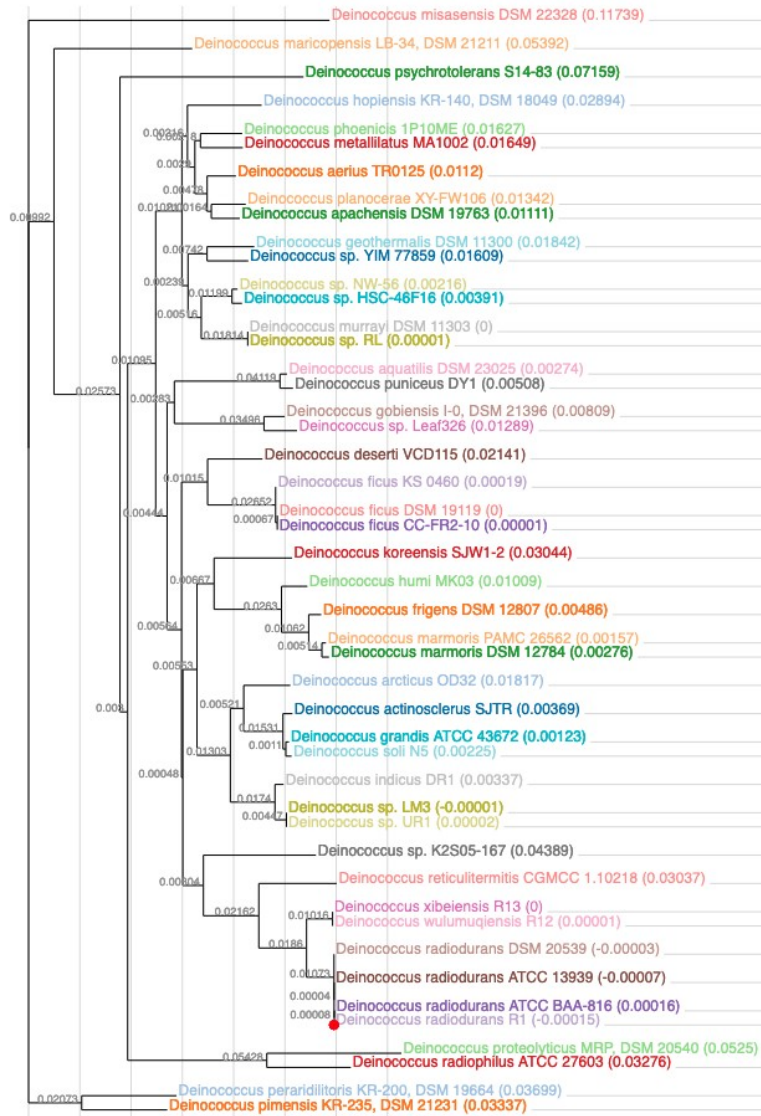
The system provides results of searches of proteomes performed against various public functional annotation sources (as described in the accompanying article by Reddy et. al.) - offering maximal opportunity to make biological inferences based on these various annotations (COG, Tigrfam, Pfam, Kegg Orthology, Superfam, etc.), which are partially overlapping, but have distinct focuses 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 (11)). 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.



**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.

### 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, an 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](#).





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

Pairwise ANI(12) (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(12).

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
<a href="#">Deinococcus radiodurans R1</a>	<a href="#">Deinococcus radiodurans DSM 20539</a>	99.998	99.9989	96.697	98.786
<a href="#">Deinococcus radiodurans R1</a>	<a href="#">Deinococcus radiodurans ATCC 13939</a>	99.9931	99.9964	96.757	98.887
<a href="#">Deinococcus radiodurans R1</a>	<a href="#">Deinococcus radiodurans ATCC BAA-816</a>	99.9873	99.9872	91.582	91.369
<a href="#">Deinococcus radiodurans R1</a>	<a href="#">Deinococcus xibeiensis R13</a>	86.7908	86.8041	65.036	66.635
<a href="#">Deinococcus radiodurans R1</a>	<a href="#">Deinococcus wulumuqiensis R12</a>	86.7801	86.7925	66.511	66.589
<a href="#">Deinococcus radiodurans R1</a>	<a href="#">Deinococcus reticulitermitis CGMCC 1.10218</a>	80.1028	80.1034	52.066	48.964
<a href="#">Deinococcus radiodurans R1</a>	<a href="#">Deinococcus gobiensis I-0. DSM 21396</a>	79.5021	79.5024	45.936	36.378
<a href="#">Deinococcus radiodurans R1</a>	<a href="#">Deinococcus sp. UR1</a>	79.2860	79.2627	23.930	22.451
<a href="#">Deinococcus radiodurans R1</a>	<a href="#">Deinococcus sp. Leaf326</a>	78.8123	78.8030	46.035	33.987
<a href="#">Deinococcus radiodurans R1</a>	<a href="#">Deinococcus phoenicis 1P10ME</a>	78.4891	78.4902	44.214	40.681
<a href="#">Deinococcus radiodurans R1</a>	<a href="#">Deinococcus sp. HSC-46F16</a>	78.4136	78.4053	41.517	42.983
<a href="#">Deinococcus radiodurans R1</a>	<a href="#">Deinococcus sp. NW-56</a>	78.3738	78.3704	41.421	38.021
<a href="#">Deinococcus radiodurans R1</a>	<a href="#">Deinococcus metallilatus MA1002</a>	78.2310	78.2281	43.801	36.367
<a href="#">Deinococcus radiodurans R1</a>	<a href="#">Deinococcus koreensis SJW1-2</a>	78.2121	78.2133	42.875	32.226
<a href="#">Deinococcus radiodurans R1</a>	<a href="#">Deinococcus actinosclerus SJTB</a>	78.1765	78.1793	43.748	36.131
<a href="#">Deinococcus radiodurans R1</a>	<a href="#">Deinococcus ficus KS 0460</a>	78.1574	78.1589	41.533	34.232
<a href="#">Deinococcus radiodurans R1</a>	<a href="#">Deinococcus ficus CC-FR2-10</a>	78.1499	78.1578	41.449	33.450
<a href="#">Deinococcus radiodurans R1</a>	<a href="#">Deinococcus ficus DSM 19119</a>	78.1489	78.1504	41.441	33.060
<a href="#">Deinococcus radiodurans R1</a>	<a href="#">Deinococcus soli N5</a>	78.0870	78.0869	39.624	40.557
<a href="#">Deinococcus radiodurans R1</a>	<a href="#">Deinococcus aerius TR0126</a>	78.0610	78.0632	42.335	31.188
<a href="#">Deinococcus radiodurans R1</a>	<a href="#">Deinococcus arcticus OD32</a>	78.0437	78.0255	41.481	34.846
<a href="#">Deinococcus radiodurans R1</a>	<a href="#">Deinococcus grandis ATCC 43672</a>	77.9951	77.9819	42.858	35.059
<a href="#">Deinococcus radiodurans R1</a>	<a href="#">Deinococcus sp. K2S05-167</a>	77.9695	78.0197	43.394	32.715
<a href="#">Deinococcus radiodurans R1</a>	<a href="#">Deinococcus indicus DR1</a>	77.9220	77.9067	43.296	32.525
<a href="#">Deinococcus radiodurans R1</a>	<a href="#">Deinococcus sp. LM3</a>	77.8555	77.8729	43.329	33.503
<a href="#">Deinococcus radiodurans R1</a>	<a href="#">Deinococcus apachensis DSM 19763</a>	77.8288	77.8535	42.077	32.313
<a href="#">Deinococcus radiodurans R1</a>	<a href="#">Deinococcus planocerae XY-FW106</a>	77.8042	77.8070	41.128	32.077
<a href="#">Deinococcus radiodurans R1</a>	<a href="#">Deinococcus murrayi DSM 11303</a>	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.

Functional content (based on CDS protein sequence assignments to COG, Pfam, KO, Tigrfam) may be compared across all genomes 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 Find Genes > Phylogentic 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 (13) (14).

Phylogenetic Profiler for Single Genes 

Sequencing Status: All Finished, Permanent Draft and Draft | Domain: Genome Cart

List  Tree  Show  Selected: 1


Search for: <enter a genome name to search>

- Deinococcus proteolyticus MRP, DSM 20540 (B) [F]
- Deinococcus psychrotolerans S14-83 (B) [F]
- Deinococcus puniceus DY1 (B) [P]
- Deinococcus puniceus DY1 (B) [F]
- Deinococcus radiodurans ATCC 13939 (B) [P]
- Deinococcus radiodurans ATCC BAA-816 (B) [F]
- Deinococcus radiodurans DSM 20539 (B) [P]
- Deinococcus radiodurans R1 (B) [F]**
- Deinococcus radiophilus ATCC 27603 (B) [P]
- Deinococcus radiopugnans DSM 12027 (B) [P]
- Deinococcus radiopugnans DY59 (B) [F]
- Deinococcus reticulitermitis CGMCC 1.10218 (B) [P]
- Deinococcus soli BE323 (B) [P]
- Deinococcus soli BE330 (B) [P]
- Deinococcus soli BE73 (B) [P]
- Deinococcus soli N5 (B) [F]
- Deinococcus soli N5 (B) [D]
- Deinococcus sp. HSC-46F16 (B) [P]
- Deinococcus sp. K2S05-167 (B) [P]
- Deinococcus sp. Leaf326 (B) [P]
- Deinococcus sp. LM3 (B) [P]

**Selected Genomes**

Step 1. Find Genes In

Deinococcus radiodurans R1 (B) [F]

Add Remove 

Step 2a. With Homologs In(max. 100) 5

- Deinococcus radiodurans ATCC 13939 (B) [P]
- Deinococcus radiodurans ATCC BAA-816 (B) [F]
- Deinococcus radiodurans DSM 20539 (B) [P]
- Deinococcus wulumuqlensis R12 (B) [P]
- Deinococcus xibeiensis R13 (B) [P]

Add Remove

Step 2b. Without Homologs In(max 100) 56

- Deinococcus sp. HSC-46F16 (B) [P]
- Deinococcus sp. K2S05-167 (B) [P]
- Deinococcus sp. Leaf326 (B) [P]
- Deinococcus sp. LM3 (B) [P]
- Deinococcus sp. LM3 (B) [F]
- Deinococcus sp. NW-56 (B) [F]
- Deinococcus sp. RL (B) [P]
- Deinococcus sp. UR1 (B) [P]
- Deinococcus sp. YIM 77859 (B) [P]
- Deinococcus yavapaiensis DSM 18048 (B) [P]

Add Remove

Submit

**Figure 6.** Screenshot of data input page for pairwise sequence based (USEARCH) comparisons of specified genomes. Tool can be found under Find Genes > Phylogenetic Profilers > Single Genes.

### 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  $\geq 97\%$  identity hits, which roughly corresponds to a species level 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](#)).

Homolog	Percent Identity	E-value	Bit Score	Genome Name	Contig Length	Contig GC	Contig Read Depth
<a href="#">Ga0194112_101864911</a>	99.58%	0.0e+00	2577	<a href="#">Freshwater microbial communities from Lake Tanganyika, Tanzania - TA2015016 Mahale Deep Cast 400m</a>	1700	0.55	67.00
<a href="#">Ga0194111_100753282</a>	99.50%	0.0e+00	2571	<a href="#">Freshwater microbial communities from Lake Tanganyika, Tanzania - TA2015033 Kigoma Deep Cast 300m</a>	2812	0.58	20.00
<a href="#">Ga0194113_100942543</a>	99.50%	0.0e+00	2571	<a href="#">Freshwater microbial communities from Lake Tanganyika, Tanzania - TA2015017 Mahale Deep Cast 200m</a>	2651	0.60	18.00
<a href="#">Ga0194110_100847092</a>	99.15%	0.0e+00	2542	<a href="#">Freshwater microbial communities from Lake Tanganyika, Tanzania - TA2015032 Kigoma Deep Cast 1200m</a>	2678	0.60	40.00
<a href="#">Ga0209023_101508251</a>	99.08%	0.0e+00	2538	<a href="#">Freshwater and sediment microbial communities from Lake Erie, Canada (SPAdes)</a>	1588	0.55	19.00
<a href="#">Ga0352974_1025421</a>	96.47%	0.0e+00	2331	<a href="#">Aquifer fluids microbial communities from Pleistocene sands, Araihasar, Bangladesh - B3 Site</a>	1917	0.57	1.00
<a href="#">Ga0209617_100604031</a>	96.40%	0.0e+00	2326	<a href="#">Freshwater microbial communities from dead zone in Lake Erie, Canada - CCB epilimnion July 2011 (SPAdes)</a>	1576	0.56	17.00
<a href="#">Ga0233424_100508132</a>	96.40%	0.0e+00	2326	<a href="#">Freshwater microbial communities from Lake Towuti, South Sulawesi, Indonesia - Watercolumn Towuti2014_125_MG</a>	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 metagenome samples available in IMG.

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  $\geq 95\%$  checkM completeness from primarily endolithic communities (Fig. 8). An 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 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.

### 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

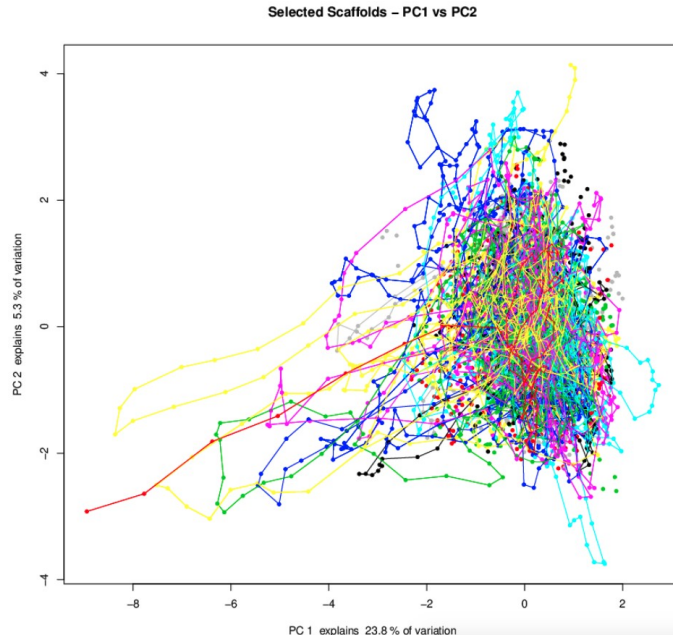
Add Genomes of Selected Bins to Cart

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

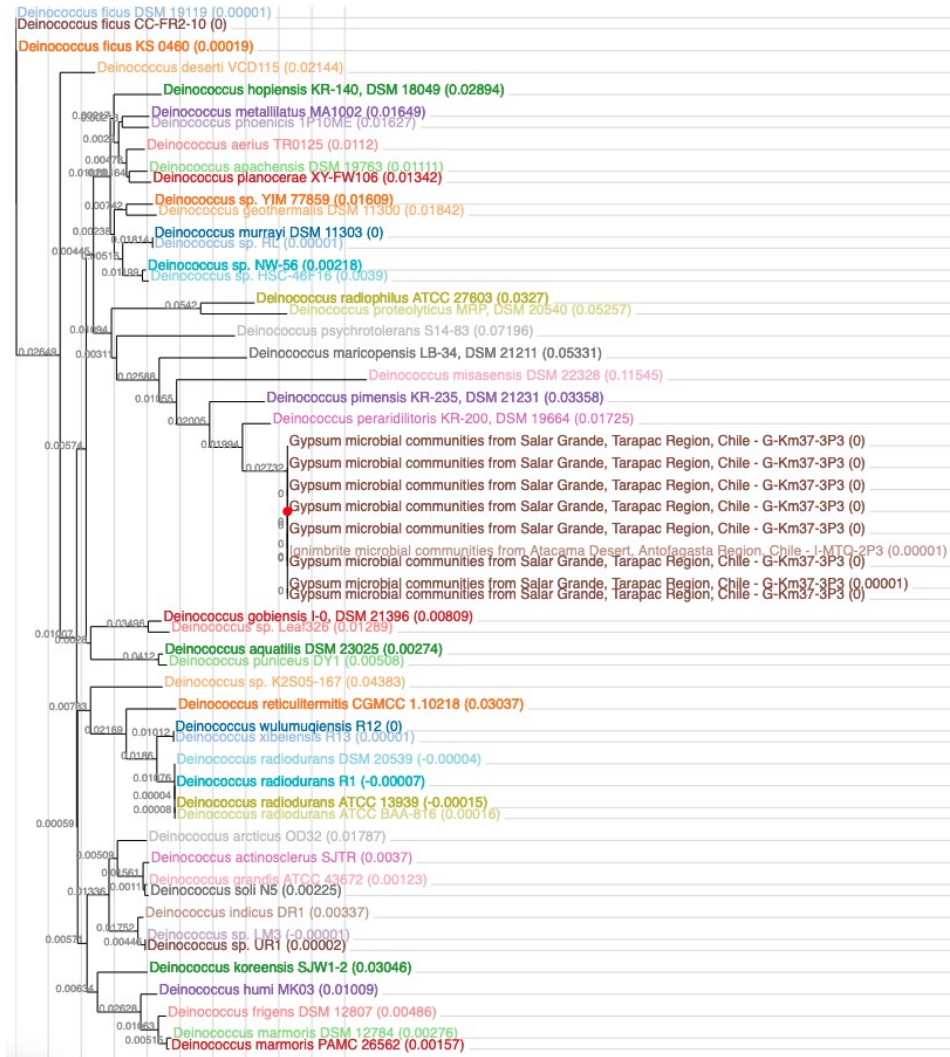
Bin ID	Genome Name	Bin Completeness	Bin Contamination	Total Number of Bases
<input type="checkbox"/> 3300039035_3	<a href="#">Halite microbial communities from Salar Grande, Tarapac Region, Chile - H-SG-2P1</a>	97.67	0.99	4116549
<input type="checkbox"/> 3300039401_2	<a href="#">Sandstone microbial communities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

**Figure 8.** High quality *Deinococcus* metagenome bins (or MAGs). High quality is delineated based on bin completeness of  $\geq 95\%$  (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.

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.

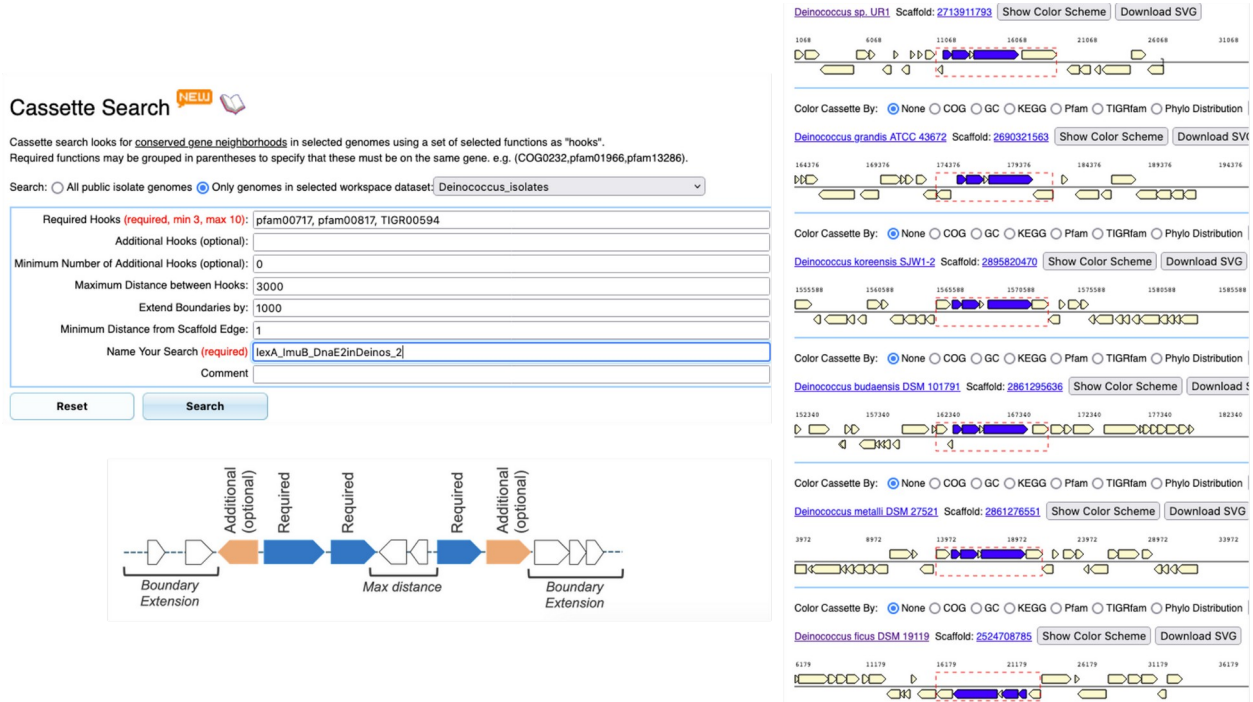


**Figure 10.** RpoB sequences from metagenome bins or MAG are added to isolate reference tree.

## 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 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 (15). We will use the IMG Cassette Search tool to survey the presence of colocalized genes

encoding this mutagenesis cassette (*lexA-imuB-dnaE* – *PLASMID* origin) across all *Deinococcus* 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)



**Figure 11. a**, cassette search input page **b**, results of cassette search can be saved as “Cassette Sets” in Workspace. Gene neighborhoods of cassettes can be view 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.

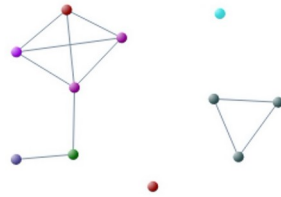
Various tools can be employed within the Cassette Workspace to assess the similarity of the 11 mutagenesis cassettes such as a function heatmap or similarity network plot (Fig. 12). These 11 cassettes organize roughly into 3 clusters and 2 singletons.

## My Workspace - Individual Cassette Set

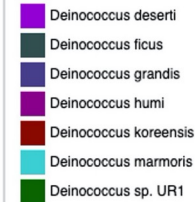
Set Name: *lexA\_ImuB\_DnaE2\_Deinos2*

Cassettes in Set Neighborhoods Function Profile Function Heatmap **Similarity Network**

Cut-off Value: 0.9



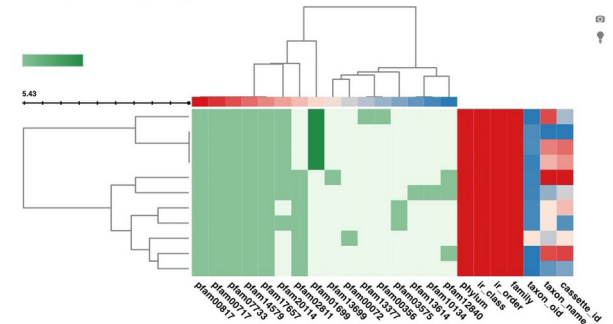
Color Nodes By: Species



Set Name: *lexA\_ImuB\_DnaE2\_Deinos2*

Cassettes in Set Neighborhoods Function Profile **Function Heatmap** Similarity Network

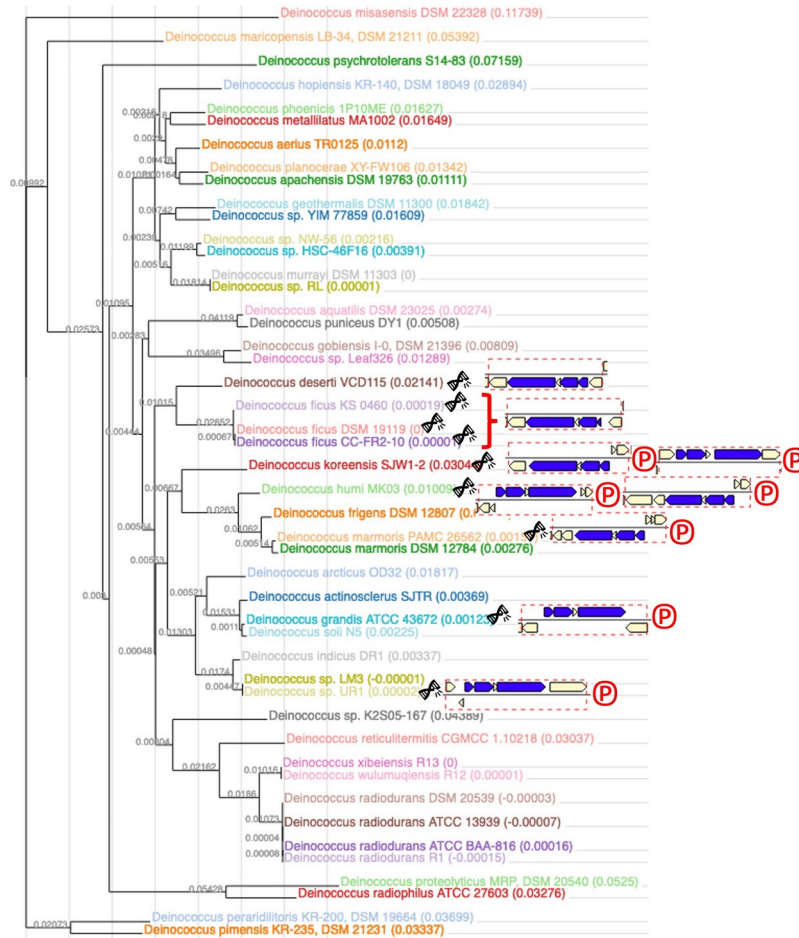
### Function Heatmap



**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.

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 (DOI: <https://zenodo.org/record/7015982>) that can predict plasmid scaffolds in draft genomes and metagenomes, and at least 7 cassettes are suggested to be plasmid-borne.



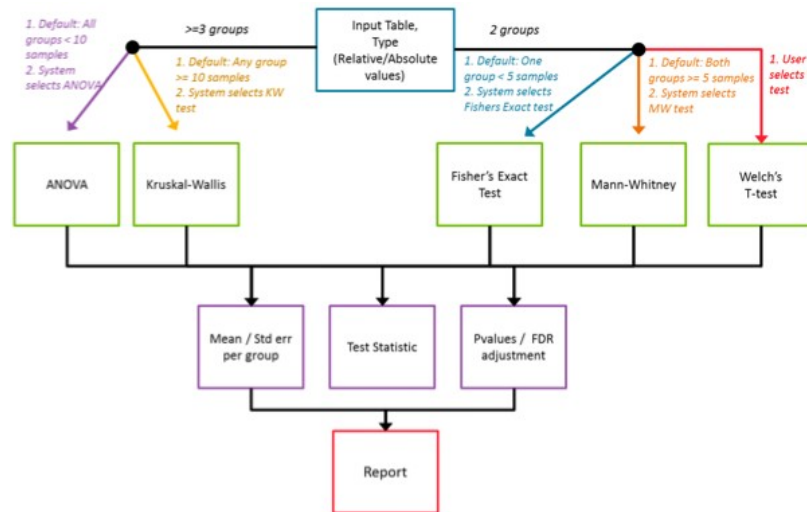


**Figure 13.** Isolate reference RpoB tree showing genomes containing TLS gene cassettes. For genomes with more than one cassette, both gene clusters are shown alongside. Putative plasmid-borne prediction is indicated by the “P.”

## Statistical Analysis Tool

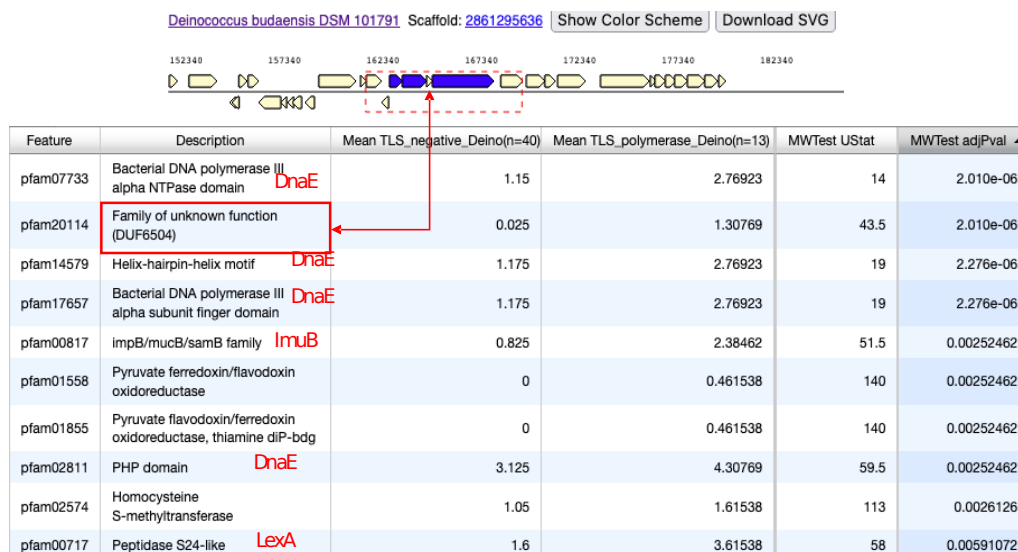
In order to see whether the presence of mutagenesis cassette is correlated with other differences in functional complement of *Deinococcus spp.* we can perform further genome comparisons by delineating two “genotype groups” – genomes with a TLS polymerase mutagenesis cassette and those without. Functional 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).



**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).

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  $\leq 0.05$ ) 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 annotation details available for DnaE protein ([example](#)).



**Figure 15.** Top 10 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.

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, mis-assembly 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 by Reddy et. al. Many genome assembly and annotation quality metrics, and other contextual data are also available to aid the 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 databases (e.g., Pfam,

KEGG, Tigrfam, COG, 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 the user interface, and in-person or virtual [workshops](#) are offered worldwide upon request.

### **Declaration of Interests**

The authors declare no competing financial interests.

### **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).

### **REFERENCES**

1. W. H. Horne *et al.*, Effects of Desiccation and Freezing on Microbial Ionizing Radiation Survivability: Considerations for Mars Sample Return. *Astrobiology* **22**, 1337-1350 (2022).
2. K. Olsson-Francis, C. S. Cockell, Experimental methods for studying microbial survival in extraterrestrial environments. *J Microbiol Methods* **80**, 1-13 (2010).
3. J. W. Lown, S. K. Sim, H. H. Chen, 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 (1978).
4. A. Krisko, M. Radman, Biology of extreme radiation resistance: the way of *Deinococcus radiodurans*. *Cold Spring Harb Perspect Biol* **5**, (2013).

5. S. Lim, J. H. Jung, L. Blanchard, A. de Groot, Conservation and diversity of radiation and oxidative stress resistance mechanisms in *Deinococcus* species. *FEMS Microbiol Rev* **43**, 19-52 (2019).
6. K. S. Makarova *et al.*, Genome of the extremely radiation-resistant bacterium *Deinococcus radiodurans* viewed from the perspective of comparative genomics. *Microbiol Mol Biol Rev* **65**, 44-79 (2001).
7. O. White *et al.*, Genome sequence of the radioresistant bacterium *Deinococcus radiodurans* R1. *Science* **286**, 1571-1577 (1999).
8. X. Q. Li, D. Du, Variation, evolution, and correlation analysis of C+G content and genome or chromosome size in different kingdoms and phyla. *PLoS One* **9**, e88339 (2014).
9. D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, G. W. Tyson, CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* **25**, 1043-1055 (2015).
10. A. M. Andersson, N. Weiss, F. Rainey, M. S. Salkinoja-Salonen, Dust-borne bacteria in animal sheds, schools and children's day care centres. *J Appl Microbiol* **86**, 622-634 (1999).
11. K. Blin *et al.*, antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. *Nucleic Acids Res* **47**, W81-W87 (2019).
12. N. J. Varghese *et al.*, Microbial species delineation using whole genome sequences. *Nucleic Acids Res* **43**, 6761-6771 (2015).
13. R. Dulermo *et al.*, Identification of new genes contributing to the extreme radioresistance of *Deinococcus radiodurans* using a Tn5-based transposon mutant library. *PLoS One* **10**, e0124358 (2015).
14. M. Jin *et al.*, The diversity and commonalities of the radiation-resistance mechanisms of *Deinococcus* and its up-to-date applications. *AMB Express* **9**, 138 (2019).
15. Y. H. Zeng, F. T. Shen, C. C. Tan, C. C. Huang, C. C. Young, 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 (2011).

