Nucleic Acids Research, 2025, **53**, gkaf329
https://doi.org/10.1093/nar/gkaf329
Data Resources and Analyses

invertiaDB: a database of inverted repeats across organismal genomes

Kimonas Provatas[1,2,†], Nikol Chantzi[1,2,†], Nafsika Amptazi[1,2], Michail Patsakis[1,2], Akshatha Nayak[1,2], Ioannis Mouratidis[1,2], Apostolos Zaravinos[3,4], Georgios A. Pavlopoulos[5], Ilias Georgakopoulos-Soares[1,2,*]

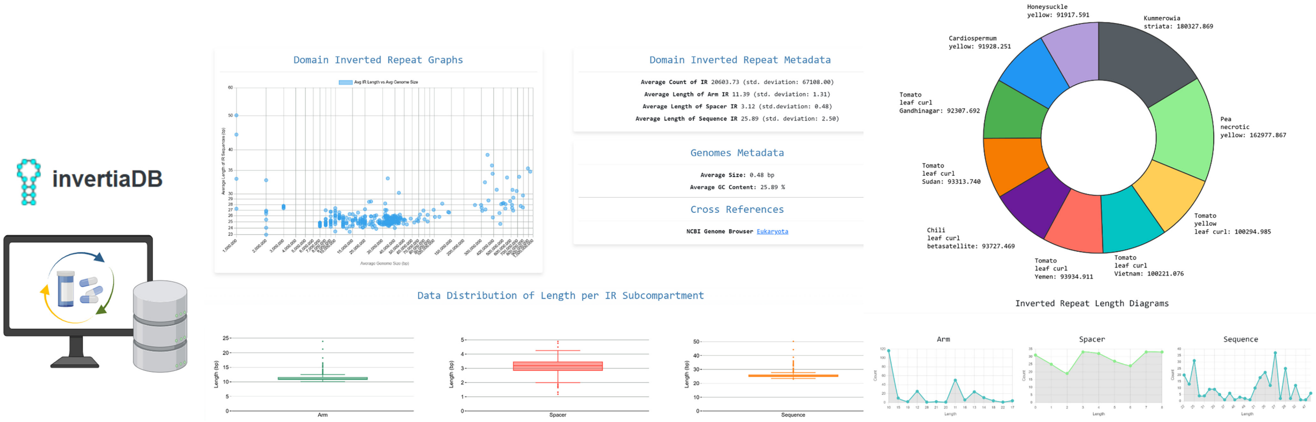
[1] Institute for Personalized Medicine, Department of Biochemistry and Molecular Biology, The Pennsylvania State University College of Medicine, Hershey, PA 17033, United States
[2] Huck Institute of the Life Sciences, Pennsylvania State University, University Park, PA, 16802, United States
[3] Department of Life Sciences, School of Sciences, European University Cyprus, Nicosia, 1516, Cyprus
[4] Cancer Genetics, Genomics and Systems Biology Laboratory, Basic and Translational Cancer Research Center (BTCRC), Nicosia, 1516, Cyprus
[5] Institute for Fundamental Biomedical Research, BSRC "Alexander Fleming", Vari 16672, Greece

[*] To whom correspondence should be addressed. Email: izg5139@psu.edu
[†] The first two authors should be regarded as Joint First Authors.

Abstract

Inverted repeats are repetitive elements that can form hairpin and cruciform structures. They are linked to genomic instability; however, they also have various biological functions. Their distribution differs markedly across taxonomic groups in the tree of life, and they exhibit high polymorphism due to their inherent genomic instability. Advances in sequencing technologies and declined costs have enabled the generation of an ever-growing number of complete genomes for organisms across taxonomic groups in the tree of life. However, a comprehensive database encompassing inverted repeats across diverse organismal genomes has been lacking. We present invertiaDB, the first comprehensive database of inverted repeats spanning multiple taxa, featuring repeats identified in the genomes of 118 101 organisms across all major taxonomic groups. For each organism, we derived inverted repeats with arm lengths of at least 10 bp, spacer lengths up to 8 bp, and no mismatches in the arms. The database currently hosts 34 330 450 inverted repeat sequences, serving as a centralized, user-friendly repository to perform searches and interactive visualizations, and download existing inverted repeat data for independent analysis. invertiaDB is implemented as a web portal for browsing, analyzing, and downloading inverted repeat data. invertiaDB is publicly available at https://invertiadb.netlify.app/homepage.html.

Graphical abstract


invertiaDB
Domain Inverted Repeat Graphs
Avg IR Length vs Avg Genome Size
Average Genomic Size (bp)
Average Length of IR Sequence (bp)
Honeysuckle yellow: 91917.591
Kummerowia striata: 180327.869
Cardiospermum yellow: 91928.251
Pea necrotic yellow: 162977.867
Tomato leaf curl Gandhinagar: 92307.692
Tomato leaf curl Sudan: 93313.740
Tomato yellow leaf curl: 100294.985
Chili leaf curl betasatellite: 93727.469
Tomato leaf curl Yemen: 93934.911
Tomato leaf curl Vietnam: 100221.076
Data Distribution of Length per IR Subcompartment
Length (bp)
Arm
Spacer
Sequence
Inverted Repeat Length Diagrams
Arm
Spacer
Sequence
Count
Length

Domain Inverted Repeat Metadata

Average Count of IR 20603.73 (std. deviation: 67108.00)
Average Length of Arm IR 11.39 (std. deviation: 1.31)
Average Length of Spacer IR 3.12 (std.deviation: 0.48)
Average Length of Sequence IR 25.89 (std. deviation: 2.50)

Genomes Metadata

Average Size: 0.48 bp
Average GC Content: 25.89 %

Cross References

NCBI Genome Browser Eukaryota

Introduction

The right-handed double helix DNA structure, also known as the canonical B-DNA structure, was originally described by Watson, Crick, Wilkins, and Franklin. Several alternative DNA conformations have since been identified [1–3]. These noncanonical DNA structures that do not conform to the standard B-DNA form are termed non-B DNA. Such conformations include hairpin and cruciform structures, which can

Received: October 5, 2024. **Revised:** April 9, 2025. **Editorial Decision:** April 10, 2025. **Accepted:** April 14, 2025

form at inverted repeat (IR) sequences [4–7]. IRs are composed of two sequence parts, one of which is the reverse complement of the other, separated by an intervening spacer sequence. IRs undergo intrastrand base pairing to form hairpin or cruciform structures, in which the two complementary arms hybridize, and the spacer loop remains single-stranded.

The formation of hairpin and cruciform structures can be facilitated by negative supercoiling, often associated with DNA transcription and replication [8, 9]. The biophysical properties of IRs, including the spacer and arm lengths and their GC content, influence the likelihood of structure formation. Increased GC content in the IR arms is linked to increased hairpin stability [5, 10, 11], whereas mismatches in the arms reduce the likelihood of structure formation [12–16]. Additionally, longer arms are linked to increased hairpin formation likelihood and stability [10], while shorter spacer lengths are favorable to hairpin and cruciform structure formation [10]. Specifically, cruciforms favor shorter spacer lengths, whereas hairpins favor short spacers but with spacer lengths above 4 bp due to steric constraints [10, 17].

IRs are over-represented in organismal genomes relative to the expected random distribution and show an inhomogeneous genomic distribution with enrichment hotspots [18]. They are particularly enriched in promoters and near transcription termination sites and can regulate gene expression [19, 20]. In prokaryotes, IRs can drive rho-independent transcription termination [21, 22]. At replication origins, IRs are involved in replication initiation [23], and in certain transposons, they are found flanking the internal sequence [24]. Additionally, several proteins can bind preferentially to hairpin and cruciform structures [25].

Hairpins and cruciforms are associated with increased genomic instability in both prokaryotes and eukaryotes [6, 15, 19, 26–35]. The human genome is depleted of perfect IRs with long arms, which are highly unstable [5, 36, 37]. IRs are recombination and rearrangement hotspots across organisms [29, 30], and artificial introduction of IRs with long arms in eukaryotic cells leads to thousands of times higher recombination and deletion rates than expected [15]. In human cancers, IRs are highly enriched in somatic mutations across mutation categories, and this excess can confound statistical models aimed at identifying driver mutations [35, 38, 39]. Specifically, off-target APOBEC mutagenesis is often directed at hairpin structures, causing an excess of mutagenesis at IRs, particularly at the single-stranded loop [5].

Multiple bioinformatic tools have been developed to systematically identify IRs [40–43]. Additionally, a previous database named Non-B DB encompasses IRs found in five organisms [44]. The recently published LIRBase is a comprehensive long IR database across 424 eukaryotic genomes and also provides information about expression levels, and serves as a valuable resource for exploring long IRs [45]. msRepDB is a database dedicated to microsatellite repeats that provides detailed annotations of repetitive elements across genomes, offering insights into the broader repeat landscapes within which IRs occur [46]. Additional resources, such as Repbase and Microsatellites Explorer, focus on repetitive DNA elements associated with transposable elements and satellite DNA across organismal genomes [47, 48].

Here, we present invertiaDB, an IR database across organismal genomes. The database currently hosts 34 330 450 IR sequences found across 118 101 organismal genomes and serves as a centralized, user-friendly web portal to perform searches and interactive visualizations, and download existing IR data for further analysis. This resource will be valuable for researchers across various disciplines, studying the functional roles and genomic instability of IRs.

Materials and methods

Data collection

The organismal genomes present in the database were downloaded from the GenBank and RefSeq databases on 21 March 2024 [49, 50] and included all complete organismal genomes. We did not consider incomplete genomes, scaffolds, and contigs in order to increase the confidence in our identified sequences. Duplicate assembly accessions were filtered out. The organism names and accession IDs are provided in Supplementary Data S1. Gene annotation files in the form of GFF files were downloaded for each genome from the same source using genome_updater version v0.6.4 from https://github.com/pirovc/genome_updater [51]. Coordinates for genic regions were derived using bash and Python scripts. The intersection with IR coordinates and the subsequent determination of the nearest genic region were extracted using BEDTools intersect and BEDTools closest commands [52]. IR densities were calculated as the ratio of total length of IR sequences to the genome size of the assembly accession, multiplied by 1 Mb. These coordinates were then examined for IR densities across organismal genomes. The organismal genome database consists of 49 197 bacteria, 67 742 viruses, 492 Eukaryota, and 687 Archaea.

Identification of IRs in organismal genomes

IRs with arm lengths $\geq$10 bp, spacer lengths <9 bp, and without mismatches in the arms were obtained from [53]. For IR detection, a modified version of the non-B gfa package was developed and wrapped in a Python program [44]. The subprocess call to the C script utilized all necessary skip flags to extract only the IR sequences present in each organismal genome, and we provide the command below:

```
./non-B_gfa/gfa -seq < input_fasta> -out < output> -
    skipAPR -skipSlipped -skipCruciform -skipTriplex -
    skipWGET -minIRrep 10 -maxIRspacer 8 -skipMR -
    skipDR -skipZ -skipGQ -skipSTR
```

A custom script was used to manually transform the data into a processable tabular format and extract the corresponding arm and spacer sequences along with their corresponding lengths based on the coordinates of the IR sequence. The IR spacer and arm thresholds align with previous observations that shorter arms and longer spacers are less prone to forming IRs and that mismatches in the hairpin arms and increased spacer lengths further reduce the probability of stable structure formation [12–16]. We systematically validated the loci and sequences of the detected IRs in order to ensure their correctness.

IR density across temperature classes

We separated bacteria strains into four mutually exclusive groups, as a function of their average optimal growth temperature (Topt_ave): psychrophiles, mesophiles, thermophiles, and hyperthermophiles, in accordance with the TEMPURA database [54]. In particular, we classified a bacterial strain as psychrophile, if the optimal growth temperature is below 20°C; mesophile, if the optimal growth temperature is at least 20°C and at most 44°C; thermophile, if it is at least 45°C and at most 79°C; and finally, hyperthermophile, if the optimal growth temperature is at least 80°C. In order to utilize the results from TEMPURA DB, we mapped the NCBI Taxonomy ID corresponding to the average optimal growth temperature to the RefSeq or GenBank assembly accession IDs, as derived from the assembly summaries of the latter genomic databases, using the species taxonomy ID column. After removing duplicated assembly accessions between RefSeq and GenBank databases, a total of 3894 bacterial genomes were analyzed using the TEMPURA database. Subsequently, we compared the derived IR densities for each individual bacterial assembly accession to the respective temperature classes. The IR density for each individual genome was calculated as the ratio of the total number of IRs to the genome size multiplied by 1 Mb. If a genome lacked IRs, we set an IR density of 0. Finally, to calculate the IR arm GC-fold enrichment for each individual bacterial strain, we computed the GC proportion of all IR arms in each strain and compared it to the GC content of the underlying genome, as derived from the assembly summaries in RefSeq and GenBank databases. For all the pairwise comparisons between the four temperature groups, we used independent *t*-tests and one-way ANOVA as provided from the SciPy Python module. All the tests were corrected for multiple hypothesis testing using the Benjamini–Hochberg procedure.

Database design and data pipelines

The backend of the application is built using Flask, a Python-based web framework, and is served through a reverse proxy to manage internal port access. The data layer leverages DuckDB in read-only mode, optimizing for fast, highly compressed, and secure online analytical processing. The DuckDB database driver operates in-process within the Flask app's memory, with all data consolidated into a single DuckDB file. Currently, the stored IRs occupy 3 GB, organized as a vectorized columnar database, accessible through SQL syntax. As of database design, metadata were extracted from the NCBI database for each of the 118 070 assembly genomes that were analyzed for IRs, and enriched with the IR filename to facilitate joint operations between the two. The term "enriched" in this context refers to the process of combining data from two separate tables, the "inverted_repeats" table and the "metadata" table, to create a single, more comprehensive dataset. Then, using joins between the metadata and the IR genomic data, the metadata table was enriched with the IR statistics (Supplementary Fig. S1). We used various scripts to curate the data and to convert them to multiple formats (Parquet, CSV, JSON, and BED), while grouping them into the three domains of life and viruses and providing them for user download.

Web interface

The frontend of invertiaDB is built using HTML, CSS, and JavaScript. The application is deployed on a standard Google Cloud Compute Engine, featuring a JavaScript-based frontend app. A comprehensive full-stack web application was developed to enable user-friendly access, in-depth analysis features, and visualization of the integrated data. To provide in-depth understanding of the IR data, dynamic and interactive graphs were also created using custom HTML and Chart.js. Additionally, pre-built components from the Bootstrap 5 and Bootstrap 5 DataTables libraries were utilized.

Database overview and functionality

Database contents and usage

invertiaDB features comprehensive integration and curation of IRs in 118 101 assemblies of organisms in the domains of life and viruses, in a user-friendly, interactive web interface (Fig. 1A and B). Out of all the examined viral genomes, 32 874 did not contain any IRs, while only 1 bacterial genome did not contain any IR sequences. Interestingly, viral genomes displayed greater variability in their genome-wide IR density, while bacterial genomes were found to have the highest average IR density. IR sequences can be filtered based on their biophysical properties, including their spacer and arm lengths, sequence motifs that the user searches for, and IR metadata such as density, the total IR length, their nucleotide composition, and genic overlaps (Fig. 1A and B). The database includes multiple features to explore, search, and download IRs from each organism in Parquet, CSV, JSON, and BED formats.

The top navigation bar is composed of six interactive pages, namely the Homepage, the Explore page, the Downloads page, the Help page, the Motif page, and the About/Contact page tabs, which enable the navigation across the different parts of the database. The Explore page is further split into the invertiaDB dataset, the Domains page, and the Organisms page, with each of these options navigating the user to subsequent pages for IR explorations. The Domains page provides aggregations of IRs by the domain of life, in Archaea, Eukaryota, bacteria, and viruses.

invertiaDB dataset page

Upon accessing the Explore → invertiaDB Dataset page, users are presented with a table of assemblies featuring advanced querying capabilities for exploring IRs across various genomes (Fig. 2A). Quick searches can be performed on NCBI Taxonomy IDs, GenBank/Reference genome accessions, or species names. Complex queries can be performed using combinations of the available metadata columns provided by the Advanced Filters feature, and users may project all columns from the NCBI database metadata. Multiple genomes can be selected for downloading IR annotations in Parquet, CSV, JSON, or BED formats. Upon clicking the assembly button that is shown as the first column, the user navigates to an analysis of the assembly displaying the IR and genome metadata (Fig. 2B). The metadata include links to the established publicly available databases of the ENA Browser [55] and the NCBI Genome Browser [56]. We generated a visualization depicting the occurrences of IRs in the organism, focusing on the arm-to-spacer ratio. This allows for an assessment of whether the structure is arm- or spacer-dominant, as some organisms exhibit longer spacer/loop lengths relative to arms/stems.

Analysis and visualization pages

The invertiaDB website features interactive bar plots, doughnut charts, tables, and drop-down menus, allowing users to select and analyze IRs across various assemblies. On the assembly analysis page on inverted repeats, the user also encounters a line chart displaying the length distribution of IRs per arm, spacer, and sequence (Fig. 3A), along with three doughnut graphs that showcase respective nucleotide compositions for the three sequences, offering insight into the structural

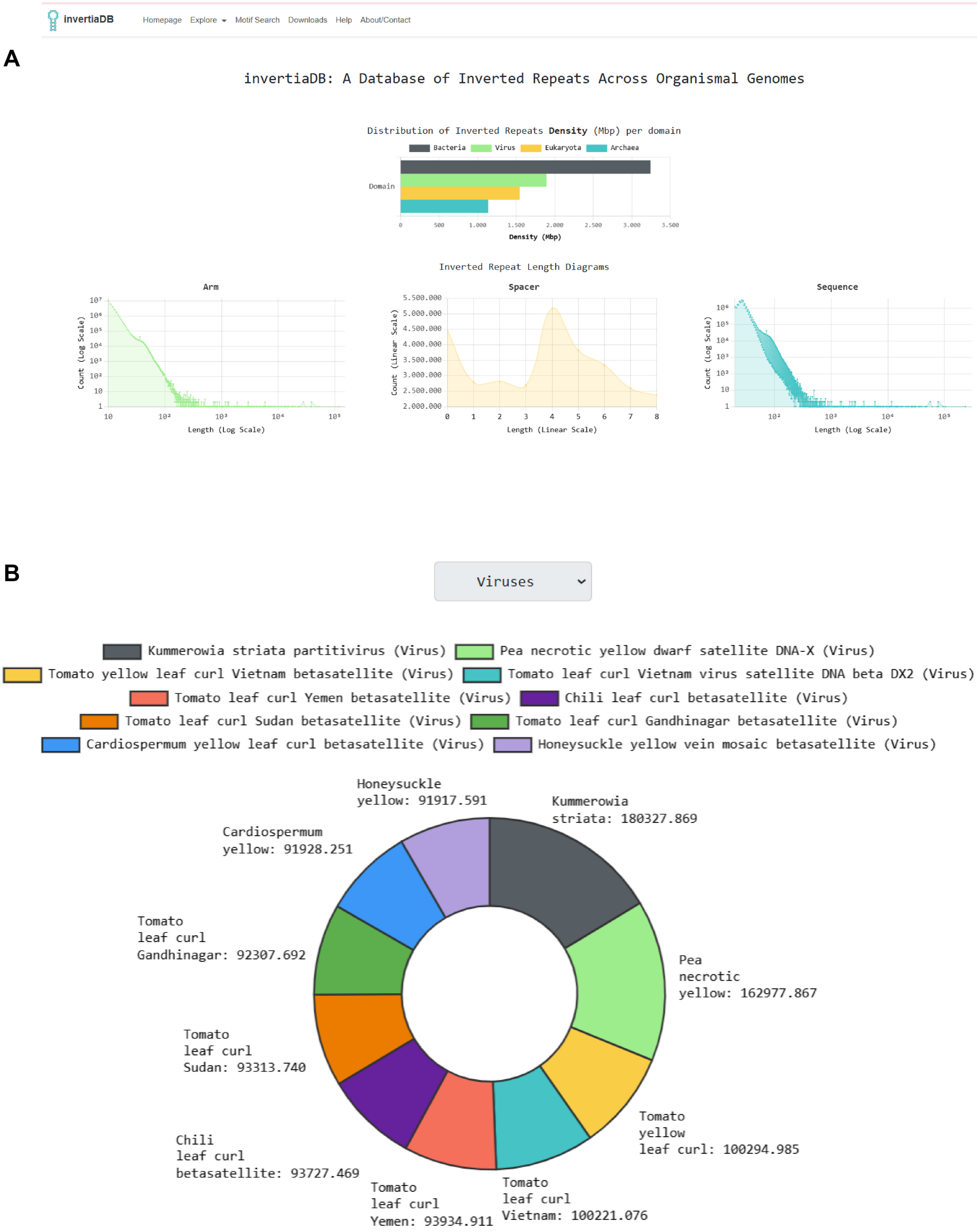
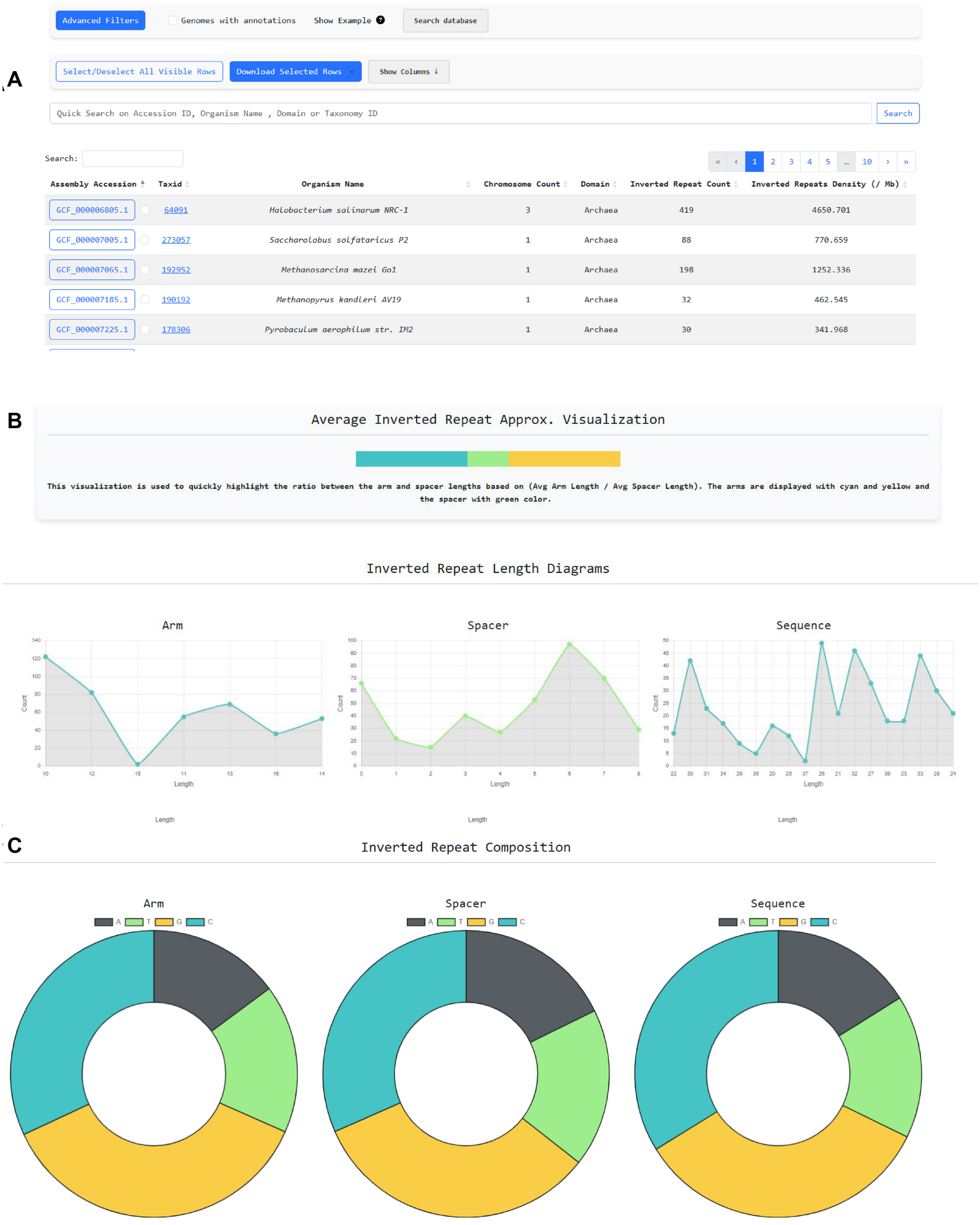

invertiaDB
Homepage
Explore
Motif Search
Downloads
Help
About/Contact
A
invertiaDB: A Database of Inverted Repeats Across Organismal Genomes
Distribution of Inverted Repeats **Density** (Mbp) per domain
Bacteria
Virus
Eukaryota
Archaea
Domain
0
500
1.000
1.500
2.000
2.500
3.000
3.500
Density (Mbp)
Inverted Repeat Length Diagrams
Arm
Count (Log Scale)
10^7
10^6
10^5
10^4
10^3
10^2
10
1
10
10^2
10^3
10^4
10^5
Length (Log Scale)
Spacer
Count (Linear Scale)
5.500.000
5.000.000
4.500.000
4.000.000
3.500.000
3.000.000
2.500.000
2.000.000
0
1
2
3
4
5
6
7
8
Length (Linear Scale)
Sequence
Count (Log Scale)
10^6
10^5
10^4
10^3
10^2
10
1
10^2
10^3
10^4
10^5
Length (Log Scale)
B
Viruses
Kummerowia striata partitivirus (Virus)
Pea necrotic yellow dwarf satellite DNA-X (Virus)
Tomato yellow leaf curl Vietnam betasatellite (Virus)
Tomato leaf curl Vietnam virus satellite DNA beta DX2 (Virus)
Tomato leaf curl Yemen betasatellite (Virus)
Chili leaf curl betasatellite (Virus)
Tomato leaf curl Sudan betasatellite (Virus)
Tomato leaf curl Gandhinagar betasatellite (Virus)
Cardiospermum yellow leaf curl betasatellite (Virus)
Honeysuckle yellow vein mosaic betasatellite (Virus)
Honeysuckle yellow: 91917.591
Kummerowia striata: 180327.869
Cardiospermum yellow: 91928.251
Tomato leaf curl Gandhinagar: 92307.692
Pea necrotic yellow: 162977.867
Tomato leaf curl Sudan: 93313.740
Tomato yellow leaf curl: 100294.985
Chili leaf curl betasatellite: 93727.469
Tomato leaf curl Yemen: 93934.911
Tomato leaf curl Vietnam: 100221.076

Figure 1. invertiaDB Homepage, navigation bar, and domain selection. (**A**) invertiaDB Homepage showing domains of life bar chart in decreasing order of IR density, and arm, spacer, and sequence length distribution in line charts. IR densities were calculated as the ratio of the total length of IR sequences to the genome size of the assembly accession, multiplied by 1 Mb. (**B**) Doughnut chart of top 10 most dense organisms in IRs for each domain of life.

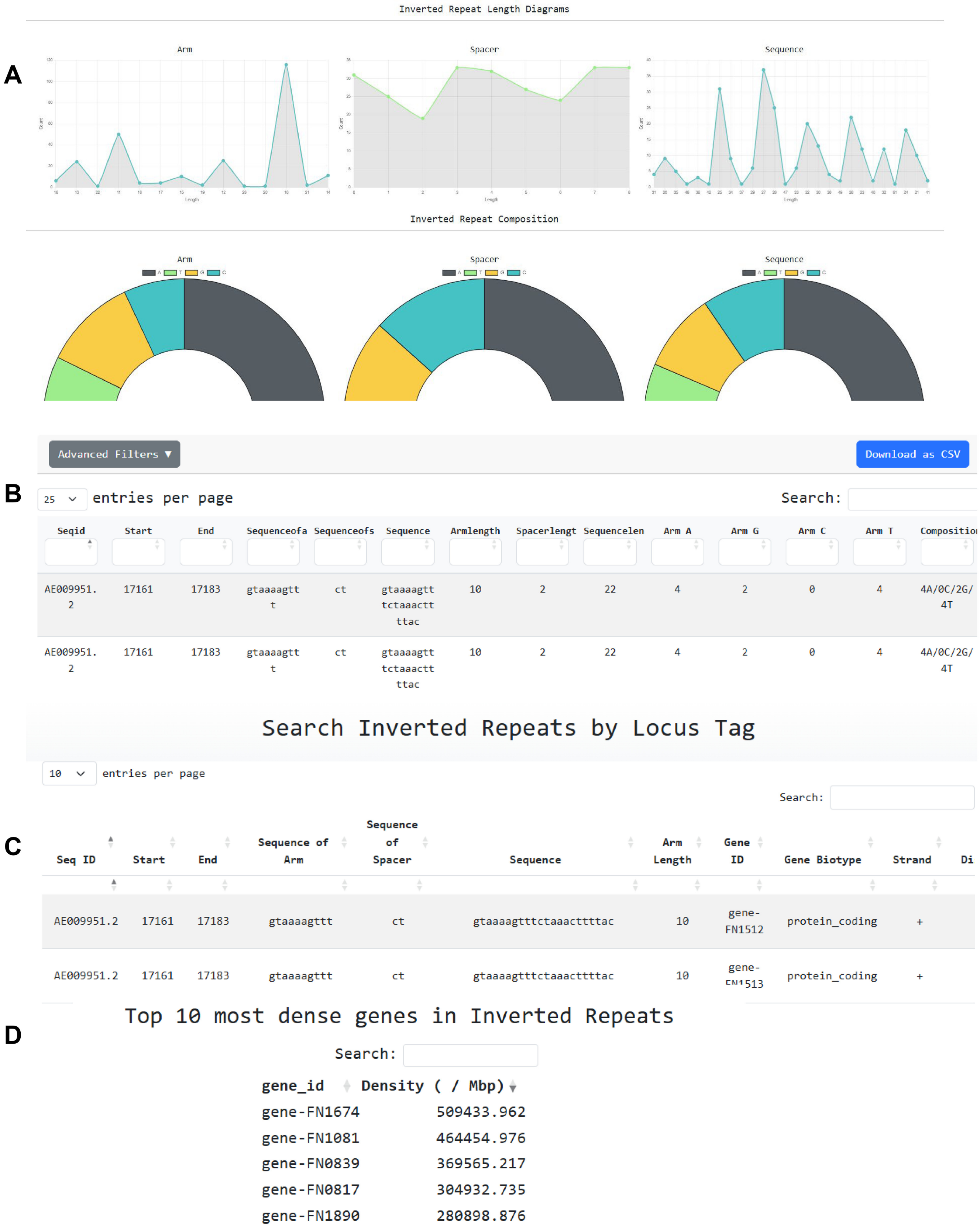

A
Advanced Filters
□ Genomes with annotations
Show Example
Search database
Select/Deselect All Visible Rows
Download Selected Rows
Show Columns ↓
Quick Search on Accession ID, Organism Name , Domain or Taxonomy ID
Search
Search:
« ‹ 1 2 3 4 5 … 10 › »
B
C
Inverted Repeat Length Diagrams
Arm
Spacer
Sequence
Count
Length
Inverted Repeat Composition
A
T
G
C

Assembly Accession	Taxid	Organism Name	Chromosome Count	Domain	Inverted Repeat Count	Inverted Repeats Density (/ Mb)
GCF_000006805.1	64091	*Halobacterium salinarum NRC-1*	3	Archaea	419	4650.701
GCF_000007005.1	273057	*Saccharolobus solfataricus P2*	1	Archaea	88	770.659
GCF_000007065.1	192952	*Methanosarcina mazei Go1*	1	Archaea	198	1252.336
GCF_000007185.1	190192	*Methanopyrus kandleri AV19*	1	Archaea	32	462.545
GCF_000007225.1	178306	*Pyrobaculum aerophilum str. IM2*	1	Archaea	30	341.968

Average Inverted Repeat Approx. Visualization

This visualization is used to quickly highlight the ratio between the arm and spacer lengths based on (Avg Arm Length / Avg Spacer Length). The arms are displayed with cyan and yellow and the spacer with green color.

Figure 2. invertiaDB dataset Explore page along with example usage and assembly analysis pages. (**A**) invertiaDB dataset Explore page showcasing Advanced Filtering, Downloading, Projecting NCBI Columns, Quick Search, and Row Filtering features. (**B**, **C**) Assembly accession analysis page, containing metadata and visualizations of genome and IR patterns.



Seqid	Start	End	Sequenceofa	Sequenceofs	Sequence	Armlength	Spacerlengt	Sequencelen	Arm A	Arm G	Arm C	Arm T	Compositio
AE009951.2	17161	17183	gtaaaagttt	ct	gtaaaagtttctaaacttttac	10	2	22	4	2	0	4	4A/0C/2G/4T
AE009951.2	17161	17183	gtaaaagttt	ct	gtaaaagtttctaaacttttac	10	2	22	4	2	0	4	4A/0C/2G/4T

Search Inverted Repeats by Locus Tag

10 entries per page

Search:

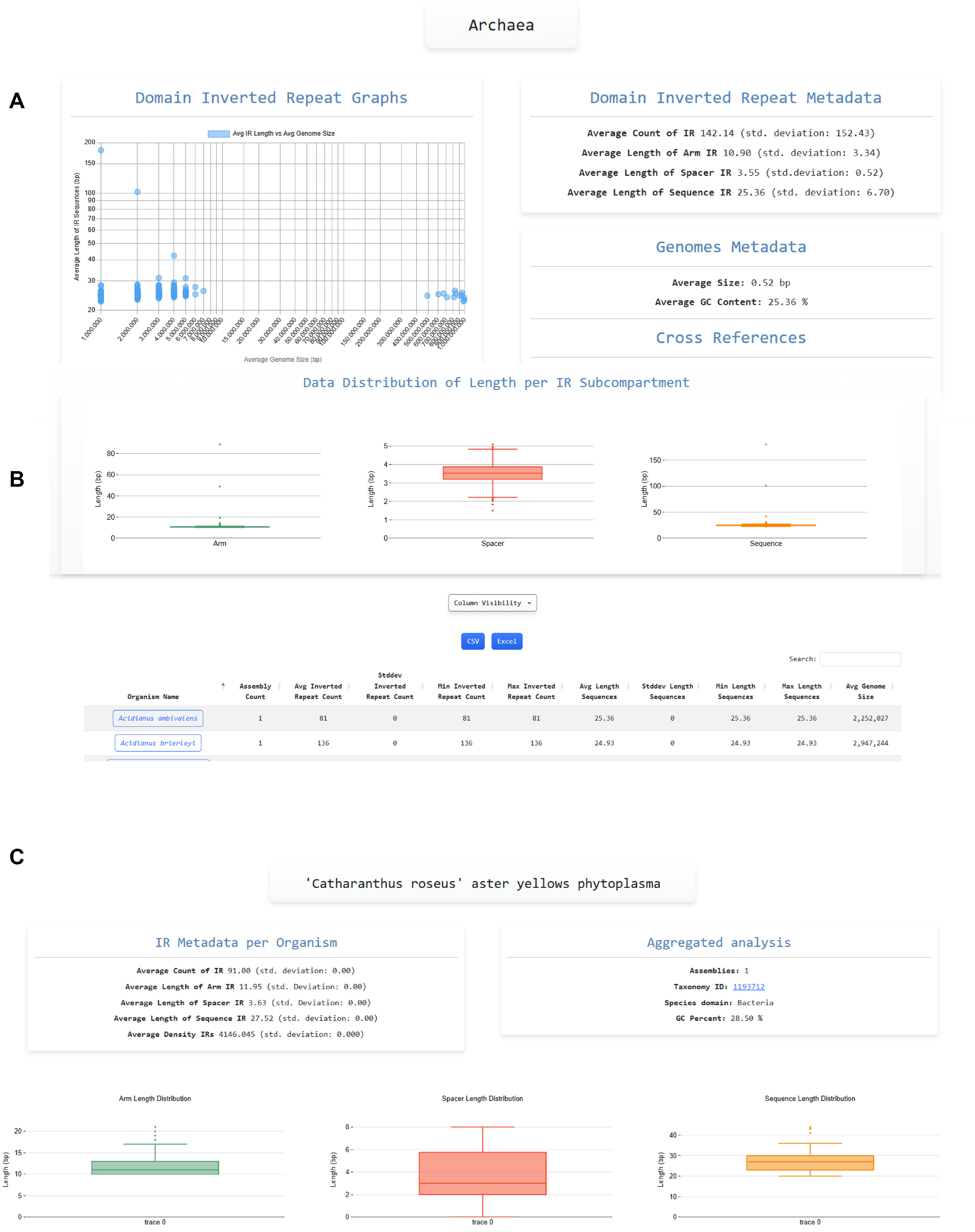
Seq ID	Start	End	Sequence of Arm	Sequence of Spacer	Sequence	Arm Length	Gene ID	Gene Biotype	Strand	Di
AE009951.2	17161	17183	gtaaaagttt	ct	gtaaaagtttctaaacttttac	10	gene-FN1512	protein_coding	+	
AE009951.2	17161	17183	gtaaaagttt	ct	gtaaaagtttctaaacttttac	10	gene-FN1513	protein_coding	+	

Top 10 most dense genes in Inverted Repeats

Search:

gene_id	Density (/ Mbp)
gene-FN1674	509433.962
gene-FN1081	464454.976
gene-FN0839	369565.217
gene-FN0817	304932.735
gene-FN1890	280898.876

Figure 3. Assembly accession analysis page showing IR graphs, raw data tables, and gene functionalities. (**A**) Line chart of length distribution of arm, spacer, and sequence along with nucleotide composition of arm, spacer, and sequence. (**B**) Table of IRs contained in the organism's assembly with the ability to fuzzy search each column. (**C**) Search IRs overlapping specific genes by gene locus tag. (**D**) Table of most dense genes in IRs by formula: total IR base pairs overlapping a gene over gene size in Mbp.



Archaea

Domain Inverted Repeat Graphs

Domain Inverted Repeat Metadata

Average Count of IR 142.14 (std. deviation: 152.43)

Average Length of Arm IR 10.90 (std. deviation: 3.34)

Average Length of Spacer IR 3.55 (std.deviation: 0.52)

Average Length of Sequence IR 25.36 (std. deviation: 6.70)

Genomes Metadata

Average Size: 0.52 bp

Average GC Content: 25.36 %

Cross References

Data Distribution of Length per IR Subcompartment

Organism Name	Assembly Count	Avg Inverted Repeat Count	Stddev Inverted Repeat Count	Min Inverted Repeat Count	Max Inverted Repeat Count	Avg Length Sequences	Stddev Length Sequences	Min Length Sequences	Max Length Sequences	Avg Genome Size
Acidianus ambivalens	1	81	0	81	81	25.36	0	25.36	25.36	2,252,027
Acidianus brierleyi	1	136	0	136	136	24.93	0	24.93	24.93	2,947,244

'Catharanthus roseus' aster yellows phytoplasma

IR Metadata per Organism

Average Count of IR 91.00 (std. deviation: 0.00)

Average Length of Arm IR 11.95 (std. Deviation: 0.00)

Average Length of Spacer IR 3.63 (std. Deviation: 0.00)

Average Length of Sequence IR 27.52 (std. deviation: 0.00)

Average Density IRs 4146.045 (std. deviation: 0.000)

Aggregated analysis

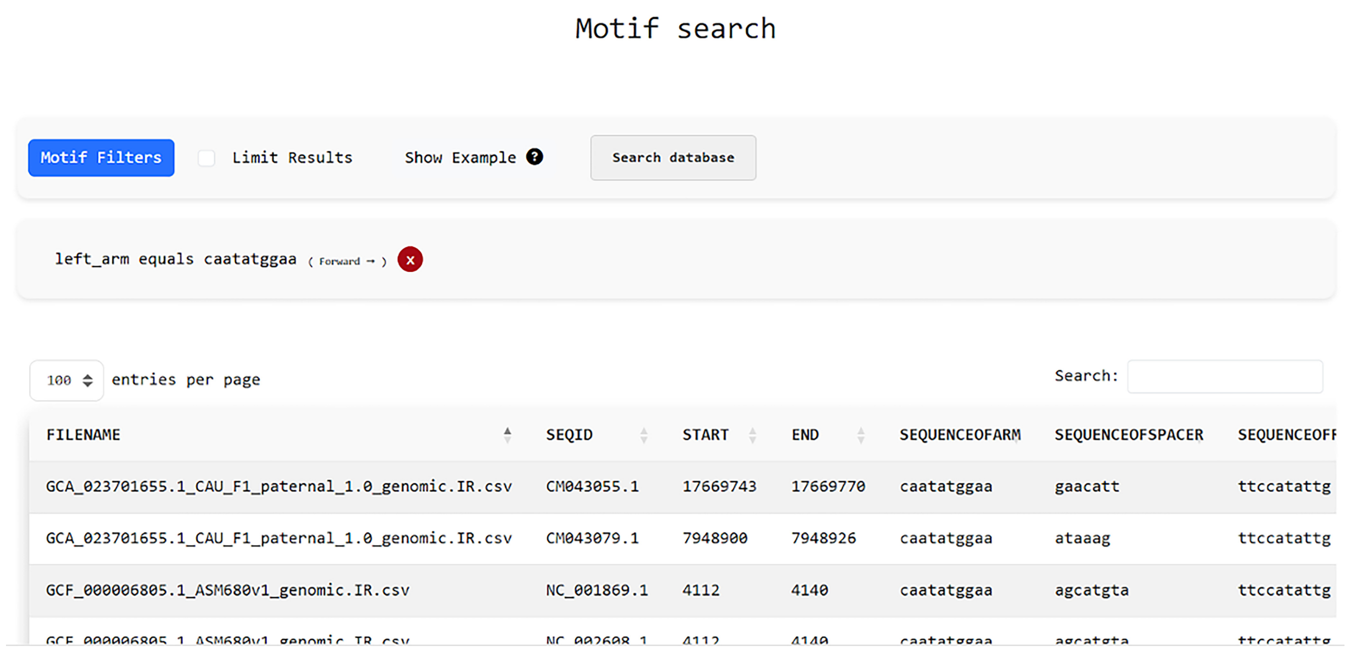
Assemblies: 1

Taxonomy ID: 1193712

Species domain: Bacteria

GC Percent: 28.50 %

Figure 4. Domain selection and domain analysis along with the organism search page and aggregated analysis page. (**A**) Analysis of the Archaea domain, where each assembly is aggregated per organism along with metadata and the ability to download. (**B**) Search unique organisms to perform aggregated analysis on their assemblies. (**C**) Aggregated analysis (metadata, IR metadata, file download, and file inspection) for specific organisms.



left_arm equals caatatggaa (Forward →) ×

100 entries per page

Search:

FILENAME	SEQID	START	END	SEQUENCEOFARM	SEQUENCEOFSPACER	SEQUENCEOFF
GCA_023701655.1_CAU_F1_paternal_1.0_genomic.IR.csv	CM043055.1	17669743	17669770	caatatggaa	gaacatt	ttccatattg
GCA_023701655.1_CAU_F1_paternal_1.0_genomic.IR.csv	CM043079.1	7948900	7948926	caatatggaa	ataaag	ttccatattg
GCF_000006805.1_ASM680v1_genomic.IR.csv	NC_001869.1	4112	4140	caatatggaa	agcatgta	ttccatattg
GCF_000006805.1_ASM680v1_genomic.IR.csv	NC_002608.1	4112	4140	caatatggaa	agcatgta	ttccatattg

Figure 5. Motif search capabilities. Example of motif search in the left arm using a single equality filter.

characteristics (Fig. 3A). A table listing the IRs can be found within the organism's assembly, with the ability to apply fuzzy search across various columns, enabling filtering of the IR data. The user can search by exact chromosome and genomic coordinate ranges, as well as by arm, spacer, sequence length, and GC percentage (Fig. 3B). Finally, there are gene-specific analyses with two functionalities: (i) the ability to search for IRs that overlap specific genes by gene locus tag (Fig. 3C) and (ii) a table identifying genes with the highest IR density, calculated by dividing the total size of IRs overlapping the gene by the gene's size in megabase pairs (Mbp) (Fig. 3D). By default, the top 10 genes with the highest intragenic IR density are displayed in IR bp per Mb.

Domains and Organisms pages

By accessing the Explore → Domains page, the user is redirected to the aggregated IR data in the three domains and viruses. When the user selects a specific domain or virus, they are presented with the statistics of average IR count and density across the assemblies, the average spacer, arm, and sequence length across assemblies and for individual assemblies in a searchable table (Fig. 4A). Upon accessing the Explore → Organisms page, the results of the invertiaDB dataset are arranged by organism name instead of assembly ID (Fig. 4B). The user can analyze and compare the IR data for multiple assemblies of the same organism when multiple assemblies are present (Fig. 4B and C). Additionally, by navigating to the individual organism's page, aggregated IR statistics and metadata are provided, and the user has the option to download these in different file formats (Parquet, CSV, JSON, or BED formats) (Fig. 4C).

Motif search options

The Motif search page enables the search of motifs within IRs or their sub-compartments. There are multiple options to customize the search; these include the following: (i) String sequence search using standard string filters. The standard string operations that can be used are equals, contains, starts with, and ends with, and they can be applied sequence-wise or granularly on specific regions such as left/right arm, spacer, and both arms. (ii) String length search implementing different comparators, including "equals," "greater than," "less than," "greater or equal," and "less or equal." These filters can also be applied on different IR regions (Fig. 5). After the search filters are applied, a metadata search is performed to fetch the different files and the unique organisms that the returned results belong to. (iii) Users can also search on the forward, reverse, and both strands of the sequences, as well for advanced filtering. (iv) There is also the added option to limit the number of results displayed.

Results

High optimal growth temperature of bacteria is linked to low IR density

We investigated the relationship between the optimal growth temperature of bacterial species and differences in IR density as well as their biophysical properties. We separated bacteria into four groups, psychrophiles, mesophiles, thermophiles, and hyperthermophiles, in accordance with the TEMPURA database [54, 57] and examined the IR density in each group. We reported that the IR density is negatively correlated with the optimal growth temperature of bacterial species, with psychrophilic bacteria having the largest number of IR sequences per Mb, with a median of 139.20 IRs per Mb, followed by mesophilic, thermophilic, and hyperthermophilic bacteria (Fig. 6A and Supplementary Table S1). To further investigate this hypothesis, we performed one-way ANOVA comparison of hairpin IR densities across the four temperature classes, which yielded statistically significant differences between the temperature groups ($P < .0001$). This result indicates that low temperatures result in increased frequency of IRs.

Next, we examined the biophysical properties of the IRs across the four temperature groups. Longer IR arm length

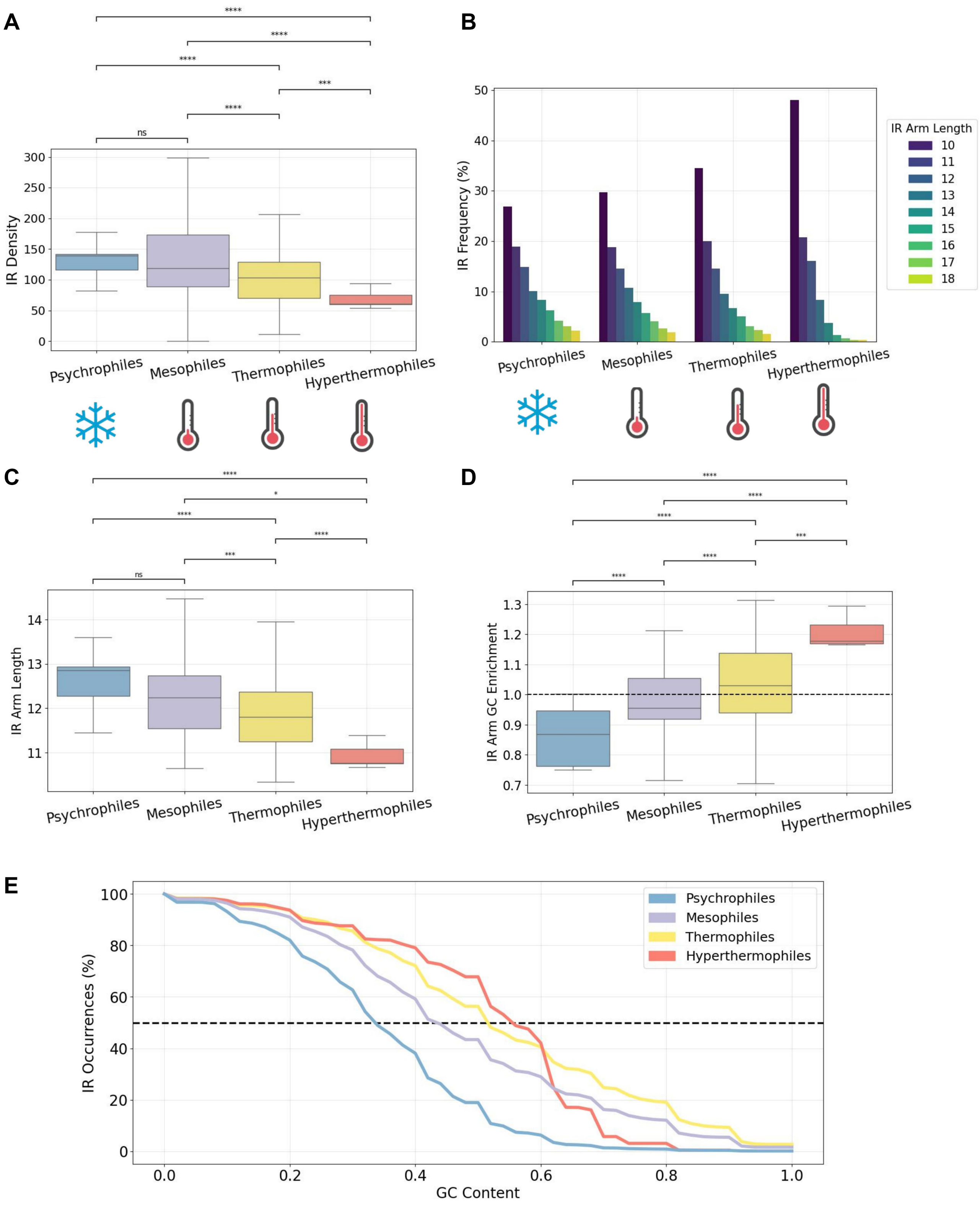


Figure 6. Relationship of hairpins to bacterial optimal growth temperature. (**A**) IR density per Mb as a function of the optimal growth temperature class of bacterial growth. (**B**) IR frequency distribution (%) across each optimal growth temperature class partitioned on IR spacer length. (**C**) Average IR arm length as a function of the optimal growth temperature class of bacterial growth. (**D**) GC enrichment of IR arm as a function of the optimal growth temperature class of bacterial growth. (**E**) Drop rate (%) of total IR sequences present in each growth temperature partition class as a function of the IR arm total GC content.

is linked to more likely hairpin formation [4, 16]. Upon examining the IR arm length proportions within each respective temperature class, we noticed that IRs with shorter arms tend to be more abundant in bacterial strains thriving in higher growth temperatures (Fig. 6B). To further evaluate this observation, we compared the distributions of IR arm length as a function of the bacteria temperature class. We report that psychrophilic bacteria exhibited a propensity for larger IR arms, with a median of 12.9 bp IR arm length, whereas hyperthermophilic bacteria had a median of 10.8 IR bp arm length, indicating that the average IR arm length is anticorrelated with optimal growth temperature (Fig. 6C). IR arms with high GC content are more likely to form stable hairpin structures [5]. We investigated the GC-content enrichment of IR arms in relationship to the GC-content background ratio in the genome of each individual bacterial species. Our results indicate that IR sequences have higher arm GC content than expected in bacteria thriving in higher temperatures, with hyperthermophilic bacteria exhibiting an average 1.15-fold enrichment, whereas mesophilic and psychrophilic bacteria displayed a below-average enrichment in GC-content arm composition, with the latter displaying an average of 0.86-fold enrichment (two-tailed *t*-test with multiple testing adjusted *P*-value; *P*-value <.001 in all comparisons) (Fig. 6D and Supplementary Fig. S2B). We conclude that higher optimal growth temperature is linked to lower IR density, and when these IRs are present in organisms that prefer higher growth temperature, they have shorter arms with higher GC content.

Discussion

Here, we have developed invertiaDB, the first centralized and comprehensive IR database spanning all major taxa across the tree of life. The website is user-friendly and interactive, and provides various features, including search and filtering tools, dynamic tables that allow for querying and sorting, visualizations, and data downloads for in-depth, independent analysis. As the number of available organismal genomes continues to grow, we plan to incorporate them in our database with regular updates.

IRs are a highly dynamic DNA element that can form hairpin and cruciform structures. They are associated with a plethora of functions, but are also linked to increased genomic instability [15, 19, 26–35]. The invertiaDB database enables the systematic examination of the functional roles of IRs, including those in gene regulation, replication, transposition, genome organization, and gene and genome evolution, among others [20, 58, 59]. The integration of biophysical properties, including the spacer and arm lengths, and the nucleotide composition of each IR allows for the study of these parameters and the impact of IR stability on its functional roles and genomic instability. Finally, our database could be utilized by researchers who are interested in examining DNA repair systems associated with hairpin and cruciform structures, which often differ between organisms belonging to different taxonomies [60, 61].

Our findings highlight a relationship between bacterial growth temperature and IR density across the studied bacterial genomes. Specifically, the reduced IR density in thermophiles and hyperthermophiles suggests a selective pressure to minimize DNA secondary structures that could destabilize genomic integrity under thermal stress. The higher GC content and shorter arm lengths observed in the IRs of high-temperature bacteria further underscore the role of sequence composition and structure in maintaining DNA stability in extreme environments. These adaptations may represent evolutionary strategies to mitigate the risk of DNA damage or replication hindrance at elevated temperatures. In contrast, the higher IR density and longer arms with lower GC enrichment in psychrophiles and mesophiles may reflect a lower selective pressure against IR formation, potentially enabling diverse regulatory roles, such as Rho-independent transcription termination [62], for IRs at lower temperatures. Overall, these findings provide new insights into how environmental conditions, such as temperature, shape genomic architecture through selective pressures on repetitive DNA elements.

One limitation of invertiaDB is excluding imperfect IRs. There are two reasons for this: (i) the search space, computational resources, and number of detected imperfect IRs across 118 101 are prohibitively large and would require large computational resources, while imperfect IRs are less likely to form hairpins and cruciforms [12, 16]. Another limitation of our study is its reliance on computational predictions without experimental validation, which may introduce biases or inaccuracies, including sequencing errors. Experimental research will be essential to validate the biological relevance and functionality of the identified IRs when utilizing our database across research projects.

Maintaining invertiaDB in sync with the continual updates to genome assemblies at NCBI poses significant challenges, including managing the integration of new assemblies and accommodating revisions to existing ones. To address these challenges, we plan to implement pipelines that retrieve and analyze updated genome data every 12 months, ensuring that the database remains comprehensive and current. Future efforts will also focus on developing a version-controlled framework to track changes, improve scalability, and support community-driven contributions to enhance the database's utility for diverse research applications.

We envision that invertiaDB will be adopted by researchers around the world to explore the roles and applications of IRs. invertiaDB will enable the systematic exploration of IRs and advance our understanding of their broader roles and impact on organismal evolution and biological functions.

Acknowledgements

Author contributions: Kimonas Provatas (Conceptualization [equal], Data curation [lead], Formal analysis [lead], Investigation [lead], Methodology [lead], Software [lead], Validation [lead], Visualization [equal], Writing—original draft [equal], Writing—review & editing [equal]), Nikol Chantzi (Conceptualization [equal], Data curation [lead], Formal analysis [lead], Investigation [lead], Methodology [lead], Software [lead], Validation [lead], Visualization [lead], Writing—original draft [equal], Writing—review & editing [equal]), Nafsika Amptazi (Data curation [equal], Formal analysis [equal], Investigation [equal], Software [equal], Visualization [lead], Writing—review & editing [equal]), Michail Patsakis (Methodology [equal], Writing—review & editing [equal]), Akshatha Nayak (Methodology [equal], Writing—review & editing [equal]), Ioannis Mouratidis (Project administration [equal], Supervision [equal], Writing—original draft [equal], Writing—review & editing [equal]), Apostolos Zaravinos (Supervision [equal], Writing—review & editing [equal]), Georgios A. Pavlopoulos (Project administration [equal], Supervision [equal], Writing—review & editing [equal]), and Ilias Georgakopoulos-Soares (Conceptualization [lead], Funding acquisition [lead], Investigation [lead], Methodology [lead], Project administration [lead], Resources [lead], Software [supporting], Supervision [lead], Visualization [supporting], Writing—original draft [lead], Writing—review & editing [lead])

Supplementary data

Supplementary data is available at NAR online.

Conflict of interest

None declared.

Funding

This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number R35GM155468. G.A.P. was funded by Fondation Sante; Onassis Foundation; Hellenic Foundation for Research and Innovation (H.F.R.I) under the call "Greece 2.0—Basic Research Financing Action," sub-action II, Grant ID: 16718-PRPFOR; Program "Greece 2.0, National Recovery and Resilience Plan," Grant ID: TAEDR-0539180.

Data availability

The invertiaDB dataset can be found in Zenodo with a stable version at https://zenodo.org/records/13856709. The GitHub code for the extraction of IRs, filtering, and analyses is found at https://github.com/Georgakopoulos-Soares-lab/invertiaDB-analysis.

References

1. Ghosh A, Bansal M. A glossary of DNA structures from A to Z. *Acta Crystallogr D Biol Crystallogr* 2003;**59**:620–6. https://doi.org/10.1107/S0907444903003251
2. Kaushik M, Kaushik S, Roy K *et al.* A bouquet of DNA structures: emerging diversity. *Biochem Biophys Rep* 2016;**5**:388–95.
3. Choi J, Majima T. Conformational changes of non-B DNA. *Chem Soc Rev* 2011;**40**:5893–909. https://doi.org/10.1039/c1cs15153c
4. Bikard D, Loot C, Baharoglu Z *et al.* Folded DNA in action: hairpin formation and biological functions in prokaryotes. *Microbiol Mol Biol Rev* 2010;**74**:570–88. https://doi.org/10.1128/MMBR.00026-10
5. Buisson R, Langenbucher A, Bowen D *et al.* Passenger hotspot mutations in cancer driven by APOBEC3A and mesoscale genomic features. *Science* 2019;**364**:eaaw2872. https://doi.org/10.1126/science.aaw2872
6. Lu S, Wang G, Bacolla A *et al.* Short inverted repeats are hotspots for genetic instability: relevance to cancer genomes. *Cell Rep* 2015;**10**:1674–80. https://doi.org/10.1016/j.celrep.2015.02.039
7. Bacolla A, Tainer JA, Vasquez KM *et al.* Translocation and deletion breakpoints in cancer genomes are associated with potential non-B DNA-forming sequences. *Nucleic Acids Res* 2016;**44**:5673–88. https://doi.org/10.1093/nar/gkw261
8. Azeroglu B, Lincker F, White MA *et al.* A perfect palindrome in the *Escherichia coli* chromosome forms DNA hairpins on both leading- and lagging-strands. *Nucleic Acids Res* 2014;**42**:13206–13. https://doi.org/10.1093/nar/gku1136
9. Rosche WA, Trinh TQ, Sinden RR. Differential DNA secondary structure-mediated deletion mutation in the leading and lagging strands. *J Bacteriol* 1995;**177**:4385–91. https://doi.org/10.1128/jb.177.15.4385-4391.1995
10. Woodside MT, Behnke-Parks WM, Larizadeh K *et al.* Nanomechanical measurements of the sequence-dependent folding landscapes of single nucleic acid hairpins. *Proc Natl Acad Sci USA* 2006;**103**:6190–5. https://doi.org/10.1073/pnas.0511048103
11. SantaLucia J Jr. A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. *Proc Natl Acad Sci USA* 1998;**95**:1460–5. https://doi.org/10.1073/pnas.95.4.1460
12. Rentzeperis D, Shikiya R, Maiti S *et al.* Folding of intramolecular DNA hairpin loops: enthalpy–entropy compensations and hydration contributions. *J Phys Chem B* 2002;**106**:9945–50. https://doi.org/10.1021/jp0260853
13. Nag DK, Petes TD. Seven-base-pair inverted repeats in DNA form stable hairpins *in vivo* in *Saccharomyces cerevisiae*. *Genetics* 1991;**129**:669–73. https://doi.org/10.1093/genetics/129.3.669
14. Nasar F, Jankowski C, Nag DK. Long palindromic sequences induce double-strand breaks during meiosis in yeast. *Mol Cell Biol* 2000;**20**:3449–58. https://doi.org/10.1128/MCB.20.10.3449-3458.2000
15. Lobachev KS, Shor BM, Tran HT *et al.* Factors affecting inverted repeat stimulation of recombination and deletion in *Saccharomyces cerevisiae*. *Genetics* 1998;**148**:1507–24. https://doi.org/10.1093/genetics/148.4.1507
16. Sinden RR, Zheng G, Brankamp RG *et al.* On the deletion of inverted repeated DNA in *Escherichia coli*: effects of length, thermal stability, and cruciform formation *in vivo*. *Genetics* 1991;**129**:991–1005. https://doi.org/10.1093/genetics/129.4.991
17. Bi X, Liu LF. DNA rearrangement mediated by inverted repeats. *Proc Natl Acad Sci USA* 1996;**93**:819–23. https://doi.org/10.1073/pnas.93.2.819
18. Strawbridge EM, Benson G, Gelfand Y *et al.* The distribution of inverted repeat sequences in the *Saccharomyces cerevisiae* genome. *Curr Genet* 2010;**56**:321–40. https://doi.org/10.1007/s00294-010-0302-6
19. Georgakopoulos-Soares I, Victorino J, Parada GE *et al.* High-throughput characterization of the role of non-B DNA motifs on promoter function. *Cell Genom* 2022;**2**:100111.
20. Brázda V, Bartas M, Lýsek J *et al.* Global analysis of inverted repeat sequences in human gene promoters reveals their non-random distribution and association with specific biological pathways. *Genomics* 2020;**112**:2772–7. https://doi.org/10.1016/j.ygeno.2020.03.014
21. Kingsford CL, Ayanbule K, Salzberg SL. Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. *Genome Biol* 2007;**8**:R22. https://doi.org/10.1186/gb-2007-8-2-r22
22. von Hippel PH. An integrated model of the transcription complex in elongation, termination, and editing. *Science* 1998;**281**:660–5. https://doi.org/10.1126/science.281.5377.660
23. Pearson CE, Zorbas H, Price GB *et al.* Inverted repeats, stem–loops, and cruciforms: significance for initiation of DNA replication. *J Cell Biochem* 1996;**63**:1–22. https://doi.org/10.1002/(SICI)1097-4644(199610)63:1%3c1::AID-JCB1%3e3.0.CO;2-3
24. Fattash I, Rooke R, Wong A *et al.* Miniature inverted-repeat transposable elements: discovery, distribution, and activity. *Genome* 2013;**56**:475–86. https://doi.org/10.1139/gen-2012-0174
25. Bowater RP, Bohálová N, Brázda V. Interaction of proteins with inverted repeats and cruciform structures in nucleic acids. *Int J Mol Sci* 2022;**23**:6171. https://doi.org/10.3390/ijms23116171
26. Gordenin DA, Lobachev KS, Degtyareva NP *et al.* Inverted DNA repeats: a source of eukaryotic genomic instability. *Mol Cell Biol* 1993;**13**:5315–22.
27. Lobachev KS, Rattray A, Narayanan V. Hairpin- and cruciform-mediated chromosome breakage: causes and

consequences in eukaryotic cells. *Front Biosci* 2007;**12**:4208–20. https://doi.org/10.2741/2381

28. Nag DK, Kurst A. A 140-bp-long palindromic sequence induces double-strand breaks during meiosis in the yeast *Saccharomyces cerevisiae*. *Genetics* 1997;**146**:835–47. https://doi.org/10.1093/genetics/146.3.835
29. Butler DK, Gillespie D, Steele B. Formation of large palindromic DNA by homologous recombination of short inverted repeat sequences in *Saccharomyces cerevisiae*. *Genetics* 2002;**161**:1065–75. https://doi.org/10.1093/genetics/161.3.1065
30. Achaz G, Coissac E, Netter P *et al.* Associations between inverted repeats and the structural evolution of bacterial genomes. *Genetics* 2003;**164**:1279–89. https://doi.org/10.1093/genetics/164.4.1279
31. Leach DR. Long DNA palindromes, cruciform structures, genetic instability and secondary structure repair. *Bioessays* 1994;**16**:893–900. https://doi.org/10.1002/bies.950161207
32. Tanaka H, Tapscott SJ, Trask BJ *et al.* Short inverted repeats initiate gene amplification through the formation of a large DNA palindrome in mammalian cells. *Proc Natl Acad Sci USA* 2002;**99**:8772–7. https://doi.org/10.1073/pnas.132275999
33. Zhou ZH, Akgūn E, Jasin M. Repeat expansion by homologous recombination in the mouse germ line at palindromic sequences. *Proc Natl Acad Sci USA* 2001;**98**:8326–33. https://doi.org/10.1073/pnas.151008498
34. Lindsey JC, Leach DR. Slow replication of palindrome-containing DNA. *J Mol Biol* 1989;**206**:779–82. https://doi.org/10.1016/0022-2836(89)90584-6
35. Georgakopoulos-Soares I, Morganella S, Jain N *et al.* Noncanonical secondary structures arising from non-B DNA motifs are determinants of mutagenesis. *Genome Res* 2018;**28**:1264–71. https://doi.org/10.1101/gr.231688.117
36. Bastos CAC, Afreixo V, Rodrigues JMOS *et al.* Concentration of inverted repeats along human DNA. *J Integr Bioinform* 2023;**20**:20220052.
37. Wang Y, Leung FCC. Long inverted repeats in eukaryotic genomes: recombinogenic motifs determine genomic plasticity. *FEBS Lett* 2006;**580**:1277–84. https://doi.org/10.1016/j.febslet.2006.01.045
38. Zou X, Morganella S, Glodzik D *et al.* Short inverted repeats contribute to localized mutability in human somatic cells. *Nucleic Acids Res* 2017;**45**:11213–21. https://doi.org/10.1093/nar/gkx731
39. Nik-Zainal S, Davies H, Staaf J *et al.* Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* 2016;**534**:47–54. https://doi.org/10.1038/nature17676
40. Brázda V, Kolomazník J, Lýsek J *et al.* Palindrome analyser—a new web-based server for predicting and evaluating inverted repeats in nucleotide sequences. *Biochem Biophys Res Commun* 2016;**478**:1739–45. https://doi.org/10.1016/j.bbrc.2016.09.015
41. Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* 2000;**16**:276–7. https://doi.org/10.1016/S0168-9525(00)02024-2
42. Ye C, Ji G, Li L *et al.* detectIR: a novel program for detecting perfect and imperfect inverted repeats using complex numbers and vector calculation. *PLoS One* 2014;**9**:e113349. https://doi.org/10.1371/journal.pone.0113349
43. Warburton PE, Giordano J, Cheung F *et al.* Inverted repeat structure of the human genome: the X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes. *Genome Res* 2004;**14**:1861–9. https://doi.org/10.1101/gr.2542904
44. Cer RZ, Donohue DE, Mudunuri US *et al.* Non-B DB v2.0: a database of predicted non-B DNA-forming motifs and its associated tools. *Nucleic Acids Res* 2013;**41**:D94–100.
45. Jia L, Li Y, Huang F *et al.* LIRBase: a comprehensive database of long inverted repeats in eukaryotic genomes. *Nucleic Acids Res* 2022;**50**:D174–82. https://doi.org/10.1093/nar/gkab912
46. Liao X, Hu K, Salhi A *et al.* msRepDB: a comprehensive repetitive sequence database of over 80 000 species. *Nucleic Acids Res* 2022;**50**:D236–45. https://doi.org/10.1093/nar/gkab1089
47. Bao W, Kojima KK, Kohany O. Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mobile DNA* 2015;**6**:11. https://doi.org/10.1186/s13100-015-0041-9
48. Provatas K, Chantzi N, Patsakis M *et al.* Microsatellites Explorer: a database of short tandem repeats across genomes. *Comput Struct Biotechnol J* 2024;**23**:3817–26. https://doi.org/10.1016/j.csbj.2024.10.041
49. O'Leary NA, Wright MW, Brister JR *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* 2016;**44**:D733–45. https://doi.org/10.1093/nar/gkv1189
50. Benson DA, Cavanaugh M, Clark K *et al.* GenBank. *Nucleic Acids Res* 2013;**41**:D36–42. https://doi.org/10.1093/nar/gks1195
51. Piro VC, Shen W. Genome Updater. 2019. https://github.com/pirovc/genome_updater, (15 November 2024, date last accessed).
52. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;**26**:841–2. https://doi.org/10.1093/bioinformatics/btq033
53. Chantzi N, Moeckel C, Chan CSY *et al.* Characterization of hairpin loops and cruciforms across 118,065 genomes spanning the tree of life. bioRxiv, https://doi.org/10.1101/2024.09.29.615628, 29 September 2024, preprint: not peer reviewed.
54. Sato Y, Okano K, Kimura H *et al.* TEMPURA: database of growth TEMPeratures of Usual and RAre prokaryotes. *Microbes Environ* 2020;**35**:ME20074. https://doi.org/10.1264/jsme2.ME20074
55. Leinonen R, Akhtar R, Birney E *et al.* The European Nucleotide Archive. *Nucleic Acids Res* 2011;**39**:D28–31. https://doi.org/10.1093/nar/gkq967
56. Pruitt KD, Tatusova T, Klimke W *et al.* NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Res* 2009;**37**:D32–6. https://doi.org/10.1093/nar/gkn721
57. Moyer CL, Eric Collins R, Morita RY. Psychrophiles and psychrotrophs. In: *Reference Module in Life Sciences*. Amsterdam, The Netherland: Elsevier, 2017.
58. Georgakopoulos-Soares I, Chan CSY, Ahituv N *et al.* High-throughput techniques enable advances in the roles of DNA and RNA secondary structures in transcriptional and post-transcriptional gene regulation. *Genome Biol* 2022;**23**:159. https://doi.org/10.1186/s13059-022-02727-6
59. Georgakopoulos-Soares I, Parada GE, Hemberg M. Secondary structures in RNA synthesis, splicing and translation. *Comput Struct Biotechnol J* 2022;**20**:2871–84. https://doi.org/10.1016/j.csbj.2022.05.041
60. Vasquez KM, Wang G. The yin and yang of repair mechanisms in DNA structure-induced genetic instability. *Mutat Res* 2013;**743–744**:118–31. https://doi.org/10.1016/j.mrfmmm.2012.11.005
61. Pitcher RS, Brissett NC, Doherty AJ. Nonhomologous end-joining in bacteria: a microbial perspective. *Annu Rev Microbiol* 2007;**61**:259–82. https://doi.org/10.1146/annurev.micro.61.080706.093354
62. Farnham PJ, Platt T. Rho-independent termination: dyad symmetry in DNA causes RNA polymerase to pause during transcription *in vitro*. *Nucleic Acids Res* 1981;**9**:563–77. https://doi.org/10.1093/nar/9.3.563

Received: October 5, 2024. **Revised:** April 9, 2025. **Editorial Decision:** April 10, 2025. **Accepted:** April 14, 2025