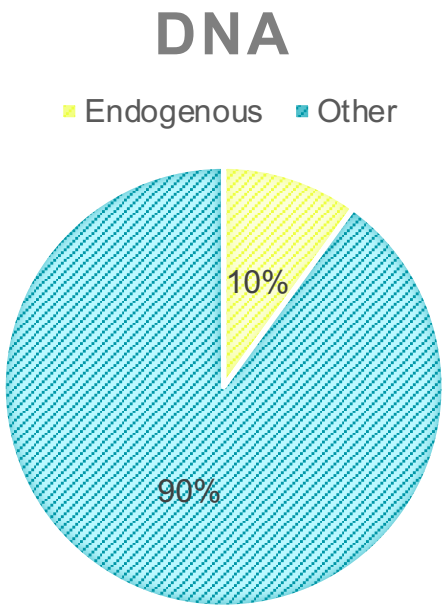
# Taxonomic classifiers

IMPRS Lecture – Maxime Borry

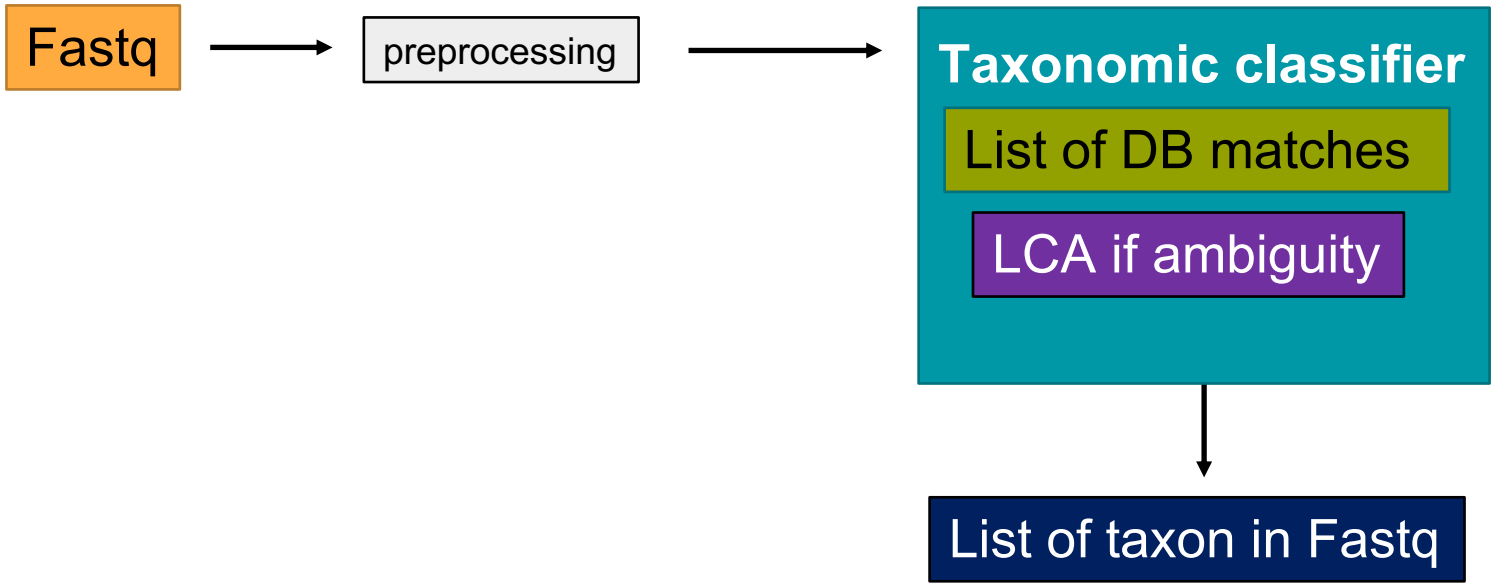
# Why do I care ?



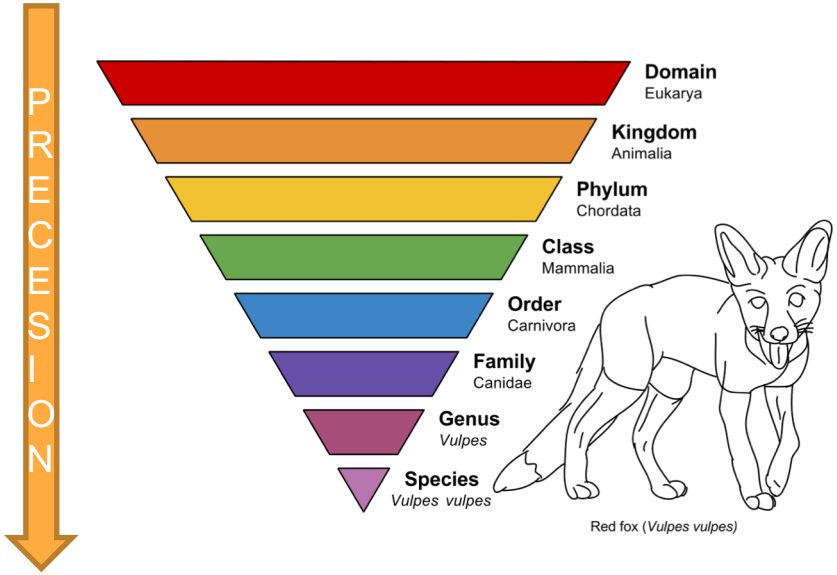

To extract information from the 90% ”other”, you need a taxonomic classifier to answer the question:

*Who is there ?*

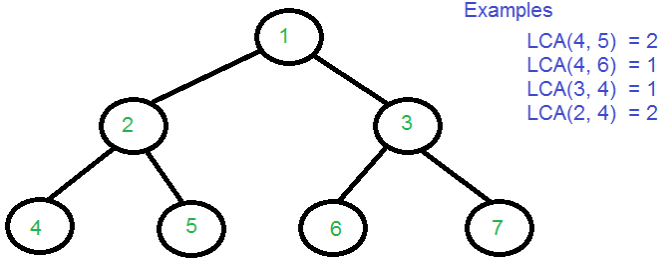
# What is a taxonomic classifier ?



# Why not species classifier ?

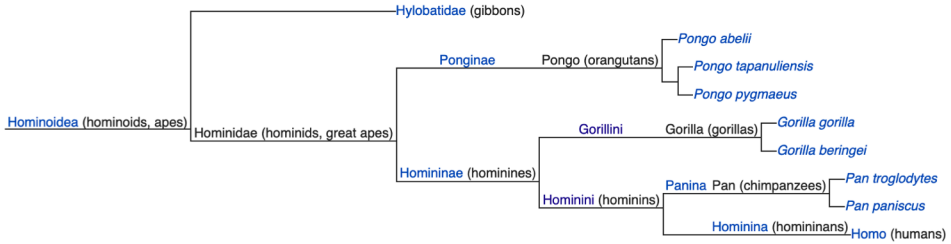



- Species level assignation is not always possible.
- Possibility of hits in different species
- Ambiguities solved by LCA (Lowest Common Ancestor) algorithm.



# LCA example

| Hit | Identity |
|---|---|
| *Pan paniscus* | 97% |
| *Pan troglodytes* | 96% |
| *Homo sapiens* | 92% |
| *Gorilla gorilla* | 87% |

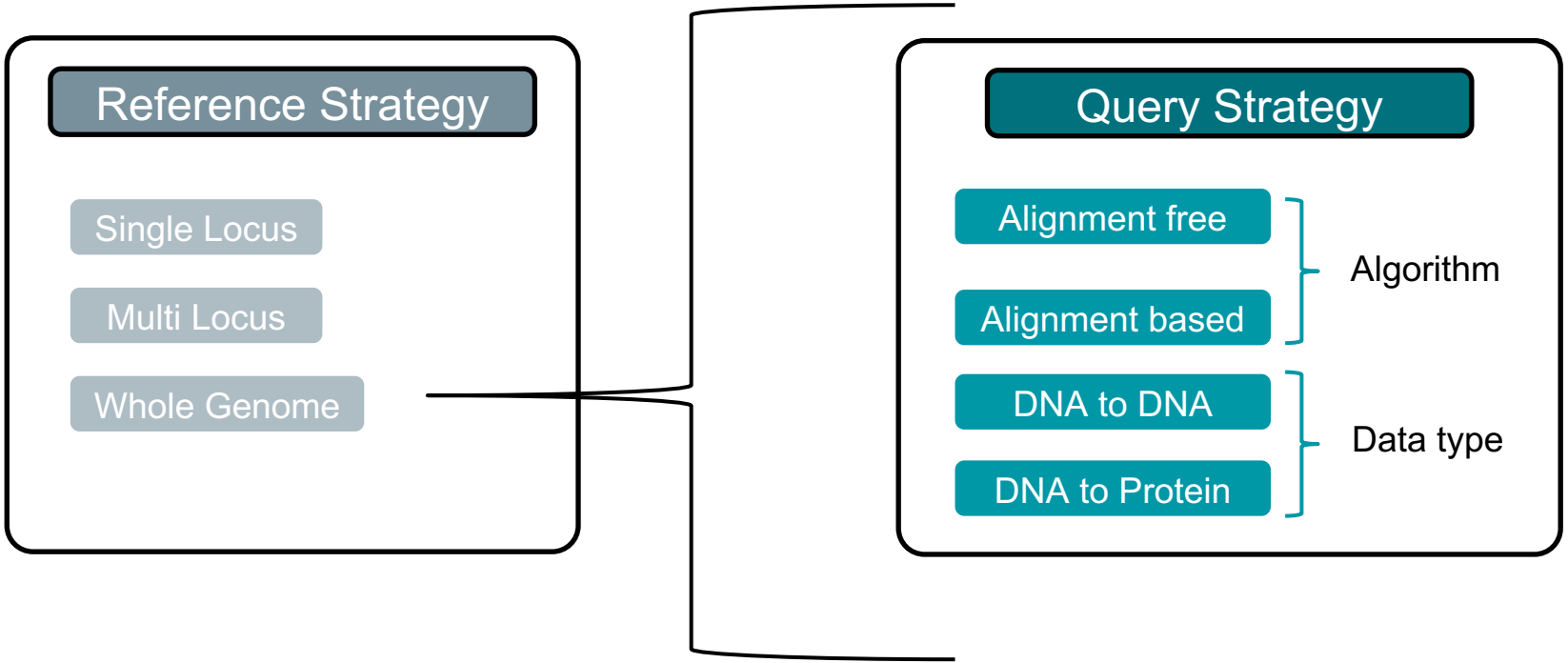



LCA 95% Pan

LCA 90% Hominini

LCA 85% Homininae

# Taxonomic classifiers overview

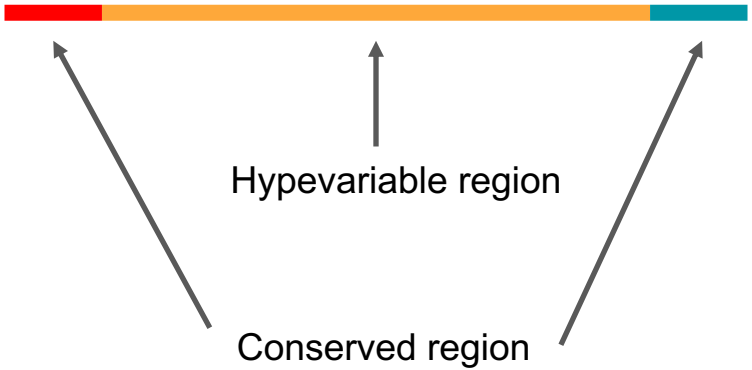


# Reference strategy: single locus

PCR WGS

(Also known as Amplicon metataxonomics, Phylotyping, Metabarcoding)

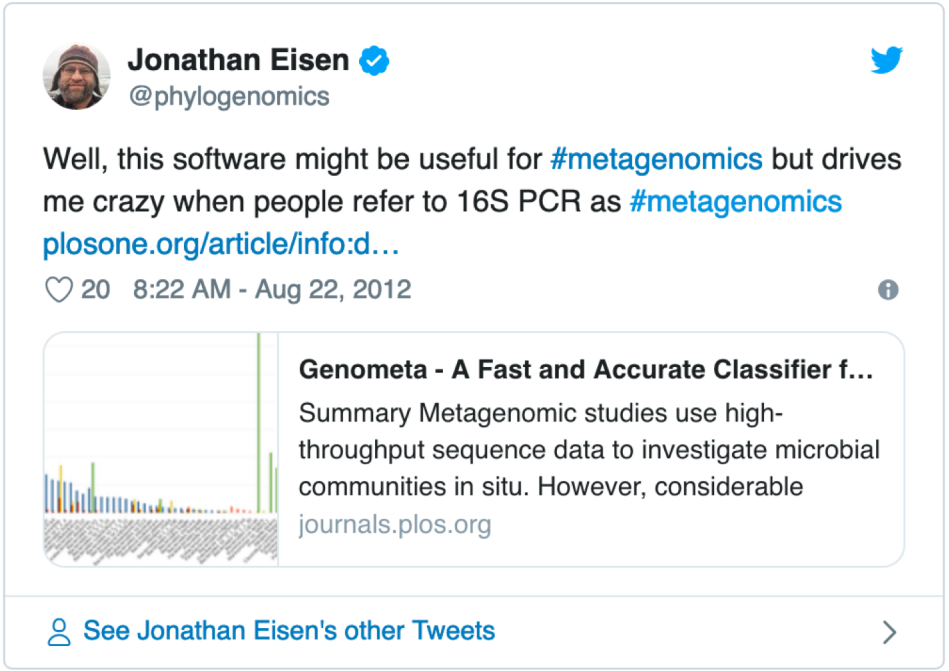
Targeted amplification and deep sequencing of clade-universal gene



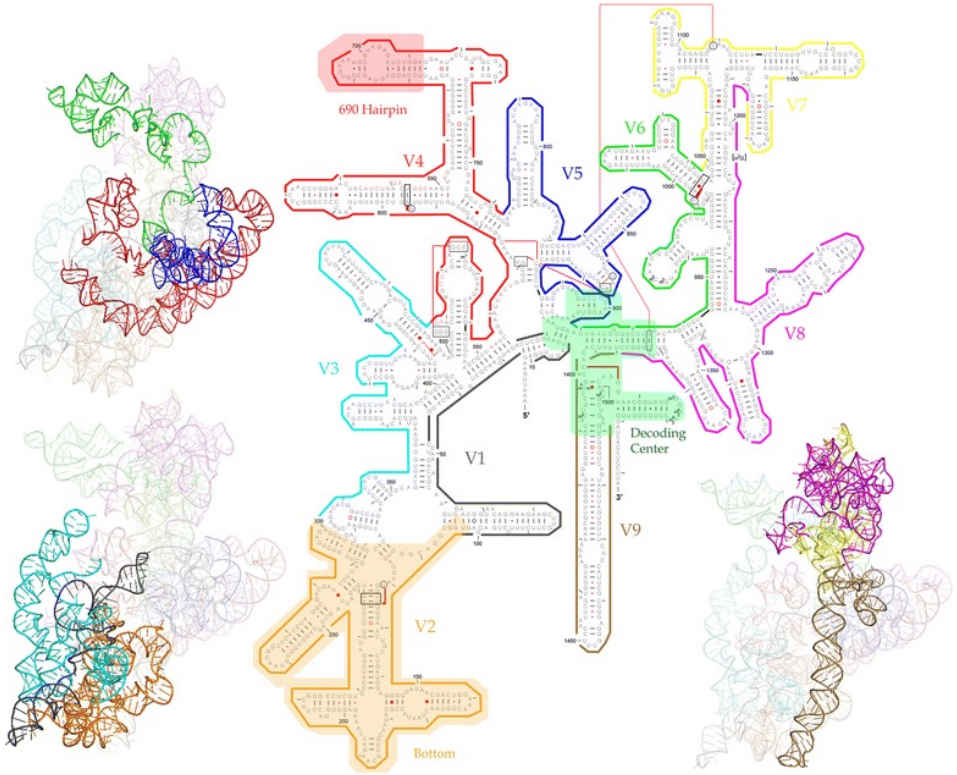

- Amplification of locus by PCR with primers targeting conserved regions or directly from WGS
- (Deep) Sequencing of amplicons
- Comparison to reference marker database

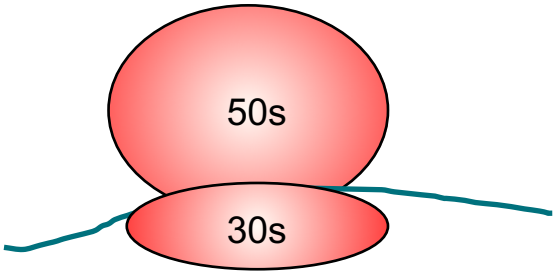
# Side note: vocabulary matters ! (to some)

**Metataxonomics vs Metagenomics**



# Single locus marker genes for bacteria: 16s




**16s rRNA**

- Part of the 30s prokaryotic ribosome subunit

- Stems are more stable -> conserved

- Loops are mutating faster -> variable

Yang, B., Wang, Y., & Qian, P. Y. (2016). Sensitivity and correlation of hypervariable regions in 16S rRNA genes in phylogenetic analysis. *BMC bioinformatics*, *17*(1), 135.

# Single locus taxon assignation

**Spoiler:** Reads are clustered by sequence identity

The example of **QIIME**: OTU picking

**OTU (Operational Taxonomic Unit):** Cluster of Organism grouped by sequence identity (usually 97%)

**De Novo OTU picking:**
Reads are clustered against one another without any external reference sequence collection

**Closed Reference OTU picking**
Reads are clustered against a reference sequence collection and any reads which do not hit a sequence in the reference sequence collection are excluded

**Open Reference OTU picking:**
Reads are clustered against a reference sequence collection and any reads which do not hit the reference sequence collection are subsequently clustered de novo.

# Single locus metaxonomics tools and databases

**Tools**

**Databases**

Qiime 2

Mothur

Dada 2

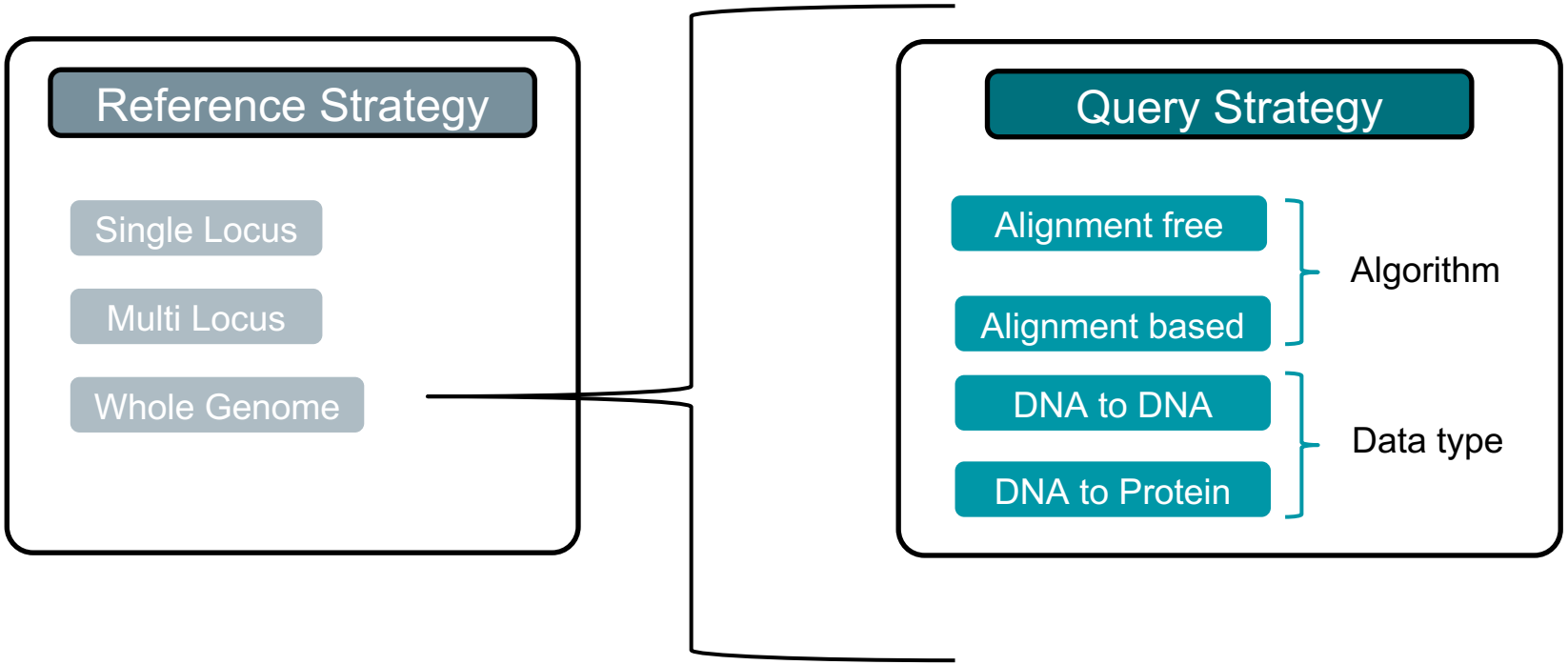
Maxime Borry

# 16s single locus limitations

- Not specific enough for sub-species classification
- Amplification bias: primer binding and GC-content
- If recovered from WGS data, very low amount (~ 0.2 - 0.6 %)

**Limitations specific to ADNA**

- Hypervariable regions length greater than typical aDNA fragment length


Reference Strategy
Single Locus
Multi Locus
Whole Genome
Query Strategy
Alignment free
Alignment based
Algorithm
DNA to DNA
DNA to Protein
Data type

# Reference strategy: multi-locus

WGS

**Use a set of single-copy housekeeping genes**

- Sequence divergence greater than 16s
- Single-copy: access to quantification

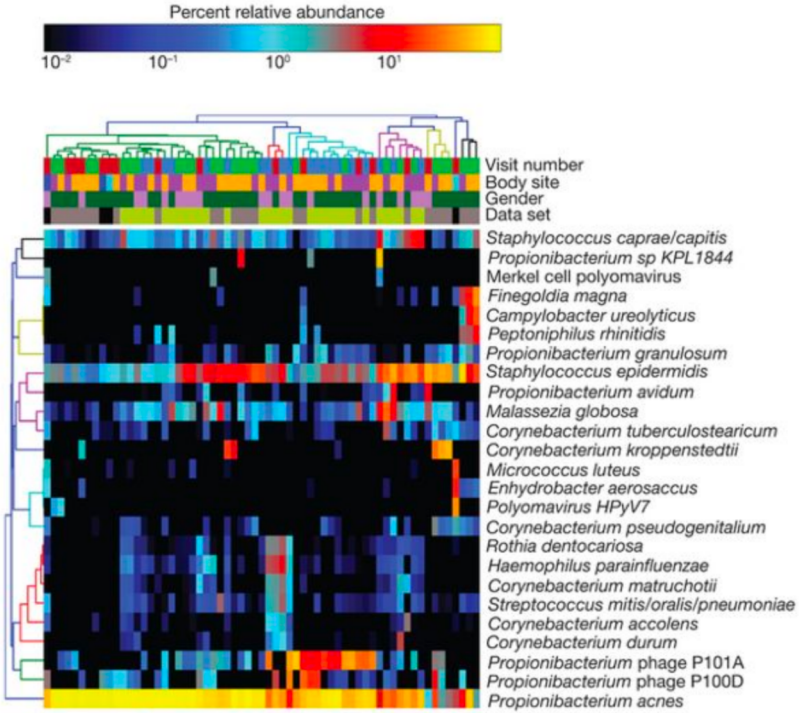
**Metaphlan 2**

- 1 million marker genes
- ~17,000 reference genomes
- ~13,500 bacterial and archaeal
- ~3,500 viral
- ~110 eukaryotic

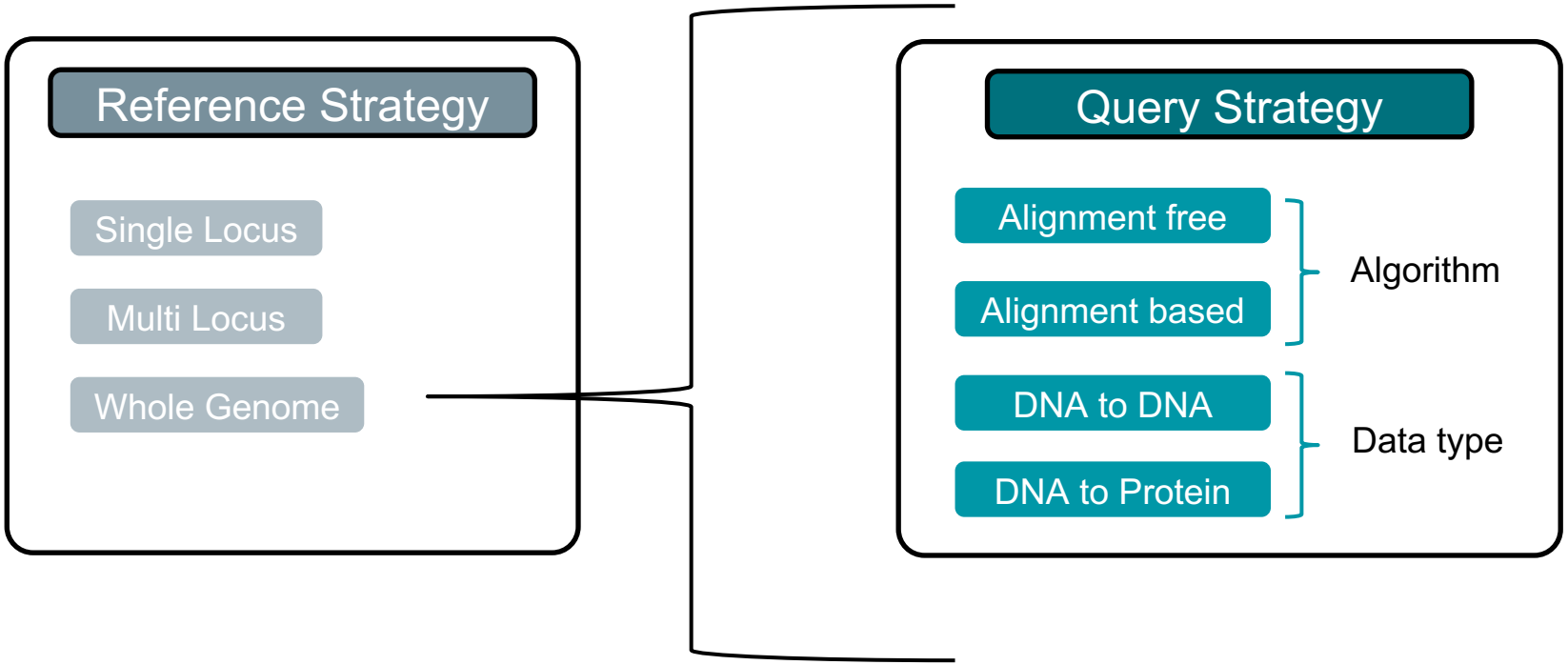
- Uses Bowtie2 for mapping

**Limitations**

- Database not updated: good specificity but poor sensibility




Truong, Duy Tin, et al. "MetaPhlAn2 for enhanced metagenomic taxonomic profiling." *Nature methods* 12.10 (2015): 902.


Reference Strategy
Single Locus
Multi Locus
Whole Genome
Query Strategy
Alignment free
Alignment based
Algorithm
DNA to DNA
DNA to Protein
Data type

# Reference strategy: whole genome

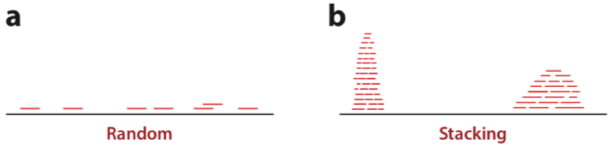
WGS

**Use entire genomes as reference database:**

- Greatest sequence diversity
- Beneficial for aDNA when only traces of ancient organism
- Need heuristic to efficiently search the reference genome database

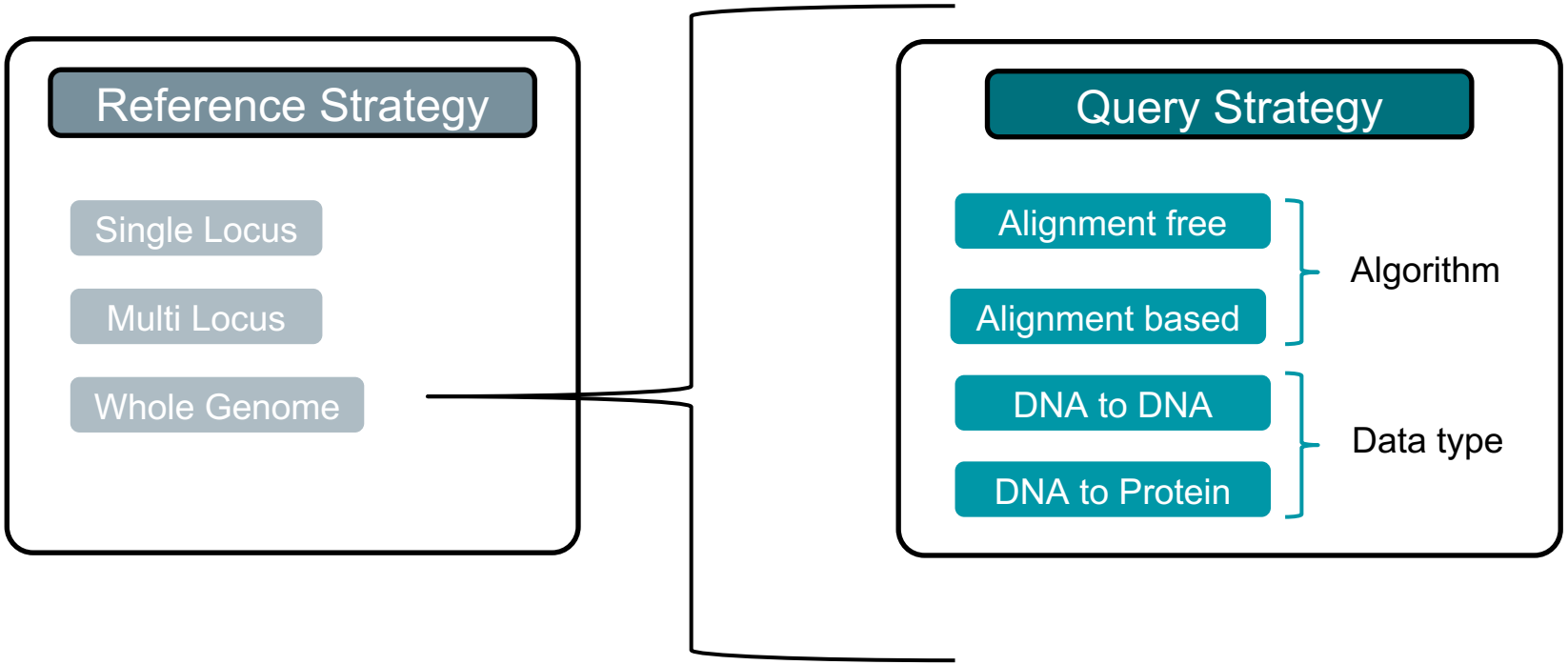
**Limitations:**

- Horizontal gene transfer, mobile elements, recombination decrease precision (b)
- Variable bacterial genome sizes skew DNA proportion estimation
- Sparse database compared to 16s lead to more false assignments




More false negative (not in DB)
More false positives (database bias) because assigned to closest relative

Warinner, Christina, et al. “A Robust Framework for Microbial Archaeology.” *Annual Review of Genomics and Human Genetics*, vol. 18, no. 1, 2017, pp. 321–56. *Annual Reviews*

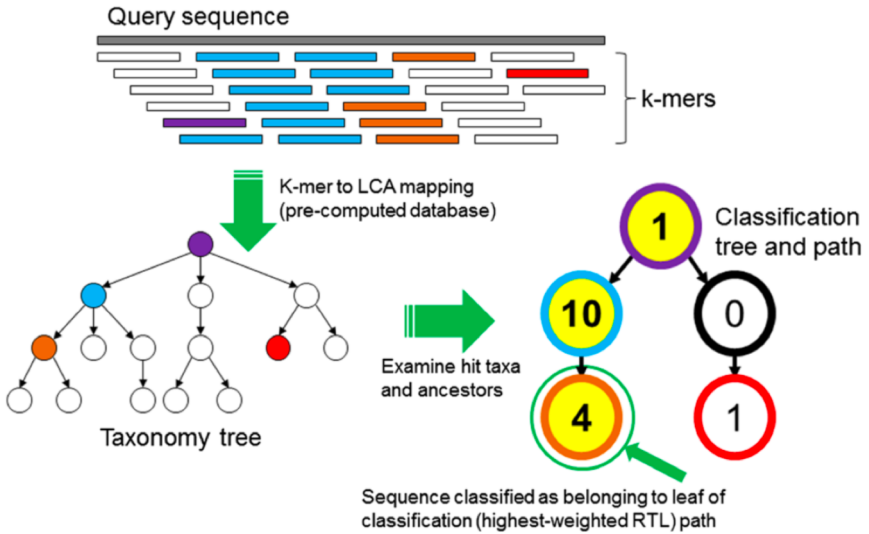

Reference Strategy
Single Locus
Multi Locus
Whole Genome
Query Strategy
Alignment free
Alignment based
Algorithm
DNA to DNA
DNA to Protein
Data type

# Query strategy: Alignment free

**Only look for exact matches between query and database**

**kmer:** substring of size k

- Slice the reference genomes and the query into kmers of fixed size and save them into hash table
- Match kmers in query and reference
- Map the kmers to the taxonomy




Wood, Derrick E., and Steven L. Salzberg. “Kraken: Ultrafast Metagenomic Sequence Classification Using Exact Alignments.” *Genome Biology*, vol. 15, no. 3, Mar. 2014, p. R46. *BioMed Central*

# Alignment free methods

**Kraken**
Taxonomic Sequence Classification System

**Kraken 2**
Taxonomic Sequence Classification System

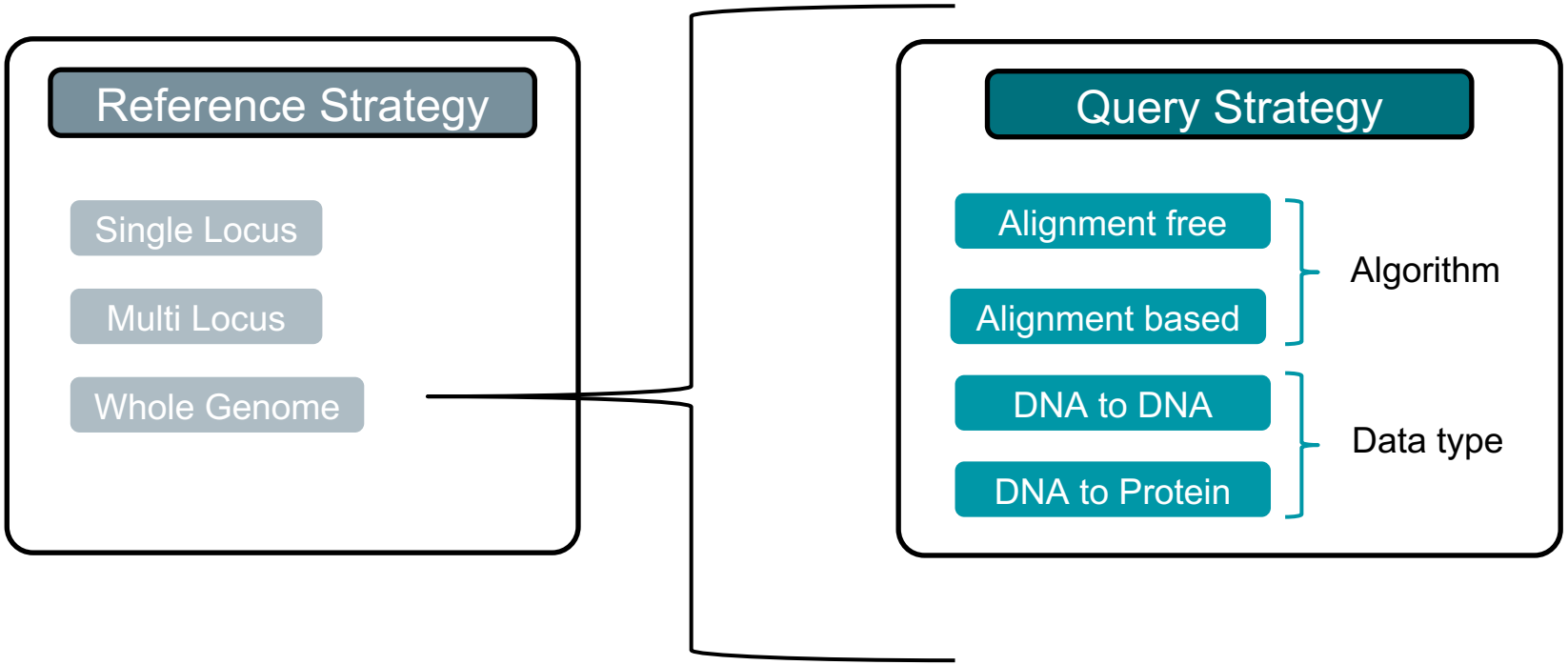
**Centrifuge**
Classifier for metagenomic sequences

**KrakenUniq: confident and fast metagenomics classification using unique k-mer counts**

**CLARK**
Fast, accurate and versatile sequence classification system

# Alignment based advantages and limitations

| Very quick |
| --- |
| - **Sensitivity and specify depend on k:**<br>- Low k: more sensitive, less specific<br>- High k: less sensitive, more specific<br><br>- **No alignment: can't check for DNA damages or functions** |


Reference Strategy
Single Locus
Multi Locus
Whole Genome
Query Strategy
Alignment free
Alignment based
Algorithm
DNA to DNA
DNA to Protein
Data type

# Query strategy: alignment based

## Local vs Global Alignments

**Local:** Start and end alignment at any location in sequence. Algorithm: *Smith-Waterman*

**Global:** Align every nucleotide in every sequence. Algorithm: *Needleman-Wunsch*

### Seed and Extend

1. Start with exact (or near exact) kmer matching – **Very fast**
2. Extend match on both sides if not too many mismatches then trigger alignment **– Slower**

# The Smith-Waterman algorithm

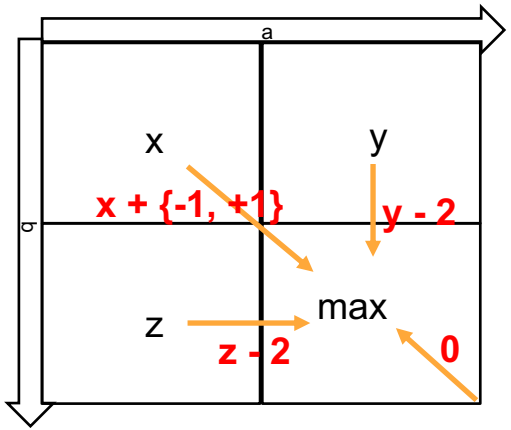
Temple Ferris Smith

Michael Waterman

Smith, T. F., & Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of molecular biology*, *147*(1), 195-197.

How to align these two sequences ? ATGC / ATCC

$s_{match}$= +1
$s_{mismatch}$ = -1
$s_{gap}$ = -2



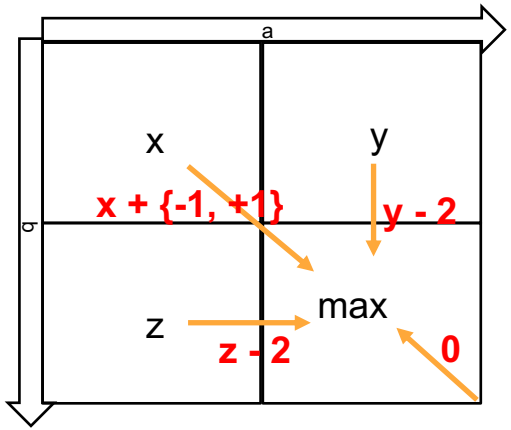

| | | $A_{i=1}$ | $T_{i=2}$ | $G_{i=3}$ | $C_{i=4}$ |
|---|---|---|---|---|---|
| | 0 | | | | |
| $A_{j=1}$ | | | | | |
| $T_{j=2}$ | | | | | |
| $C_{j=3}$ | | | | | |
| $C_{j=4}$ | | | | | |

# The Smith-Waterman algorithm

**Scoring**

$s_{match} = +1$
$s_{mismatch} = -1$
$s_{gap} = -2$

**Rule**




**Matrix**

| | | $A_{i=1}$ | $T_{i=2}$ | $G_{i=3}$ | $C_{i=4}$ |
|---|---|---|---|---|---|
| | 0 | 0 | 0 | 0 | 0 |
| $A_{j=1}$ | 0 | 1 | 0 | 0 | 0 |
| $T_{j=2}$ | 0 | 0 | 2 | 0 | 0 |
| $C_{j=3}$ | 0 | 0 | 0 | 1 | 0 |
| $C_{j=4}$ | 0 | 0 | 0 | 0 | 2 |

# The Smith-Waterman algorithm

Traceback from (all) max scores, back to cell that gave maximum until reaching a 0

| | | $A_{i=1}$ | $T_{i=2}$ | $G_{i=3}$ | $C_{i=4}$ |
|---|---|---|---|---|---|
| | 0 | 0 | 0 | 0 | 0 |
| $A_{j=1}$ | 0 | 1 | 0 | 0 | 0 |
| $T_{j=2}$ | 0 | 0 | 2 | 0 | 0 |
| $C_{j=3}$ | 0 | 0 | 0 | 1 | 0 |
| $C_{j=4}$ | 0 | 0 | 0 | 0 | 2 |

```
ATGC
||*|
ATCC
```

# Your turn !

| | | G | T | T | G | A | C |
|---|---|---|---|---|---|---|---|
| | 0 | | | | | | |
| G | | | | | | | |
| T | | | | | | | |
| T | | | | | | | |
| A | | | | | | | |
| C | | | | | | | |

# Solution

| | | G | T | T | G | A | C |
|---|---|---|---|---|---|---|---|
| | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| G | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| T | 0 | 0 | 2 | 1 | 0 | 0 | 0 |
| T | 0 | 0 | 1 | 3 | 1 | 0 | 0 |
| A | 0 | 0 | 0 | 1 | 2 | 2 | 0 |
| C | 0 | 0 | 0 | 0 | 0 | 0 | 3 |

**Solution 1**

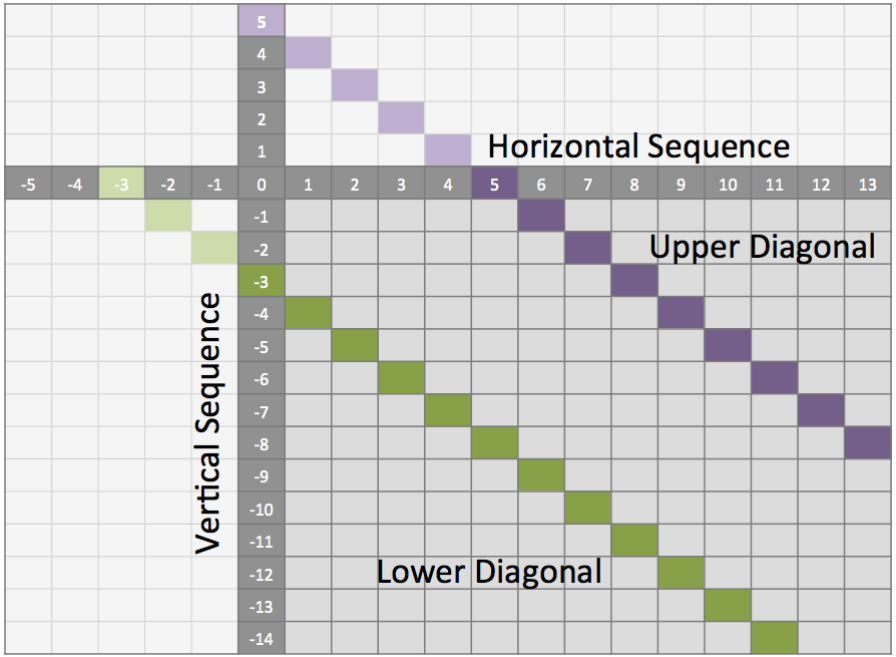
```
GTT
|||
GTT
```

**Solution 2**

```
GTTGAC
|||  ||
GTT-AC
```

**Interactive Smith-Waterman:**
rna.informatik.uni-freiburg.de/Teaching/index.jsp?toolName=Smith-Waterman

# Optimization: banded alignments

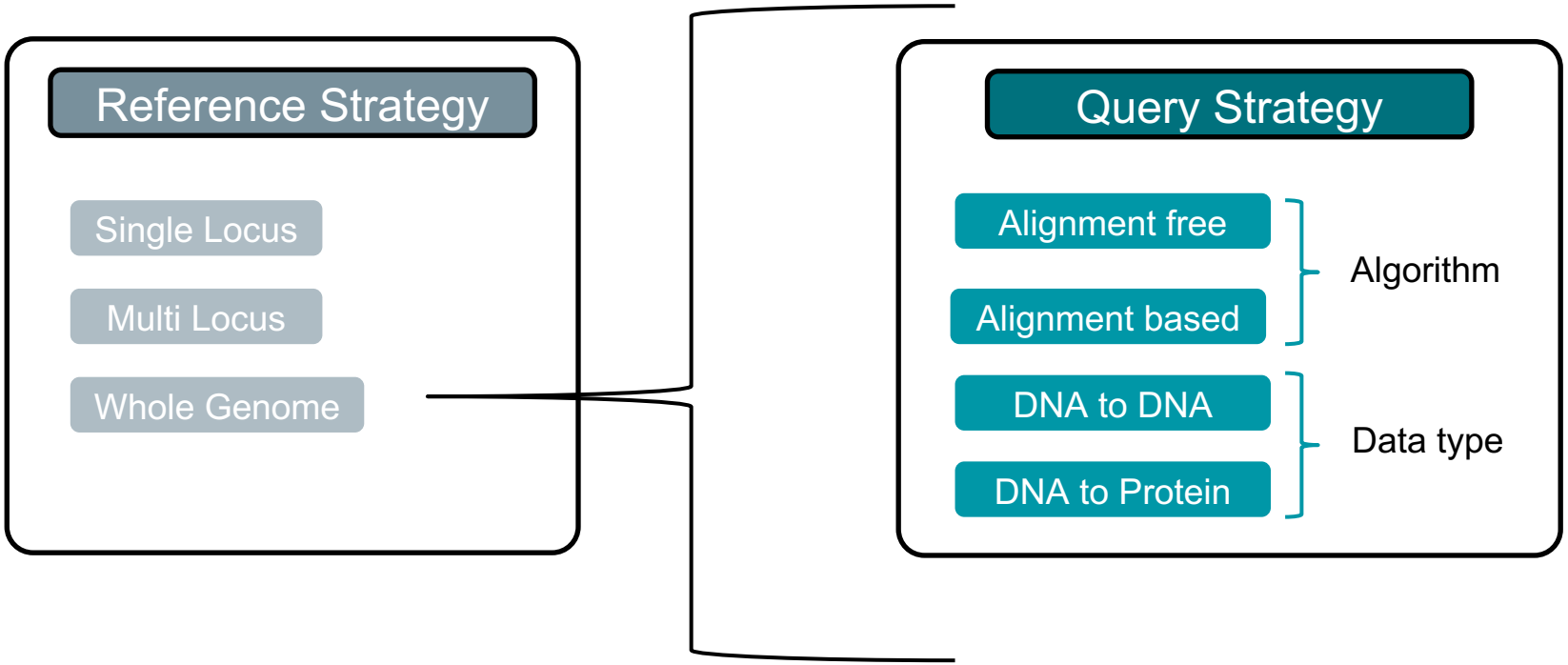



Döring, A., Weese, D., Rausch, T., & Reinert, K. (2008). SeqAn an efficient, generic C++ library for sequence analysis. *BMC bioinformatics*, *9*(1), 11.

# Alignment based: advantages and limitations

**Alignment:**
**- Check for DNA damages**
**- Check for functions**

**Much (much) slower**

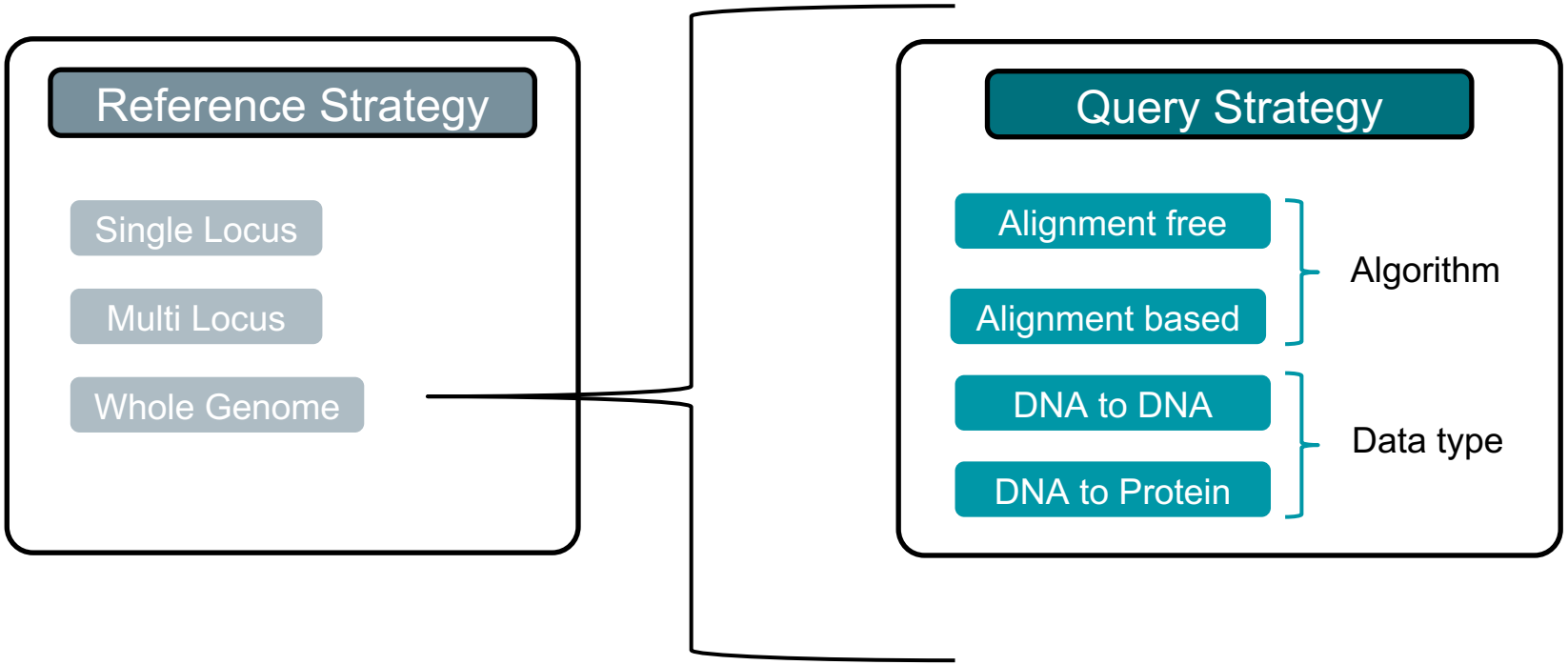

Reference Strategy
Single Locus
Multi Locus
Whole Genome
Query Strategy
Alignment free
Alignment based
Algorithm
DNA to DNA
DNA to Protein
Data type

# DNA based alignment

MALT: **ME**GAN **AL**ignment **T**ool:

- Near exact seeds (spaced seeds)
- Mismatch tolerant extension (x-drop)
- Smith-Waterman (semi-global: read is aligned end-to-end)
- LCA


Herbig, A., Maixner, F., Bos, K. I., Zink, A., Krause, J., & Huson, D. H. (2016). *MALT: Fast alignment and analysis of metagenomic DNA sequence data applied to the Tyrolean Iceman.*


Reference Strategy
Single Locus
Multi Locus
Whole Genome
Query Strategy
Alignment free
Alignment based
Algorithm
DNA to DNA
DNA to Protein
Data type

# Protein based alignment

```
A +1
T -1 +1
G -1 -1 +1
C -1 -1 -1 +1
   A  T  G  C
```



| | Ala | Arg | Asn | Asp | Cys | Gln | Glu | Gly | His | Ile | Leu | Lys | Met | Phe | Pro | Ser | Thr | Trp | Tyr | Val |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Ala** | 4 | | | | | | | | | | | | | | | | | | | |
| **Arg** | −1 | 5 | | | | | | | | | | | | | | | | | | |
| **Asn** | −2 | 0 | 6 | | | | | | | | | | | | | | | | | |
| **Asp** | −2 | −2 | 1 | 6 | | | | | | | | | | | | | | | | |
| **Cys** | 0 | −3 | −3 | −3 | 9 | | | | | | | | | | | | | | | |
| **Gln** | −1 | 1 | 0 | 0 | −3 | 5 | | | | | | | | | | | | | | |
| **Glu** | −1 | 0 | 0 | 2 | −4 | 2 | 5 | | | | | | | | | | | | | |
| **Gly** | 0 | −2 | 0 | −1 | −3 | −2 | −2 | 6 | | | | | | | | | | | | |
| **His** | −2 | 0 | 1 | −1 | −3 | 0 | 0 | −2 | 8 | | | | | | | | | | | |
| **Ile** | −1 | −3 | −3 | −3 | −1 | −3 | −3 | −4 | −3 | 4 | | | | | | | | | | |
| **Leu** | −1 | −2 | −3 | −4 | −1 | −2 | −3 | −4 | −3 | 2 | 4 | | | | | | | | | |
| **Lys** | −1 | 2 | 0 | −1 | −3 | 1 | 1 | −2 | −1 | −3 | −2 | 5 | | | | | | | | |
| **Met** | −1 | −1 | −2 | −3 | −1 | 0 | −2 | −3 | −2 | 1 | 2 | −1 | 5 | | | | | | | |
| **Phe** | −2 | −3 | −3 | −3 | −2 | −3 | −3 | −3 | −1 | 0 | 0 | −3 | 0 | 6 | | | | | | |
| **Pro** | −1 | −2 | −2 | −1 | −3 | −1 | −1 | −2 | −2 | −3 | −3 | −1 | −2 | −4 | 7 | | | | | |
| **Ser** | 1 | −1 | 1 | 0 | −1 | 0 | 0 | 0 | −1 | −2 | −2 | 0 | −1 | −2 | −1 | 4 | | | | |
| **Thr** | 0 | −1 | 0 | −1 | −1 | −1 | −1 | −2 | −2 | −1 | −1 | −1 | −1 | −2 | −1 | 1 | 5 | | | |
| **Trp** | −3 | −3 | −4 | −4 | −2 | −2 | −3 | −2 | −2 | −3 | −2 | −3 | −1 | 1 | −4 | −3 | −2 | 11 | | |
| **Tyr** | −2 | −2 | −2 | −3 | −2 | −1 | −2 | −3 | 2 | −1 | −1 | −2 | −1 | 3 | −3 | −2 | −2 | 2 | 7 | |
| **Val** | 0 | −3 | −3 | −3 | −1 | −2 | −2 | −3 | −3 | 3 | 1 | −2 | 1 | −1 | −2 | −2 | 0 | −3 | −1 | 4 |

Same strategy as DNA based alignment, but using Amino acid substitution matrix.

**BLOSUM: BLO**cks **SU**bstitution **M**atrix

**See also:** Optimization with reduced alphabet (grouping amino acids by chemical properties)

# Protein based alignment tools

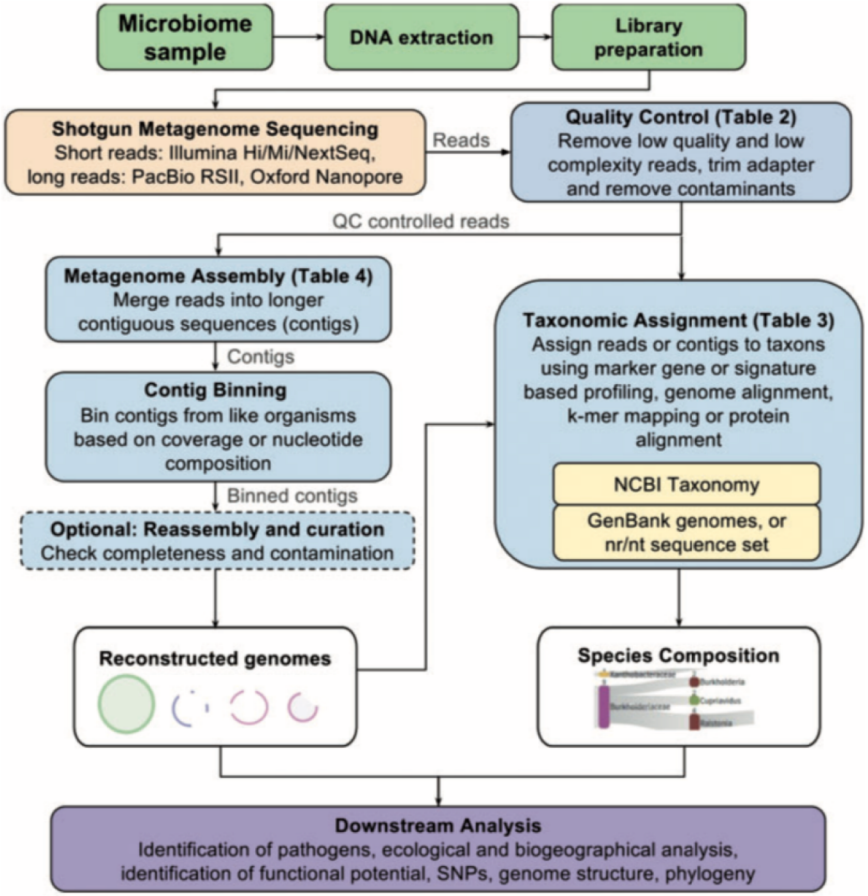
**MALT:**
Protein to Protein Database
DNA to Protein Database

**Diamond:**
Protein to Protein Database
DNA to Protein Database
No LCA (Megan)


No DNA damage estimation with protein alignment

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature methods*, *12*(1), 59.

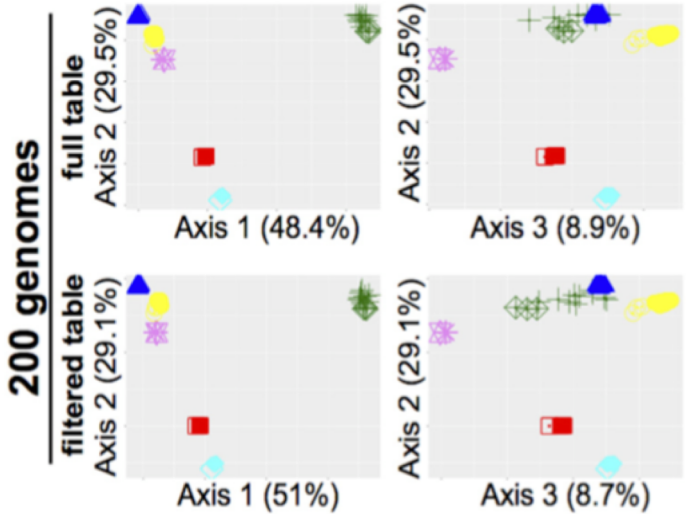
# Other strategies

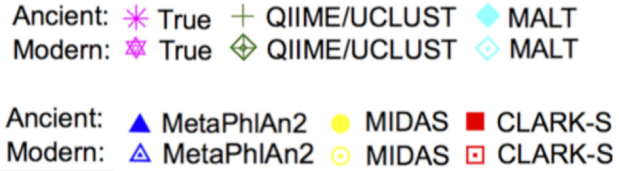



Breitwieser, F. P., Lu, J., & Salzberg, S. L. (n.d.). A review of methods and databases for metagenomic classification and assembly. *Briefings in Bioinformatics*.

# Taxonomic classifiers for aDNA

# Damage isn't really an issue



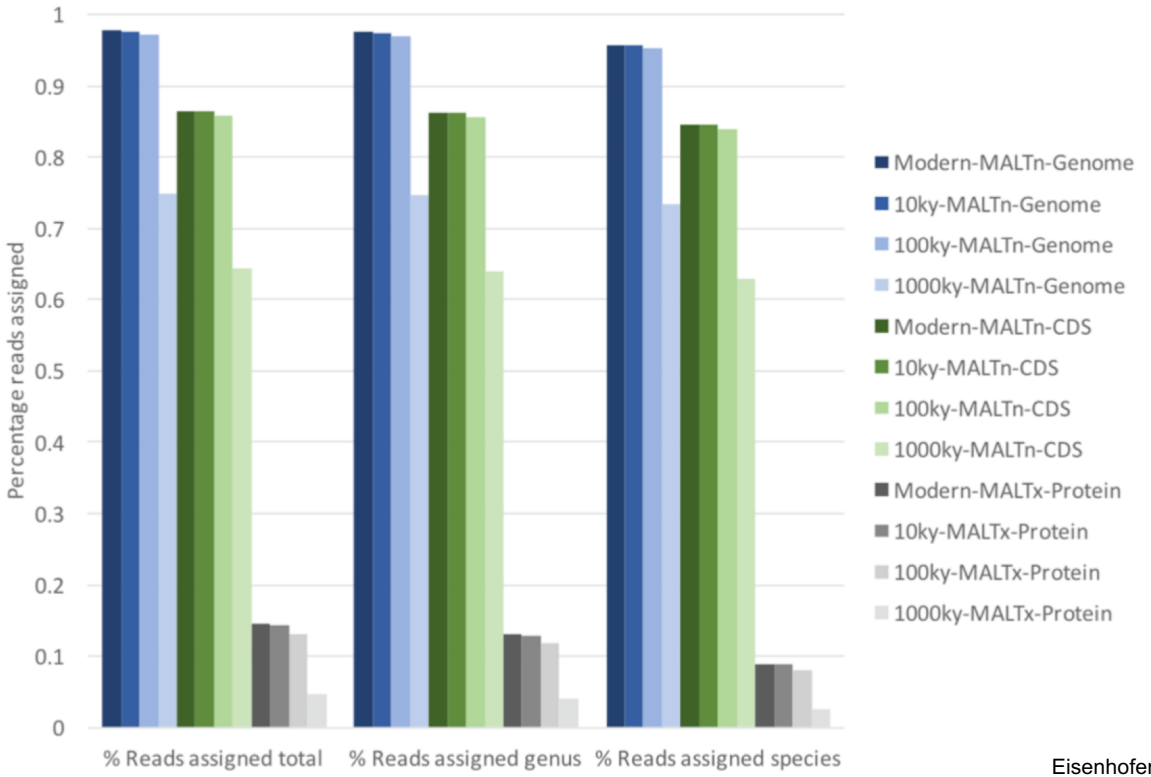




PCoA of simulated community composition with Weighted-Unifrac Distance

**aDNA damage doesn't really affect inferred community composition**

Velsko, I. M., Frantz, L. A. F., Herbig, A., Larson, G., & Warinner, C. (2018). Selection of Appropriate Metagenome Taxonomic Classifiers for Ancient Microbiome Research. *MSystems*, *3*(4).

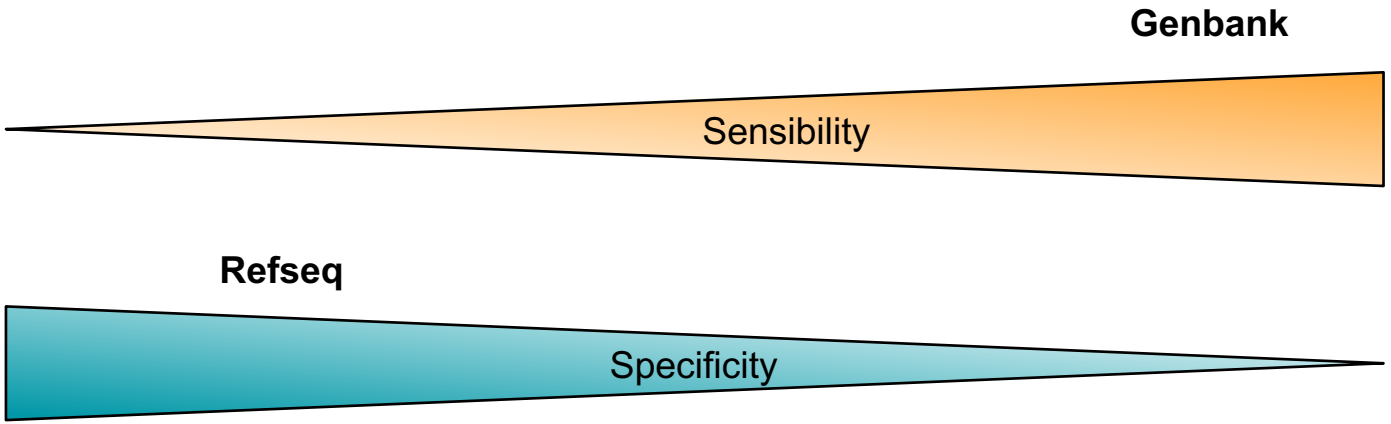
# The choice of database




- DNA database is better for short aDNA (<60 bp) than protein databases
- Translation of shorter sequences gives too small peptides to align
- **Database composition matters**

Eisenhofer, R., & Weyrich, L. S. (2019). Assessing alignment-based taxonomic classification of ancient microbial DNA. *PeerJ*, *7*, e6594.

# Genbank vs Refseq




Breitwieser, F. P., Lu, J., & Salzberg, S. L. (n.d.). A review of methods and databases for metagenomic classification and assembly. *Briefings in Bioinformatics*.

# The definition of species

- **Taxonomic classifiers work with Taxonomy**
- Wanna be a species ? Better be culturable !
- Not culturable but otherwise well characterized : included as *Candidatus Genus species*
- Otherwise: **unclassified bacteria**

Lineage: Cellular Organisms; Bacteria; unclassified bacteria
-> Problem with LCA

Breitwieser, F. P., Lu, J., & Salzberg, S. L. (n.d.). A review of methods and databases for metagenomic classification and assembly. *Briefings in Bioinformatics*.

# Species vs ANI

# 15%

- ANI: Average Nucleotide Identity
- **Gold standard:** 95% ANI for species
- 15% of Genomes in Genbank from the same species have ANI < 93%

Rosselló-Móra, R., & Amann, R. (2015). Past and future species definitions for Bacteria and Archaea. *Systematic and Applied Microbiology*, *38*(4), 209-216.

# Choosing the right tool

**First define the question:**

Am I only asking **who is there** ?

Yes: Locus based or Alignment free method

No: Whole genome DNA alignment based method

Do I have a lot of computing time to spare (and/or not so many samples at once) ?:

Yes: Whole genome DNA alignment based method

No: Locus based or Alignment free method

Do I have WGS sequencing data ?

Yes: Prefer avoiding single locus (open debate)

No: Single locus it is

Do I have long fragments and feel adventurous ?

Yes: Try whole genome protein alignment based method, or even Assembly

No: Stick to the rest


Worst protein alignment ever (score = 2)