ZygosDB: An efficient read-only database for Genome-Wide Association Studies (GWAS)

Author: Nick van Luijk[1] Supervisor: Niccolo Tesi[1] Responsible Professor: Marcel Reinders[1]
[1]TU Delft N.vanLuijk@student.tudelft.nl, N.Tesi@tudelft.nl, M.J.T.Reinders@tudelft.nl

1 INTRODUCTION & OBJECTIVE

- **Genome-Wide Association Studies (GWAS)**: Identify associations between genetic variants and traits.
- **snpXplorer**: Website for visualising SNPs (single-base mutations) and SVs (multi-base mutations) in GWAS summaries.
- **Tabix**: Popular genomic indexing tool, currently in use by snpXplorer.

- **Challenge**: Querying genomic datasets is slow.
- **Solution**: Developed a custom specialized read-only database optimized for query throughput.

How can an efficient read-only database be designed and implemented for querying chromosomal, positional data?

2 METHOD

1. **Implement database in Rust**
   - Custom binary format.
   - Store all data sequentially.
   - Encode positions using variable-length integers.
   - Indices using B-trees for fast lookup.
2. **Optimise query throughput**
   - Compress database using Gzip or LZ4.
   - Resolve queries in parallel using multiple threads.
3. **Measure performance**

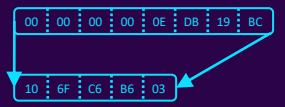

Figure 1: Variable-length integer encoding

Figure 2: An example B-tree.

3 RESULTS

1

When returning small numbers of results, ZygosDB with Gzip compression is the fastest. As the window size increases, query throughput increases as well, with a not compressed ZygosDB database returning **up to 7 million rows per second**.

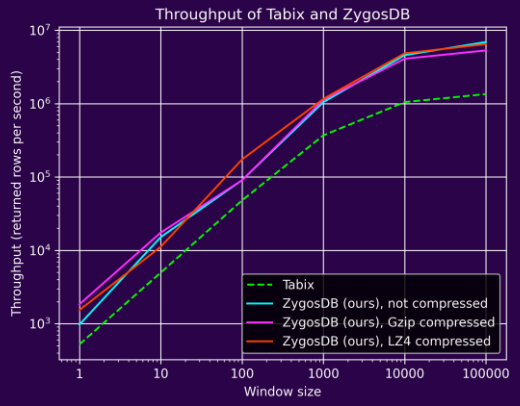

Figure 3: The query throughput, measured in the number of rows returned per window size, for both Tabix and our database.

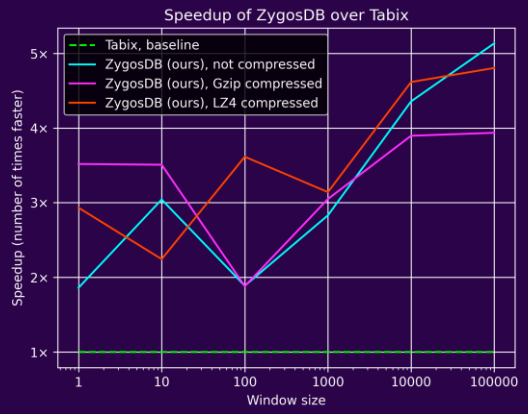

Figure 4: The speedup of ZygosDB over Tabix.

2

In a **worst-case scenario**, ZygosDB is approximately **2 times faster** than Tabix. By increasing the window size, ZygosDB can reach throughputs that are **over 5 times faster** than Tabix.

3

The maximum query thoughput is reached when using **3-5 threads**. Using **more threads decreases performance**, eventually reaching throughputs lower than when using a single thread.

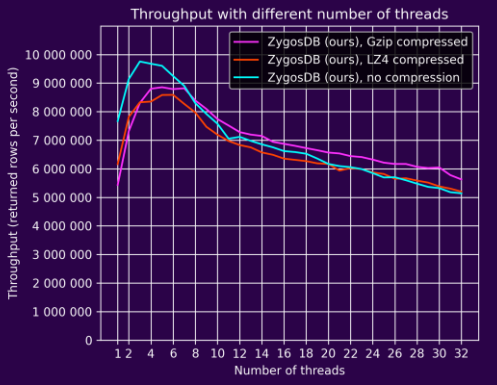

Figure 5: The multi-threaded query throughput with a window size of100 000, measured with thread counts of 1-32.

4 CONCLUSION & FUTURE WORK

ZygosDB, our read-only database, specialised for storing and querying genomic datasets, has a **2-5 times higher query throughput** than Tabix.

Optimise further using:
- Profiling
- Specialised column types

Available at: https://github.com/TechnologicNick/zygos_db