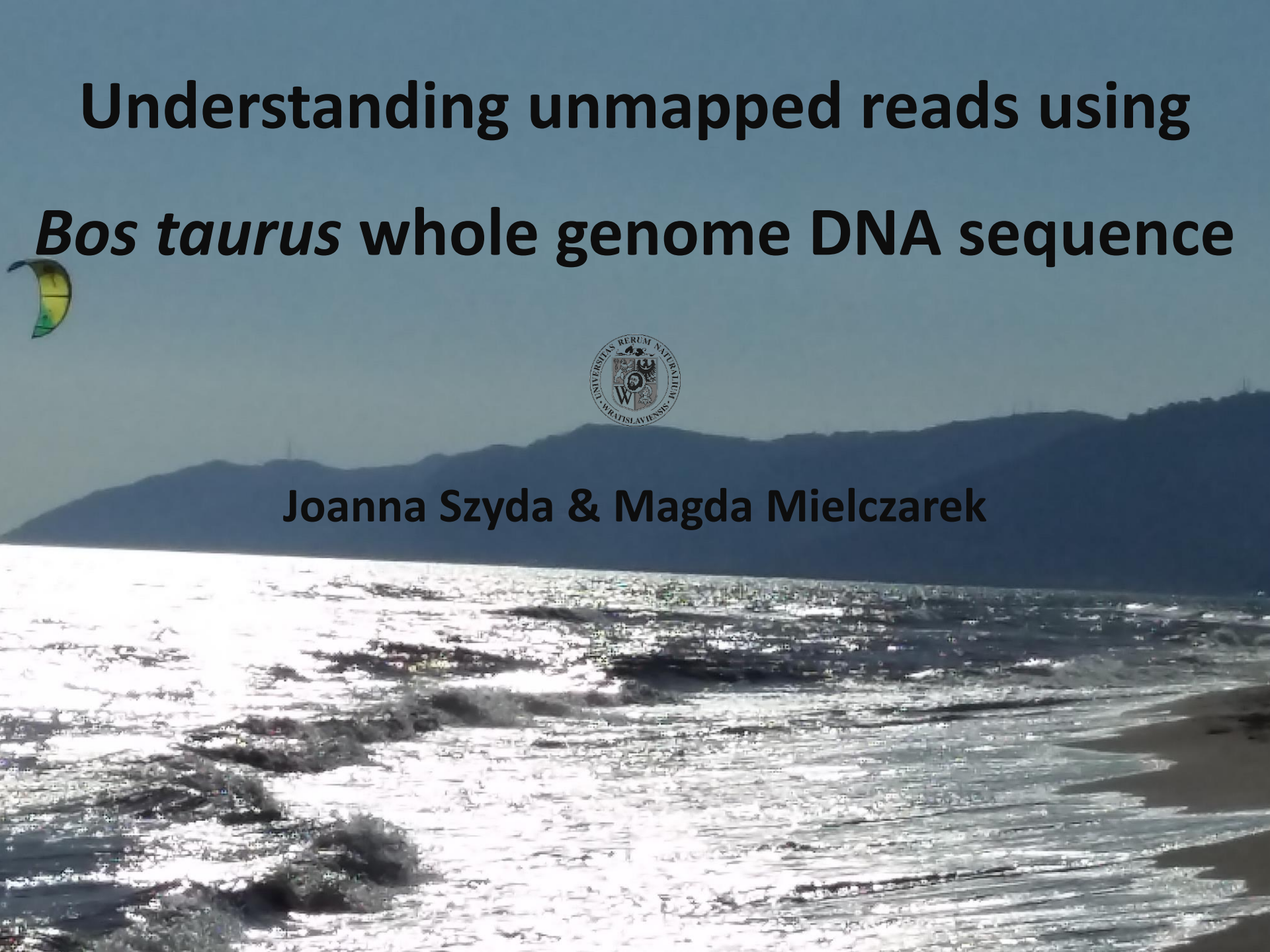
Understanding unmapped reads using *Bos taurus* whole genome DNA sequence



Joanna Szyda & Magda Mielczarek

Outline

genomes

reads **not mapped** to UMD3.1.1

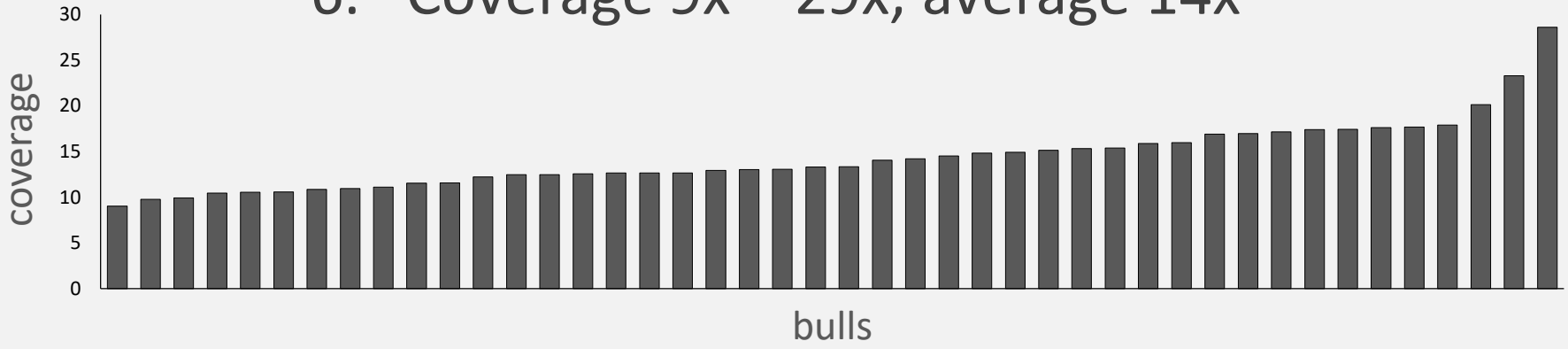
- both read pairs unmapped
- read mapped ⟷ read unmapped

annotated reads
HIT

still unmapped reads
noHIT

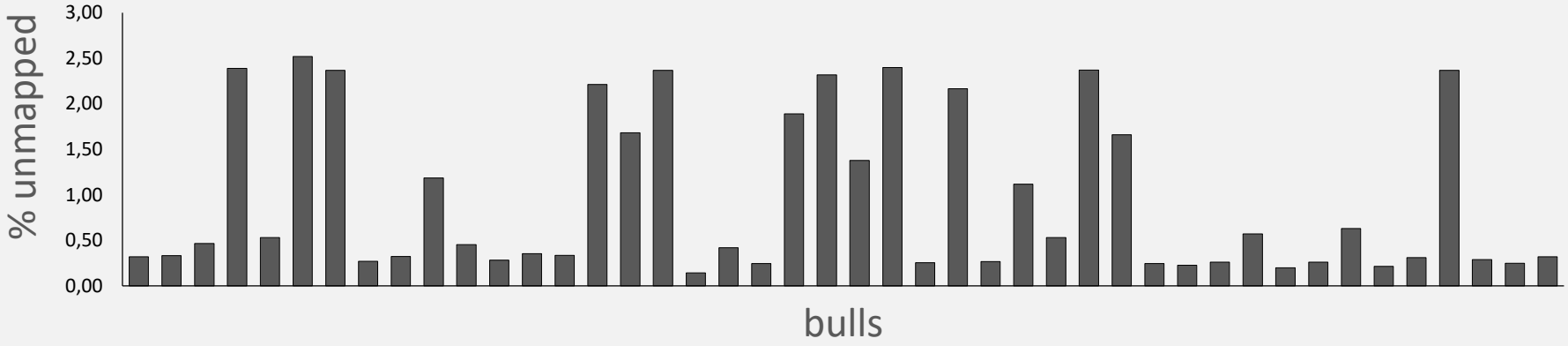
Genomes

1. Whole genome DNA sequence
2. 44 Brown Swiss Bulls
3. Gene2Farm
4. Illumina HiSeq 2000
5. Pair-end, 101 bp long
6. Coverage 9x – 29x, average 14x



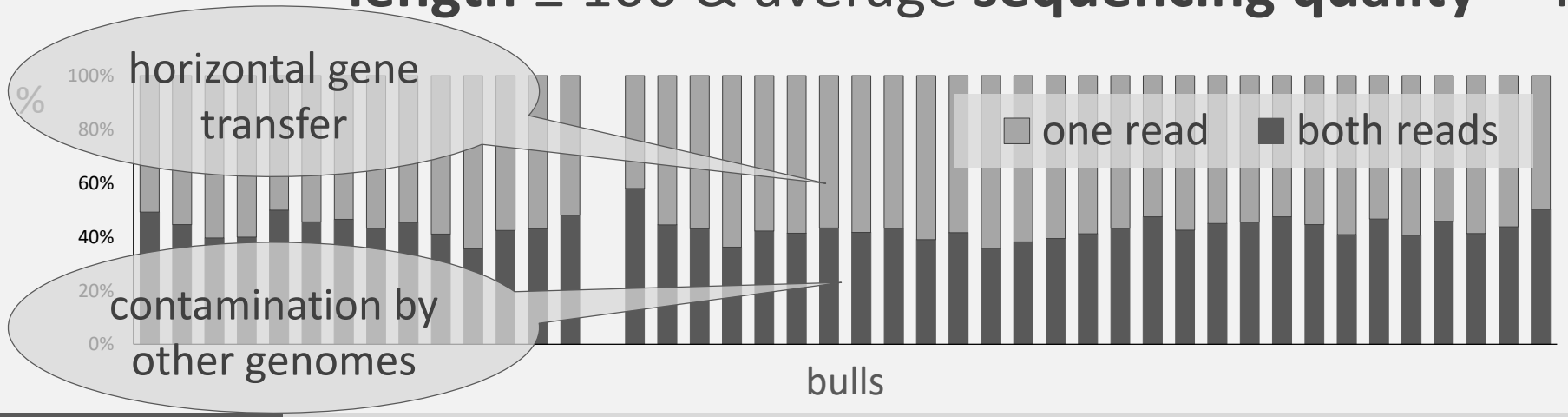
Pipeline

Alignment to UMD3.1.1 → BWAmem



Unmapped read selection →

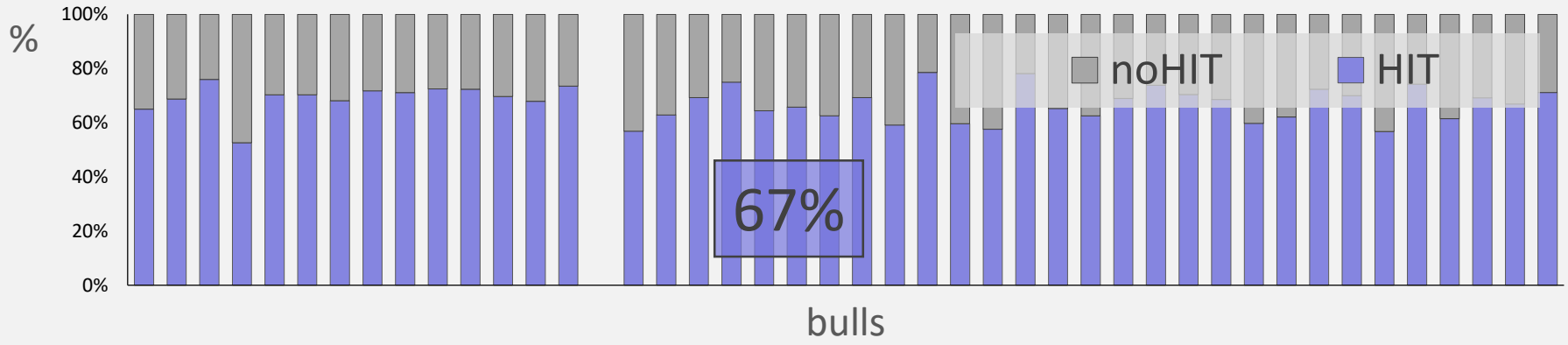
length ≥ 100 & average **sequencing quality** = 40



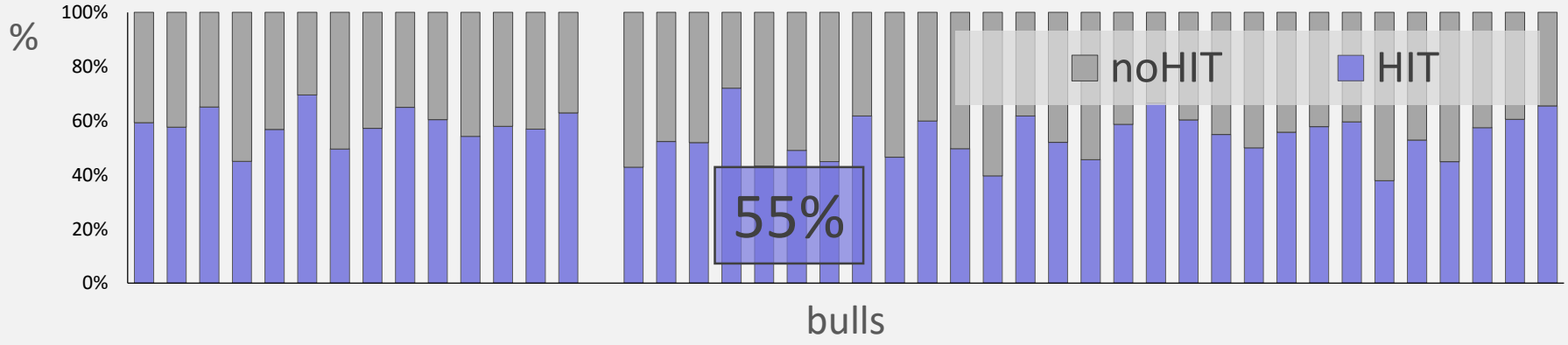
Pipeline

Annotation to **RefSeq** → BLASTn

both reads

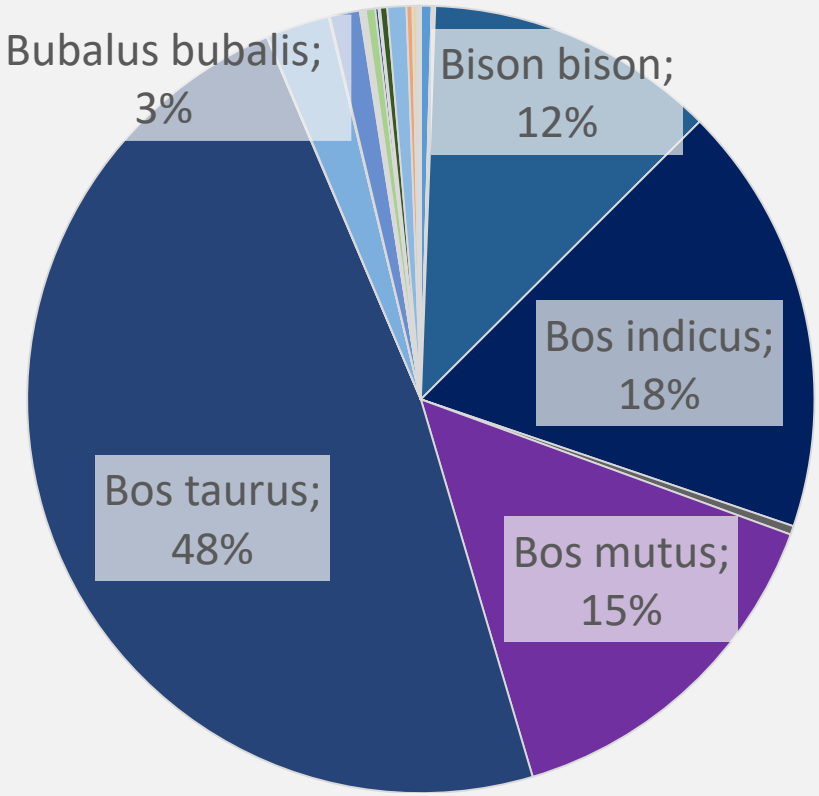


one read

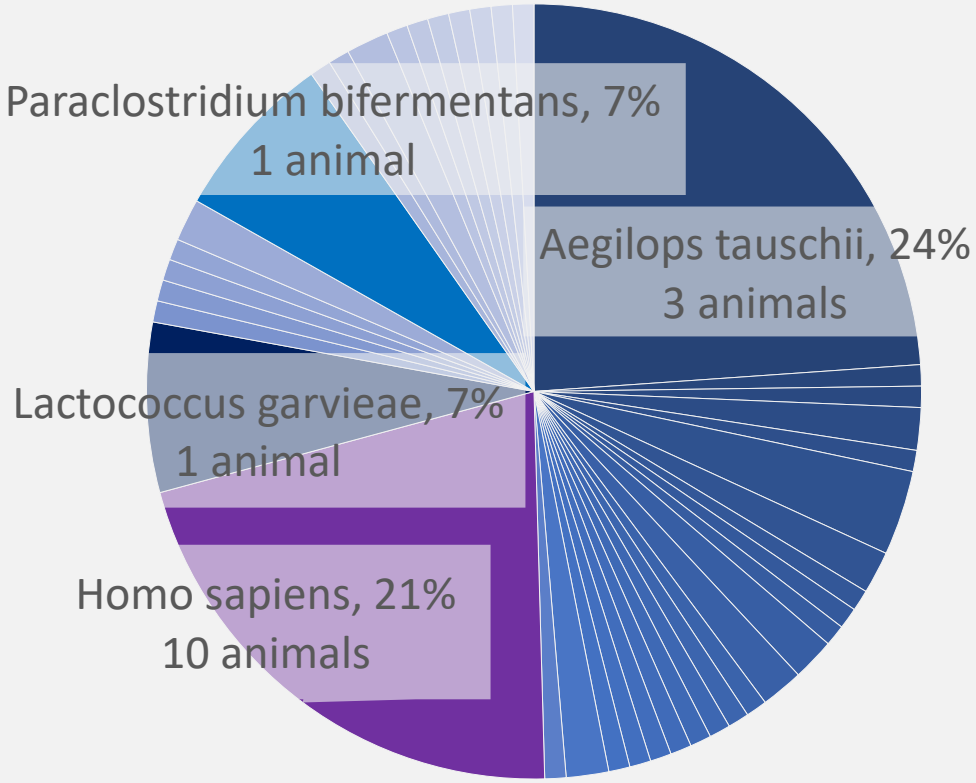


HIT reads annotation → RefSeq

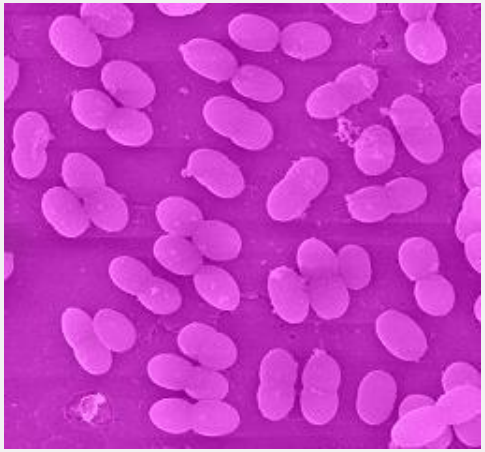
both reads → all species

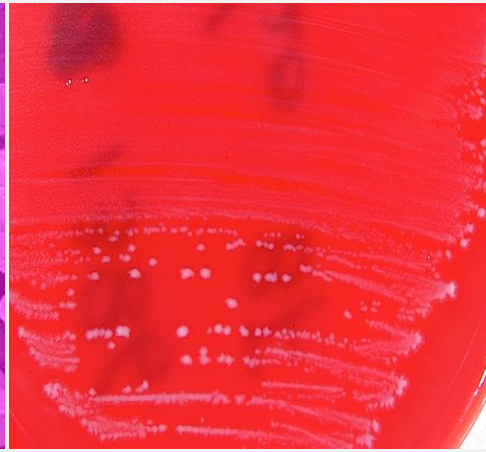


both reads → non ruminant



HIT read annotation → RefSeq





Homo sapiens

- 10 animals
- 21%
- Common lab contamination

nypost.com

Aegilops tauschii

- 3 animals
- 24%
- Many TE

www.chungvisinh.com

Lactococcus garvieae

- 1 animal
- 7%
- Potential mastitis pathogen

www.chungvisinh.com

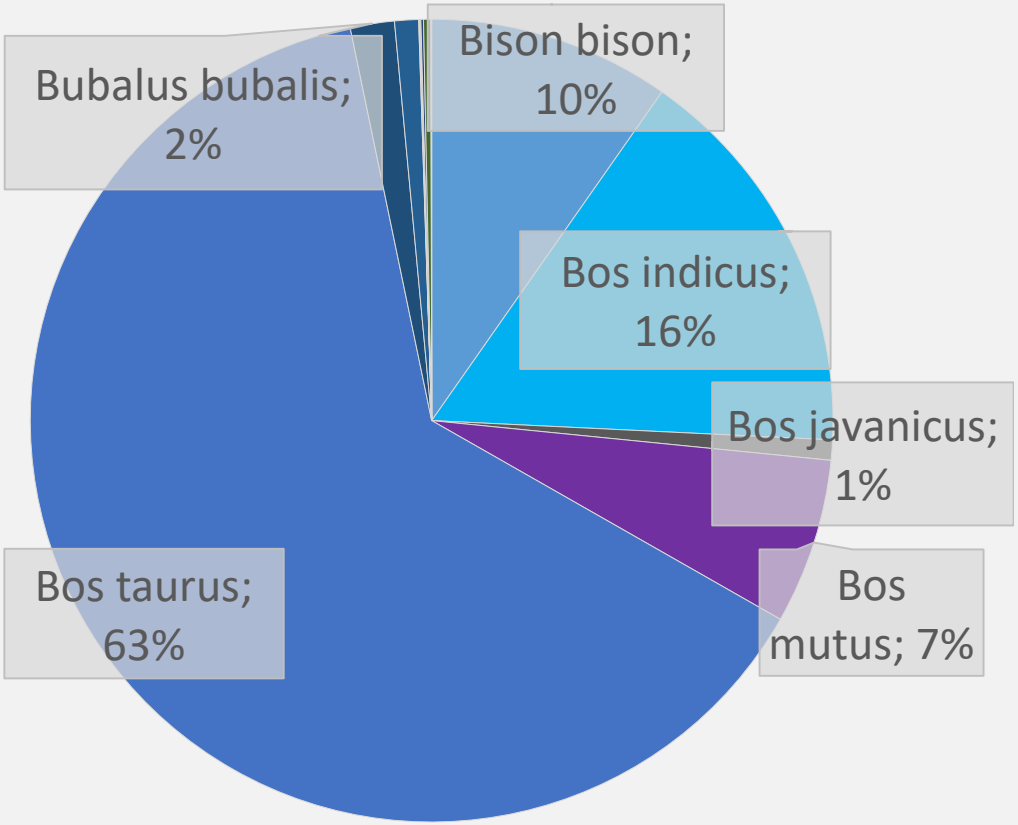
Paraclostridium bifermentans

- 1 animal
- 7%
- UC pathosis in mice

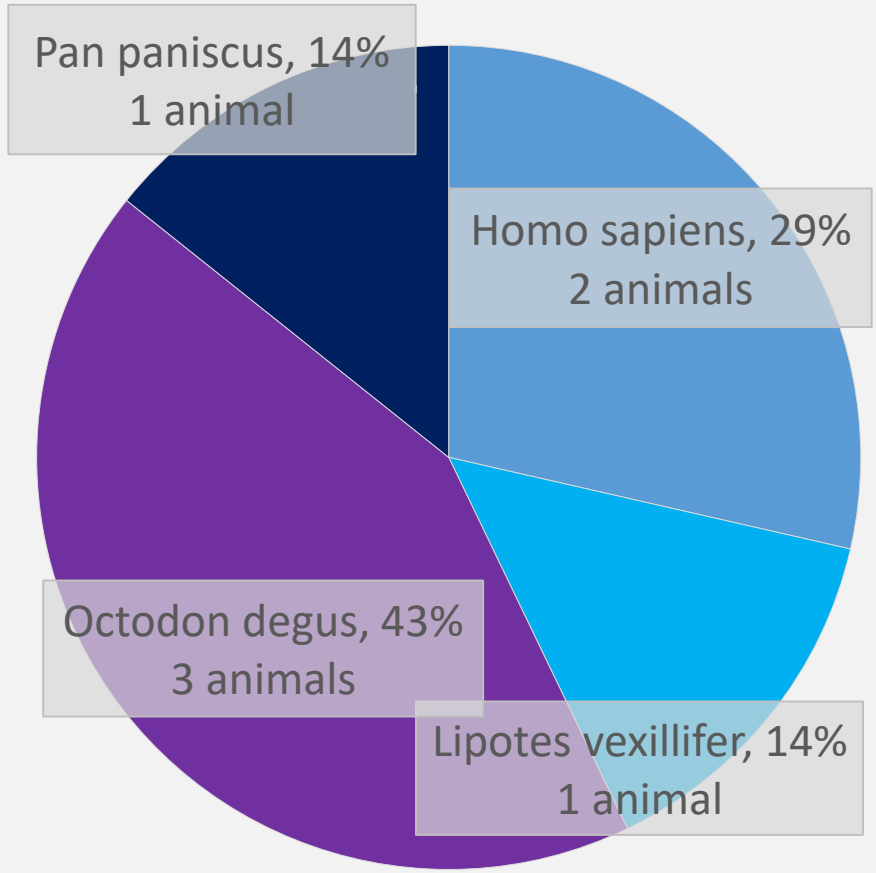
www.jcm.riken.go.jp

HIT read annotation → RefSeq

one read → all species



one read → non ruminant



HIT read annotation → RefSeq

Octodon degus

- 3 animals
- 43%
- Shares the same habitat … in South America

nypost.com

Homo sapiens

- 2 animals
- 29%
- Common lab contamination

Max Planck Institute

Pan paniscus

- 1 animal
- 14%
- Contamination with Homo sapiens

biodiversityconservation blog.wordpress.com

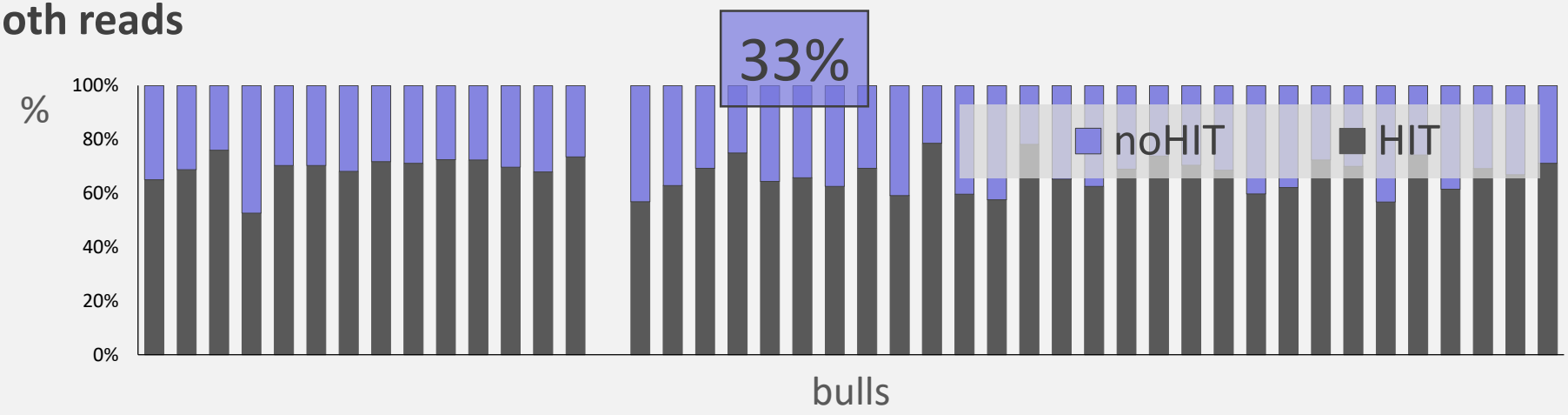
Lipotes vexillifer

- 1 animal
- 14%
- Sequence similarity to selected cattle genes

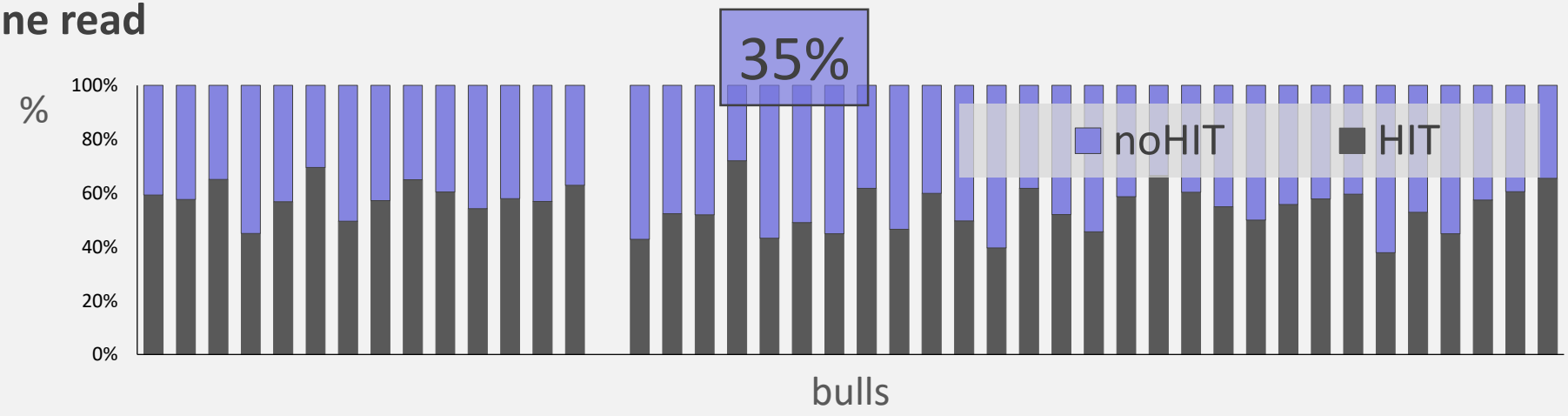
Still unmapped → noHIT

Annotation to **RefSeq** → BLASTn

both reads



one read

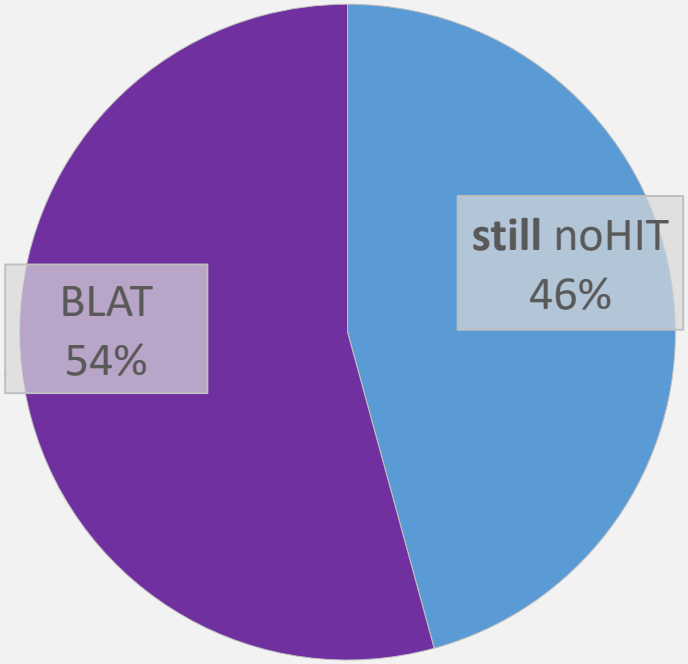


Searching further ……

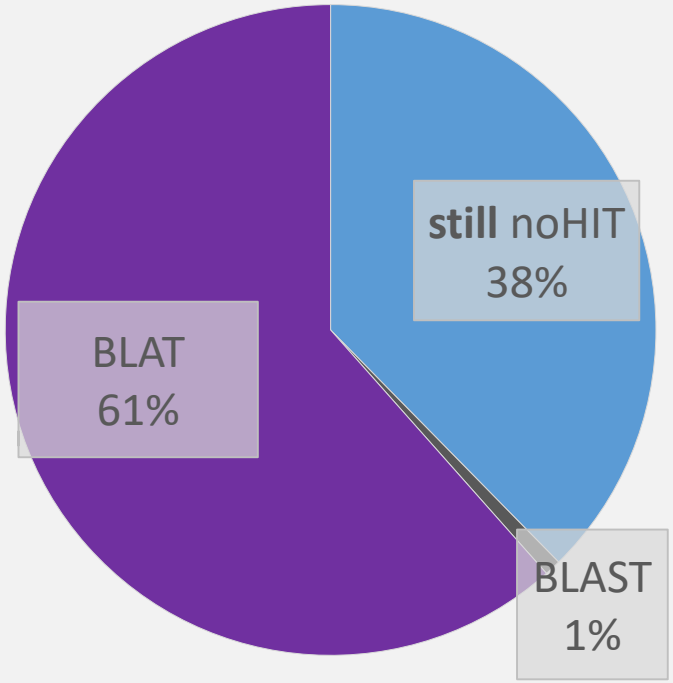
Annotation to **Nucleotide db** → BLASTn

Annotation to UMD3.1.1 → **BLAT**

one read



both reads



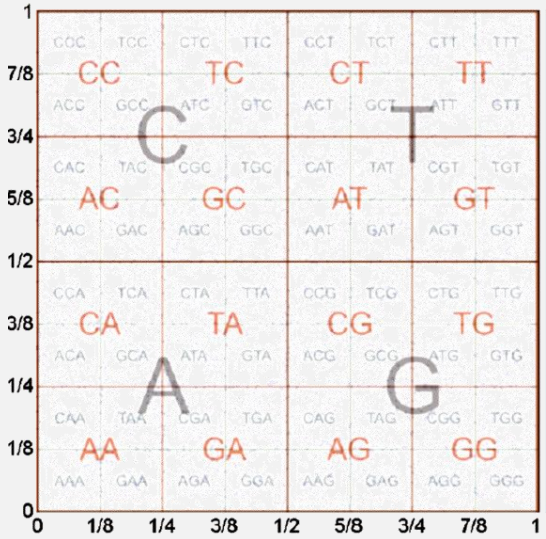
BLAST → Hordeum vulgare (2/21), Bos mutus (1/1), Ovis canadensis (1/1)

BLAT → unassigned scaffolds

Still no HIT …… pattern mining

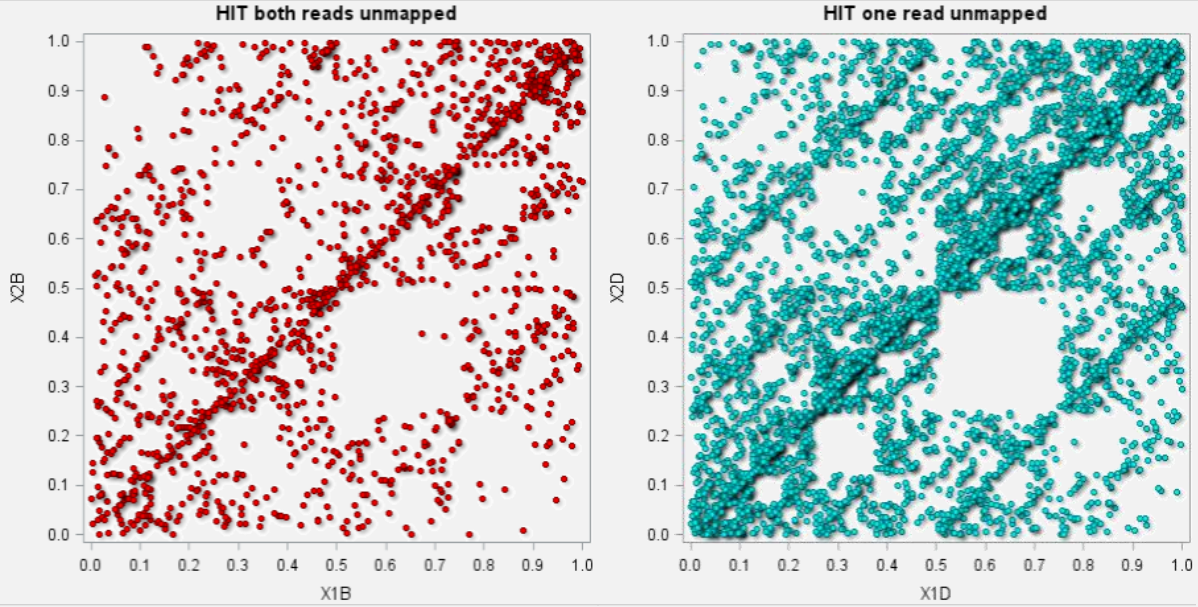
Sequence pattern visualisation →Chaos Game Representation

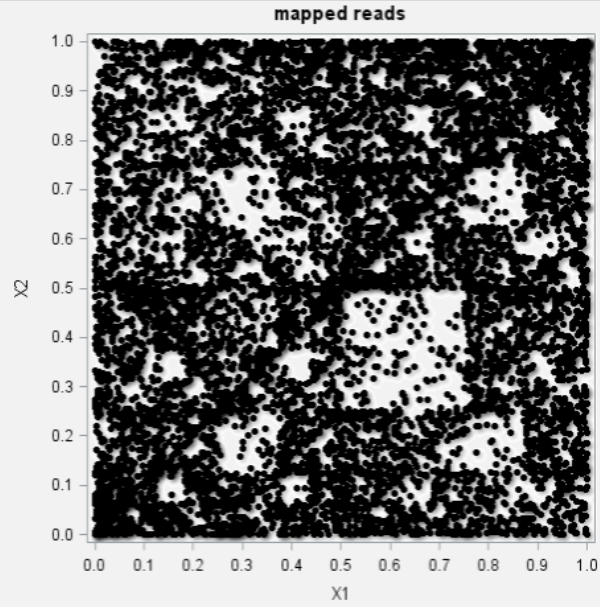
- define sequence coordinates

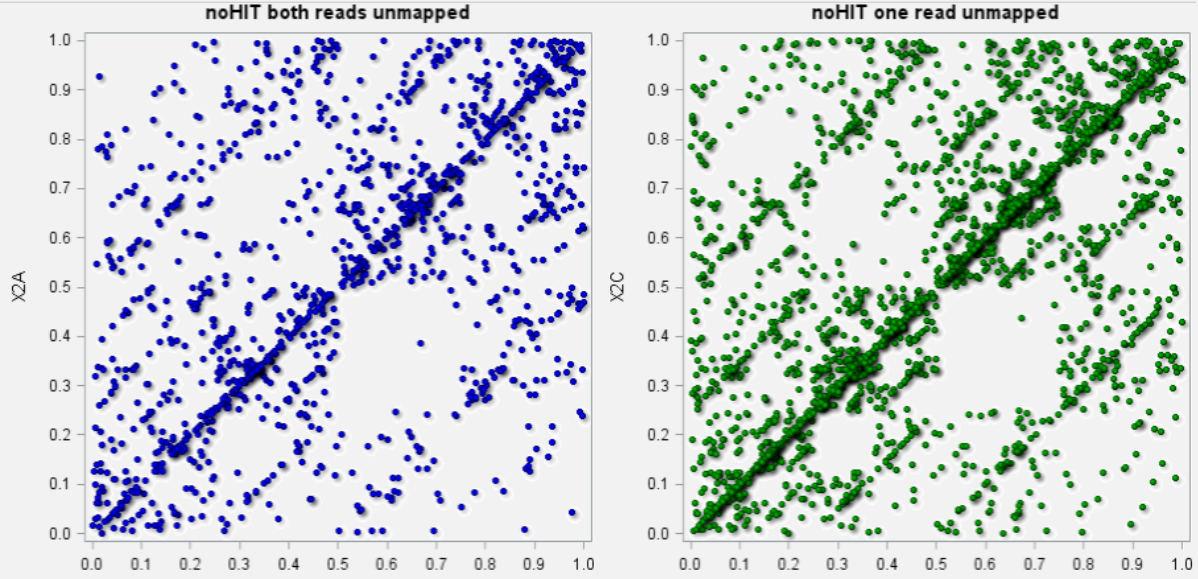


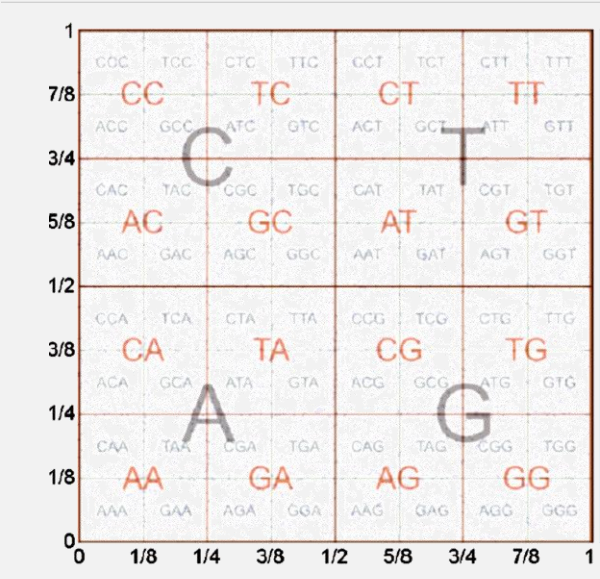
- $$\begin{cases} x_0 = \left(\frac{1}{2}, \frac{1}{2}\right) \\ x_i = x_{i-1} + \frac{1}{2}(\gamma_i - x_{i-1}), i = 1, \ldots, N \end{cases} \quad \gamma_i = \begin{cases} (0,0)\ if\ A \\ (0,1)\ if\ C \\ (1,0)\ if\ G \\ (1,1)\ if\ T \end{cases}$$

Still no HIT …… pattern mining





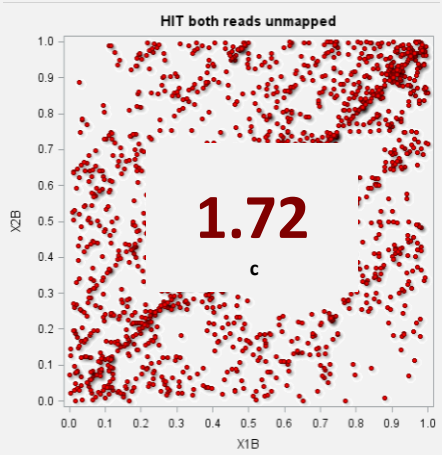


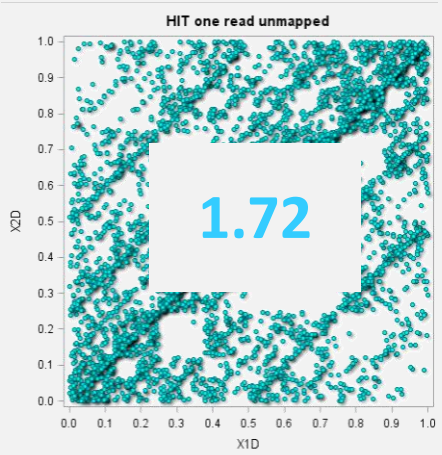


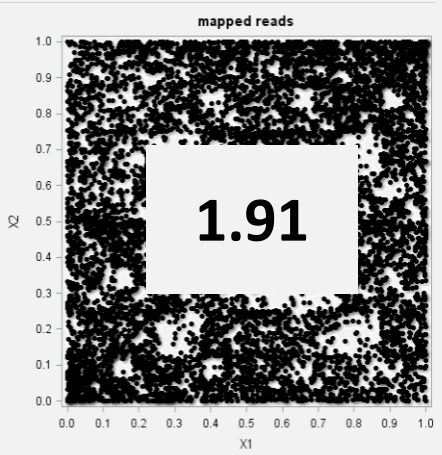
Still no HIT …… sequence complexity

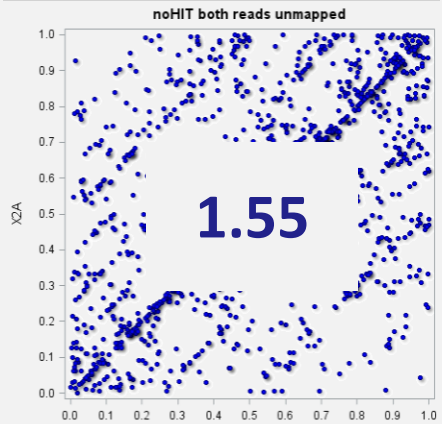
Sequence complexity →Shannon entropy score

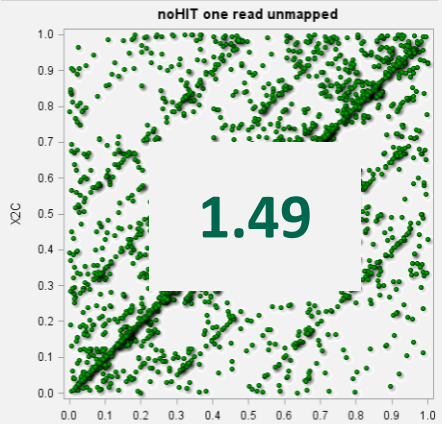
- $H = -\sum_{i=1}^{4}[p_i \, log_2(p_i)] \quad i \in \{A, C, G, T\}$











noHIT

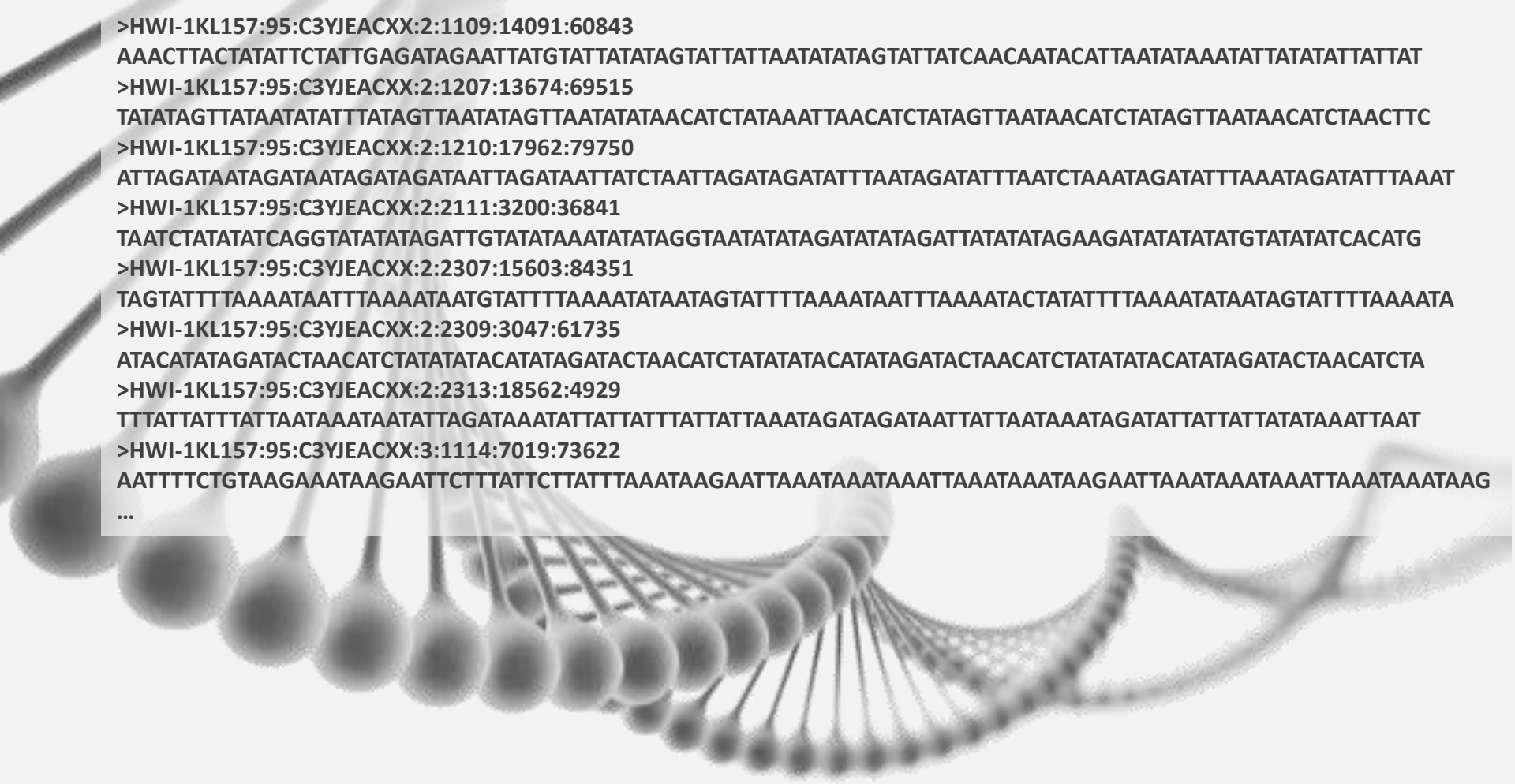
↓

lower complexity

Conclusions

1. **Some** of the unmapped reads contain biological information
   - Evidence of sample contamination
   - Evidence on pathogen infection
   - No evidence of horizontal gene transfer

2. **Some** of the unmapped reads → unknown origin
   - Lower sequence complexity
   - Individual genetic variation
   - Imperfect reference genome: gaps, individual variation



>HWI-1KL157:95:C3YJEACXX:2:1109:14091:60843
AAACTTACTATATTCTATTGAGATAGAATTATGTATTATATAGTATTATTAATATATAGTATTATCAACAATACATTAATATAAATATTATATATTATTAT
>HWI-1KL157:95:C3YJEACXX:2:1207:13674:69515
TATATAGTTATAATATATTTATAGTTAATATAGTTAATATATAACATCTATAAATTAACATCTATAGTTAATAACATCTATAGTTAATAACATCTAACTTC
>HWI-1KL157:95:C3YJEACXX:2:1210:17962:79750
ATTAGATAATAGATAATAGATAGATAATTAGATAATTATCTAATTAGATAGATATTTAATAGATATTTAATCTAAATAGATATTTAAATAGATATTTAAAT
>HWI-1KL157:95:C3YJEACXX:2:2111:3200:36841
TAATCTATATATCAGGTATATATAGATTGTATATAAATATATAGGTAATATATAGATATATAGATTATATATAGAAGATATATATATGTATATATCACATG
>HWI-1KL157:95:C3YJEACXX:2:2307:15603:84351
TAGTATTTTAAAATAATTTAAAATAATGTATTTTAAAATATAATAGTATTTTAAAATAATTTAAAATACTATATTTTAAAATATAATAGTATTTTAAAATA
>HWI-1KL157:95:C3YJEACXX:2:2309:3047:61735
ATACATATAGATACTAACATCTATATATACATATAGATACTAACATCTATATATACATATAGATACTAACATCTATATATACATATAGATACTAACATCTA
>HWI-1KL157:95:C3YJEACXX:2:2313:18562:4929
TTTATTATTTATTAATAAATAATATTAGATAAATATTATTATTTATTATTAAATAGATAGATAATTATTAATAAATAGATATTATTATTATATAAATTAAT
>HWI-1KL157:95:C3YJEACXX:3:1114:7019:73622
AATTTTCTGTAAGAAATAAGAATTCTTTATTCTTATTTAAATAAGAATTAAATAAATAAATTAAATAAATAAGAATTAAATAAATAAATTAAATAAATAAG
…

Thank you for attention

RefSeq vs Nucleotide

1. **RefSeq**
   - non-redundant, well-annotated set of sequences, including genomic DNA, transcripts
   - provides a stable reference for genome annotation, gene identification, gene characterization, mutation and polymorphism analysis, etc.
   - curated
2. **Nocleotide**
   - collection of sequences from several sources, including GenBank, RefSeq, TPA and PDB
   - Collection including International Nucleotide Sequence Database Collaboration, comprising the DNA DataBank of Japan, the European Nucleotide Archive, and GenBank
   - Not curated

BLAST vs BLAT

1. **BLAST** (Basic Local Alignment Search Tool)
   - finds regions of **local similarity** among nucleotide sequences
   - Indexes the query sequence → scans a data base (RefSeq / Nucleotide)
   - searching more distantly related sequences
2. **BLAT** (BLAST-like Alignment Tool)
   - needs an **exact** or **nearly-exact** match to find a hit
   - Indexes the data base → scans the query sequence
   - fast
3. For short sequences not much difference expected