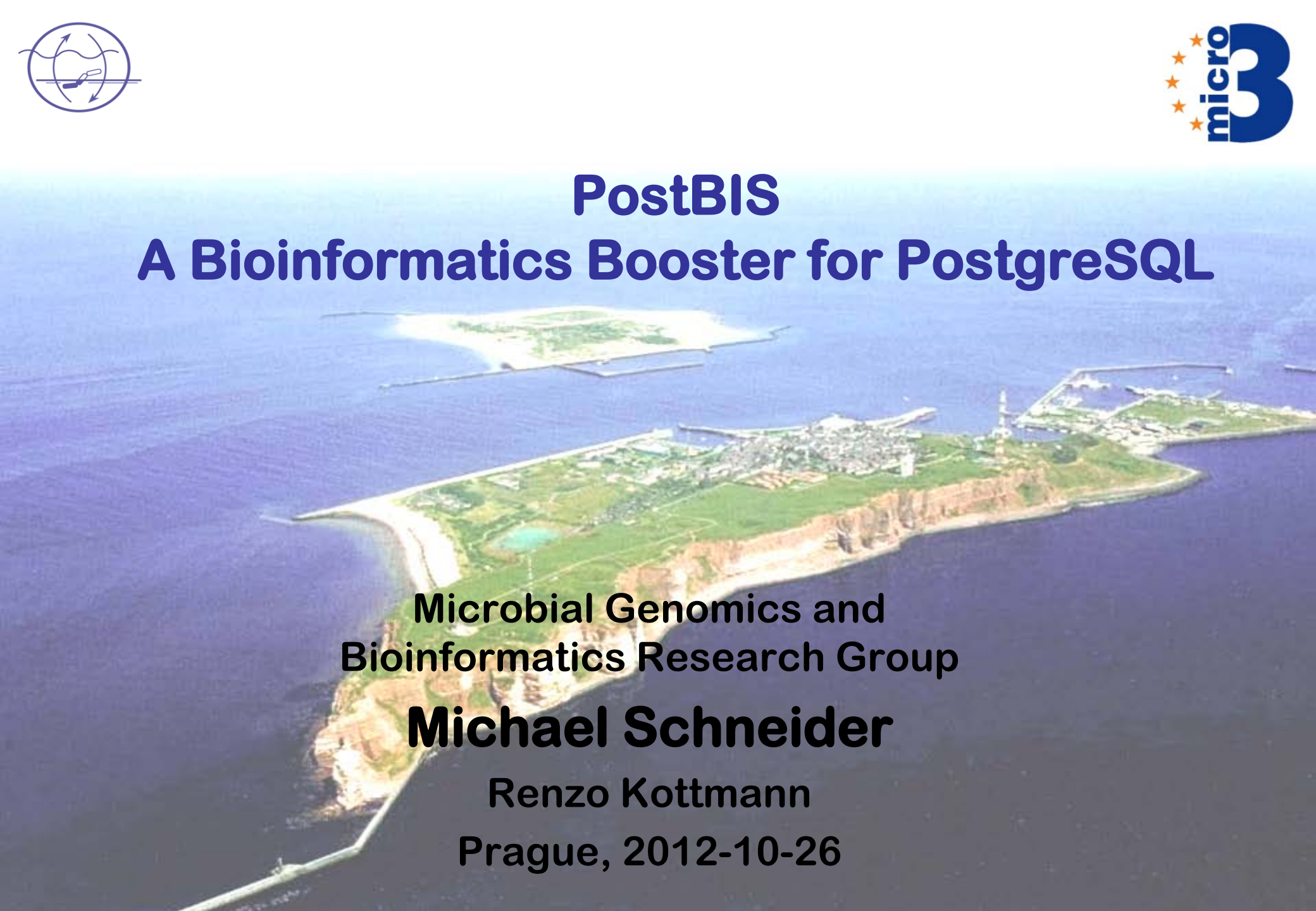


PostBIS
A Bioinformatics Booster for PostgreSQL

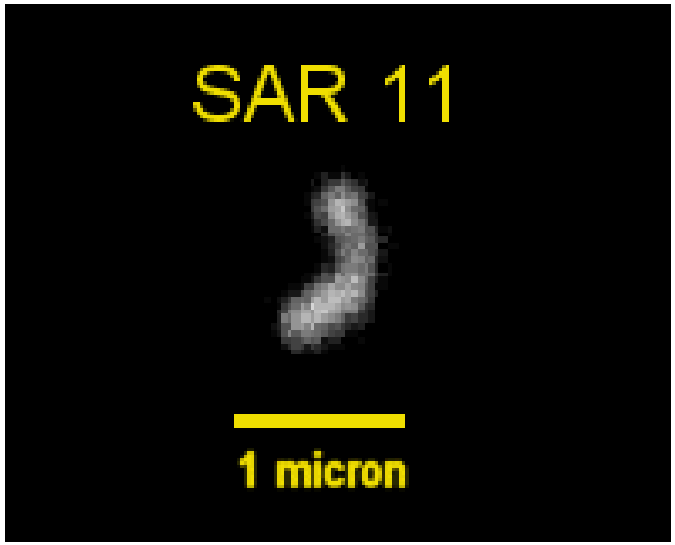
Microbial Genomics and
Bioinformatics Research Group

Michael Schneider

Renzo Kottmann

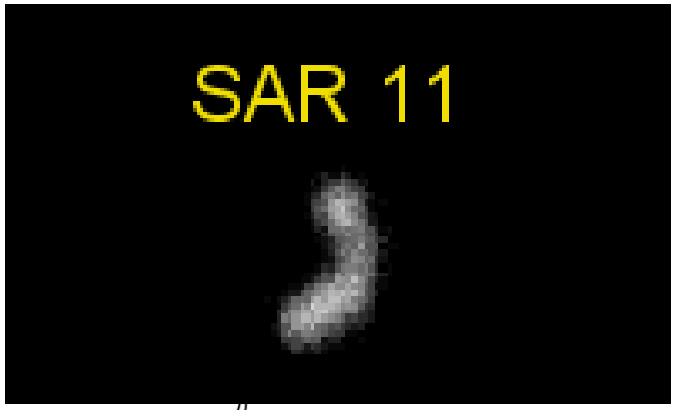
Prague, 2012-10-26

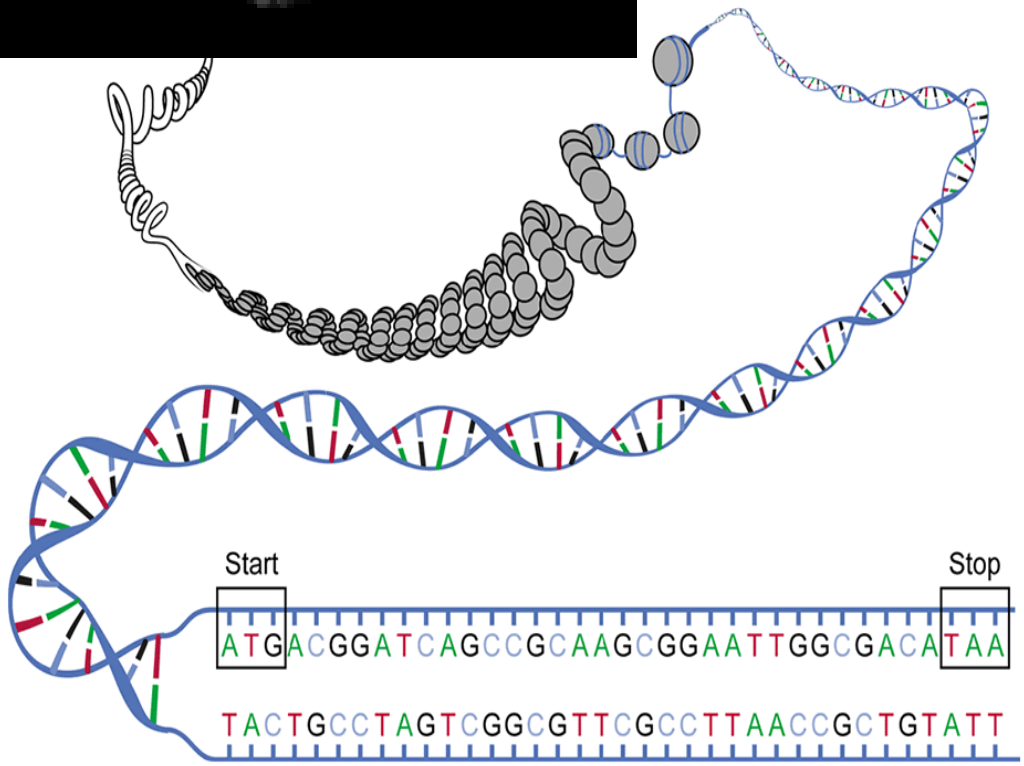
Marine Microbiologie – Ecologically Important



- 1 million bacterial cells/cm^3 ocean water
- In total 10^{30}
  - More than stars in universe
- ½ of the world wide oxygen production
- ½ of the earth biomass
- The weight of
  > 240 billion elephants

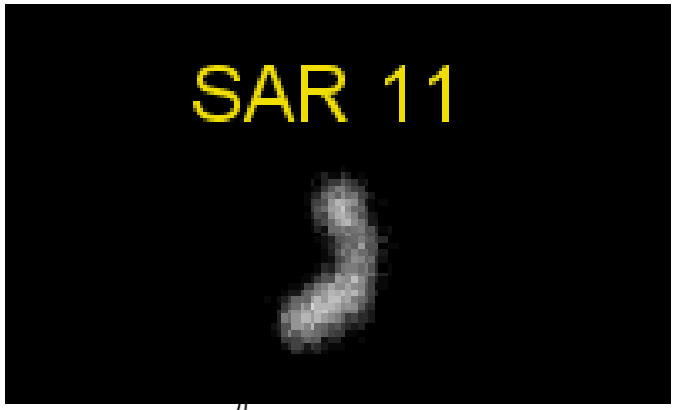
Single Bacterial Genomes





- All heredity is encoded in the genomes of cells
- Sequencing of thousands of genomes:
  - Each ~ 5MB

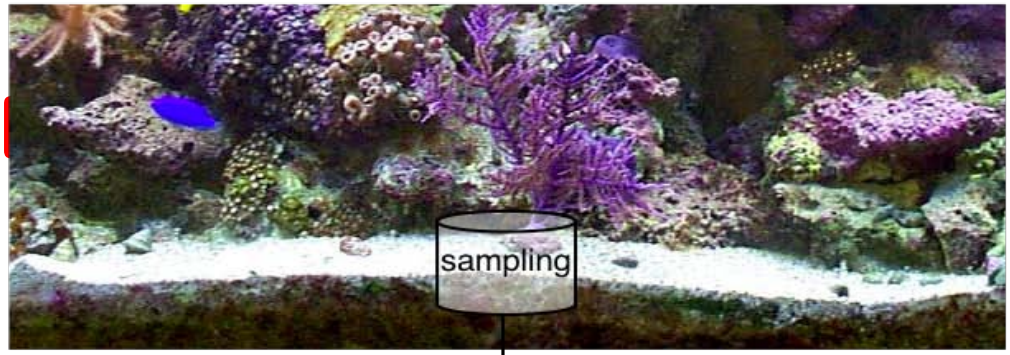
Single Bacterial Annotation



- Each gene sequence needs analysis
  - Which sequences are similar to current one?
  - What is the function?

Gene:Oxy1

Metagenomics



culturable microorganisms (~1%)

not yet cultured microorganisms (~99%)

isolation, cultivation

DNA-extraction

genomes

DNA-extraction, shotgun-sequencing

fragmentation

whole genome sequence

sequencing

cloning

insertion into vectors

genomic fragments

sequence of a genomic fragment (~1/100 of whole genome)

cultivation on plates

vectors carrying inserts

The Sequence Data

Single genomes

- Sequencing of thousands of genomes each:
  - 1 long sequence
  - ~ 5MB
  - ~ 5000 genes/genome

Metagenomes

- Sequencing of thousands of sample each:
  - Millions of short sequences
  - < 1 KB
  - Millions of genes/ metagenome

Standard bioinformatic query

Give me all sequences which encode gene OXY1

Ecological Perspective

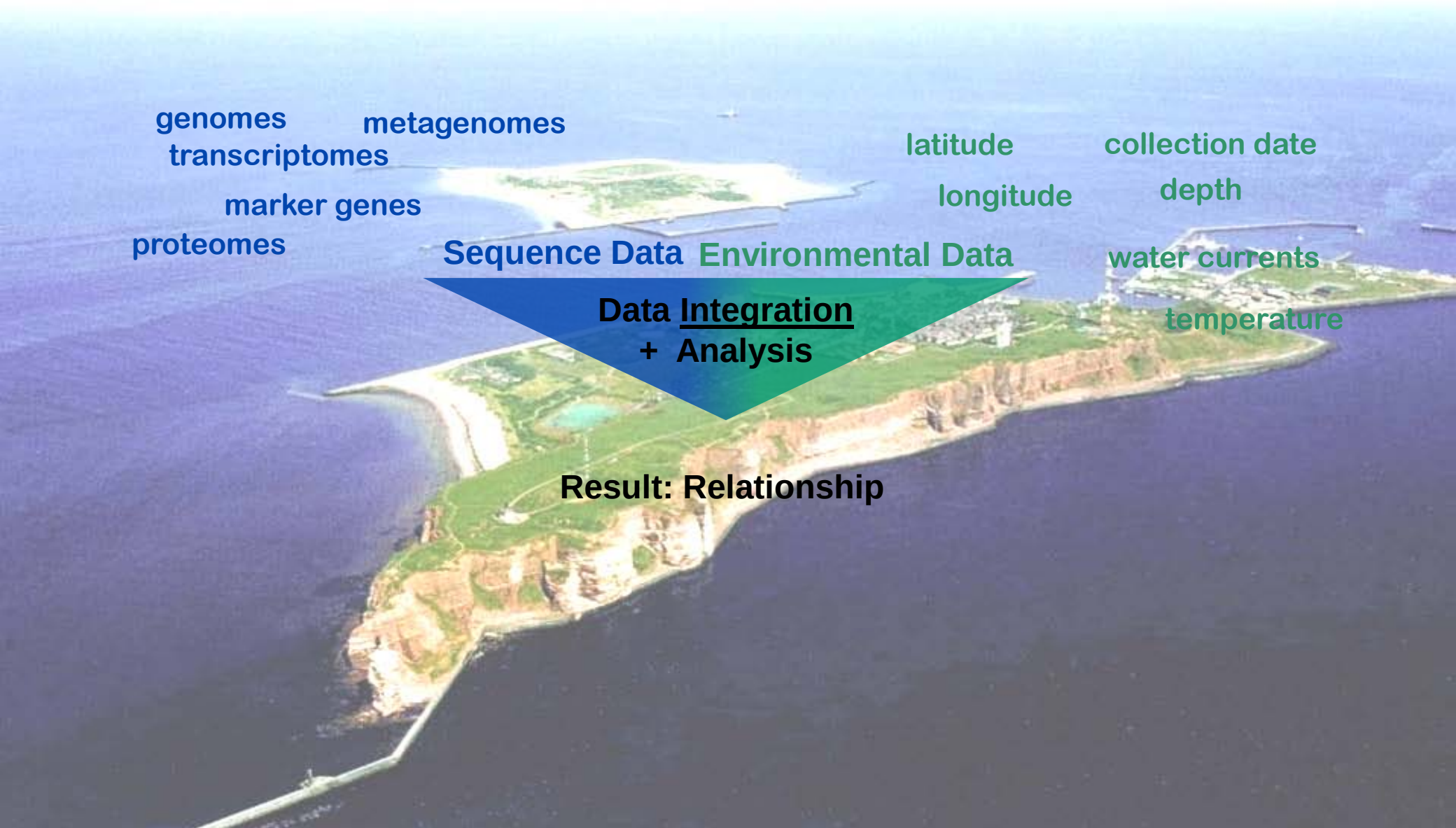
From where ??

Ecological Perspective

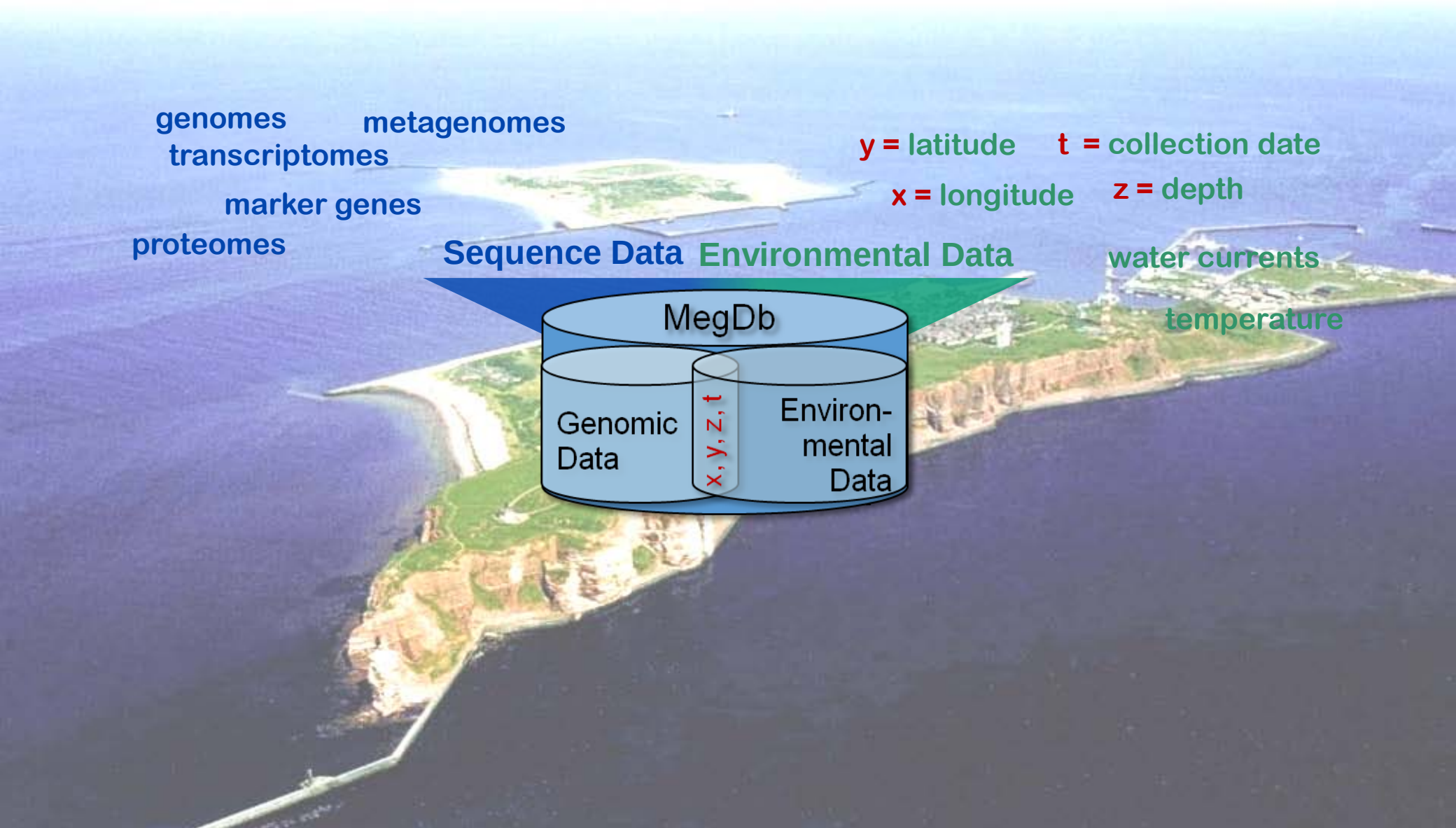
From where ??

Give me all sequences which encode gene OXY1 and were found at Helgoland roads at a depth deeper 50 m.

Data Integration

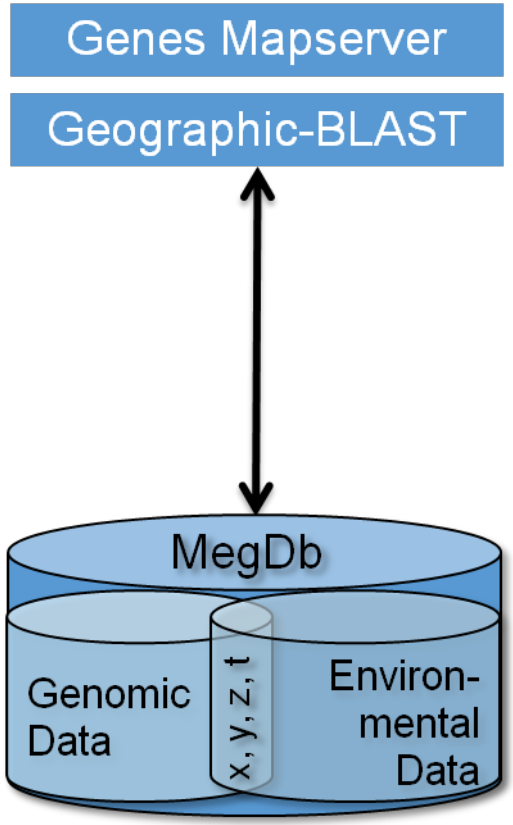


Data Integration: Geo-referencing



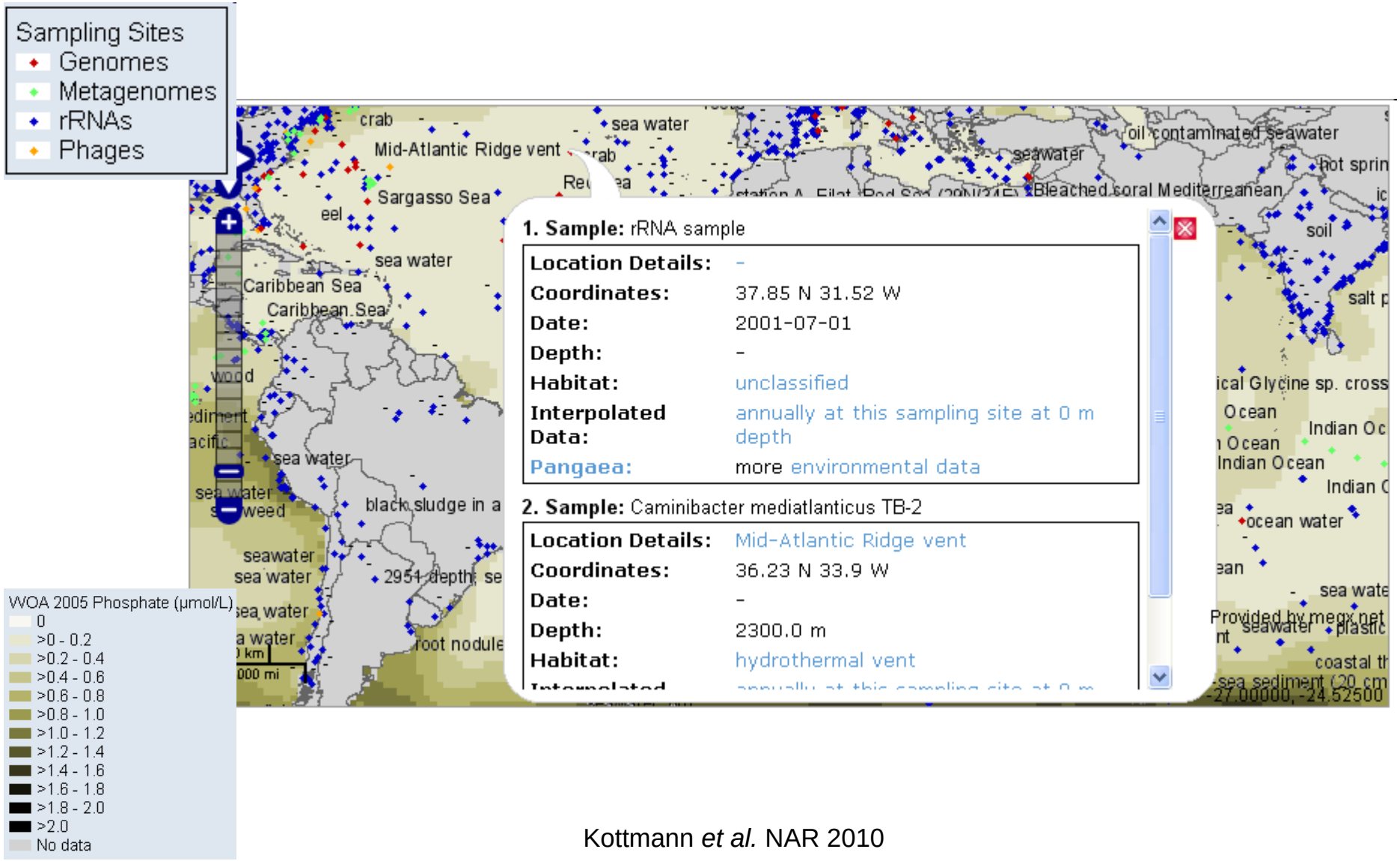
Megx.net:
Data Portal for Microbial Ecological GenomiX

- Solely based on Open Source Software
  - Database: PostgreSQL
    - PostGIS extension (geo-spatial data)
  - Web-Server:
    - Apache
    - UMN Mapserver
  - Web-client
    - OpenLayers



Kottmann *et al.* NAR. 2010

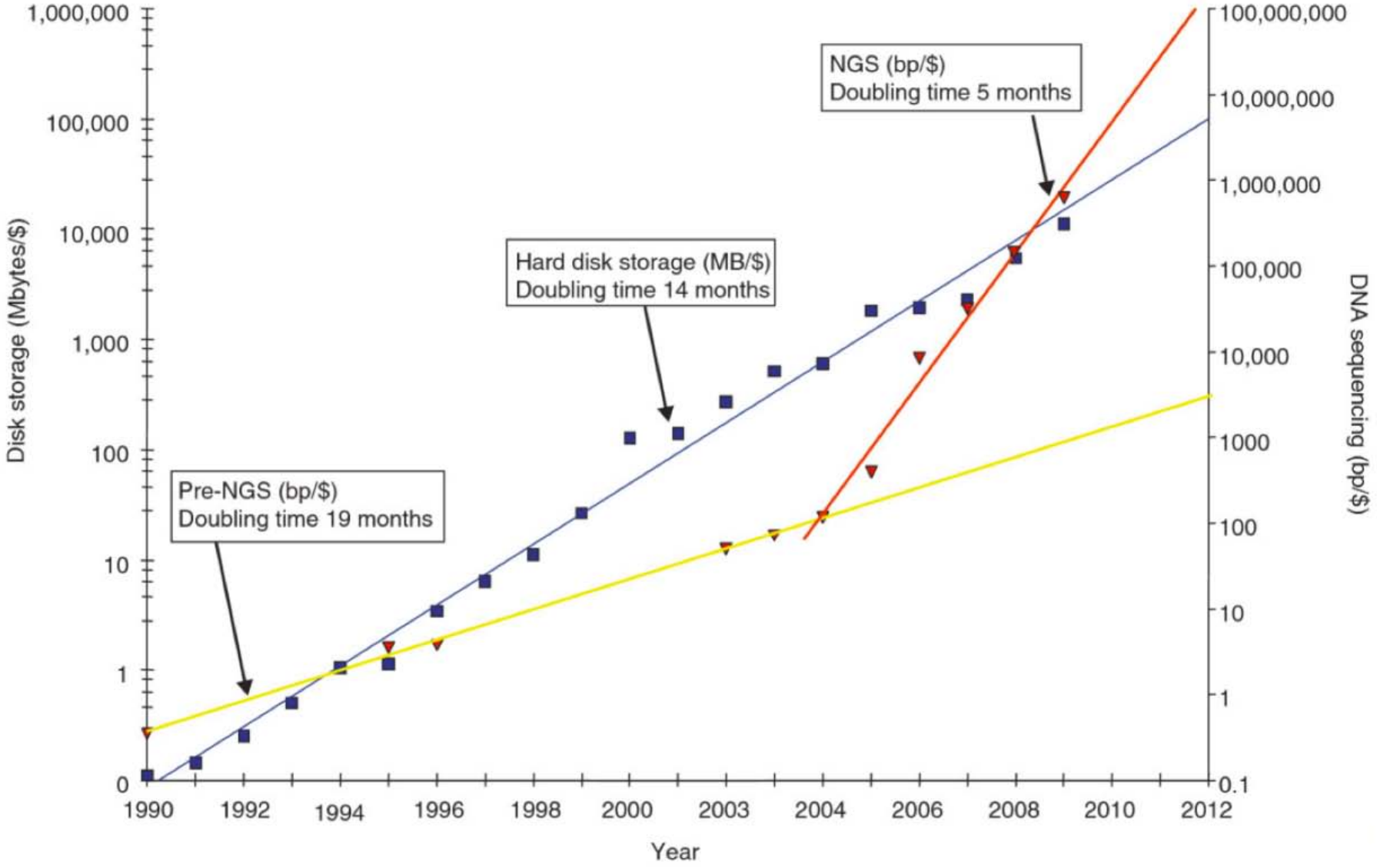
Who is out there and where?
(in terms of sequenced genomes, metagenomes and key genes)



Kottmann *et al.* NAR 2010

Nice, nice
BUT
where is the
problem????

NextGen Sequencing a Game-Changer



Lincoln Stein

More efficient ways to store Sequence Data needed

http://www.microb3.eu

Biodiversity. Bioinformatics. Biotechnology.

http://twitter.com/Micro_B3

- All Bioinformatics moves from flat files to NOSQL (MongoDB)
- We want to stay with Postgreee's great features:
  - Range types
  - JSON
  - hstore
  - PostGIS
  - Performance (shared_buffer_cache)
  - extensibility

PostBIS

- What is biological sequence data?
- How does PostgreSQL compression work?
- How does PostgreSQL compression perform on biological sequence data?
- How does PostBIS compression work?
- How does PostBIS perform in comparison to PostgreSQL and other approaches?
- What can we do with PostBIS?
- What do we want to do with PostBIS in the future?

What is biological sequence data?

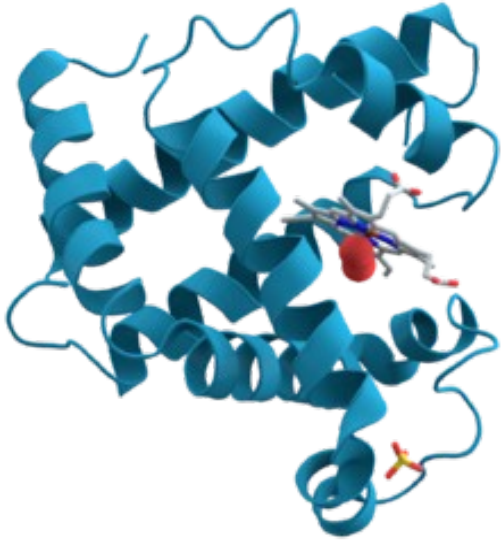
Genomic DNA

- Stores hereditary information
- Encodes information as a sequence of 4 different bases:
  - Adenine, Thymine, Cytosine, Guanine
    Example: `ACGATCGACTGAC`
- Alphabet size = 4, up to 15
- Lengths between few thousands and billions
- Genomic DNA can be repetitive

What is biological sequence data?

Short Sequences

- Short read DNA
  - From 50 to 10,000 bases long
- RNA
  - Similar to short read DNA
- Protein
  - Alphabet of 20 to 23!
  - At maximum thousands long



What is biological sequence data?

Alignments

- Method to find and display
  - Similarities
  - Differences
- Example:
  - Compare ACGATCGACGCAT with ACGAAAGACGA

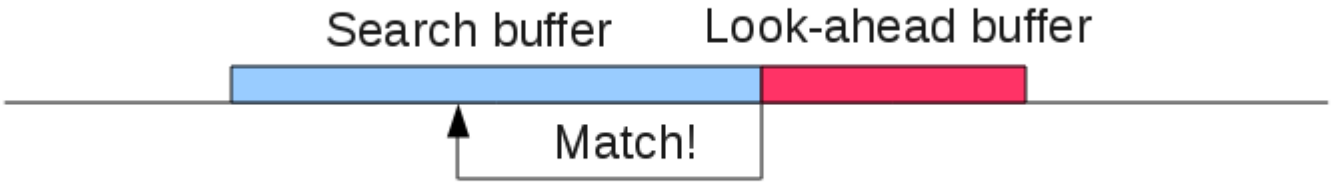
    ACGATC--GACGCAT

    ACGA--AAGACG-A-
- Length depends on:
  - Underlying sequences
  - Their similarity
- Long stretches of gap symbols

How does PostgreSQL compression work?

Lempel-Ziv PostgreSQL Variant

- Maintains a sliding window



- Finds match between
  - Prefix of look-ahead buffer
  - Substring starting in search buffer
- Encodes matches with 2 or 3-byte tokens
- No match → Standard encoding
- Termination conditions
  - Short than 32 character
  - Compression less than 25%
  - No match within first KB

How does PostgreSQL compression perform on biological sequence data?

- Entropy = average information content per character
  - Lower bound for compression
- Natural Text?
- Genomic DNA
  - ~one third → fair compression
- Short DNA, RNA, Protein
  - Not at all → no compression
- Alignments
  - Often:

    Down to entropy → very good compression
  - Sometimes:

    less

How does PostBIS compression work?

1. Run-Length Encoding

   TCGAAAAAAAAAGCTAG

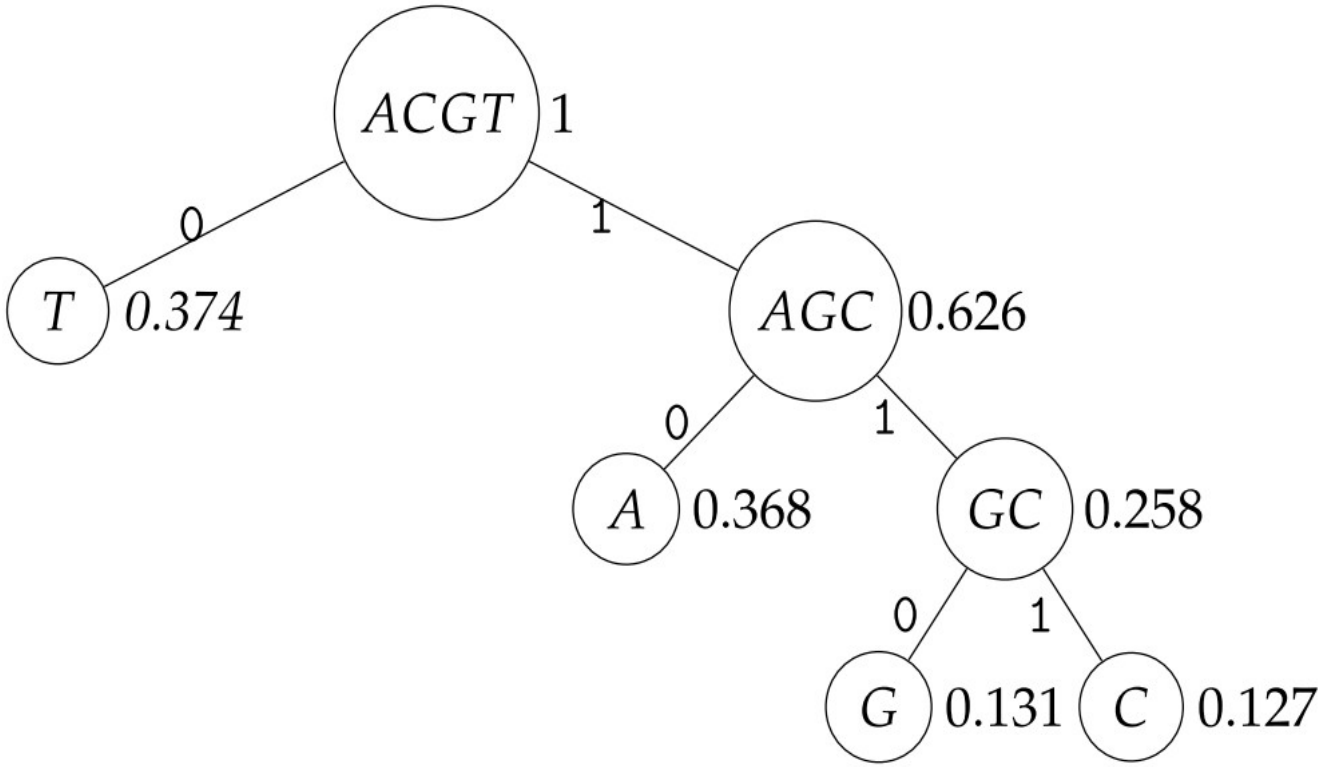
   TCGr8AGCTAG

2. Huffman codes
3. Rare Symbol Swapping

How does PostBIS compression work?

Huffman codes

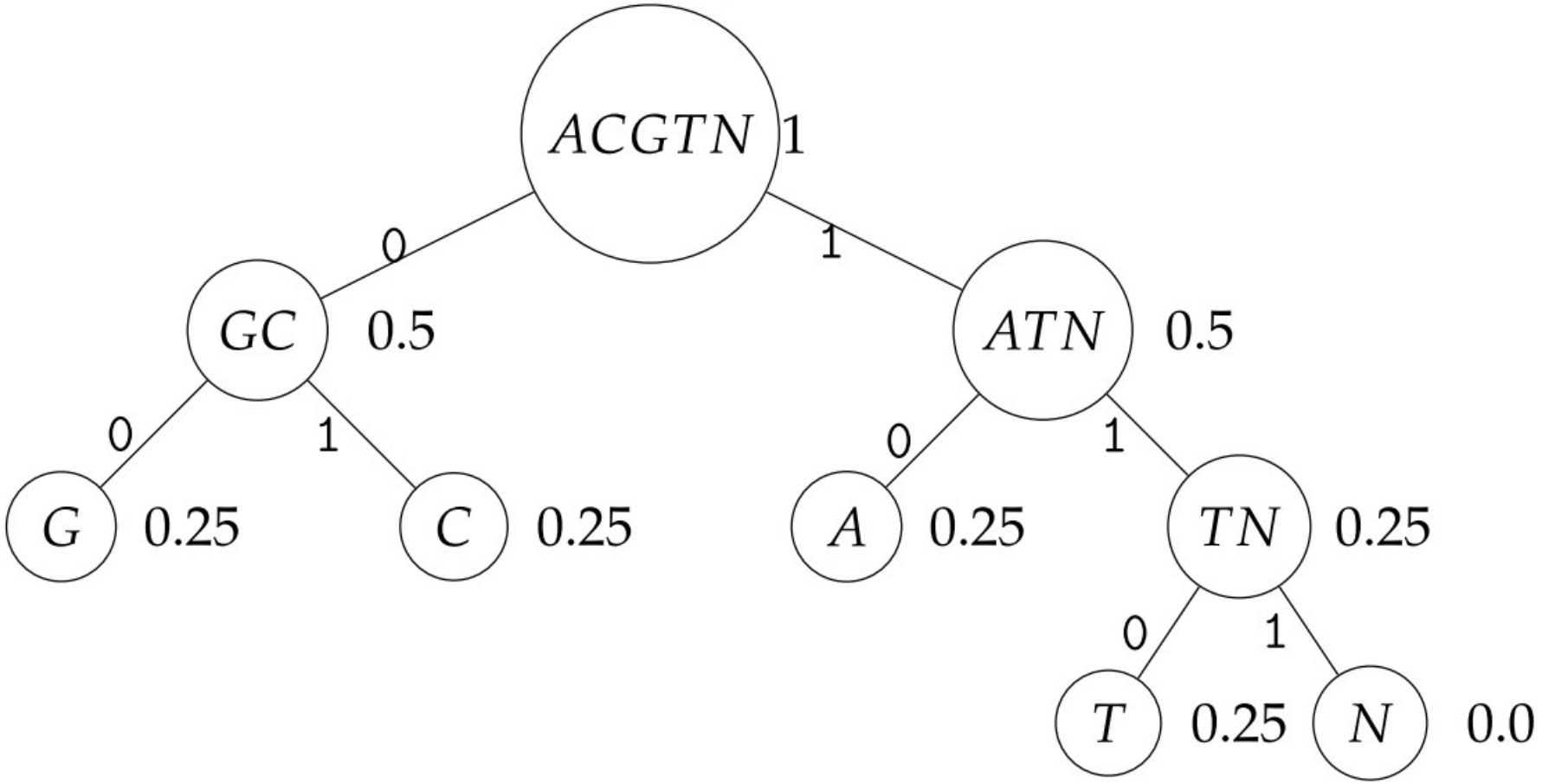
- Reduced alphabet
- Assign short codewords to frequent symbols



How does PostBIS compression work?

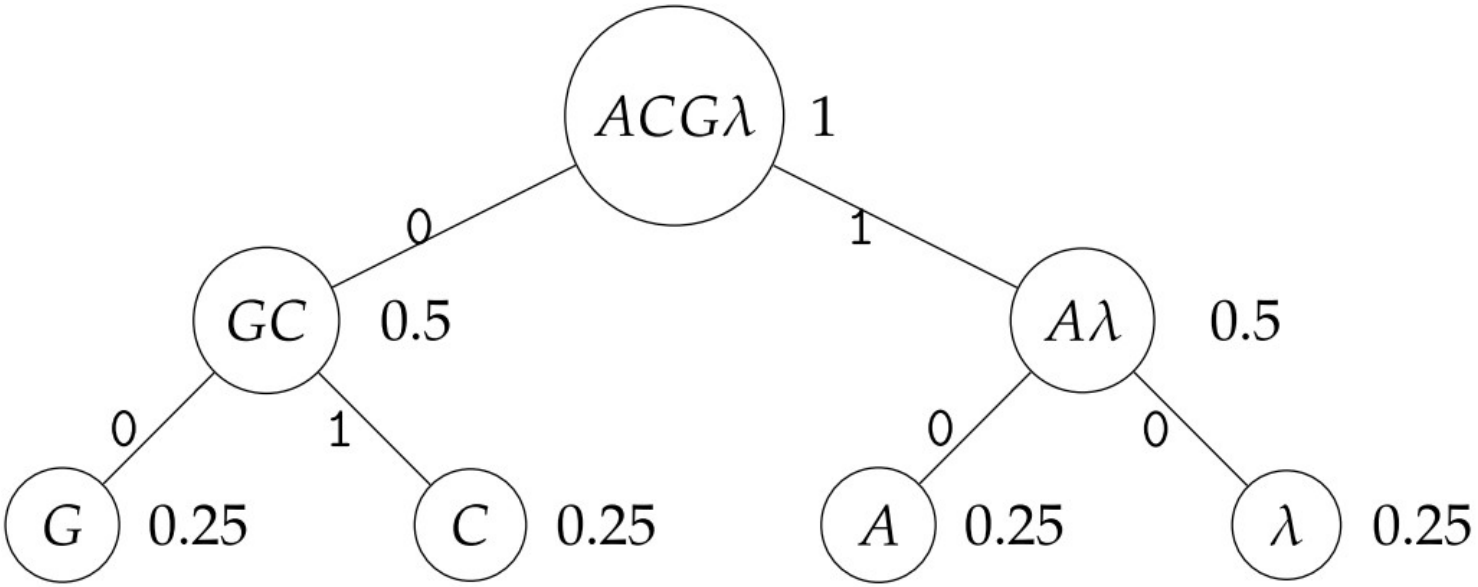
Rare Symbol Swapping

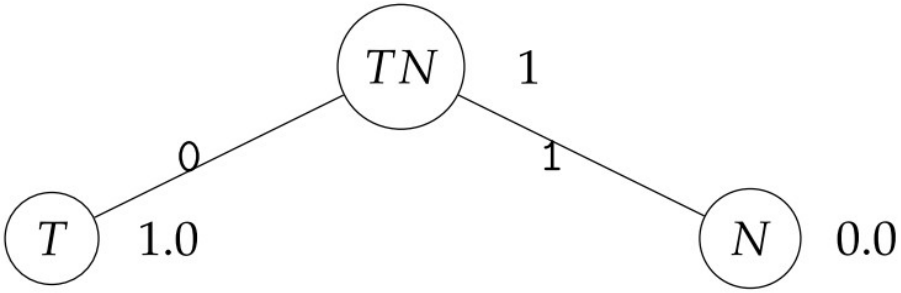
- On DNA, Redundancy of 0.25 = 12.5% possible!



How does PostBIS compression work?

Rare Symbol Swapping



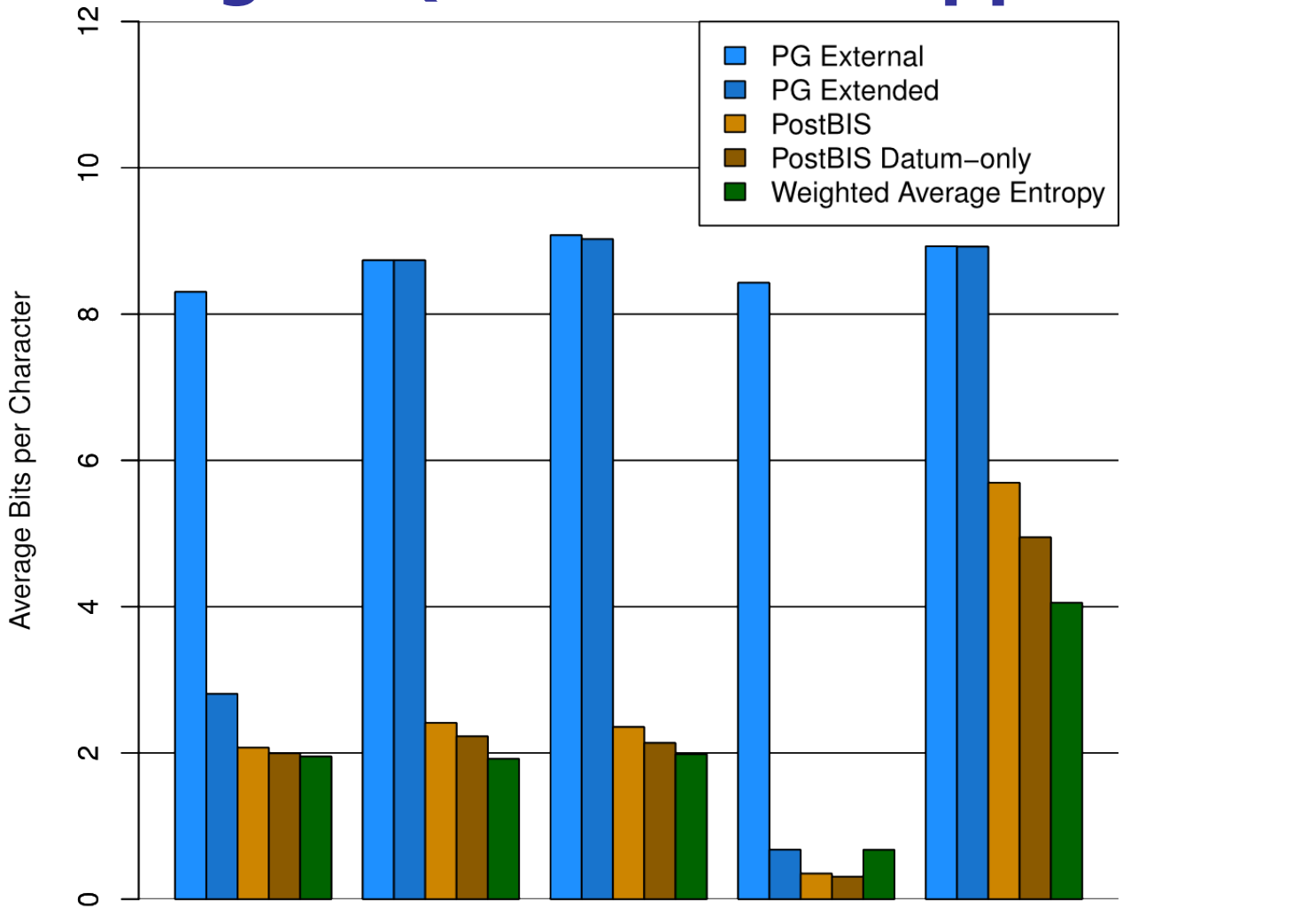


- Lower Limit of Redundancy = 0.000003815

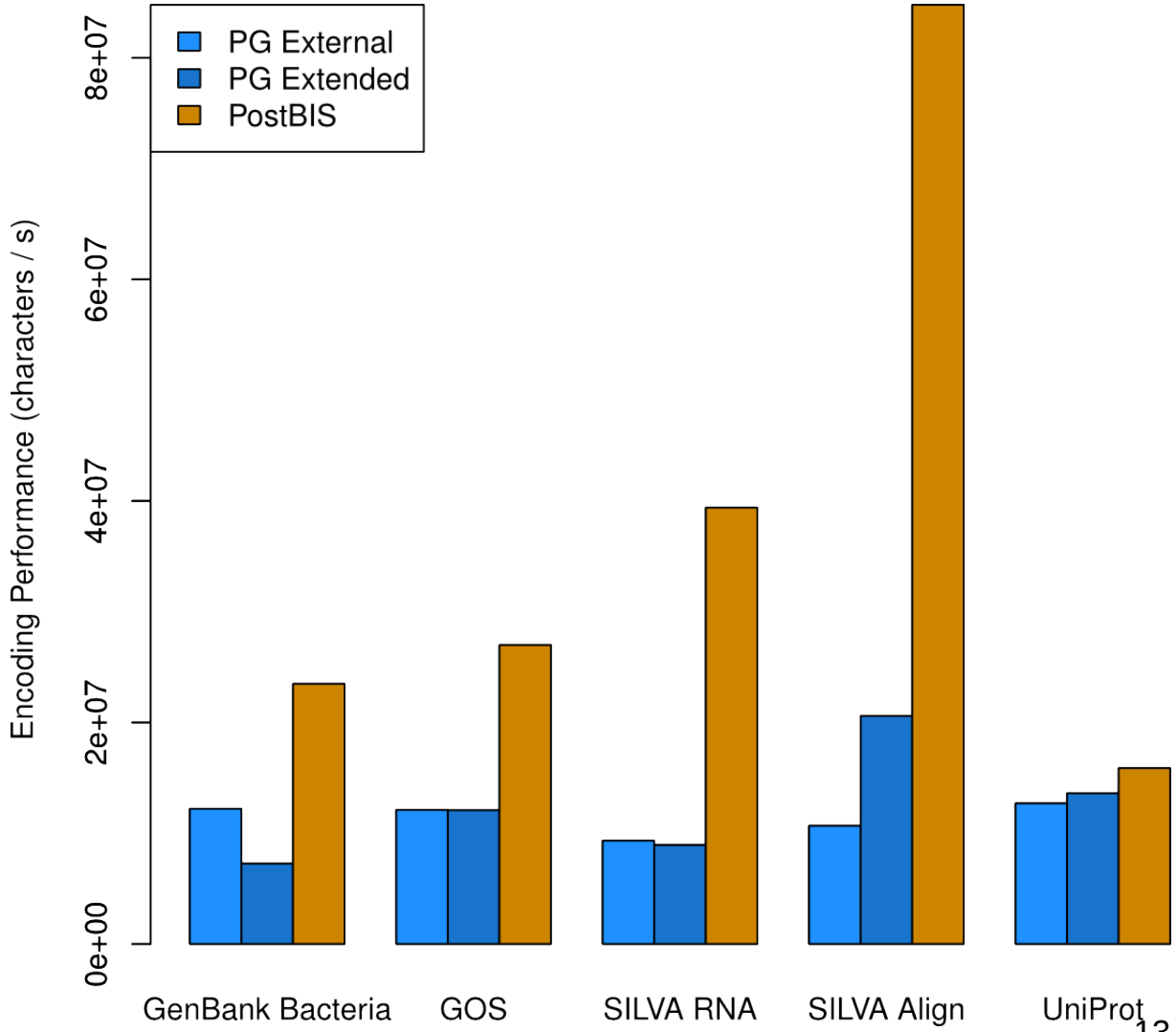
How does PostBIS compression work?

- New data types:
  - DNA_SEQUENCE
  - RNA_SEQUENCE
  - AA_SEQUENCE
  - ALIGNED_DNA_SEQUENCE
  - ALIGNED_RNA_SEQUENCE
  - ALIGNED_AA_SEQUENCE
  - Type modifiers:
    - CASE_SENSITIVE / CASE_INSENSITIVE
    - FLC / IUPAC / ASCII
    - SHORT / DEFAULT / REFERENCE (only DNA)

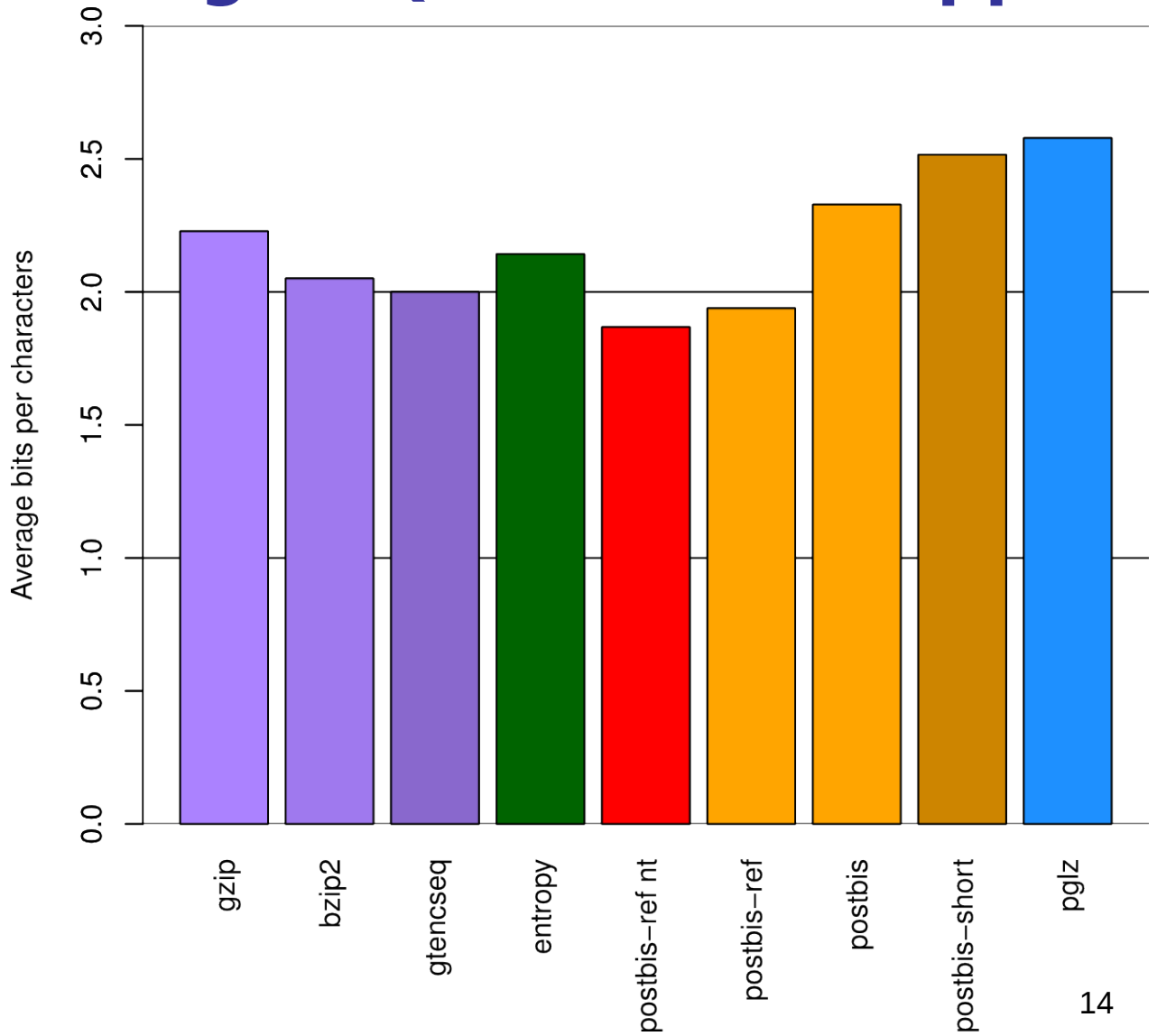
How does PostBIS perform in comparison to PostgreSQL and other approaches?



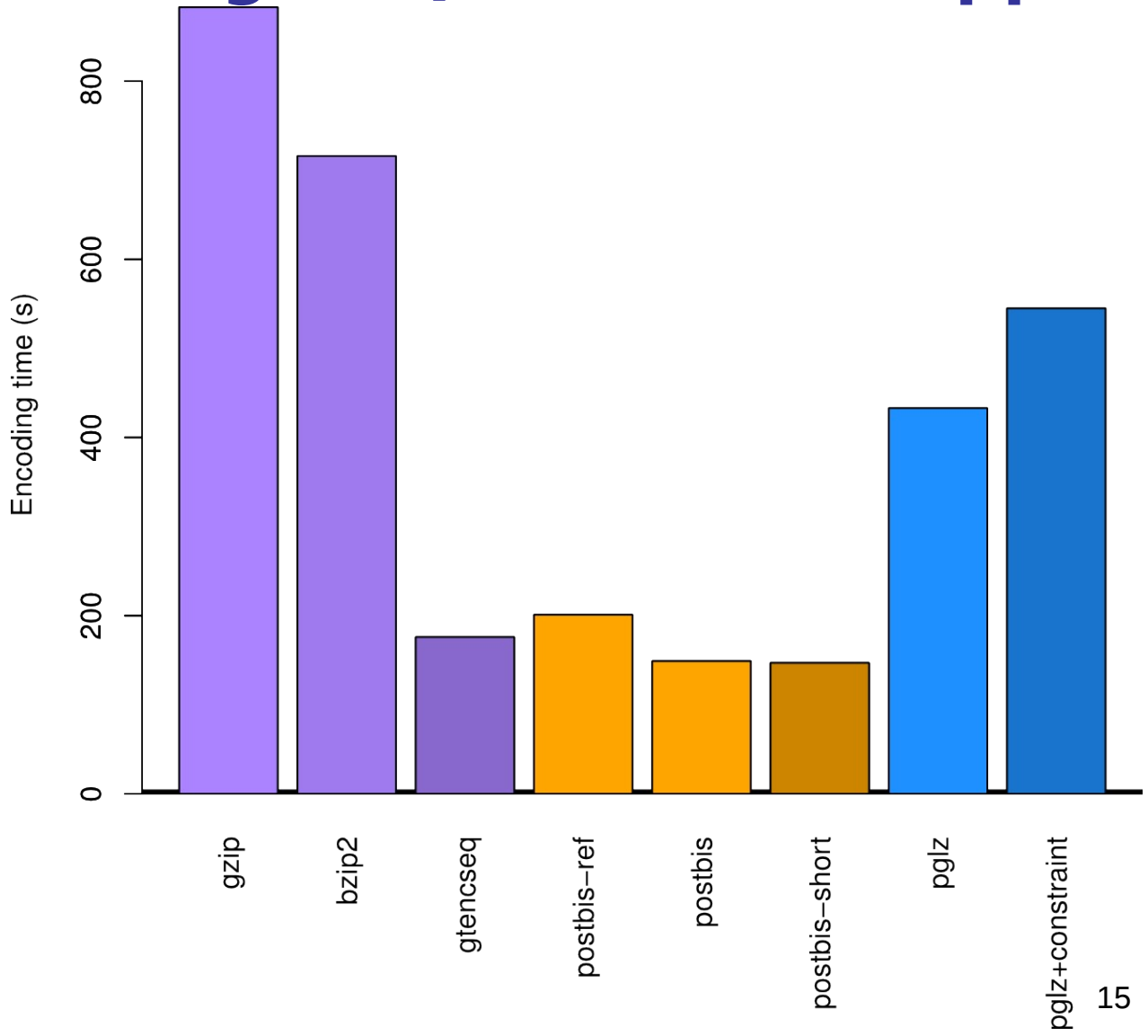
How does PostBIS perform in comparison to PostgreSQL and other approaches?



How does PostBIS perform in comparison to PostgreSQL and other approaches?



How does PostBIS perform in comparison to PostgreSQL and other approaches?



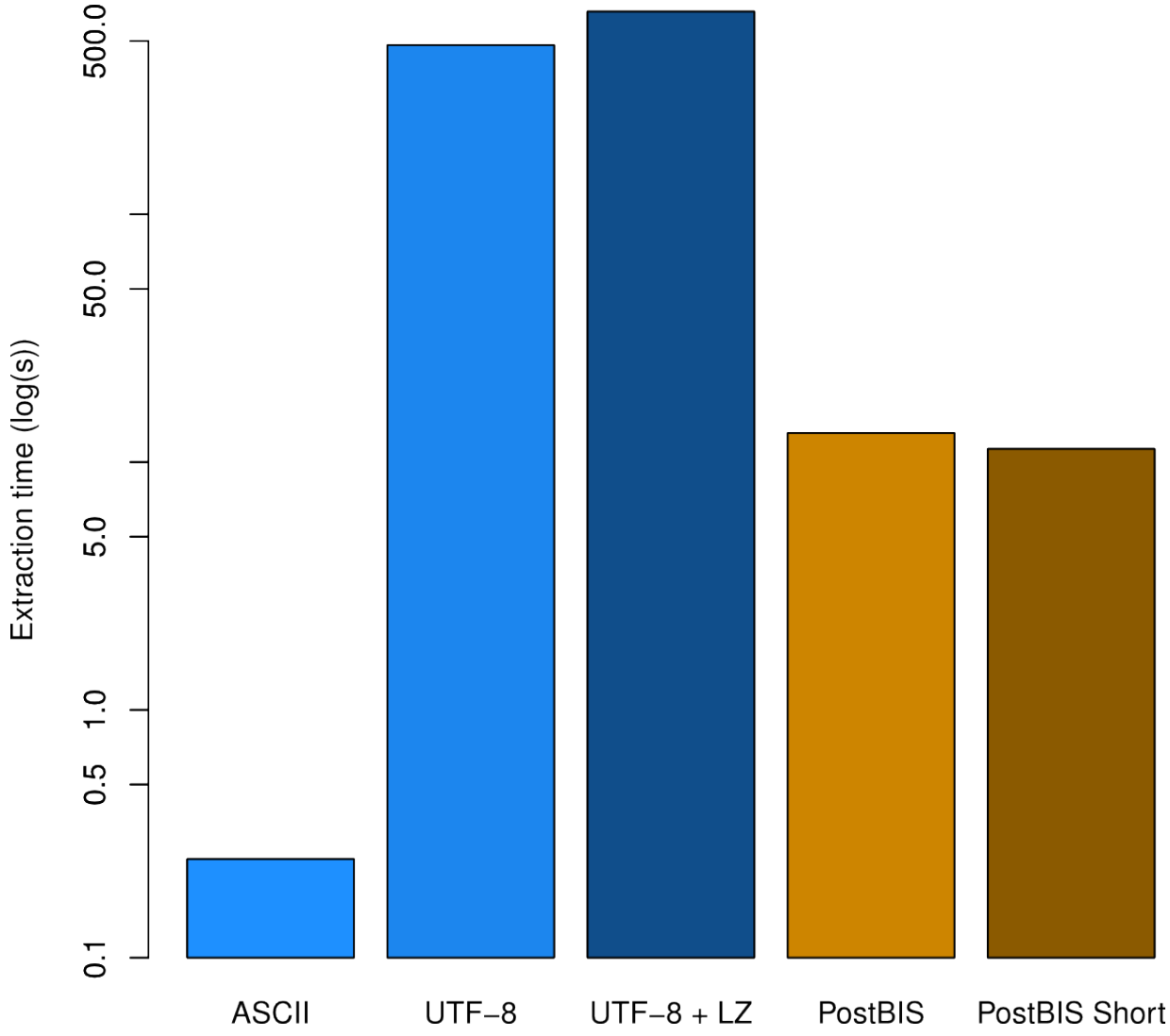
What can we do with PostBIS?

- Sequences in database, now what!?
- Doing Bioinformatics is flat file based
  - Select subset in database
  - Export sequences to flat-file
  - Do bioinformatics with command-line tool
  - Parse output
  - Import output to database
- Use-Cases:
  - tRNAscan
  - Gene extraction

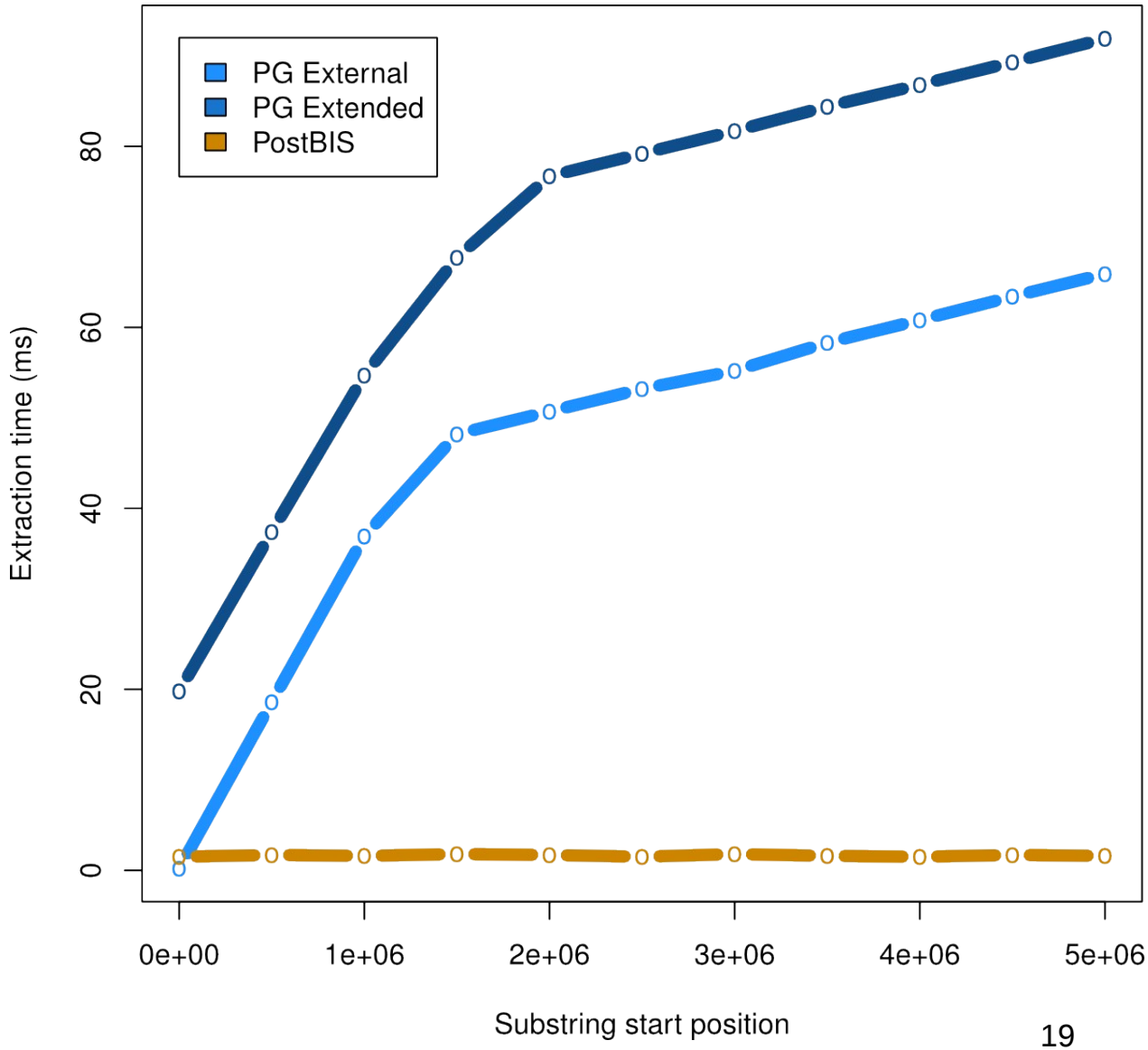
What can we do with PostBIS?

```
CREATE TABLE human_genome (
    sequence dna_sequence(reference),
    chromosome text
    );
SELECT trna(sequence, chromosome)
    INTO human_genes
    FROM human_genome;
SELECT substr(a.sequence, b.start_pos, b.len)
FROM
    human_genome AS a
    INNER JOIN
    human_genes AS b
    ON a.chromosome = b.chromosome;
```

Substring performance



Substring Performance



What do we want to do with PostBIS in the future?

- Reference-based compression
- Reference-based heuristic approximative full-text search
  - Compressive BLAST
- NN-searches
- FDWs for relevant file formats
- Adapt existing tools

Thank you for your attention!

Tips, Comments and Questions will be appreciated!

Please give feedback at

http://2012.pgconf.eu/feedback/