Metagenomic Analysis and Development of Comprehensive Gene Sequences for Bacteria across Public Genomic Databases

Diya Sreedhar[1,2], Nitesh Sharma[1], Serghei Mangul[1]

*1 - Titus Family Department of Clinical Pharmacy, **USC** Alfred E. Mann School of Pharmacy and Pharmaceutical Sciences, Bridge Institute, University of Southern California, Los Angeles, CA , USA*

2 – Troy High School, Fullerton, CA, USA

Bridge UnderGrad Science (BUGS) ***Summer Research Program***

Abstract

Metagenomics is the study of the structure and function of entire nucleotide sequences, typically from a specific community of microorganisms such as bacteria. The most popular bacterial genome datasets available for research and analysis are – the **Bacterial and Viral Bioinformatics Resource Center (BV-BRC)**, the **National Center for Biotechnology Information Reference Sequence Database (NCBI RefSeq)**, and **Ensembl Genomes**.

A key challenge in sourcing data from these public databases is that not all bacterial species are commonly available across all the datasets, and the gene sequencing data for common bacteria may not be of the same completeness and quality.

In this study, we conducted a novel, deep analysis of the genome databases, identified the common bacteria strains available in all the databases, and extracted the most complete nucleotide sequences across all the databases for the common bacteria strains.

To the best of our knowledge, no prior research has ventured into this domain, making our findings truly pioneering. We present a unified, meticulously curated, and thoroughly annotated amalgamation of the three databases, offering researchers an unparalleled resource of the highest data quality for investigating bacterial strains. The potential benefits of integrating these datasets can facilitate cross-species comparison, support drug discovery efforts and enable the discovery of new therapeutic targets and biomarkers.

Future work will focus on developing an automated mechanism to maintain the completeness and accuracy of these sequences, ensuring a sustained level of data quality over time. Careful consideration of data compatibility and mapping of gene identifiers will be crucial to ensuring the reliability of the resulting unified dataset.

Objective/Hypothesis

Is the integration of data from RefSeq, Patric, and Ensembl feasible, enabling the creation of a comprehensive and unified genomic resource for enhanced analysis and insights into gene annotations and functional characteristics?

Methods

Patric Dataset Analysis:

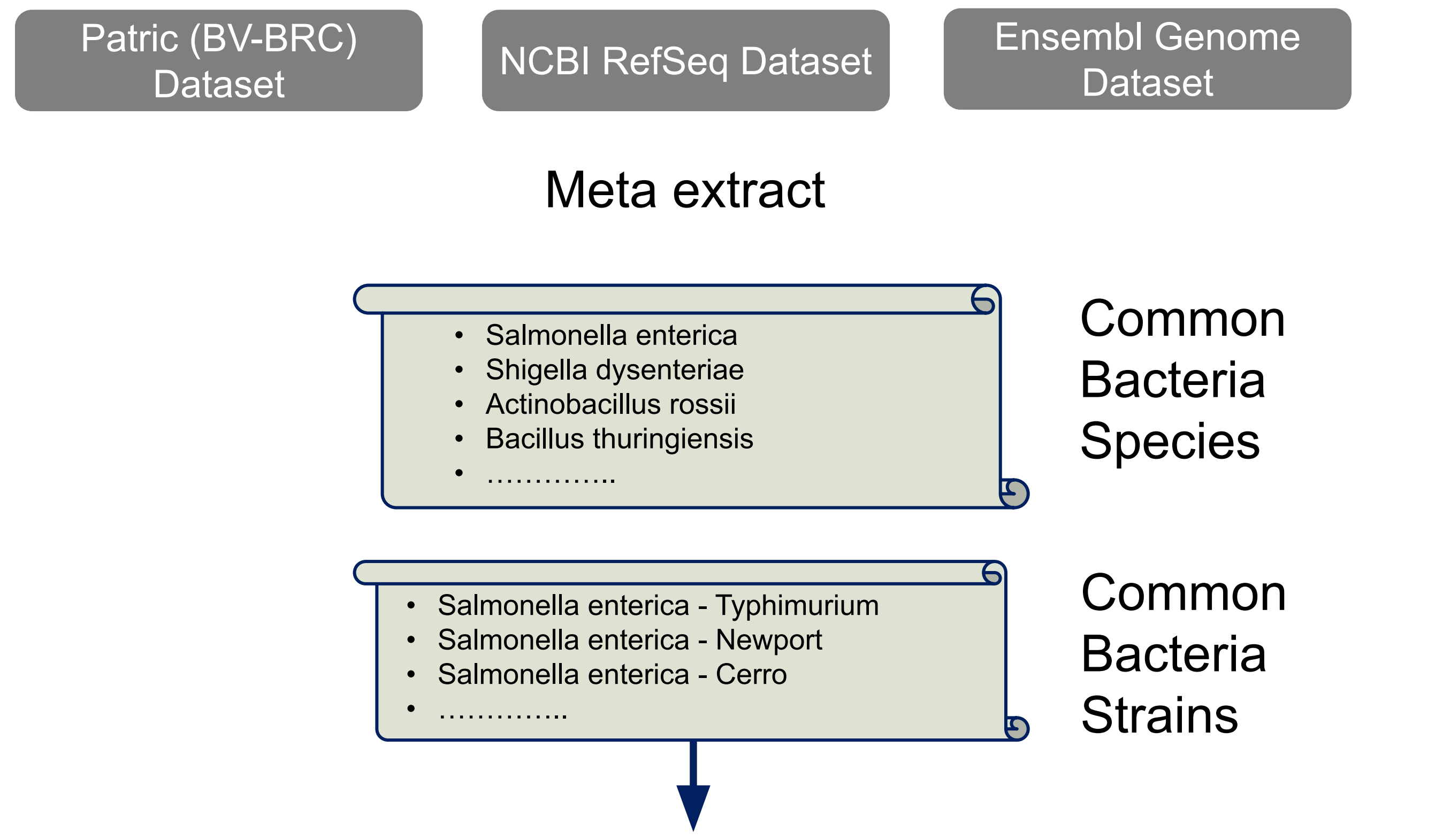
- Extract bacteria names from the genome_lineage file using Python Pandas.
- Read each FASTA file and extract genome record using Biopython package.
- Save genome record in target file for comparative analysis.

RefSeq Dataset Analysis:

- Each bacteria has multiple files, and each file contains multiple “records” for the bacteria, with its description
- Use Gzip and Biopython SeqIO to open each FASTA file. Extract the first record and its description. Parse the bacteria name from the description line.
- Save bacteria record in a target file for comparative analysis.

Ensembl Dataset Analysis:

- The Ensembl dataset has 184 “collections” of bacteria (bacteria_0_collection, bacteria_1_collection …. bacteria_183_collection)
- Use Python to list the collection folders and the subfolder names. The subfolder names are the bacteria names.



Patric (BV-BRC) Dataset

NCBI RefSeq Dataset

Ensembl Genome Dataset

Meta extract

- Salmonella enterica
- Shigella dysenteriae
- Actinobacillus rossii
- Bacillus thuringiensis
- …………..

Common Bacteria Species

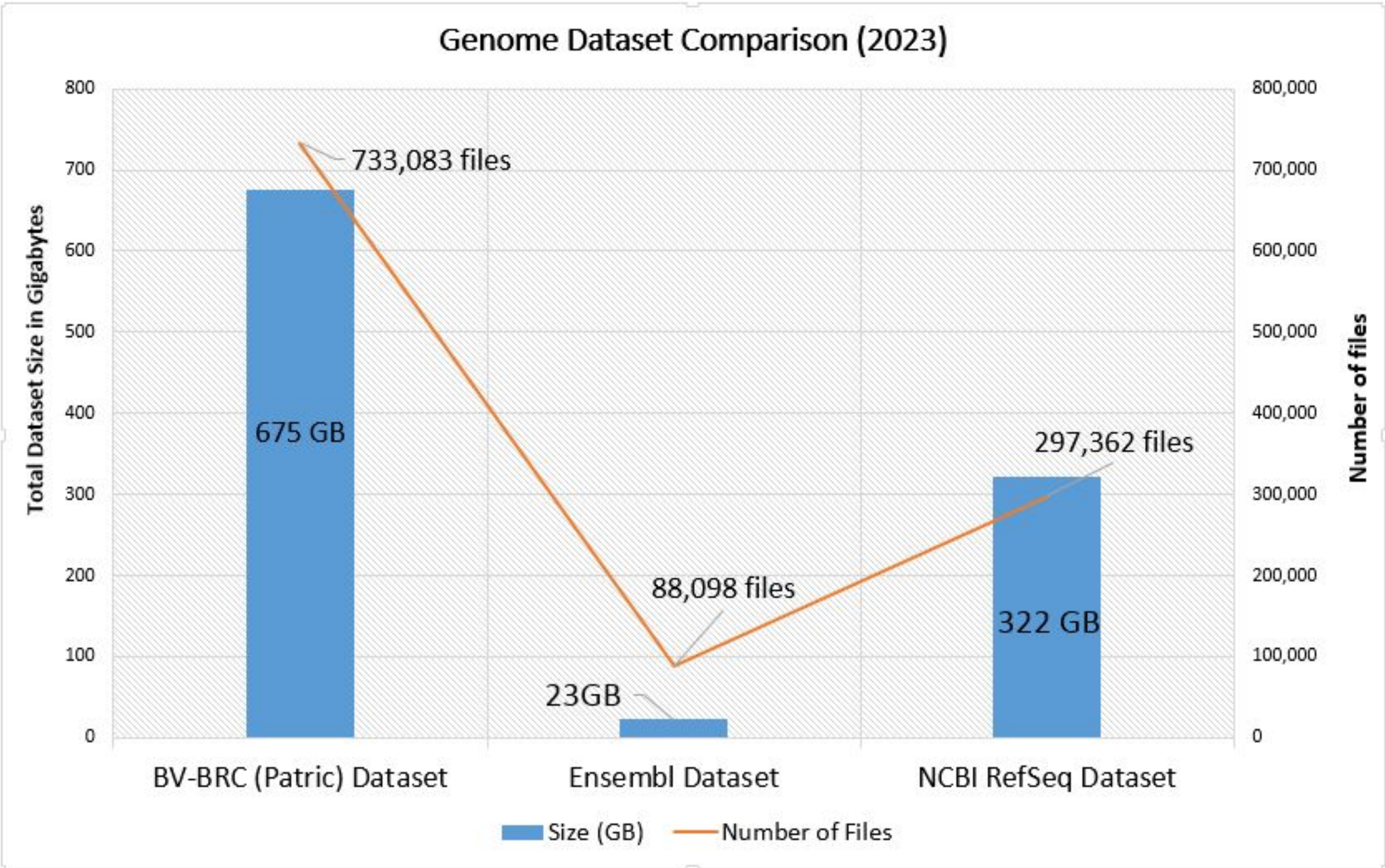
- Salmonella enterica - Typhimurium
- Salmonella enterica - Newport
- Salmonella enterica - Cerro
- …………..

Common Bacteria Strains

```
N31402_contig_10 dna:supercontig
supercontig:Salmonella_enterica_CVM_N31402-SQ_v1.0:N31402_contig_10:1:568462:1 REF
Seq('GTATCGCTCTGCCACTGCCTGCAACGCGACAATAACGGGTTCGTCACGCGTGGT...CCG')
568462
```

Genome Sequence

Dataset Characteristics



Data Preprocessing and Main Record Selection

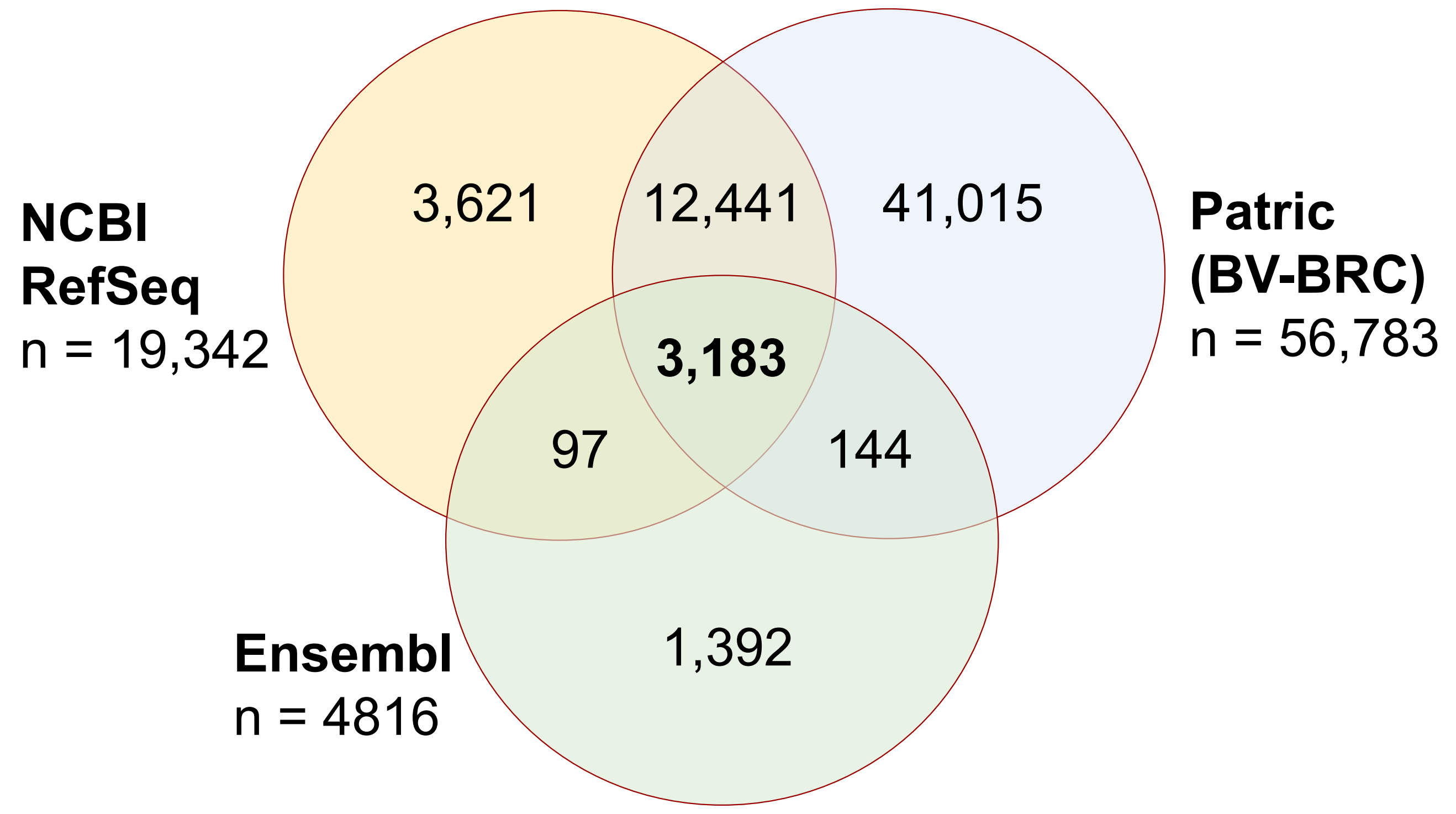
Patric (BV-BRC)	**Ensembl Genomes**	**NCBI-RefSeq**
Salmonella enterica Str. Senftenberg	Salmonella enterica Str. Seftenberg	Salmonella enterica Str. Senftenberg
Sub-strain: CVM_N31402	Sub-strain: CVM_N31402	Sub-strain: CVM_N31402
Contig chunks: 58	Contig chunks: 76	**Contig chunks: 92**

We pick the most up to date and complete genome sequence

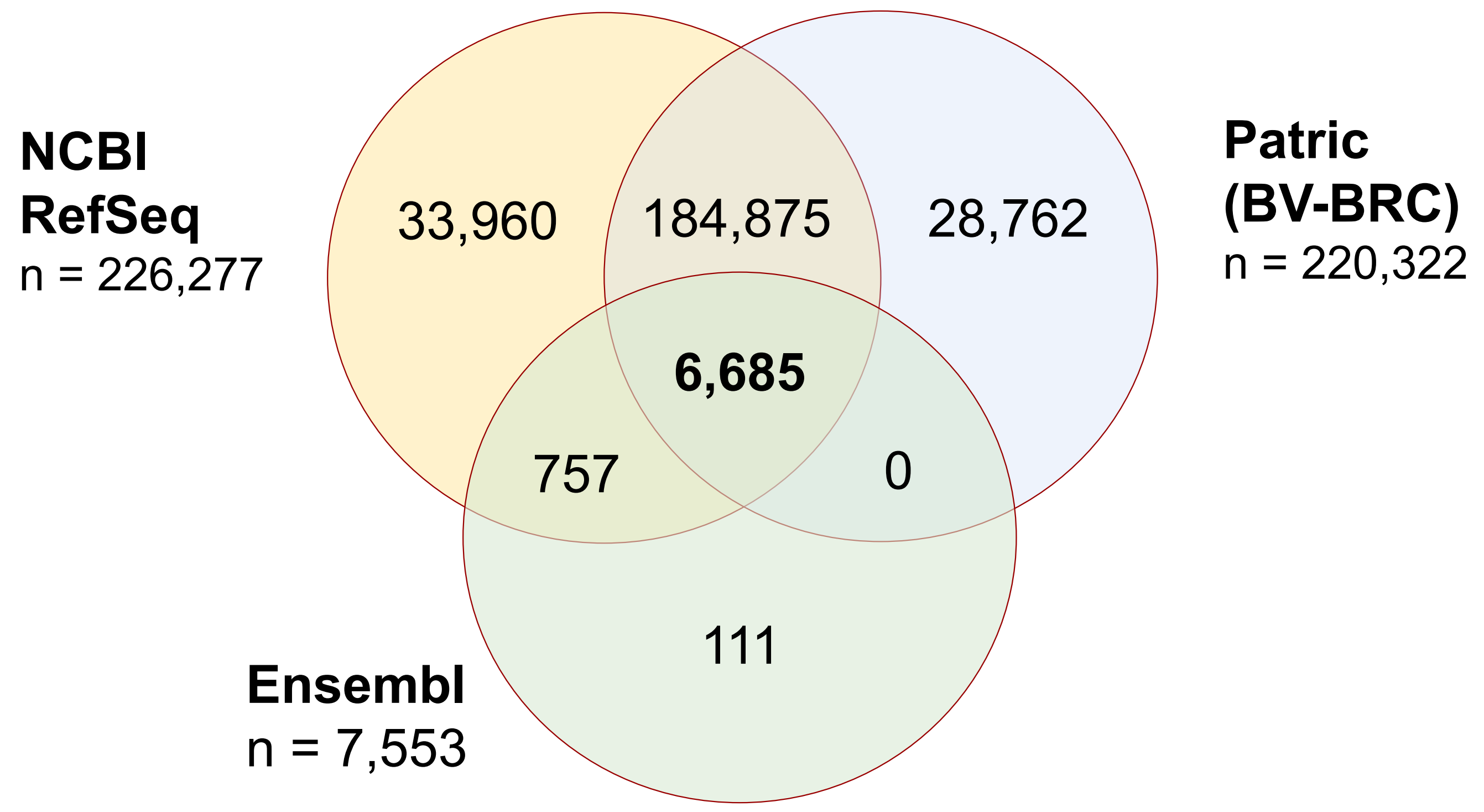
```
N31402_contig_01 dna:supercontig
supercontig:Salmonella_enterica_CVM_N31402-SQ_v1.0:N31402_contig_01:1:568462:1 REF
Seq('GTATCGCTCTGCCACTGCCTGCAACGCGACAATAACGGGTTCGTCACGCGTGGT...CCG')
Sequence Length: 568462

….
…..
…..

N31402_contig_92 dna:supercontig
supercontig:Salmonella_enterica_CVM_N31402-SQ_v1.0:N31402_contig_92:1:44589:1 REF
Seq('CGCCCCAGGTTGATTTGCTGGAGATGGCCGCTCCCGTAGTACAGGTACCGCAGC...CAT')
Sequence Length: 44589
```

Results

Bacterial species that are common across the three datasets



Bacterial sub-species, serovars, strains and chromosomes for the 3,183 common bacterial species



Summary

NCBI Refseq provides a curated, stable reference for genome annotation, gene identification, gene characterization, expression studies, and comparative analyses.

Patric focuses on bacterial genomic data and includes unidentified, uncultured and candidate species, leading to a larger dataset size

Ensembl includes a much smaller subset of bacteria compared to the other two datasets as it is focused on eukaryotic genomes and less focused on bacteria.

All three databases store information in different formats and conventions. A significant amount of effort and advanced Python coding was spent in normalizing the datasets for comparative analysis.

This study shows that there is a potential for a consolidated database comprised of RefSeq, Patric and Ensembl datasets for bacterial strains. This presents a promising avenue for researchers to leverage a more extensive and integrated genomic dataset, that could lead to valuable discoveries and a deeper understanding of the complexities of gene functions and biological processes. The availability of such a unified resource can expedite research and advance studies related to genomics, functional genomics, comparative genomics and systems biology.

Acknowledgements

Thank you to my mentor Mr. Nitesh Sharma for his guidance throughout the project and Dr. Serghei Mangul for the amazing opportunity to conduct research in his lab. My sincere appreciation to USC BUGS Jr. program team and Angie Walker for this research opportunity.