Searching in nucleotide archives at petabase scale with MetaGraph

ETH zürich
BIOMEDICAL INFORMATICS
@m_karasikov

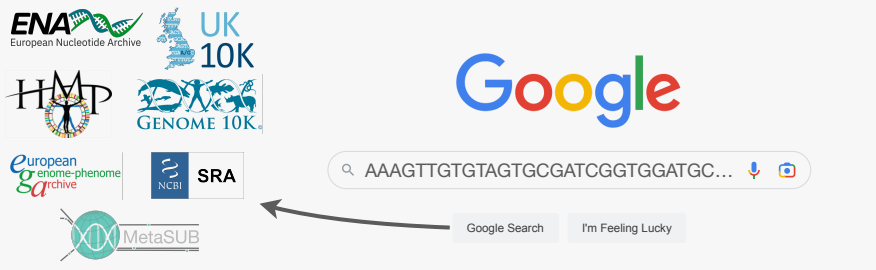
Mikhail Karasikov*, *Harun Mustafa**, *Daniel Danciu, Marc Zimmermann, Marek Kokot, Christopher Barber, Gunnar Rätsch, André Kahles*

MetaGraph

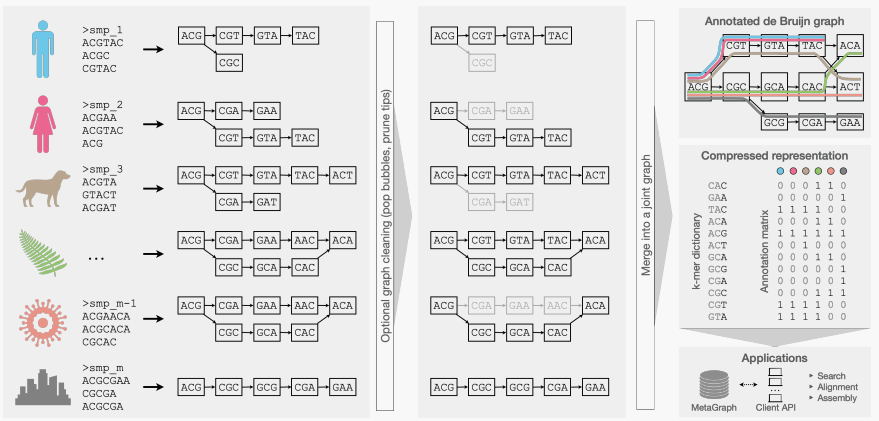
metagraph.ethz.ch/search

Try it out!

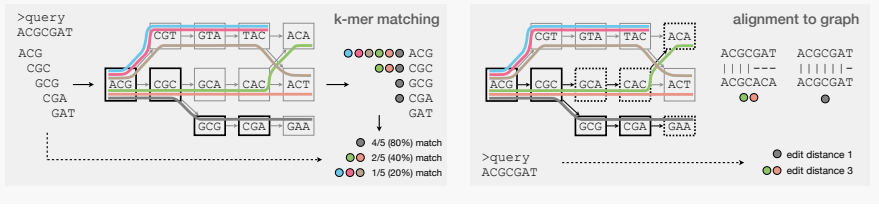
1. Wouldn't it be cool if we could **"google" in sequence archives?**



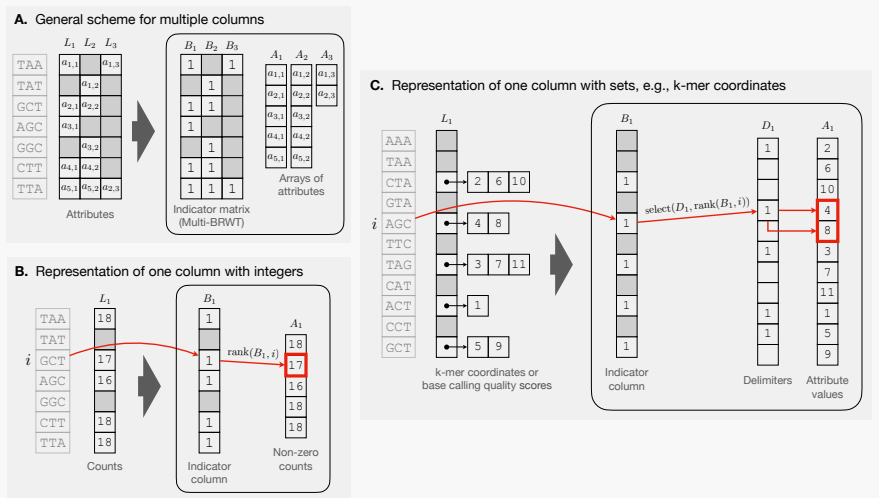
2. **Compressed and lossless** k-mer set representation



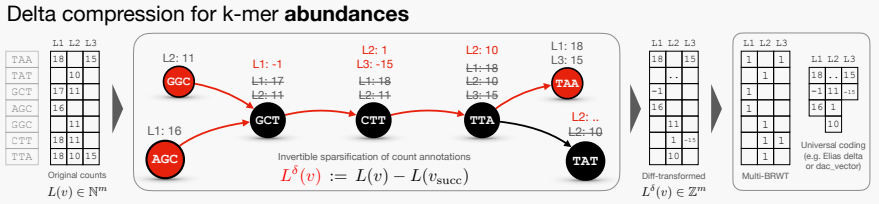
3. Search via k-mer matching or **alignment** with sub-*k* seeding



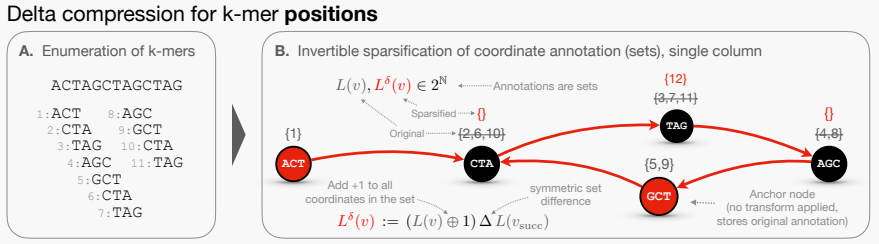
4. Representing **non-binary** attributes of k-mers



Delta compression for k-mer **abundances**



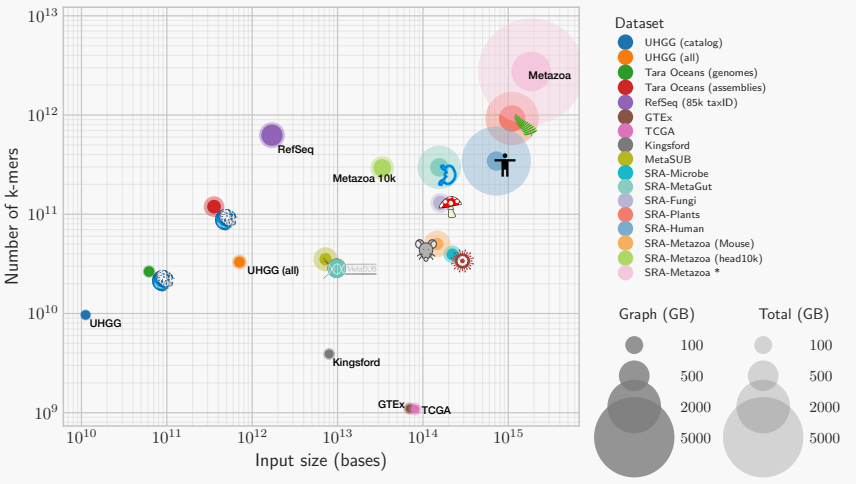
Delta compression for k-mer **positions**



Main features of MetaGraph

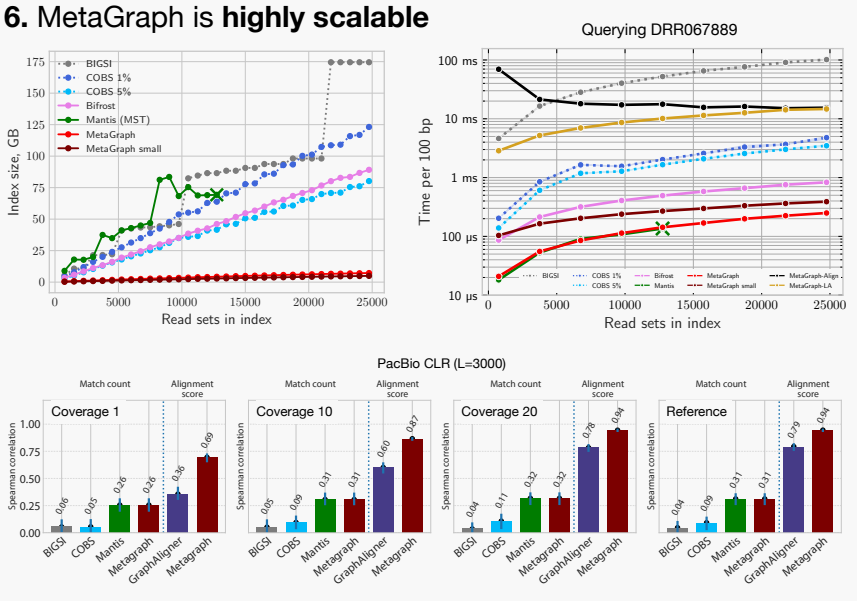
- **Large-scale indexing** of sequences
- **Python API** for querying MetaGraph in server mode
- Encoding **k-mer abundances** and **k-mer coordinates** in input
- **Sequence alignment** against very large annotated graphs

5. Indexed **data sets**



Data set	Tbp	Input (gz)	# k-mers	# labels	Index size	Ratio
UHGG (all)	0.71	206.0 GB	$33.0 \cdot 10^9$	286,997	27.3 GB	7.6×
Tara Oceans with coord.	0.06	17.9 GB	$26.5 \cdot 10^9$	34,815	14.6 GB	1.2×
Tara Oceans (assemblies)	0.36	106.8 GB	$119.4 \cdot 10^9$	318,205,057	124.1 GB	0.9×
RefSeq with coord.	1.70	502.4 GB	$626.2 \cdot 10^9$	85,375	508.9 GB	1.0×
Kingsford with counts	8.0	2.9 TB	$3.9 \cdot 10^9$	2,652	20.9 GB	138×
GTEx	70.0	40.0 TB	$1.1 \cdot 10^9$	9,759	8.4 GB	4,742×
TCGA	81.2	65.0 TB	$1.1 \cdot 10^9$	11,095	11.1 GB	5,831×
MetaSUB	7.2	5.5 TB	$35.2 \cdot 10^9$	4,220	206.4 GB	27×
SRA-MetaGut	155.8	86.0 TB	$296.9 \cdot 10^9$	242,619	1,111.3 GB	77×
SRA-Microbe	221.1	170.0 TB	$39.5 \cdot 10^9$	446,506	65.5 GB	2,595×
SRA-Fungi	160.2	80.0 TB	$129.7 \cdot 10^9$	121,900	108.1 GB	740×
SRA-Plants	1,109.2	575.9 TB	$923.4 \cdot 10^9$	531,736	1,844.1 GB	312×
SRA-Human	725.4	345.7 TB	$343.9 \cdot 10^9$	436,502	3,402.1 GB	102×
SRA-Metazoa (Mouse)	146.6	61.3 TB	$50.2 \cdot 10^9$	57,938	291.6 GB	210×
SRA-Metazoa (10k)	33.4	16.5 TB	$293.3 \cdot 10^9$	10,000	192.7 GB	86×
SRA-Metazoa*	1,856.8	925.3 TB	$2749.8 \cdot 10^9$	797,883	8,858.8 GB	104×

6. MetaGraph is **highly scalable**



7. MetaGraph API

```
In [1]: from metagraph.client import GraphClient

        SRV = "metagraph.ethz.ch"
        PORT = 12345
        g1 = GraphClient(SRV, PORT, api_path="/metasub")
        g2 = GraphClient(SRV, PORT, api_path="/refseq")

In [2]: query = "GGCTAACTACGTGCCAGCAGCCGCGGTAATAC"
        g1.search(query, align=True)

Out [2]:
```

	sample	sequence	score
0	SRR2201245	GGCTAACTACGTGCCAGCAGCCGCGGTAATAC	64
1	ERR1732568	GGCTAACTACGTGCCAGCAGCCGCGGTAATAC	64
2	ERR847096	GGCTAACTACGTGCCAGCAGCCGCGGTAATAC	64

8. MetaGraph Online