Phages and DuckDB: The mighty duo against the giants

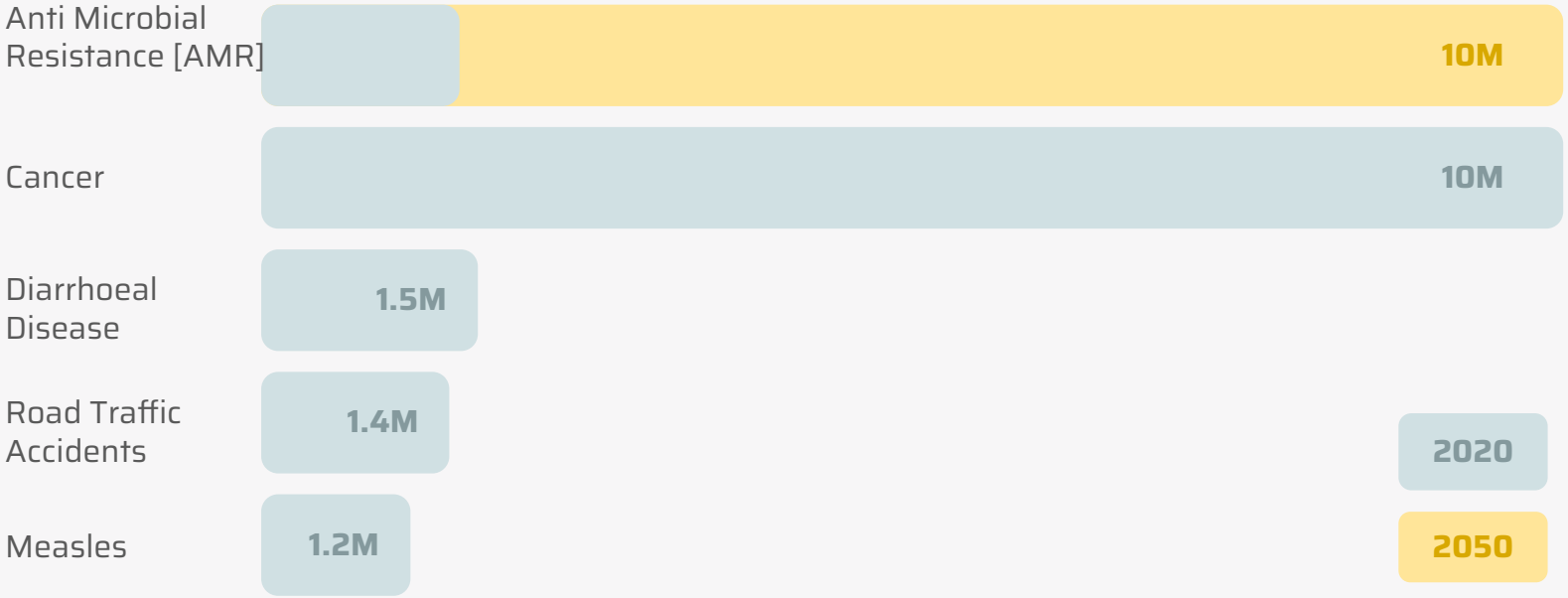
Virginie Grosboillot
University of Ljubljana

Herminio Vazquez
ETH Zurich

DuckDB Conference #7 | Amsterdam, Netherlands | June 24, 2026

Risks

Predicted mortality from antimicrobial resistant infections versus today's common causes of death



Source: https://www.statista.com/chart/3095/drug-resistant-infections/

Phages

Most abundant organism in the planet

Are the virus of of bacteria and its natural predator.

Coolest head out of a sci-fi movie

Phage-therapy presents an alternative to traditional antibiotic treatments

Bioinformatics Challenges

Lab Management

Experiment Design

Protocol Effectiveness

Working with Living Things

Reproducibility

Configuration Management

Statistical Significance

Valuable Inferences

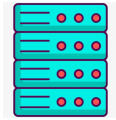
Data Management

Manual Data Entry

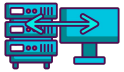
Data Quality

Data Storage

File Formats



Big Data



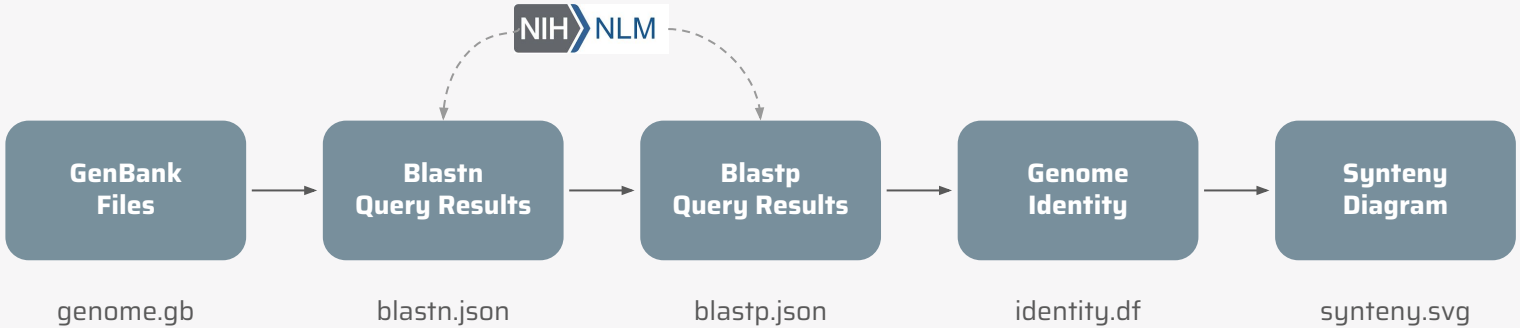
Local and HPC

Analysis at Scale

Data Challenges

Synphage

Phage genome analytics



DuckDB

genome

blastn

blastp

identity

synteny

Solution Space

Id	Class	Bioinformatics challenge	Duckdb solution	Gain
1	Read	File-based workloads	Wild card readers	From 10 minutes to 12 seconds
2	Compute	Python Loops	Lateral joins	From 20 minutes to 1 second
3	Parse	Deeply nested json	Schema inference and sampling	Remove HITL
4	Compute	Large string gene sequences	Larger than memory compute	Remove Cluster and/or JVM
5	Compute	Wide dataframes and column gymnastics	Macros and column expressions	Simplicity and readability

```sql
WITH blast AS (
   SELECT unnest(BlastOutput2) AS item FROM read_json_auto('*')
),
search AS (
   SELECT
       item->>'$.report.program' AS program,
       item->>'$.report.version' AS version,
       item->>'$.report.reference' AS reference,
       item->>'$.report.search_target' AS search_target,
       item->>'$.report.results.search' AS search,
       item->'$.report.results.search.hits' AS hits
   FROM blast
   WHERE json_array_length(hits) > 0
)
SELECT
   program, version, reference, search_target,
   search->>'$.query_id' AS query_id,
   search->>'$.query_title' AS query_title,
   search->>'$.query_len' AS query_len,
   lat.*,
   round((identity/align_len)*100,3) as percentage_of_identity
FROM search s,
LATERAL (SELECT * FROM blast_top_hit(s.hits)) lat;
```

1. wildcards
2. no loops
3. no human in the loop
4. large strings
5. macros

Solution Space

Id	Class	Bioinformatics challenge	Duckdb solution	Gain
6	Analysis	GC content in genome	Histogram	Prevent context switching
7	Scalability	Volume of sequences to compare	A folder is a database	File/Table format support
8	Portability	Reproducible science	Single binary	Cross-platform compatibility
9	Performance	Time scarcity	Lean and fast engine	Quality of life and user experience
10	Community	Support decay	Release cadence and extensions	Continuity

```
WITH base AS (
       SELECT UNNEST(seq) as nucleotide
       FROM (SELECT STR_SPLIT(cds_extract,'') as seq FROM cds LIMIT 1)
)
FROM HISTOGRAM(base, nucleotide);
```

6 histogram

bin varchar	count uint64	bar varchar
A	63	
C	28	
G	35	
T	51	

```
WITH blastn AS (
      SELECT * exclude(query_id, query_strand) from (
            SELECT COLUMNS('^(query|source)_.*$'), percentage_of_identity
            FROM read_parquet('synphage.parquet')
            )
     ),
     locus AS (SELECT * FROM read_parquet('locus_*.parquet')),
     query_side AS (SELECT COLUMNS('(\w*)') AS 'query_\1' FROM locus),
     source_side AS (SELECT COLUMNS('(\w*)') AS 'source_\1' FROM locus),
     joined_locus AS (
            SELECT B.*, A.* EXCLUDE(query_key) FROM query_side B
            LEFT JOIN blastn A USING(query_key)
     )
SELECT C.*, D.* exclude(source_key) FROM joined_locus C
LEFT JOIN source_side D
USING(source_key)
```

5 column expression

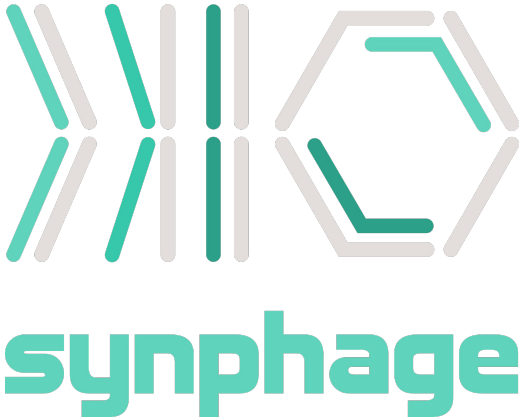
7 scalable and fast analytics

9 query performance

Every minute you save waiting for an analytical workload is a minute closer to **a publication or a discovery.**

Is a minute between you **and your next grant**

Project



Tell us about your DuckDb experience in bioinformatics and grab a sticker!

Virginie Grosboillot, Anna Dragoš,
Bioinformatics Advances, 2024
https://doi.org/10.1093/bioadv/vbae126

Acknowledgements

https://github.com/vestalisvirginis/synphage

BF UNIVERSITY OF LJUBLJANA
Biotechnical Faculty

European Research Council
Established by the European Commission

PHAGECONTROL
Dragoš Lab

DuckLabs