**What Makes Extremophiles So Extreme?**

**A Comparative Genomic Analysis of the Extremophilic Bacterium *Deinococcus radiodurans***

**Mae Bothoff-Shanahan**

**Abstract**

This study investigates the comparative genomics of extremophiles, particularly *Deinococcus radiodurans,* exploring horizontal gene transfer (HGT) with other extremophiles like *Thermus thermophilus*, other *Dienococcus* species, and *Thermococcus gammatolerans*. Advanced bioinformatic tools were employed to analyze genomic data, aiming to unravel the genetic adaptations for survival in harsh environments.

The results revealed significant gene conservation and diversity among extremophiles. Genes like RecA displayed high BLAST scores and low e-values across species, indicating a high degree of conservation and their role in genomic integrity in extreme conditions. The presence of similar genes in diverse bacteria such as *Pseudomonas* and *Aeromonas* suggests widespread genetic exchange, possibly via HGT.

Additionally, significant BLAST hits within the plasmids of *Deinococcus proteolyticus* imply that certain genes may be part of mobile genetic elements facilitating HGT. The study identified two distinct gene groups: highly conserved genes likely shared through HGT, and less conserved genes indicating unique evolutionary adaptations

In conclusion, the research highlights the complex landscape of gene conservation, adaptation, and potential HGT among extremophiles, contributing to the understanding of their genetic resilience in extreme environments. These findings underscore the importance of HGT in microbial evolution, particularly among extremophiles, highlighting the importance for future research in extremophile biology.

**Introduction**

**Background**

The study of extremophiles has gained attention in the field of genomics due to their remarkable adaptations for survival. Extremophiles are microorganisms, often single-celled, that thrive in extreme and harsh environments which would be lethal to most life forms. These environments encompass extreme temperature ranges (thermophiles and psychrophiles), high radiation levels (radiophiles), acidic or alkaline conditions (acidophiles and alkaliphiles), high salinity (halophiles), and vacuums like outer space [1]. Extremophiles have evolved a diverse array of molecular mechanisms to withstand these extreme conditions.

For example, one especially hardy extremophile is *D. radiodurans*, often referred to as "Conan the Bacterium" due to its legendary resistance to ionizing radiation and desiccation (drying out due to lack of water). *D. radiodurans* can endure radiation doses thousands of times higher than what would be lethal to humans, as well as extreme desiccation and oxidative stress. Researchers have identified several genes that are responsible for DNA repair, antioxidation, and radiation resistance mechanisms through comparative genomics, shedding light on the genetic adaptations that allow *D. radiodurans* to thrive in such challenging environments[2]. Details about those genes, later used as target genes in this study, are as follows;

* PprA (Protein protecting DNA during radiation A): PprA is a central regulator in *D. radiodurans* that plays a crucial role in DNA repair and radiation resistance. It helps protect and repair damaged DNA.
* RecA: The RecA protein is involved in homologous recombination, a DNA repair process that is particularly efficient in *D. radiodurans*. It helps in repairing double-strand breaks in DNA.
* DdrA, DdrB, and DdrC: These genes are part of the *Deinococcal* DNA damage response (Ddr) system and are involved in DNA repair and radiation resistance. They help repair DNA damage and maintain genomic integrity.
* Ku and LigD: *D. radiodurans* possesses a non-homologous end-joining (NHEJ) DNA repair pathway mediated by Ku and LigD proteins. This pathway is important for repairing double-strand breaks in DNA. It’s also known as NADH-quinone oxidoreductase subunit N.
* Peroxiredoxins (Dps and Bcp): These enzymes help protect the bacterium from the damaging effects of radiation-induced reactive oxygen species.
* DR 0423 (PprI): Pprl, also known as IrrE, is another protein involved in the radiation resistance of *D. radiodurans*. It interacts with PprA and participates in the regulation of DNA repair processes.
* Ssb (Single-stranded DNA binding protein): This protein helps protect single-stranded DNA during DNA repair processes, ensuring its stability.
* PolA (DNA polymerase I): It is involved in DNA replication and repair and plays a role in maintaining the integrity of the genome.
* ThyA (Thymidylate Synthase): In *D. radiodurans*, the gene responsible for thymidylate synthase, which is involved in DNA synthesis and repair, is DR2630 (COG0207). It is worth noting that this gene seems to be acquired from a different source compared to its counterpart in *T. thermophilus*.
* Purine-Nucleoside Phosphorylase: The gene responsible for purine-nucleoside phosphorylase, an enzyme involved in purine metabolism, is DR2166 (COG0813) in *D. radiodurans*. The gene TTC1070 (COG0813) in *T. thermophilus* appears to be involved in a similar process.

Comparative genome analysis is a fundamental approach in genomics that involves the systematic comparison of the genetic material of different organisms to identify similarities, differences, and evolutionary relationships[3]. This approach enables researchers to determine the genetic basis of specific traits, such as adaptations or unique characteristics. In the context of computer science, comparative genomics relies on a range of programming languages and software tools. Commonly used programming languages include Python and R, which can manipulate and analyze genomic data efficiently. Additionally, software packages like BLAST, Galaxy, the IMG/M System[4], and MEGA X[5] have all been utilized for sequence alignment, identification, and phylogenetic analysis.

Comparative genomic methods have proven to be valuable in unraveling the genetic underpinnings of extremophile survival strategies and their adaptations from one extremophile species to another[6]. This particularly applies to horizontal gene transfer. Horizontal gene transfer (HGT) is a pivotal process in the evolution of extremophiles. It involves the exchange of genetic material across different species through three main mechanisms: transformation (uptake of environmental DNA), transduction (transfer via viruses), and conjugation (direct cell-to-cell transfer). HGT accelerates adaptation by allowing extremophiles to rapidly acquire genes essential for survival in harsh conditions, such as genes for heat-shock proteins or enzymes functioning under extreme stress[7]. This process not only enhances individual survival but also contributes to the genetic diversity and resilience of microbial communities in extreme habitats, enabling a dynamic response to environmental challenges.

**Significance**

Comparative genome analysis allows researchers to unravel the genetic adaptations that enable extremophiles to thrive in environments that would be deadly to most life forms. By studying the genomes of these extremophiles, science gains a deeper understanding of the specific genes and mechanisms responsible for their extraordinary resilience[6]. This knowledge not only advances our understanding of fundamental biology but also has potential practical applications like biotechnology and medicine.

Knowledge gained from studying extremophiles can potentially have far-reaching implications for society. For instance, insights into radiation resistance mechanisms, as seen in *Deinococcus radiodurans*, may have applications in cancer research and treatment[8]. Understanding how extremophiles repair DNA damage caused by radiation and how they acquired those genes could potentially inform strategies for improving the radiation resistance of normal human cells, reducing side effects in radiation therapy, and enhancing cancer treatment outcomes.

.

The primary objective of this research project was to confirm the occurrence of horizontal gene transfer in the extremophilic bacterium *D. radiodurans*. Specifically, the aim was to validate the hypothesis that *D. radiodurans* has acquired critical genes related to stress response and radiation resistance through horizontal gene transfer from other extremophiles, with a focus on its close relative, *Thermus thermophilus,* three other *Deinococcus* species, and a distant extremophile *Thermococcus gammatolerans*. By conducting a comparative genomic analysis using advanced bioinformatic tools and statistical tests, significant sequence similarities between *D. radiodurans* and its close relatives were identified. The project's significance lies in contributing to our understanding of the genetic mechanisms underlying the exceptional adaptability of *D. radiodurans* to extreme environments, which can have implications for biotechnology, medicine, and the existing knowledge of evolutionary processes.

The results of this study revealed notable patterns in the genomic relationships among extremophiles, highlighting the role of horizontal gene transfer (HGT). Key genes like RecA showed significant sequence similarities across various extremophiles, indicating a shared evolutionary background potentially driven by HGT. This was particularly evident in *D. radiodurans* when compared with species like *T. thermophilus* and other *Deinococcus* species, as seen in high BLAST scores and low e-values. The analysis also pointed to gene conservation and adaptation with essential genes for survival displaying high conservation across species, while others exhibited variability, suggesting species-specific evolutionary changes. The presence of similar genes in diverse bacteria like *Pseudomonas* and *Aeromonas* suggests wider genetic exchanges across different environmental niches. Additionally, the role of plasmids in *D. proteolyticus* highlighted the significance of mobile genetic elements in extremophiles’ genetic evolution.

Overall, the study underscores the complexity of genetic adaptation in extremophiles and the pivotal role of HGT in their evolution, offering insights that could support future research and applications in diverse scientific fields.

**Related Work**

Omelchenko et al. conducted a comparative genomics analysis of the thermophile *T. thermophilus* and the radiation-resistant mesophile *D. radiodurans*. By reconstructing the evolutionary histories of each lineage, they found extensive horizontal gene transfer contributed to the thermophilic adaptation of *T. thermophilus* and stress resistance of *D. radiodurans*[6]. This study provided key genomic insights into the distinct extremophile traits of these bacteria. However, further experimental validation is needed to characterize the functional roles of proposed horizontally transferred genes. This research served as the inspiration for this project, with the aim of confirming and expanding upon these horizontal gene transfers.

Liu et al. expanded on the taxonomy of an extremophilic red alga through comparative genomic analysis combined with morphological and physiological data. This multifaceted evidence supported the designation of a new genus *Cyanidiococcus*, with genomic characteristics distinct from the closely related *Galdieria* and *Cyanidioschyzon* genera[5]. While Liu's work provided valuable taxonomic clarity, it underscored the need for additional examinations of diverse strains to validate the proposed genus. This study exemplifies the importance of multidisciplinary approaches in extremophile research, with genomics playing a pivotal role.

Ellington et al. identified specific DNA repair and antioxidant genes enabling extreme UV radiation tolerance in stratospheric bacteria through comparative genomics and directed evolution experiments. This study effectively linked certain genes to increased UV resistance phenotypes[1]. However, a limitation of the study was the investigation of only a limited number of bacterial strains, warranting further genomic analyses across a more extensive range of extremophile isolates.

These studies collectively highlight the potential of comparative genomics in determining the genetic underpinnings of extremophile traits, whether in the context of taxonomy, evolutionary origins, or specific adaptive genes. A common theme among these studies is the necessity to expand genomic analyses across more strains and establish links between proposed genomic determinants and functional phenotypes, which is what will be attempted in this project to a lesser degree.

**Methodology**

**Data Collection**

**Genomic Data Retrieval:** The genomic data for the extremophile *Deinococcus radiodurans* was sourced from GenBank using individual RefSeq assembly numbers for each chromosome and plasmid due to Biopython's limitations with the primary assembly number GCA\_020546685.1. These numbers include NZ\_CP038663.1 and NZ\_CP038664.1 for chromosomes I and II, and NZ\_CP038666.1 and NZ\_CP038665.1 for plasmids pCP1 and pMP1, respectively. The same method was used to import the genome of *Thermus thermophilus* with NC\_006461.1 for the chromosome and NC\_006463.1 and NC\_006462.1 for plasmids pTT8 and pTT27. Additionally, genomic data for three other *Deinococcus* species was imported via the same method using their respective assembly numbers: *Deinococcus deserti* (NC\_012526.1 - NC\_012528.1), *Deinococcus proteolyticus* (NC\_015161.1 - NC\_015163.1), and *Deinococcus geothermalis* (NC\_008025.1, NC\_008010.2, and NC\_009939.1). For *Thermococcus gammatolerans*, a single assembly number was used (NC\_012804.1).

**Script Functionality and Error Handling:** The code for this process utilizes the BioPython library for retrieving and handling genomic data from GenBank. The purpose of this script is to programmatically access and download genomic data for the extremophiles listed above from GenBank, which can then be used for further analysis in bioinformatics research. The script is designed to handle errors and retry fetching the data.

**Initial Setup and Configuration**

**Importing Libraries and Setting Up Email:** The script starts with the importation of essential BioPython modules to facilitate bioinformatics tasks. It includes Entrez, for database access, SeqIO for sequence data manipulation, NCBIWWW and NCBIXML for BLAST operations, and the time module to introduce delays as needed. Additional libraries, such as pandas for data manipulation, scipy.stats for statistical analysis, matplotlib and seaborn for visualization, and threading with queue for concurrent execution, are loaded. Tabulate, re, and Counter are utilized for table formatting, regular expression operations, and tallying elements, respectively. An email address is provided to comply with NCBI's usage policy.

**Data Retrieval and Processing**

**Function to Retrieve Genomic Data (retrieve\_genomic\_data):** This function fetches genomic data from the GenBank database using a provided accession number. It attempts to download the data up to three times, parsing the retrieved GenBank files into record objects. Should the attempt fail, it waits two seconds before retrying.

**Genomic Data Retrieval Setup:** Accession numbers for various organisms and their genomic components are stored in a dictionary. The script iterates through these numbers, calls retrieve\_genomic\_data for each one, and aggregates the records into all\_organism\_sequences, categorizing them by organism.

**BLAST Analysis and Gene Sequence Extraction:**

**BLAST Execution Function (perform\_blast):** This function orchestrates a BLAST search on a given sequence against a specified database and organism. It employs threading to prevent the search from exceeding the timeout limit. The results are fetched via a queue system to facilitate the asynchronous operation.

**Gene Sequence Extraction Function (extract\_gene\_sequence):** Here, the script searches for a specific gene within the genomic records, using either gene names or locus tags. Successful finds are written to a file and the gene sequence is returned.

**BLAST Results Parsing Function (parse\_blast\_results):** This function processes the BLAST results, extracting the title, score, e-value, identities, and alignment length from each alignment, and compiles them into a list of dictionaries for further analysis.

**Display Function for BLAST Results (display\_blast\_results):** This utility function prints out the BLAST results for each gene in a readable format, detailing the alignment's score, e-value, identities, and length.

**Gene Identification and BLAST Result Management**

**Target Gene Dictionary Setup (target\_genes):** A dictionary is established, mapping gene names to their respective locus tags for *D. radiodurans*. It is utilized to identify and work with target genes within the organism's genomic data.

**BLAST Results Storage (blast\_results):** An empty dictionary is prepared to store the BLAST results for each target gene post-analysis.

**Gene Sequence Extraction and BLAST Analysis:** The script iterates over the target genes, extracting their sequences and conducting BLAST searches. The results are either directly parsed or, upon failure, subjected to a retry mechanism.

**BLAST Results Retry Function (retry\_blast\_for\_gene):** When initial BLAST searches fail, this function is invoked to attempt a retry, subsequently parsing and storing any successful results.

**Data Analysis and Visualization**

**Saving BLAST Results (save\_blast\_results\_to\_csv):** This function consolidates the BLAST results into a pandas DataFrame and saves it as a CSV file.

**Filtering and Displaying BLAST Results:** The script loads the BLAST results from a CSV file, filters them based on an e-value threshold, and saves significant hits. It then sorts and displays the top significant results in a formatted table.

**Organism Chart Creation:** The script extracts species names from the BLAST titles, counts occurrences, and generates a bar chart displaying the frequency of organisms in the BLAST results.

**Data Visualization**

**Visualization of BLAST Scores:** Histograms and boxplots are generated to visualize the distribution of BLAST scores and compare them across different genes, aiding in the interpretation of the results.

All code is available to the public on the Github repository for this project, located at <https://github.com/Cordeeceps/Extremophiles>.

**Results**

The comparative genomic analysis yielded significant insights for different target genes across different species. The results from the BLAST analysis for these genes highlight key metrics crucial for understanding genetic similarities, outlined below:

**Title:** The title in a BLAST result provides information about the matching sequence. It typically includes the GenBank identifier, the type of nucleotide or protein sequence, and the source organism. For example, a title could read, "gi|1012291606|gb|CP014867.1| Organism X, complete genome," indicating the specific sequence in the GenBank database that matches the query.

**Score:** The BLAST score quantifies the similarity between the query sequence and the matching sequence in the database. Higher scores indicate a greater degree of similarity. These scores are context-specific and vary based on the sequences compared, the database used, and search parameters. Absolute score values are less important than their relative significance in a given search context, typically assessed alongside E-values, with lower E-values indicating more statistically significant alignments. In this analysis, the gene ThyA had a high BLAST score of 603.0, suggesting a strong sequence similarity.

**E-value:** The E-value represents the number of matches one can expect to find by chance when searching a database of a particular size. E-values closer to zero indicate a more significant match, suggesting that the similarity is less likely to have occurred by random chance. For ThyA, an E-value of 4.90599e-152 signifies a highly significant match.

**Identities:** This metric refers to the number of exact matches in the nucleotide or amino acid sequence alignment between the query and the database sequence. A higher number of identities generally indicates closer evolutionary relationships or functional similarity. For instance, ThyA showed 597 identities in an alignment length of 794.

**Alignment Length:** This refers to the total number of nucleotides or amino acids compared in the alignment. Longer alignments with more identities are typically indicative of more extensive and significant similarities. The alignment length for ThyA was 794, which, along with a high number of identities, underscores the strong similarity.



***Thermus thermophilus* Results:**

**Table 1: Top 10 Significant BLAST Hits for *T. thermophilus*.** The table lists the genes from the BLAST query against *D. radiodurans* target genes with their corresponding highest-scoring homologs in various bacteria. Notably, sequences from *Pseudomonas aeruginosa* appear predominantly, suggesting a closer evolutionary relationship or functional similarity in the context of the genes analyzed.

A screenshot of a computer screen

Description automatically generated

The BLAST analysis of *Thermus thermophilus* resulted in a total of 1451 hits, with 331 hits having significant matches (e-values equal to or greater than 0.05). In addition to several high scoring matches within the *Thermus* genus, significant matches were found with a diverse array of common soil and aquatic bacteria, including *Pseudomonas, Azospirillum, Geobacter, Aeromonas*, and more.

**RecA Gene:** This gene showed high similarity across multiple bacterial species. The most significant matches were with *Thermus* species (e.g., *Thermus oshimai* JL-2, *Thermus thermophilus* strains) with the highest score of 597 and e-values as low as 2.89642e-150, suggesting strong genetic conservation.

**ThyA Gene**: This gene exhibited high conservation, particularly with *Pseudomonas* species. The highest score observed was 603, with e-values reaching 4.906e-152, indicating strong conservation across species.

**PNPase Gene:** This gene displayed a wide range of BLAST scores (from 136 to 250), with e-values ranging from 1.55e-56 to 1.36e-25. This range suggests diverse genetic similarities and potential functional adaptations across different species.

**PolA Gene (DNA Polymerase I):** Scores for this gene ranged from 112 to 371, with e-values from 1.65118e-88 to 9.15664e-35, reflecting both conserved functions and sequence diversity. Significant matches were found with various species, including *Thermus oshimai* JL-2 and *Thermus thermophilus* strains.

**Other Genes:** Genes like DdrA, DdrC, Ku, Dps, and IrrE also showed significant matches but generally with lower scores and higher e-values compared to RecA, ThyA, and PolA.

A white background with black and white lines

Description automatically generated with medium confidence

**Figure 1: Boxplot of BLAST Scores for T. thermophilus Across Different Genes.** The boxplot displays the range and distribution of BLAST scores obtained from comparing genes from *D. radiodurans* against the *T. thermophilus* genome. The genes 'ThyA' and 'PprA' show notably higher median scores and variability, suggesting a higher degree of similarity and conservation with the corresponding genes in *T. thermophilus* compared to the other genes analyzed. Of note, RecA has the highest median BLAST score.

A graph with blue bars

Description automatically generated

**Figure 2: Histogram of the Frequency Distribution of BLAST Scores for *T. thermophilus*.** This histogram presents the frequency of BLAST scores achieved when querying a set of genes from *D. radiodurans* against the *T. thermophilus* genome. Notable peaks suggest common score ranges, indicating clusters of genes with similar levels of similarity to the *T. thermophilus* genome.

A graph of a number of organisms

Description automatically generated

**Figure 3: Frequency of Organisms Identified in *T. thermophilus* BLAST Results.** The bar chart illustrates the count of different organisms that were identified as top hits in the BLAST analysis, using a selection of genes from *D. radiodurans* against *T. thermophilus*. The chart highlights the prevalence of specific organisms, such as *Pseudomonas aeruginosa*, which appears most frequently in the results.

**Deinococcus deserti Results**

**Table 2: Top Significant BLAST Results for *D. deserti*.** The table presents the highest scoring hits from the BLAST analysis of selected genes from *D. radiodurans* against the genomic sequences of *D. deserti*. The table showcases the notable similarity between the genes of *D. radiodurans* and those of *P. salivibrio*, as evidenced by the high scores and low e-values.

A screen shot of a computer

Description automatically generated

The BLAST analysis of *D. deserti* resulted in a total of 235 hits, with 5 hits having significant matches (e-values equal to or greater than 0.05). A high score (242.0) and a very low e-value (approximately 1.58e-57) indicate a strong similarity between the RecA gene in *D. radiodurans* and the sequence from *Pontimonas salivibrio*, one of the organisms within the dataset of *D. deserti.* The score (38.0) and e-value (0.021) of the DPS gene are relatively low, indicating a less strong similarity compared to RecA. DNA polymerase I (PolA) has two separate alignments with scores of 76.0 and 59.0 and e-values of 4.65e-12 and 3.57e-07 respectively suggesting a significant similarity.

A white rectangular object with green lines

Description automatically generated with medium confidence

**Figure 4: BLAST Score Comparison by Gene for *D. deserti*.** This boxplot visualizes the distribution of BLAST scores for three targeted genes from *D. radiodurans* when compared against *D. deserti*. RecA shows a narrower interquartile range, indicating less variance in BLAST scores, while PolA displays a broader score distribution, suggesting greater variability in sequence alignment quality.

A graph of a number of bars

Description automatically generated

**Figure 5: Histogram of BLAST Scores for *D. deserti*.** This histogram displays the frequency distribution of BLAST scores for *D. deserti* across a set of genes BLASTed against *D. radiodurans*. The scores are predominantly clustered around 100, with fewer occurrences as scores increase, indicating a concentration of moderate sequence similarity and fewer instances of high similarity scores.

***Deinococcus geothermalis* Results**

**Table 3: Top Significant BLAST Results for *D. geothermalis*.** This table presents the highest scoring BLAST hits when querying *D. radiodurans* target genes against the *D. geothermalis* genome, with incidental matches to other species sequences present within the queried genome database.

A screenshot of a computer program

Description automatically generated

The BLAST analysis of *Deinococcus geothermalis* resulted in a total of 18 hits, with 4 being identified as significant matches (e-values equal to or less than 0.05). All significant matches were found with the same mRNA sequence from *Heterobasidion irregulare*, an organism known for its wood decay capabilities.

A diagram with a bar graph

Description automatically generated with medium confidence

**Figure 6: BLAST Score Distribution for Selected Genes in *D. geothermalis*.** The boxplot displays the distribution of BLAST scores for the genes Dps, IrrE, and PolA when *D. radiodurans* target genes were used for BLAST analysis against the *D. geothermalis* genome.

A graph with blue lines

Description automatically generated

**Figure 7: Histogram of BLAST Scores for *D. geothermalis*.** This histogram illustrates the frequency distribution of BLAST scores obtained from querying *D. radiodurans* target genes against the *D. geothermalis* genome.

***Deinococcus proteolyticus* results**

**Table 4: Top 10 Significant BLAST Results for *D. proteolyticus*.** This table summarizes the highest scoring alignments obtained from a BLAST search using selected genes from *D. radiodurans* against the *D. proteolyticus* genome.

A screenshot of a computer program

Description automatically generated

The BLAST analysis of *Deinococcus proteolyticus* resulted in a total of 1616 hits, with 66 being identified as significant matches (e-values equal to or less than 0.05). The RecA gene from *D. proteolyticus* has a considerably high BLAST score (1298.0) with an e-value of 0.0, indicating an extremely significant match, likely representing a very close or identical sequence in the database. The genes PolA and ThyA also show highly significant matches (scores of 1255.0 and 1013.0, respectively).

Other genes like PprA, DdrA, DdrB, and DdrC show moderate to low BLAST scores, suggesting less conservation or partial matches with sequences in the database. Additionally, there are multiple hits to sequences within plasmids of *D. proteolyticus.*

A graph with a green and blue box

Description automatically generated with medium confidence

**Figure 8: BLAST Score Distribution for Selected Genes in *D. proteolyticus*.** This boxplot displays the range of BLAST scores for genes from *D. radiodurans* when queried against the *D. proteolyticus* genome, highlighting variations in sequence similarity across different genes.

A graph with blue squares

Description automatically generated

**Figure 9: Histogram of BLAST Scores for *D. proteolyticus*.** This histogram illustrates the frequency distribution of BLAST scores obtained from querying selected genes of *D. radiodurans* against the *D. proteolyticus* genome. The majority of scores cluster at the lower end of the score range, indicating a varying degree of homology, with a significant number of high-scoring alignments, suggesting regions of strong sequence conservation.

***Thermococcus gammatolerans* results**

**Table 5:** Summary of Top Significant BLAST Hits for *T. gammatolerans*. The table lists the most significant matches when genes from *D. radiodurans* were used as queries in a BLAST search against the *T. gammatolerans* genome.

A screen shot of a computer

Description automatically generated

The BLAST analysis for *Thermococcus gammatolerans* generated a total of 179 hits, among which 4 were identified as significant matches (e-values equal to or less than 0.05). All significant matches were found with sequences from *Thermococcus gammatoleran*s.

A white rectangular object with black lines

Description automatically generated

**Figure 10: BLAST Score Distribution by Gene for *T. gammatolerans*.** This boxplot visualizes the range and distribution of BLAST scores obtained when genes from *D. radiodurans* were used to query the *T. gammatolerans* genome.

A graph with blue bars

Description automatically generated

**Figure 11: Histogram depicting the distribution of BLAST scores for *Thermococcus gammatolerans.*** The frequency of BLAST scores resulting from the comparison of *D. radiodurans* target genes against the *T. gammatolerans* genome is shown.

**Discussion**

***Thermus thermophilus:***

The BLAST analysis of *T. thermophilus* demonstrates significant gene conservation within the *Thermus* genus, evident from high-scoring matches with near-zero e-values, especially for the RecA gene. This finding, illustrated in Figure 1, underscores the gene's critical role in DNA repair and its evolutionary significance. It aligns with the findings of Omelchenko et. al, which also emphasized extensive horizontal gene transfer (HGT) in *T. thermophilus*, particularly from thermophilic bacteria and archaea.[6]

The high conservation of the RecA gene, with a high median BLAST score, mirrors the Omelchenko et. al’s emphasis on conventional DNA-repair systems shared between *T. thermophilus* and *D. radiodurans.* This is further illustrated in Figure 1, where RecA shows a notable median BLAST score. This study also discovered a high conservation for the ThyA gene, crucial for DNA synthesis, as evidenced by high BLAST scores (up to 603) and very low e-values (as low as 4.906e-152), predominantly with *Pseudomonas* species. This finding, consistent with less score variability seen in Figure 1, resonates with the paper's analysis of shared stress response and repair mechanisms between these species.

In contrast, this study reveals a broader range of scores for PNPase, indicating diverse genetic similarities and potential functional adaptations across species depicted in Figure 2's bimodal distribution of BLAST scores. This diversity suggests unique evolutionary pressures faced by *T. thermophilus*, echoing Omelchenko et. al’s observation of gene acquisition from other thermophiles.

The DNA polymerase I gene (PolA) in the analysis displayed significant score variability (scores ranging from 371 to 136, e-values from 1.65e-88 to 1.36e-25), reflecting conserved core functions amid sequence diversity, likely due to species-specific environmental adaptations. This variability, highlighted in Figure 1, may indicate adaptations unique to *T. thermophilus*, consistent with the suggestion in Omelchenko et. al’s paper of mesophilic common ancestors and distinct traits acquired via HGT.

The observation of lower BLAST scores for genes like DdrA, DdrC, Ku, DPS, and IrrE aligns with the finding that *D. radiodurans* has several unique repair proteins absent in *T. thermophilus*.[6] This implies different evolutionary trajectories for these organisms, supported by the bimodal distribution of BLAST scores in Figure 2, signifying distinct gene groups and evolutionary paths.

Figure 3 further contextualizes these findings by showing a high frequency of hits for *Pseudomonas aeruginosa*, indicating gene conservation or similarity. The frequent hits with species like *Aeromonas hydrophila* and *Thermus thermophilus* support the sharing of fundamental genetic elements across species, underscoring the complex nature of HGT.

***Deinococcus deserti:***

The BLAST analysis of *D. deserti* presents a view of gene conservation, as illustrated in the histogram displaying BLAST scores (**Fig. 5**). This histogram shows a bimodal distribution, with peaks at both high and low ends. Scores around 250, such as the high score of 242.0 for the RecA gene, indicate robust conservation across various species, underscoring the gene’s importance in maintaining genome integrity. [9] This high level of conservation is further emphasized by the unusual presence of three distinct RecA genes in *D. deserti,* a rarity among bacteria that highlights its critical role in extreme radiation tolerance​​. [10] In contrast, lower scores near 50, exemplified by the Dps score of 38.0, suggest genes with more divergent evolutionary paths, possibly due to specific adaptations or lesser importance in stress-related DNA protection. This variability in conservation, highlighted in both Figure 4 and Figure 5, reflects the dynamic and complex nature of sequence conservation in bacteria, influenced by a range of functional necessities and evolutionary pressures.

PolA, vital for DNA replication and repair, displayed two alignments with significant scores of 76.0 and 59.0, and e-values of 4.65e-12 and 3.57e-07 respectively (**Fig. 4**). These results indicate conserved core functions despite sequence variations, likely influenced by different evolutionary pressures on various protein domains.

All identified sequences originated from *Pontimonas salivibrio*, suggesting a possible shared ecological niche or recent common ancestor with *D. radiodurans*. This connection raises questions about potential ecological or evolutionary links.

***Deinococcus geothermalis:***

The genomic analysis of *D. geothermalis* compared to *D. radiodurans* revealed BLAST hits against *Heterobasidion irregulare* mRNA sequences that were part of the *D. geothermalis* dataset. These hits, totaling 18 with 4 significant matches (e-values ≤ 0.05) against a single *H. irregulare* mRNA sequence, had BLAST scores ranging from 26 to 32 and e-values between 0.00048 and 0.03258 (**Fig. 6**). This pattern of weak similarity suggests a nuanced relationship between the two organisms, which could be indicative of shared conserved domains, potential shared metabolic pathways or stress responses, or anomalies in the genomic database. [[11]](https://www.zotero.org/google-docs/?zr8pJB) The limited yet notable similarity with *H. irregulare* raises questions about the complexity of horizontal gene transfer and the challenges in interpreting genomic data, especially in extremophiles like *Deinococcus* species. While these findings from *D. geothermalis* are not conclusive, they add an additional dimension to the study, underscoring the intricacies of microbial genetics and evolutionary biology.

This similarity is further illustrated in Figure 6's boxplot, where the median scores for *D. geothermalis* genes cluster at the lower end, indicating a general trend towards low similarity. The histogram (**Fig. 7**) reinforces this, showing a concentrated range of low BLAST scores, characteristic of weak homology. Such a consistent pattern of low scores across the dataset suggests that the genes in *D. radiodurans* have limited commonality with those in *D. geothermalis*, possibly due to significant evolutionary divergence or the specificity of the selected genes for BLAST analysis.

Significantly, all notable hits align with sequences from *H. irregulare*, furthering the evidence supporting a specialized or limited similarity with *D. geothermalis*. This finding, however, is based on a small dataset and a narrow range of similarities, making definitive conclusions about gene conservation or functional homology challenging. Further investigation is needed to deepen the understanding of these potential relationships.

***Deinococcus proteolyticus:***

A standout observation was the high conservation of the RecA gene, pivotal in DNA repair, as shown in Figure 8. The RecA gene registered an exceptionally high BLAST score (1298.0) with an e-value of 0.0, emphasizing its fundamental role in genomic maintenance and resilience in extreme environments. [[9]](https://www.zotero.org/google-docs/?YQeEUe) Genes such as PprA, DdrA, DdrB, and DdrC presented multiple hits with moderate BLAST scores, suggesting the preservation of conserved domains within the *Deinococcus* genus. These genes, related to DNA repair and stress response, demonstrate evolutionary pressures to maintain specific functional domains while allowing other regions to diverge.

Further, the analysis highlighted the significant role of plasmids in *D. proteolyticus*, with a substantial number of BLAST hits (1617 total hits) to sequences within these extrachromosomal elements. This suggests that plasmids may harbor genes that confer adaptive traits, potentially playing a role in horizontal gene transfer. [[12]](https://www.zotero.org/google-docs/?9nzyiH)

The PolA and ThyA genes, essential for DNA replication and repair, showed highly significant matches (scores of 1255.0 and 1013.0, respectively), echoing their conservation observed in *T. thermophilus*. Similarly, the PNPase gene, involved in RNA metabolism, displayed a notable BLAST score (732.0), indicating its significant conservation in *D. proteolyticus*.

The distribution of BLAST scores, as depicted in Figure 8 and Figure 9, reveals a spectrum of conservation and evolutionary adaptation among these genes. While RecA, PolA, and ThyA show high conservation, likely necessary for extremophile survival, genes like DdrA, DdrB, and DdrC exhibit a broader range of scores, reflecting conservation of specific functional domains amid sequence variation.

***Thermococcus gammatolerans:***

Despite sharing extremophilic traits, the limited similarities in certain genes between *T. gammatolerans* and *D. radiodurans* highlight the selective adaptation and evolutionary paths unique to each organism, shaped in part by the complex dynamics of HGT in extremophiles. [[13]](https://www.zotero.org/google-docs/?KjkdXv) The BLAST analysis reveals moderate gene conservation, particularly noted in the Ku gene, which shows a notable BLAST score of 48.0 and an e-value of 0.000114441 (Figs. 10 & 11). This finding suggests a significant, yet not profound, sequence similarity. The genes RecA and PNPase display more moderate similarities, with BLAST scores of 39.0 and 38.0, respectively, and e-values near the 0.05 threshold, suggesting partial sequence alignments.

The presence of homologous genes linked to stress response in these extremophiles, both known for their radiation resistance, hints at potential parallels in their DNA repair mechanisms and energy metabolism. This observation aligns with the expectations of shared stress-response strategies, potentially facilitated by historical horizontal gene transfer (HGT) events or convergent evolution.

Figure 10's boxplot and Figure 11's histogram analysis further illustrate this relationship, showing a consistent yet moderate level of sequence similarity across the examined genes. Most BLAST hits score between 38 and 48, indicating some genetic overlap but also significant differences. These differences could be pivotal in the unique survival strategies of these extremophiles in their respective environments.

The BLAST search results, revealing only four significant matches within *T. gammatolerans* itself, imply limited similarity with the genes from *D. radiodurans*. This suggests that while there may be some shared genetic elements, the two organisms have likely evolved distinct adaptations to their extreme habitats. The similarities are not extensive enough to denote a high degree of genetic overlap, emphasizing the unique and selective nature of any HGT events that might have occurred.

**Results Significance:**

This study highlights the potential for horizontal gene transfer (HGT) as evidenced by the high degree of similarity in essential genes like RecA across different species. These similarities, extending to diverse bacteria such as *Pseudomonas* and *Aeromonas*, suggest genetic exchanges that might have occurred beyond closely related species. The research also uncovers a complex pattern of gene conservation and adaptation among extremophiles. Key survival genes, especially those involved in DNA repair, show high conservation across species, while others like Dps and PNPase exhibit variability, pointing to species-specific evolutionary changes. Additionally, the significant role of plasmids in genetic diversity and the bimodal distribution of BLAST scores further underline the presence of both highly conserved genes, possibly shared through HGT, and less conserved genes, reflecting unique adaptations. These findings collectively provide critical insights into the genetic resilience of extremophiles in harsh environments and the mechanisms driving their genetic evolution.

**Project Limitations:**

**Potential Database Anomalies**: The study encountered instances, such as the BLAST hits against *H. irregulare* in the *D. geothermalis* analysis, that may indicate database anomalies (e.g., misannotation or contamination). These instances highlight the limitations of reliance on existing genomic databases and the need for cautious interpretation of BLAST results.

**Partial Sequence Similarities:** In several cases, the BLAST analyses revealed only partial sequence similarities (short alignment lengths), limiting the ability to draw comprehensive conclusions about full-length gene homology. This limitation underscores the need for further detailed genomic studies.

**Limited Scope of Gene Selection**: The research focused on a select group of genes, which may not represent the entire genomic diversity of the studied extremophiles. This limitation suggests that additional genes and genomic regions should be explored for a more complete understanding of the genetic landscape.

**Challenges in Interpreting Evolutionary Relationships:** While the study provides insights into potential HGT events and gene conservation, interpreting these findings to delineate clear evolutionary relationships remains challenging. The evolutionary distance between the species, coupled with the variability in genetic conservation, calls for more extensive genomic analyses and functional studies.

**Specificity to Extremophiles Studied:** The conclusions drawn are specific to the extremophiles studied and may not be universally applicable to all extremophilic organisms. Each extremophile's unique adaptations to its specific environment necessitate organism-specific studies.

**Future Directions:**

**Expanding the Gene Pool:** Future studies should include a broader range of genes and genomic regions to capture the full spectrum of genetic diversity in extremophiles. Analyzing a more extensive set of genes could reveal additional instances of HGT and offer deeper insights into the evolutionary history and adaptation mechanisms of these organisms.

**Functional Genomic Studies:** To complement the comparative genomic analysis, functional genomic studies, such as gene expression profiling under various stress conditions, could be valuable. These studies would help expand on the roles of conserved genes in extremophiles' survival and adaptation to extreme environments.

**Advanced Bioinformatic Analysis**: Employing more advanced bioinformatic tools and algorithms could refine the analysis of genomic data. Techniques such as whole-genome alignment and phylogenetic analysis could provide a more detailed understanding of the evolutionary relationships between extremophiles and the extent of horizontal gene transfer.

**Investigating Mobile Genetic Elements:** Given the significant role of plasmids observed in this study, a focused investigation into the mobile genetic elements of extremophiles could be revealing. Such research might uncover the mechanisms behind the transfer of advantageous traits among extremophiles, particularly through horizontal gene transfer.

**Cross-Species Genomic Comparisons:** Broadening the range of species included in the comparative genomic analysis, especially incorporating more diverse extremophilic organisms, would enhance our understanding of gene conservation and diversity across a wider spectrum of extreme environments.

**Summary:**

Comparative genomic analysis across various extremophilic organisms, with a focus on *D. radiodurans*, has revealed intriguing insights into the potential horizontal gene transfer (HGT) among extremophiles. Key findings from the BLAST analyses indicate a pattern of gene conservation and diversity with certain genes showing high degrees of similarity across different species. Notably, genes like RecA, known for their roles in DNA repair, showcased high BLAST scores and very low e-values in multiple species. This suggests that such genes are conserved due to their crucial roles in maintaining genomic integrity, a vital trait in extremophiles facing harsh environmental conditions. [[9]](https://www.zotero.org/google-docs/?AW9iIL)

The presence of genes such as ThyA and PolA (essential in DNA synthesis and repair) across different extremophilic species also points to their conservation. However, the degree of similarity varied, with some genes like Dps and PNPase showing lower scores and greater variability, suggesting species-specific adaptations or evolutionary divergence.

The occurrence of high-scoring matches with diverse bacterial species, including *Pseudomonas* and *Aeromonas*, in the context of genes from *D. radiodurans*, raises the possibility of HGT. Such cross-species similarities could be indicative of shared housekeeping genes or adaptive genes beneficial in specific environmental niches. [[13]](https://www.zotero.org/google-docs/?MOgJop) The presence of these genes in unrelated species or unique environmental contexts supports the hypothesis of HGT events.

Additionally, the pattern of hits observed in the BLAST results, such as the multiple hits within the plasmids of *D. proteolyticus*, suggests that some of the target genes may be located on mobile genetic elements, which are often involved in HGT. [[12]](https://www.zotero.org/google-docs/?nfsa0Z) This further underscores the potential role of HGT in spreading advantageous traits among extremophiles. The bimodal distribution of BLAST scores in some analyses, with peaks at low and high score ranges, hints at the existence of two distinct groups of genes: those that are highly conserved and potentially shared via HGT, and those that are less conserved, reflecting unique evolutionary paths and adaptations.

In conclusion, the genomic comparisons across different extremophilic organisms suggest a complex landscape of gene conservation, adaptation, and potential HGT. The study provides potential evidence for the role of HGT in shaping the genomes of extremophiles, contributing to their ability to thrive in extreme environments. These findings highlight the significance of HGT in microbial evolution when present, particularly among extremophiles, and underline the need for further research into the mechanisms and impact of HGT in extremophile biology.

**Bibliography**

[[1] A. J. Ellington, T. J. Schult, C. R. Reisch, and B. C. Christner, “Comparative Genomics and Directed Evolution Reveal Genetic Determinants of Extreme UVC Radiation Tolerance in Bacteria Recovered from the Stratosphere.” bioRxiv, p. 2023.03.27.534493, Mar. 28, 2023. doi: 10.1101/2023.03.27.534493.](https://www.zotero.org/google-docs/?6NCJym)

[[2] K. S. Makarova](https://www.zotero.org/google-docs/?6NCJym) *[et al.](https://www.zotero.org/google-docs/?6NCJym)*[, “Genome of the Extremely Radiation-Resistant Bacterium Deinococcus radiodurans Viewed from the Perspective of Comparative Genomics,”](https://www.zotero.org/google-docs/?6NCJym) *[Microbiol. Mol. Biol. Rev.](https://www.zotero.org/google-docs/?6NCJym)*[, vol. 65, no. 1, pp. 44–79, Mar. 2001, doi: 10.1128/mmbr.65.1.44-79.2001.](https://www.zotero.org/google-docs/?6NCJym)

[[3] J. Armstrong, I. T. Fiddes, M. Diekhans, and B. Paten, “Whole-Genome Alignment and Comparative Annotation,”](https://www.zotero.org/google-docs/?6NCJym) *[Annu. Rev. Anim. Biosci.](https://www.zotero.org/google-docs/?6NCJym)*[, vol. 7, pp. 41–64, Feb. 2019, doi: 10.1146/annurev-animal-020518-115005.](https://www.zotero.org/google-docs/?6NCJym)

[[4] I.-M. A. Chen](https://www.zotero.org/google-docs/?6NCJym) *[et al.](https://www.zotero.org/google-docs/?6NCJym)*[, “The IMG/M data management and analysis system v.6.0: new tools and advanced capabilities,”](https://www.zotero.org/google-docs/?6NCJym) *[Nucleic Acids Res.](https://www.zotero.org/google-docs/?6NCJym)*[, vol. 49, no. D1, pp. D751–D763, Oct. 2020, doi: 10.1093/nar/gkaa939.](https://www.zotero.org/google-docs/?6NCJym)

[[5] S.-L. Liu, Y.-R. Chiang, H. Yoon, and H.-Y. Fu, “Comparative Genome Analysis Reveals Cyanidiococcus gen. nov., A New Extremophilic Red Algal Genus Sister to Cyanidioschyzon (Cyanidioschyzonaceae, Rhodophyta),”](https://www.zotero.org/google-docs/?6NCJym) *[J. Phycol.](https://www.zotero.org/google-docs/?6NCJym)*[, vol. 56, Jul. 2020, doi: 10.1111/jpy.13056.](https://www.zotero.org/google-docs/?6NCJym)

[[6] M. V. Omelchenko](https://www.zotero.org/google-docs/?6NCJym) *[et al.](https://www.zotero.org/google-docs/?6NCJym)*[, “Comparative genomics of Thermus thermophilus and Deinococcus radiodurans: divergent routes of adaptation to thermophily and radiation resistance,”](https://www.zotero.org/google-docs/?6NCJym) *[BMC Evol. Biol.](https://www.zotero.org/google-docs/?6NCJym)*[, vol. 5, p. 57, Oct. 2005, doi: 10.1186/1471-2148-5-57.](https://www.zotero.org/google-docs/?6NCJym)

[[7] J. V. Etten and D. Bhattacharya, “Horizontal Gene Transfer in Eukaryotes: Not if, but How Much?,”](https://www.zotero.org/google-docs/?6NCJym) *[Trends Genet.](https://www.zotero.org/google-docs/?6NCJym)*[, vol. 36, no. 12, pp. 915–925, Dec. 2020, doi: 10.1016/j.tig.2020.08.006.](https://www.zotero.org/google-docs/?6NCJym)

[[8] I. Maqbool, M. Sudharsan, G. Kanimozhi, S. T. Alrashood, H. A. Khan, and N. R. Prasad, “Crude Cell-Free Extract From Deinococcus radiodurans Exhibit Anticancer Activity by Inducing Apoptosis in Triple-Negative Breast Cancer Cells,”](https://www.zotero.org/google-docs/?6NCJym) *[Front. Cell Dev. Biol.](https://www.zotero.org/google-docs/?6NCJym)*[, vol. 8, p. 707, Jul. 2020, doi: 10.3389/fcell.2020.00707.](https://www.zotero.org/google-docs/?6NCJym)

[[9] S. Karlin and L. Brocchieri, “Evolutionary conservation of RecA genes in relation to protein structure and function,”](https://www.zotero.org/google-docs/?6NCJym) *[J. Bacteriol.](https://www.zotero.org/google-docs/?6NCJym)*[, vol. 178, no. 7, pp. 1881–1894, Apr. 1996, doi: 10.1128/jb.178.7.1881-1894.1996.](https://www.zotero.org/google-docs/?6NCJym)

[[10] A. De Groot](https://www.zotero.org/google-docs/?6NCJym) *[et al.](https://www.zotero.org/google-docs/?6NCJym)*[, “Alliance of Proteomics and Genomics to Unravel the Specificities of Sahara Bacterium Deinococcus deserti,”](https://www.zotero.org/google-docs/?6NCJym) *[PLoS Genet.](https://www.zotero.org/google-docs/?6NCJym)*[, vol. 5, no. 3, p. e1000434, Mar. 2009, doi: 10.1371/journal.pgen.1000434.](https://www.zotero.org/google-docs/?6NCJym)

[[11] N. Shah, M. G. Nute, T. Warnow, and M. Pop, “Misunderstood parameter of NCBI BLAST impacts the correctness of bioinformatics workflows,”](https://www.zotero.org/google-docs/?6NCJym) *[Bioinformatics](https://www.zotero.org/google-docs/?6NCJym)*[, vol. 35, no. 9, pp. 1613–1614, May 2019, doi: 10.1093/bioinformatics/bty833.](https://www.zotero.org/google-docs/?6NCJym)

[[12] E. Harrison and M. A. Brockhurst, “Plasmid-mediated horizontal gene transfer is a coevolutionary process,”](https://www.zotero.org/google-docs/?6NCJym) *[Trends Microbiol.](https://www.zotero.org/google-docs/?6NCJym)*[, vol. 20, no. 6, pp. 262–267, Jun. 2012, doi: 10.1016/j.tim.2012.04.003.](https://www.zotero.org/google-docs/?6NCJym)

[[13] C. A. Fuchsman, R. E. Collins, G. Rocap, and W. J. Brazelton, “Effect of the environment on horizontal gene transfer between bacteria and archaea,”](https://www.zotero.org/google-docs/?6NCJym) *[PeerJ](https://www.zotero.org/google-docs/?6NCJym)*[, vol. 5, p. e3865, Sep. 2017, doi: 10.7717/peerj.3865.](https://www.zotero.org/google-docs/?6NCJym)