International Journal on Science and Technology (IJSAT)

E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Bacterial Genome Annotation of Escherichia coli

Niranjana Sreenivasan[1], Moulya R Gowda[2], M D Yaana Muthamma[3], Shivandappa Mr[4], Divyashree Kori[5], Navya N[6], Taru Das[7]

[1, 2, 3, 4, 6, 7]Biotechnology Department, R V College of Engineering, India
[5]Artificial Intelligence and Machine Learning Department, Dr. Ambedkar Institute of Technology, India

Abstract

Annotating the bacterial genome is essential for comprehending the genetic foundations of an organism's functions and behaviors. Escherichia coli serves as a well-researched model organism, and its genome annotation has been thoroughly examined. The E. coli genome includes four thousand two hundred and eighty eight protein-coding genes, of which thirty eight percent lack a defined function. Comparative genomics uncovers widespread and narrowly situated gene families, along with paralogous protein families, including the eighty ABC transporters. The genome is structured according to replication direction and includes insertion sequence elements, remnants of phages, and areas of atypical composition that suggest genome plasticity. Sophisticated bioinformatics tools and pipelines, including Galaxy, Prokka, Beav, and coliBASE, have been created to aid in E. coli genome annotation, offering insights into gene functionality, genome evolution, and species variation. These tools allow scientists to investigate the intricate biology of E. coli and its related species, promoting progress in areas like microbiology, genetics, and biotechnology.

Keywords: Bacteria, genome, annotation, Escherichia coli, Galaxy

I. Introduction

Genome sequencing is essentially a molecular scientific technique that can identify the exact arrangement of bases or nucleotides in an organism's DNA. These discoveries enable genomic medicine to interpret a genome's information as guidelines for an organism's development, operation, and well-being.Genome sequencing empirically focuses on identifying the base sequence (adenine [A], thymine [T], cytosine [C], and guanine [G]) that constitutes that DNA molecule in order to ascertain the function of those genes, gene variation, and disease risk.

A. *Whole Genome Sequencing*

WGS uses DNA sequencing to identify all or almost all of a genome's nucleotides. WGS provides detailed information about an organism's genetic composition, which is very useful when studying intricate traits and diseases.

B. *Next-Generation Sequencing*

This quick DNA sequencing of enormous amounts was made practical due to the present NGS methodology. NGS is quicker, less expensive, and usually regarded as less costly than the Sanger sequencing process and other antiquated methods.

C. *Targeted Sequencing*

This technique focuses on the particular regions of interest, in this case genes thought to be connected to particular diseases or characteristics, and offers quicker and less expensive sequencing. However, this kind of approach yields a lot of valuable information.

D. *Techniques in Genome Sequencing*

The procedures or techniques used in genome sequencing include:

- Extraction: By removing DNA from the organism being studied, this initiates the sequencing process.
- The term "library preparation" describes the process of preparing sequenced DNA, which includes converting RNA to DNA, breaking DNA up into smaller fragments, and adding specific sequences that the sequencer can recognize.
- Sequencing: This involves making a sequence decision after sequencing a prepared and verified DNA library.

Genome annotation encompasses two primary objectives:

- Identification of Elements: This involves locating genes, regulatory regions, and other functional elements in a DNA sequence.
- Assigning Functions: Following identification, the next objective is to determine what these elements do-that is, their biological functions or roles.

This process transforms a raw sequence of DNA into a more understandable format that scientists can use to analyze and understand effectively.

E. Focal Points of Genome Annotation

Genome annotation mainly deals with several aspects:

- Structural Annotation : In this type, one predicts the locations of genes and other important features such as regulatory motifs and non-coding regions. A variety of techniques involving computational algorithms and machine learning models is often used to accomplish the above.
- Functional Annotation: After the structural elements have been identified, researchers want to know their functions. It compares the sequences found with those already known in the databases for inferences on the biological role from homology or contextual information within the genome such as adjacency of genes.
- Process Annotation: This part aims at creating relationships among various genomic elements; often it rebuilds metabolic pathways and provides understanding of the regulation mechanisms at a molecular level.

A prokaryotic cell is a simple, unicellular organism that lacks a nucleus and membrane-bound organelles, distinguishing it from eukaryotic cells. Its genetic material, composed of a single circular DNA molecule, is located in the nucleoid region, which is not enclosed by a membrane. The cell is surrounded by a rigid cell wall, primarily made of peptidoglycan in bacteria, providing structural support and protection. A plasma membrane beneath the cell wall regulates the transport of substances in and out of the cell. Ribosomes, though smaller than those in eukaryotic cells, facilitate protein synthesis. Some prokaryotic cells possess flagella for motility and pili for adhesion to surfaces and other cells.

Prokaryotic cells are classified into two domains: Bacteria and Archaea. Examples include Escherichia coli, which inhabits the intestines and aids in digestion; Streptococcus pneumoniae, which can cause respiratory infections; and Cyanobacteria, photosynthetic bacteria that contribute to oxygen production and nitrogen fixation in aquatic ecosystems. Prokaryotic cells play essential roles in nutrient cycling, biotechnology, and medicine.

A cell from the bacterial domain is a bacterial cell if it is unicellular and prokaryotic, meaning that such cells lack membrane-bound organelles or an obvious nucleus. Instead, a nucleoid in a bacterial cell is where its genetic material can be found. The general makeup of bacteria is a cell membrane, a firm cell wall, and often a protective outer capsule. Some bacteria have appendages like tails that allow them to move. Bacteria also carry extra genes such as antibiotic resistance. Plasmids are small, circular DNA molecules that are found in bacteria. Bacteria are capable of thriving in many environments including soil, water, and the human body due to their small size and flexibility. Some examples include Escherichia coli (E. coli).

Escherichia coli (E. coli) is a bacillus bacterium that has probably developed in tandem with its hosts, residing in the intestines of humans and various warm-blooded creatures for millions of years. This enduring connection indicates a profoundly rooted mutualistic relationship. In 1885, Theodor Escherich, a pediatrician from Germany, reached a significant milestone by isolating and identifying this bacterium, which he originally referred to as Bacterium coli commune. Escherich's extensive research focused on examining the bacterial populations found in infants' intestines, exploring how these microbial communities change following birth and evaluating their possible impact on health concerns in infants. His research, which included meticulous observations and thorough experimentation, established the groundwork for our comprehension of the intricate connections between humans and their gut microbiota, profoundly influencing the domains of microbiology and medicine. E. coli bacteria mostly live in the digestive system of humans and animals. Nonetheless, because of fecal pollution, they can be present in different environments. Typical sources of E. coli consist of tainted food like undercooked ground beef, unpasteurized milk and dairy items, along with contaminated vegetables, especially leafy greens. Water tainted with contaminants, whether for recreation or drinking, may also contain E. coli. Interacting with animals, such as having direct contact with infected individuals or their waste, as well as engaging with contaminated areas or objects, can result in exposure. Transmission between individuals can happen because of poor hygiene practices, particularly the absence of handwashing.

II. LITERATURE REVIEW

We did a literature review of about 35 papers. Here are the following papers. Some of the papers are as follows:

A. “Beav: a bacterial genome and mobile element annotation pipeline”

This article, released in mSphere on 28th August 2024, presents Beav, an automated system aimed at improving the annotation of bacterial genomes along with their mobile genetic components. Created by scientists at Oregon State University, Beav combines various databases and analytical tools to offer detailed annotations of genes, regulatory elements, and non-coding regions. An essential characteristic of Beav is its capability to detect different mobile genetic elements, including plasmids, transposons, and prophages, which are vital for comprehending horizontal gene transfer and the propagation of antibiotic resistance. The user-friendly interface of the pipeline allows researchers with different levels of bioinformatics knowledge to access it easily. Validation research has shown that Beav provides high precision in gene prediction and successfully detects a wide variety of mobile genetic elements, performing similarly to or better than current annotation tools. The writers propose that Beav can enhance comprehension of bacterial genetics, evolution, and the processes driving gene mobility.

B. *"Annotation of bacterial and archaeal genomes: enhancing precision and uniformity"*

This article, released in Chemical Reviews in August 2007, discusses the difficulties related to annotating bacterial and archaeal genomes. Precise genome annotation is crucial for comprehending the functional abilities and ecological functions of microorganisms. The authors explore methods to improve the accuracy and consistency of genome annotations, highlighting the significance of standardized techniques and the incorporation of different computational tools. They underscore the necessity for uniform annotation methods to enhance comparative genomics and to improve our comprehension of microbial biology. The document acts as an essential tool for scholars looking to enhance genome annotation methods and highlights the importance of precision and uniformity in genomic research.

C. *"Genome Annotation"*

This article, released in the Encyclopedia of Life Sciences in 2001, offers an extensive summary of the techniques and importance of annotating genomic sequences. The authors explore the method of detecting and tagging functional components in a genome, including genes, regulatory areas, and non-coding regions. They highlight the significance of using both computational and experimental methods to attain precise annotations. The article also emphasizes the difficulties linked to genome annotation, such as predicting gene functions and combining various data types. Through clarifying these concepts, the authors intend to educate researchers on the essential importance of genome annotation in comprehending biological systems and promoting genomic research.

D. *"Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences"*

The 2010 publication named "Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences" presents Galaxy, an accessible, online platform aimed at tackling issues in computational biology. The authors emphasize concerns about the accessibility and reproducibility of computational analyses in the life sciences, pointing out that numerous researchers encounter challenges in using intricate computational tools and confirming that their analyses can be consistently replicated. Galaxy tackles these issues by offering an easy-to-use interface that enables users to conduct intricate analyses without requiring deep programming skills. It autonomously monitors and oversees data lineage, guaranteeing that the sources and changes of data are thoroughly recorded. Moreover, Galaxy provides "Galaxy Pages," interactive online documents that allow researchers to share their entire computational analyses in a clear manner. By incorporating these features, Galaxy improves the accessibility, reproducibility, and transparency of computational research in genomics and related disciplines.

III. Features of a bacterial E.coli cell

Cell shape and size– The shape of bacteria E. coli is bacillus which means it has a rod shape, and on average it is 1-2 µm long and 0.5 µm wide.

Gram-Negative Structure – It has an outer membrane containing lipopolysaccharides (LPS) and a thin peptidoglycan layer to protect and also to regulate immune response.

Cell Membranes – E. coli is a gram negative bacteria so it has two membranes, the inner (cytoplasmic) and the outer which are responsible for regulating the entry and exit of the cell, importing the necessary nutrients and exporting the waste products.

Motility – The chemotaxis induced by peritrichous flagella makes the majority of E. coli strains motile. However, there are some non-motile strains.

Fimbriae and Pili – These bacterial structures resembling hair-like structures enable the bacteria to adhere to surfaces and host tissues. Sex pili are involved in the bacterial conjugation process, which is the transfer of genetic material from one bacterium to another. Genetic Material – E. coli has a single circular chromosome, and extra genes like the one encoding for antibiotic resistance are present on plasmids.

Facultative Anaerobe — Bacteria can survive in either an aerobic or an anaerobic environment, which means it could potentially adapt to vary environments and use either respiration or fermentation.

Reproduction –E. coli bacteria use a process of asexual reproduction known as binary fission to reproduce themselves in a very short time.

IV. SOFTWARES

A. Galaxy

Galaxy is a powerful, widely used tool in bioinformatics for analyzing and managing genome scale data. An open source, web based environment which means that complex computational workflows can be conducted without having to learn how to program. Analysis, sequence alignment, variant calling, metagenomics, transcriptomics and functional annotation are among the many bioinformatics applications it supports. It facilitates straightforward processing of high-throughput sequencing data by integrating tools, such as Bowtie, BWA, HISAT2, GATK and Prokka. Raw sequencing reads can be analyzed, mutations identified, gene expression comparisons made and phylogenetic analysis performed. It also guarantees reproducibility by enabling the sharing and reusing of workflows. FASTQ, BAM, VCF and GFF3 are multiple file formats that Galaxy supports, which facilitates data interoperability and compatibility with other bioinformatics tools. Local installations to cloud based implementations makes it suitable for small scale studies to large collaborative projects. In genomics, personalized medicine, microbiology and evolutionary biology, bioinformatics analyses are thus streamlined by Galaxy.

B. Prokka

Prokka is a powerful command-line software tool for the rapid annotation of prokaryotic genomes, including Escherichia coli. It simplifies the process of genome annotation by integrating various tools that exist into one, so researchers can identify and label genomic features such as protein-coding regions and RNA genes efficiently. Prokka annotates a draft bacterial genome in 10 minutes on a standard desktop computer, and it gives output files in several formats, such as GenBank, EMBL, and GFF. This makes it useful for handling large datasets generated by high-throughput sequencing technologies. The software develops a two-stage annotation process for the protein-coding regions: first, it detects coding sequences using the Prodigal algorithm. Based on similarity to known proteins in established databases, it predicts functionality. Prokka is open-source and available under the GPLv2 license so it is accessible for integration into various computational pipelines and genomic analyses.

C. BLASTN

BLASTN is to compare a nucleotide sequence query with a collection of nucleotide sequences. This characteristic helps researchers identify similarities between fragments, which can reveal potential evolutionary evolution, gene activity, or even gene alterations. Essential Elements of BLASTN Goal: Finally, BLASTN is used to perform a local alignment, which looks for sequences that are comparable to an input sequence in a local database. This advances our knowledge of gene activity and interspecies relationships.

Heuristic Approach: BLASTN speeds up the search in addition to making other minor enhancements by using heuristic techniques. Although it is faster than other approaches, it is nevertheless preferred for the majority of big genomic databases, even though it will undoubtedly lose some accuracy when compared to exhaustive methods like the Smith-Waterman algorithm.

Output: A report containing helpful information like alignment and statistical significance for the sequences that were found to be similar to the query is given to the user after they finish a BLASTN search.

Uses: BlastN will determine how closely a query sequence resembles known sequences in the database that may be able to provide insight into the function, structure, and evolution of the query sequence.

Gene finding: To ascertain whether genes are included in the query sequence, a genomic sequence can be compared to a database of known genes.

Phylogenetic analysis: The program can reconstruct the phylogenetic relationships by detecting comparable sequences from various organisms.

Genome assembly: BlastN can be used to identify overlapping sequences between different genomic segments in order to assemble the genome.

Working: Entering a sequence to query with: A nucleotide sequence may be entered into the computer either as DNA or RNA.

D. Bakta

Bakta is a command-line software tool designed for the rapid and standardized annotation of bacterial genomes and plasmids. It features a taxon-independent approach, making it suitable for a wide range of bacterial species, including those from metagenomic datasets. Bakta employs an alignment-free sequence identification method, allowing it to annotate typical bacterial genomes in about 10 minutes and plasmids in seconds to minutes. The software provides extensive database cross-references to well-established databases like RefSeq and UniProt, enhancing functional annotation. Results can be exported in various formats, including GFF3 and JSON. Bakta is implemented in Python 3, compatible with MacOS and Linux, and is freely available under the GPLv3 license, with a web-based version also accessible for users preferring a graphical interface. It is particularly useful for researchers conducting high-throughput genomic studies or working with less-characterized metagenomic assembled genomes (MAGs).

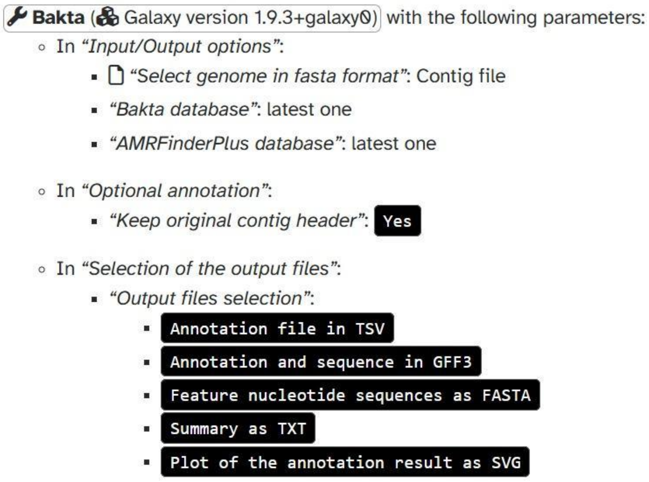
E. Beav

Beav: A command line tool that streamlines and automates bacterial genome and mobile genetic element annotation. It is also a comprehensive genome annotation pipeline for bacteria and associated mobile genetic elements. The Beav pipeline incorporates many annotation tools, which automate the process of running, parsing and combining the obtained results into a single-easy to read output. Beav contains many elements that enhance the annotation of plant-associated microbes, including genes and regulatory elements important to phytopathogens and mutualistic symbionts. The Agrobacterium specific pipeline identifies the presence of oncogenic Ti and Ri plasmids and classifies them under a published scheme. This is also responsible for annotating Ti/Ri plasmid-specific regulatory elements and for reporting the taxonomic classification of the input strain under the Agrobacterium biovar/genomospecies scheme. Beav generates a separate plot to visualize oncogenic Ti/Ri plasmids, only if present.

V. METHODOLOGY

This study employed a comprehensive bioinformatics approach to analyze the genome sequence of E. coli. The methodology consisted of three primary steps. Initially, the FASTA file of the genome sequence was downloaded from the National Center for Biotechnology Information (NCBI) GenBank database. The file, titled "sequence.fasta", contained the complete genome sequence of E. coli.

- The downloaded file was then imported into the Galaxy platform, a web-based bioinformatics framework, to facilitate further analysis. Subsequently, the contigs were annotated using Bakta, a prokaryotic genome annotation tool. This annotation enabled the identification of functional elements within the genome, including coding sequences, non-coding RNAs, and regulatory regions.
- The annotated genome sequence provided valuable insights into the genetic makeup of E. coli, allowing for a deeper understanding of its physiological and pathological processes. The results of this study have important implications for future research into E. coli biology and its applications in biotechnology and medicine.



Bakta (Galaxy version 1.9.3+galaxy0) with the following parameters:

- In *"Input/Output options"*:
  - *"Select genome in fasta format"*: Contig file
  - *"Bakta database"*: latest one
  - *"AMRFinderPlus database"*: latest one
- In *"Optional annotation"*:
  - *"Keep original contig header"*: `Yes`
- In *"Selection of the output files"*:
  - *"Output files selection"*:
    - `Annotation file in TSV`
    - `Annotation and sequence in GFF3`
    - `Feature nucleotide sequences as FASTA`
    - `Summary as TXT`
    - `Plot of the annotation result as SVG`

Fig. 1 Galaxy workflow code

VI. RESULTS AND DISCUSSION

The circular genome of *Escherichia coli* is approximately 4.6 million base pairs long, organized as a single DNA molecule. This structure facilitates efficient replication, with the origin of replication (oriC) and terminus (ter) marking the start and end of DNA synthesis. The genome is compacted into a nucleoid through supercoiling and protein binding. Circular genome plots visually represent features like GC skew, gene locations, and functional elements. This organization enhances genetic stability and regulation, making *E. coli* an essential model organism for studying prokaryotic genetics and cellular processes.

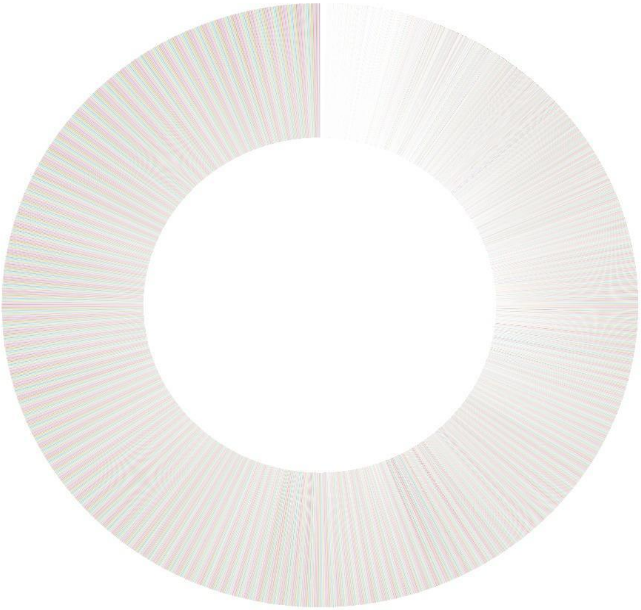


Fig. 2 Circular Genome Plot

The Bakta circular genome plot for *Escherichia coli* provides a comprehensive visualization of its genomic features. This plot typically displays the circular chromosome's structure, highlighting key elements such as the origin of replication (oriC), terminus (ter), and various functional genes. Additionally, it often includes GC skew and G+C content, which are essential for understanding replication dynamics. The Bakta tool enhances the analysis of genomic data by allowing researchers to visualize gene locations, regulatory elements, and other genomic annotations in a user-friendly format, facilitating insights into the organization and function of the *E. coli* genome.

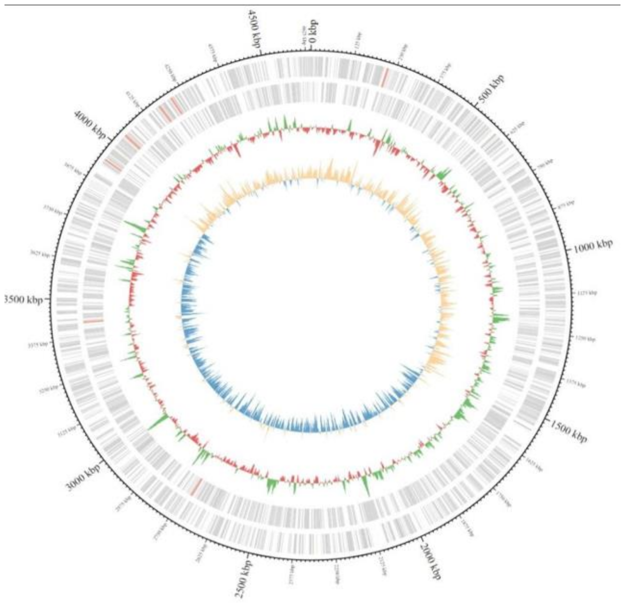
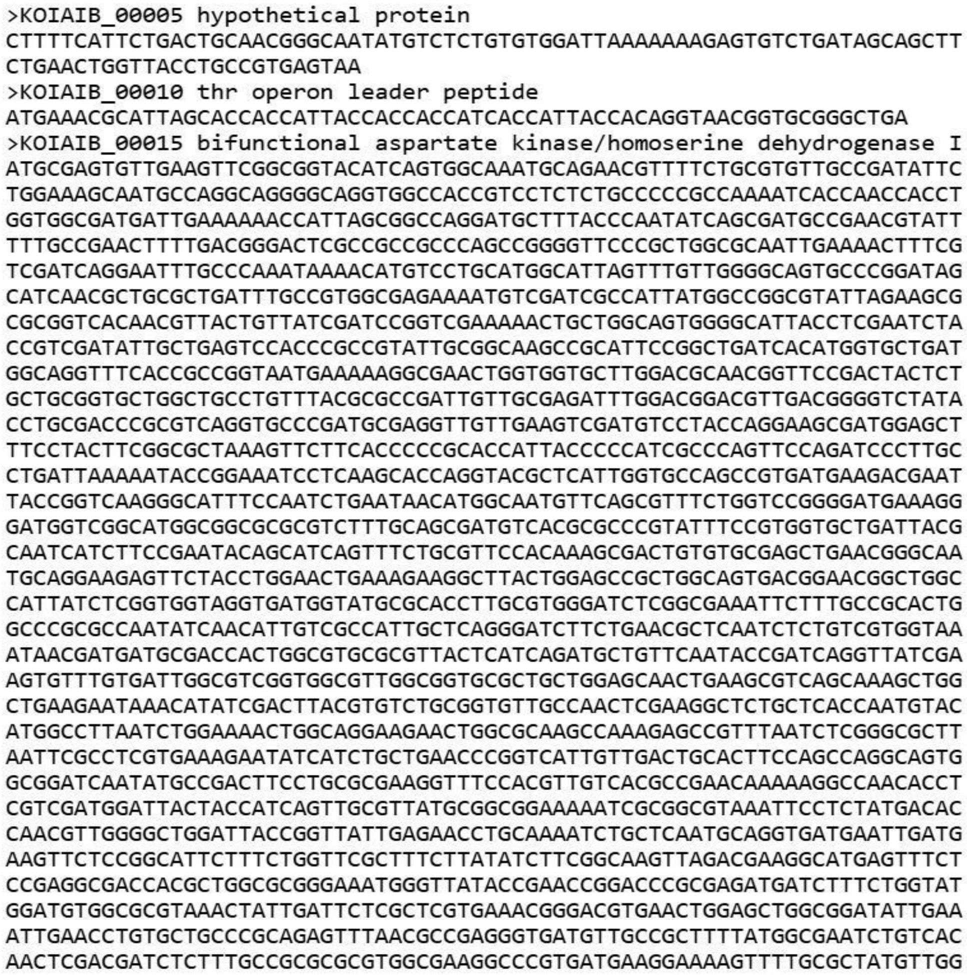


Fig. 3 Circular Bakta Plot of the Escherichia coli K-12 genome

The annotated genome sequence of *Escherichia coli* includes comprehensive details about protein-coding and non-coding genes, splice variants, and RNA sequences. For instance, the *E. coli* K-12 MG1655 strain has a circular genome of approximately 4.6 million base pairs, with 4,278 protein-coding genes and 186 RNA genes.



```
>KOIAIB_00005 hypothetical protein
CTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGCTT
CTGAACTGGTTACCTGCCGTGAGTAA
>KOIAIB_00010 thr operon leader peptide
ATGAAACGCATTAGCACCACCATTACCACCACCATCACCATTACCACAGGTAACGGTGCGGGCTGA
>KOIAIB_00015 bifunctional aspartate kinase/homoserine dehydrogenase I
ATGCGAGTGTTGAAGTTCGGCGGTACATCAGTGGCAAATGCAGAACGTTTTCTGCGTGTTGCCGATATTC
TGGAAAGCAATGCCAGGCAGGGGCAGGTGGCCACCGTCCTCTCTGCCCCCGCCAAAATCACCAACCACCT
GGTGGCGATGATTGAAAAAACCATTAGCGGCCAGGATGCTTTACCCAATATCAGCGATGCCGAACGTATT
TTTGCCGAACTTTTGACGGGACTCGCCGCCGCCCAGCCGGGGTTCCCGCTGGCGCAATTGAAAACTTTCG
TCGATCAGGAATTTGCCCAAATAAAACATGTCCTGCATGGCATTAGTTTGTTGGGGCAGTGCCCGGATAG
CATCAACGCTGCGCTGATTTGCCGTGGCGAGAAAATGTCGATCGCCATTATGGCCGGCGTATTAGAAGCG
CGCGGTCACAACGTTACTGTTATCGATCCGGTCGAAAAACTGCTGGCAGTGGGGCATTACCTCGAATCTA
CCGTCGATATTGCTGAGTCCACCCGCCGTATTGCGGCAAGCCGCATTCCGGCTGATCACATGGTGCTGAT
GGCAGGTTTCACCGCCGGTAATGAAAAAGGCGAACTGGTGGTGCTTGGACGCAACGGTTCCGACTACTCT
GCTGCGGTGCTGGCTGCCTGTTTACGCGCCGATTGTTGCGAGATTTGGACGGACGTTGACGGGGTCTATA
CCTGCGACCCGCGTCAGGTGCCCGATGCGAGGTTGTTGAAGTCGATGTCCTACCAGGAAGCGATGGAGCT
TTCCTACTTCGGCGCTAAAGTTCTTCACCCCCGCACCATTACCCCCATCGCCCAGTTCCAGATCCCTTGC
CTGATTAAAAATACCGGAAATCCTCAAGCACCAGGTACGCTCATTGGTGCCAGCCGTGATGAAGACGAAT
TACCGGTCAAGGGCATTTCCAATCTGAATAACATGGCAATGTTCAGCGTTTCTGGTCCGGGGATGAAAGG
GATGGTCGGCATGGCGGCGCGCGTCTTTGCAGCGATGTCACGCGCCCGTATTTCCGTGGTGCTGATTACG
CAATCATCTTCCGAATACAGCATCAGTTTCTGCGTTCCACAAAGCGACTGTGTGCGAGCTGAACGGGCAA
TGCAGGAAGAGTTCTACCTGGAACTGAAAGAAGGCTTACTGGAGCCGCTGGCAGTGACGGAACGGCTGGC
CATTATCTCGGTGGTAGGTGATGGTATGCGCACCTTGCGTGGGATCTCGGCGAAATTCTTTGCCGCACTG
GCCCGCGCCAATATCAACATTGTCGCCATTGCTCAGGGATCTTCTGAACGCTCAATCTCTGTCGTGGTAA
ATAACGATGATGCGACCACTGGCGTGCGCGTTACTCATCAGATGCTGTTCAATACCGATCAGGTTATCGA
AGTGTTTGTGATTGGCGTCGGTGGCGTTGGCGGTGCGCTGCTGGAGCAACTGAAGCGTCAGCAAAGCTGG
CTGAAGAATAAACATATCGACTTACGTGTCTGCGGTGTTGCCAACTCGAAGGCTCTGCTCACCAATGTAC
ATGGCCTTAATCTGGAAAACTGGCAGGAAGAACTGGCGCAAGCCAAAGAGCCGTTTAATCTCGGGCGCTT
AATTCGCCTCGTGAAAGAATATCATCTGCTGAACCCGGTCATTGTTGACTGCACTTCCAGCCAGGCAGTG
GCGGATCAATATGCCGACTTCCTGCGCGAAGGTTTCCACGTTGTCACGCCGAACAAAAAGGCCAACACCT
CGTCGATGGATTACTACCATCAGTTGCGTTATGCGGCGGAAAAATCGCGGCGTAAATTCCTCTATGACAC
CAACGTTGGGGCTGGATTACCGGTTATTGAGAACCTGCAAAATCTGCTCAATGCAGGTGATGAATTGATG
AAGTTCTCCGGCATTCTTTCTGGTTCGCTTTCTTATATCTTCGGCAAGTTAGACGAAGGCATGAGTTTCT
CCGAGGCGACCACGCTGGCGCGGGAAATGGGTTATACCGAACCGGACCCGCGAGATGATCTTTCTGGTAT
GGATGTGGCGCGTAAACTATTGATTCTCGCTCGTGAAACGGGACGTGAACTGGAGCTGGCGGATATTGAA
ATTGAACCTGTGCTGCCCGCAGAGTTTAACGCCGAGGGTGATGTTGCCGCTTTTATGGCGAATCTGTCAC
AACTCGACGATCTCTTTGCCGCGCGCGTGGCGAAGGCCCGTGATGAAGGAAAAGTTTTGCGCTATGTTGG
```

Fig. 4 Sequence of Annotated Genome

VII. CONCLUSION

Genome annotation is a critical step in understanding the genetic composition, functional elements, and evolutionary adaptations of bacterial species. In this study, the annotation of *Escherichia coli* was conducted using a combination of bioinformatics tools, including Galaxy, Prokka, Bakta, and Beav, to systematically identify coding sequences, regulatory elements, and mobile genetic components. The findings provide a comprehensive overview of the genomic organization of *E. coli*, elucidating its structural and functional characteristics.

The analysis highlights the circular genome architecture, gene distribution, and replication dynamics, contributing to a deeper understanding of *E. coli*’s metabolic pathways and regulatory networks. The integration of high-throughput sequencing data with automated annotation pipelines demonstrates the effectiveness of computational approaches in enhancing the accuracy and efficiency of genome annotation.

These results underscore the significance of genome annotation in microbial genomics, offering insights into bacterial physiology, genetic evolution, and potential applications in biotechnology and medicine. Continued advancements in annotation methodologies will further refine our ability to decode bacterial genomes, facilitating research in genetics, microbiology, and related fields.

REFERENCES

[1] Jewell M. Jung, Arafat Rahman, Andrea M. Schiffer, and Alexandra J. Weisberg, "Beav: A bacterial genome and mobile element annotation pipeline", Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR 97331, USA

[2] Ross Overbeek, Daniela Bartels, Veronika Vonstein, and Folker Meyer, "Annotation of Bacterial and Archaeal Genomes: Improving Accuracy and Consistency", Fellowship for Interpretation of Genomes, Burr Ridge, Illinois 60527, The Computation Institute, University of Chicago, Chicago, Illinois 60637, and Mathematics and Computer Science, Argonne National Laboratory, Argonne, Illinois 60439

[3] Carlos A. Ruiz-Perez, Roth E. Conrad and Konstantinos T. Konstantinidis, "MicrobeAnnotator: a user-friendly, comprehensive functional annotation pipeline for microbial genomes", School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, 30332, USA, Ocean Science and Engineering, School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, 30332, USA, School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, 30332, USA. kostas.konstantinidis@gatech.edu, Ocean Science and Engineering, School of Biological Sciences, Georgia Institute of Technology, Atlanta, GA, 30332, USA. kostas.konstantinidis@gatech.edu, School of Civil and Environmental Engineering, Georgia Institute of Technology, Atlanta, GA, 30332, USA. kostas.konstantinidis@gatech.edu, Center for Bioinformatics and Computational Genomics, Georgia Institute of Technology, Atlanta, GA, 30332, USA. kostas.konstantinidis@gatech.edu

[4] Briallen Lobb, Benjamin Jean-Marie Tremblay, Gabriel Moreno-Hagelsieb and Andrew C. Doxey, "An assessment of genome annotation coverage across the bacterial tree of life"

[5] Tatiana Tatusova, Michael DiCuccio, Azat Badretdin, Vyacheslav Chetvernin, Eric P Nawrocki, Leonid Zaslavsky, Alexandre Lomsadze, Kim D Pruitt, Mark Borodovsky, James Ostell, "NCBI prokaryotic genome annotation pipeline", National Center for Biotechnology Information, U.S. National Library of Medicine, Bethesda, MD 20894, USA, Wallace H. Coulter Department of Biomedical Engineering, Georgia Tech, Atlanta, GA 30332, USA, Wallace H. Coulter Department of Biomedical Engineering, Georgia Tech, Atlanta, GA 30332, USA School of Computational Science and Engineering, Georgia Tech, Atlanta, GA 30332, USA borodovsky@gatech.edu.

[6] Sinyeon Kim, Haeyoung Jeong, Eun-Youn Kim, Jihyun F Kim, Sang Yup Lee, Sung Ho Yoon, "Genomic and transcriptomic landscape of Escherichia coli BL21(DE3)", Department of Bioscience and Biotechnology, Konkuk University, Seoul 05029, Republic of Korea, Infectious Disease Research Center, Korea Research Institute of Bioscience and Biotechnology (KRIBB), Daejeon 34141, Republic of Korea, School of Basic Sciences, Hanbat National University, Daejeon 34158, Republic of Korea, Department of Systems Biology and Division of Life Sciences, Yonsei University, Seoul 03722, Republic of Korea, Metabolic and Biomolecular Engineering National Research Laboratory, Department of Chemical and Biomolecular Engineering (BK21 Plus Program), BioProcess Engineering Research Center, Center for Systems and Synthetic Biotechnology, and Institute for the BioCentury, KAIST, Daejeon 34141, Republic of Korea

[7] Torsten Seemann, "Prokka: rapid prokaryotic genome annotation", Victorian Bioinformatics Consortium, Monash University, Clayton 3800 and Life Sciences Computation Centre, Victorian Life Sciences Computation Initiative, Carlton 3053, AustraliaVictorian Bioinformatics Consortium,

Monash University, Clayton 3800 and Life Sciences Computation Centre, Victorian Life Sciences Computation Initiative, Carlton 3053, Australia.

[8] Nicholas Beckloff, Shawn Starkenburg, Tracey Freitas, Patrick Chain, “Bacterial genome annotation”, Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM, USA

[9] Briallen Lobb, Benjamin Jean-Marie Tremblay, Gabriel Moreno-Hagelsieb, Andrew C Doxey, “An assessment of genome annotation coverage across the bacterial tree of life”, Department of Biology, University of Waterloo, 200 University Avenue West, Waterloo, ON N2L 3G1, Canada, Department of Biology, Wilfrid Laurier University, 75 University Avenue West, Waterloo, ON, Canada

[10] American Academy of Microbiology Colloquia Reports, “An Experimental Approach to Genome Annotation”, This report is based on a colloquium sponsored by the American Academy of Microbiology held July 19-20, 2004, in Washington, DC

[11] Felipe Marques de Almeida, Tatiana Amabile de Campos, Georgios Joannis Pappas Jr, “Scalable and versatile container-based pipelines for de novo genome assembly and bacterial annotation”, Programa de Pós-graduação em Biologia Molecular, Universidade de Brasília, Brasília, FD, 70910-900, Brazil, Departamento de Biologia Celular, Universidade de Brasília, Brasília, DF, 70910-900, Brazil, Programa de Pós-graduação em Biologia Microbiana, Universidade de Brasília, Brasília, DF, 70910-900, Brazil

[12] Yibo Dong, Chang Li, Kami Kim, Liwang Cui, Xiaoming Liu, “Genome annotation of disease-causing microorganisms”, College of Public Health, University of South Florida, Tampa, FL, USA, Division of Infectious Disease and International Medicine, Department of Internal Medicine, Morsani College of Medicine, University of South Florida, Tampa, FL, USA

[13] Ryan R Wick, “Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads”, PLoS Comput Biol