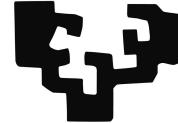




eman ta zabal zazu



Universidad  
del País Vasco

Euskal Herriko  
Unibertsitatea



---

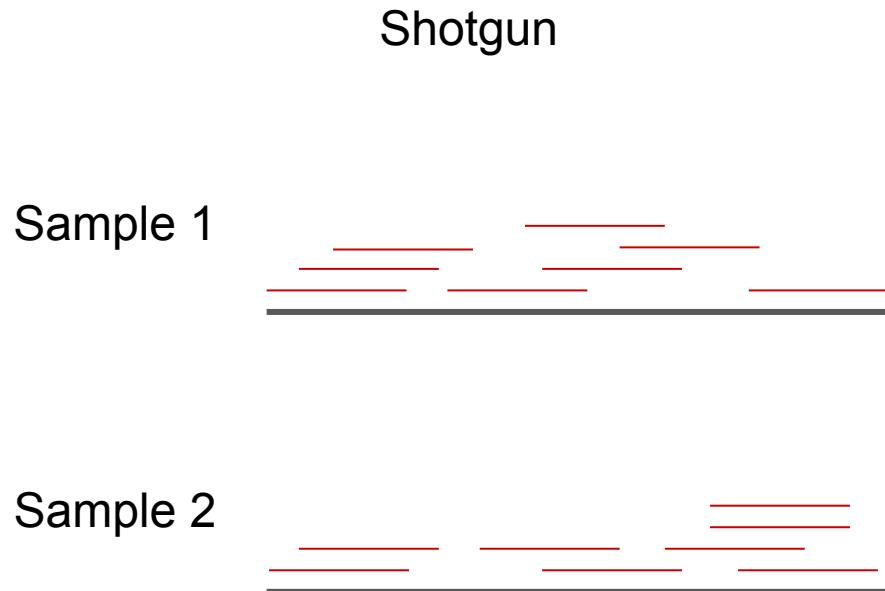
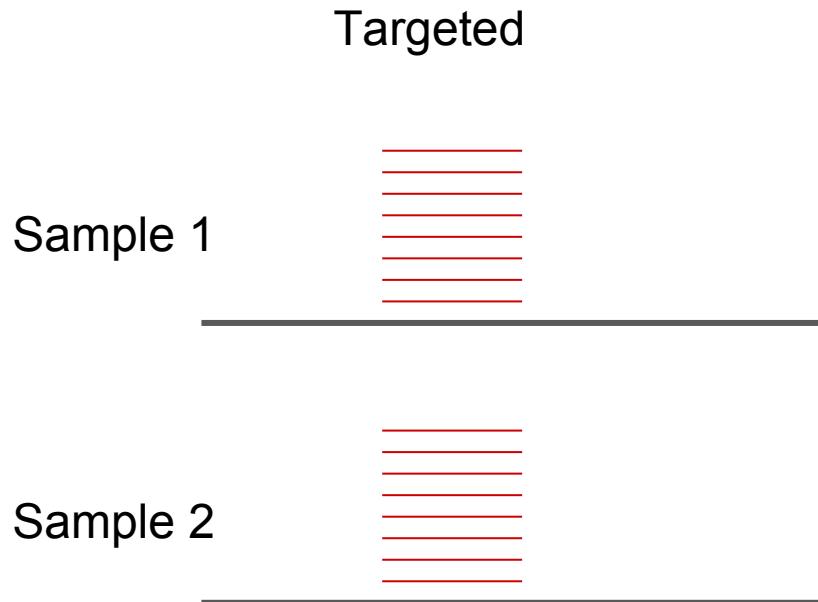
NATURAL HISTORY MUSEUM OF DENMARK  
UNIVERSITY OF COPENHAGEN

# Genomics of Marine Resources Metagenomics

2018 May 14-17

Antton Alberdi - [antton.alberdi@smn.ku.dk](mailto:antton.alberdi@smn.ku.dk)  
Assistant Professor - University of Copenhagen

## 2 Metabarcoding



## Metabarcoding vs. Shotgun sequencing



### Metabarcoding

Small fraction of total DNA

Taxonomic identification

Biases

No quantitative

No functional profiling

### Metagenomics

(Almost) no biases

Quantitative

Functional identification

Large fraction of total DNA

No quantitative

### 3 Metagenomics

## Metabarcoding

Amplicon sequences



Sequence filtering



OTUs



Taxonomic  
assignment

### 3 Metagenomics

## Metabarcoding

Amplicon sequences



Sequence filtering



OTUs



Taxonomic  
assignment



Functional  
assignment

<http://picrust.github.io/picrust/>

## Metabarcoding

Amplicon sequences

Sequence filtering

OTUs

Taxonomic  
assignment

## Metagenomics

Shotgun sequences

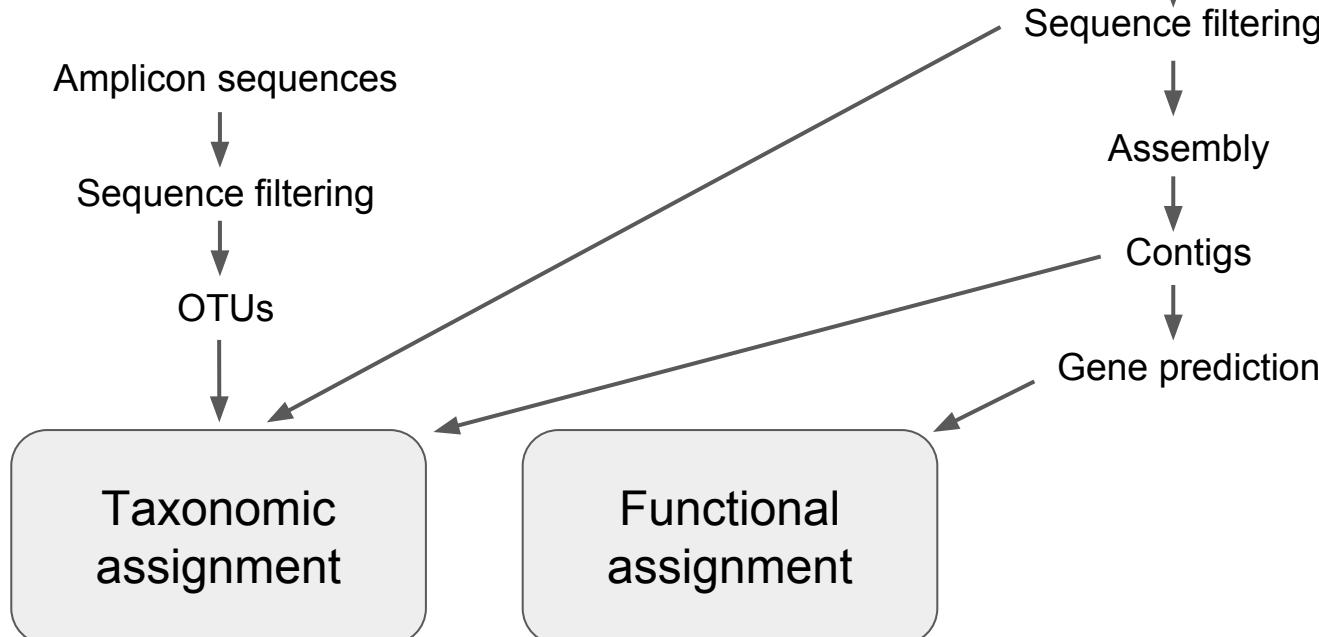
Sequence filtering

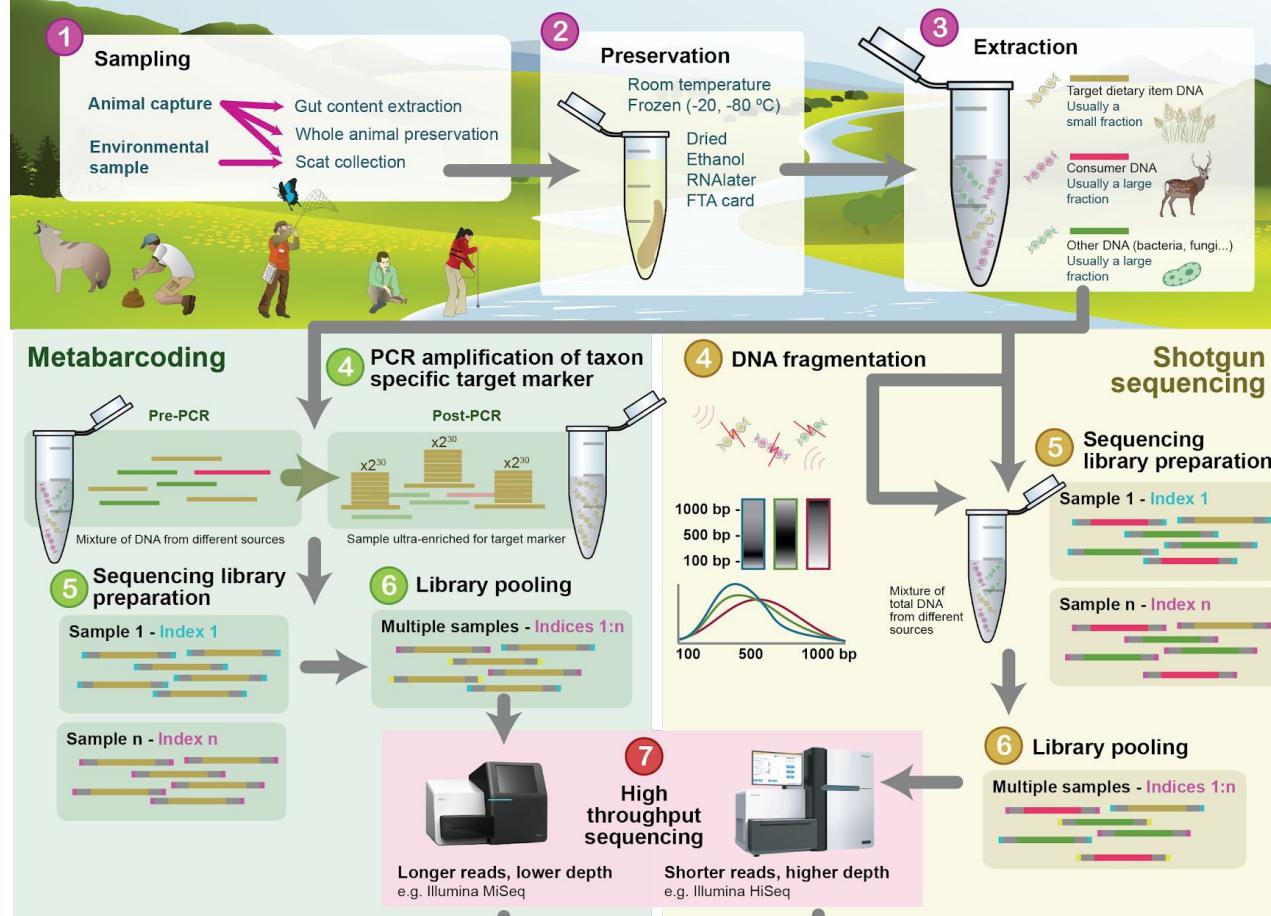
Assembly

Contigs

Gene prediction

Functional  
assignment





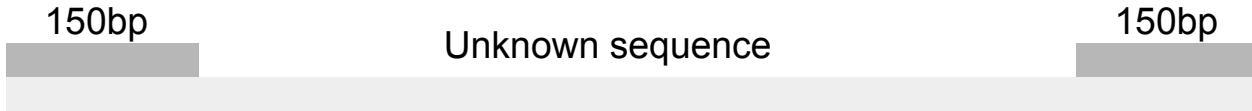
### 3 Metagenomics

200 million reads  
SR or PE  
50bp - 150 bp



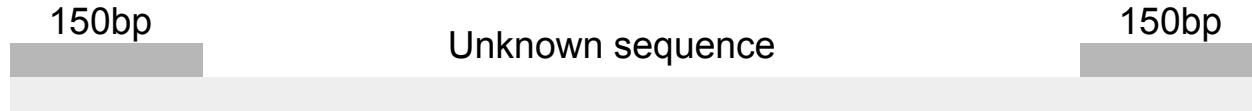
### 3 Metagenomics

## Long DNA fragments



### 3 Metagenomics

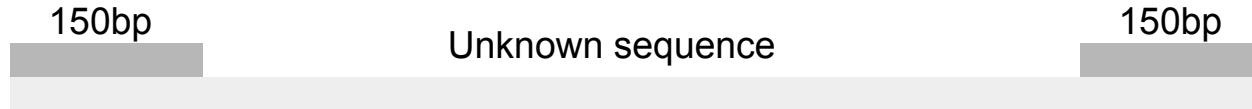
## Long DNA fragments



Useful for de novo genome sequencing  
Not useful for shotgun metagenomics

### 3 Metagenomics

#### Long DNA fragments



#### Short DNA fragments



### 3 Metagenomics

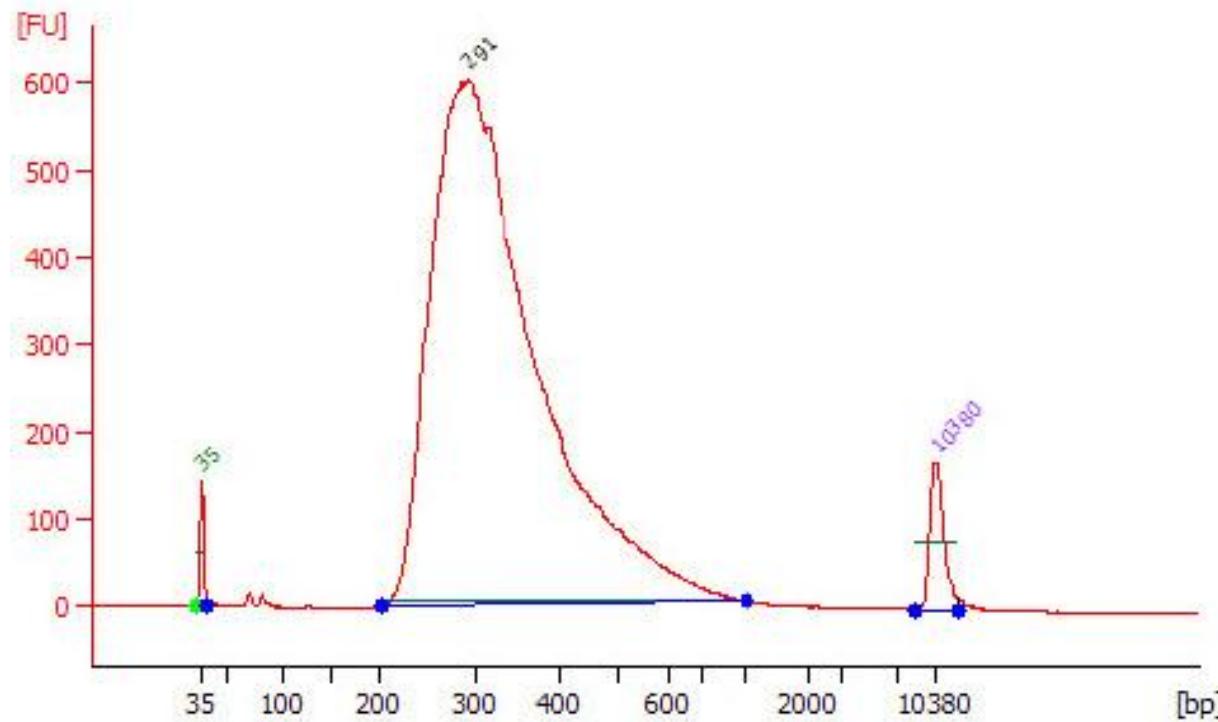
## Sonication



Covaris

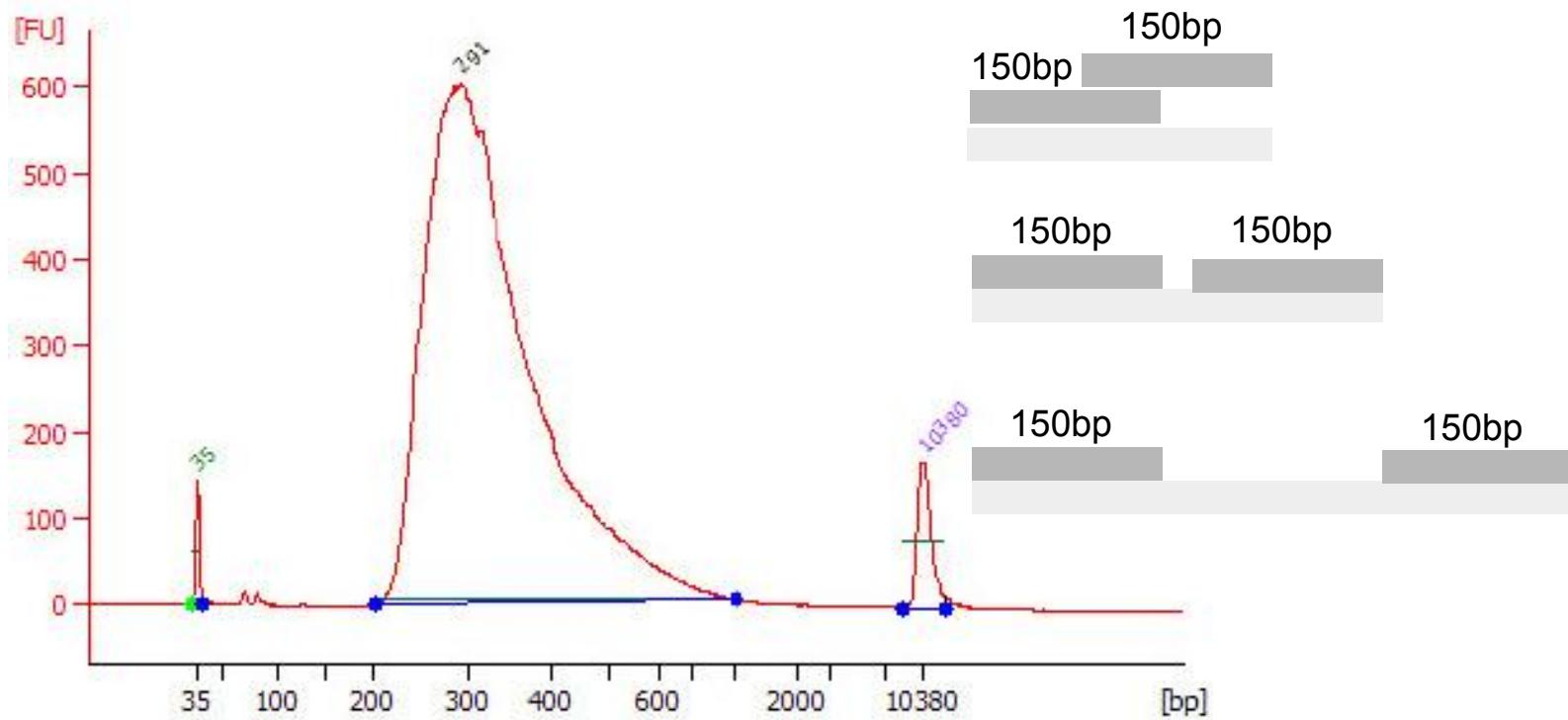


Bioruptor

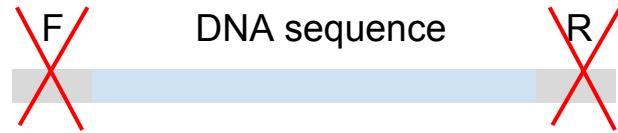
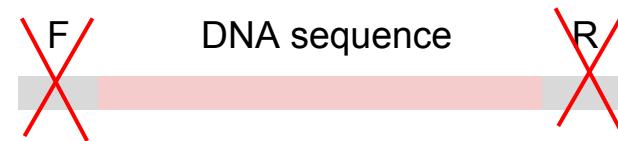
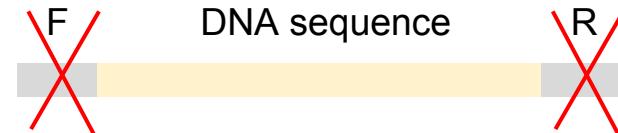


## 3

## Metagenomics



### 3 Metagenomics



DNA sequence



DNA sequence



DNA sequence



Ligate adaptors  
DNA Ligase

### 3 Metagenomics

1 base

1,000 bases

1,000,000 bases

1,000,000,000 bases

1,000,000,000,000 bases

1 kilobase

1 megabase

1 gigabase

1 terabase

### 3 Metagenomics

1 base

1,000 bases

1,000,000 bases

1,000,000,000 bases

1,000,000,000,000 bases

1 kilobase

1 megabase

1 gigabase

1 terabase

Human genome ≈ 3 gigabases

Mouse genome ≈ 2.8 gigabases

### 3 Metagenomics



Metagenomics: 5-8 GB

### 3 Metagenomics



Metagenomics: 5-8 GB

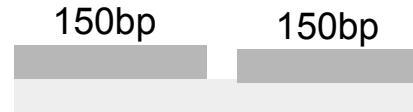
5,000,000,000 bases

### 3 Metagenomics

## Metagenomics: 5-8 GB

5,000,000,000 bases

150PE = 300 bases/read

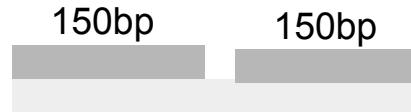


### 3 Metagenomics

## Metagenomics: 5-8 GB

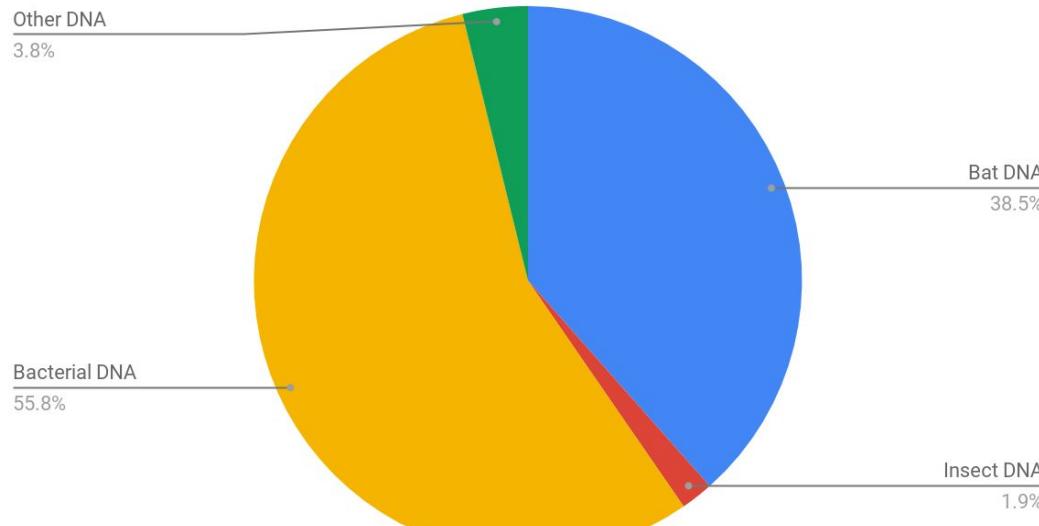
5,000,000,000 bases

150PE = 300 bases/read



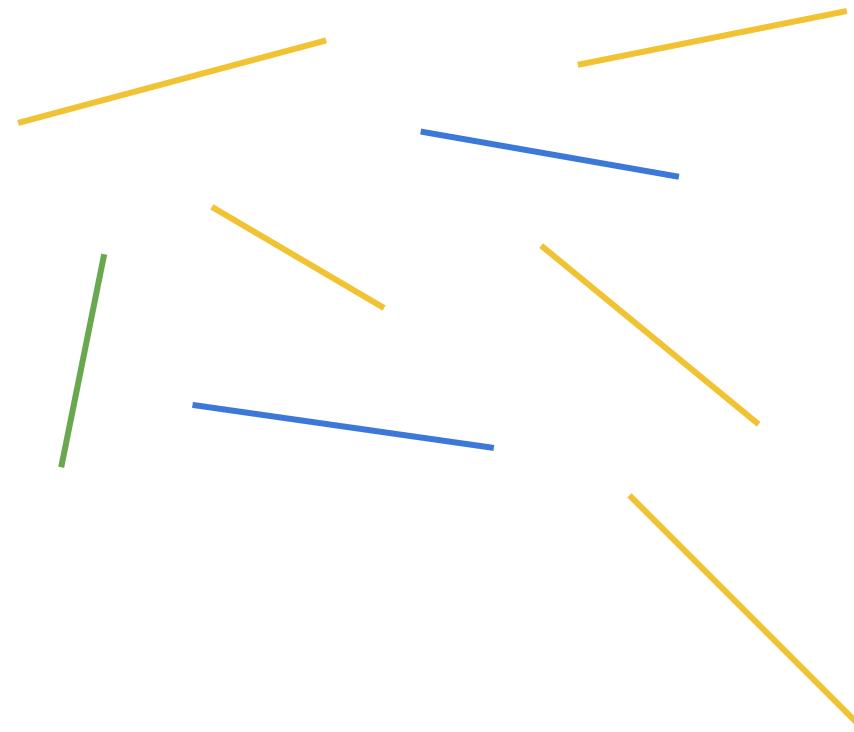
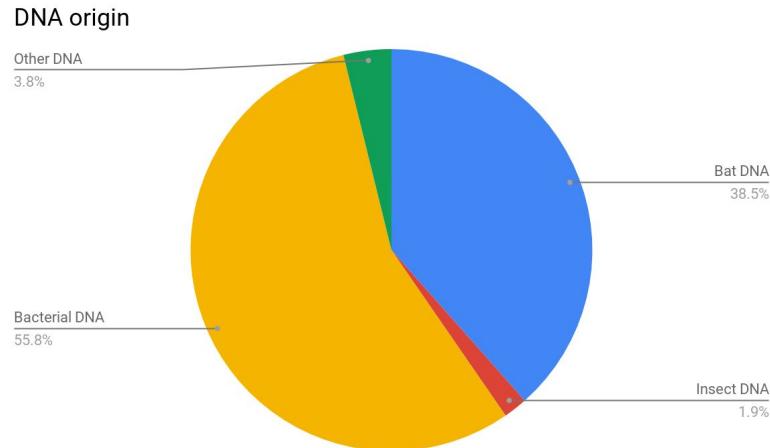
$$5,000,000,000 / 300 = \mathbf{16,666,666 \text{ reads}}$$

## DNA origin



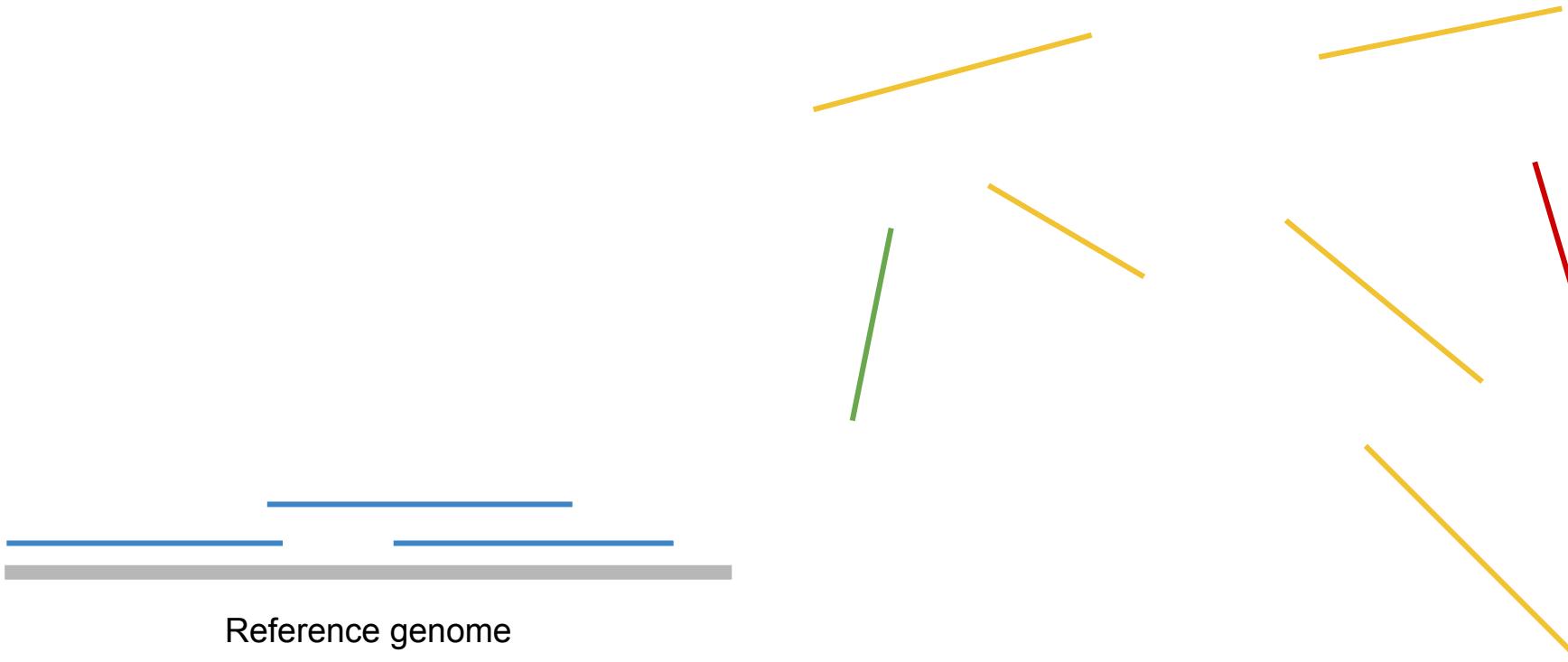
Bat fecal sample

### 3 Metagenomics



## 3

# Metagenomics

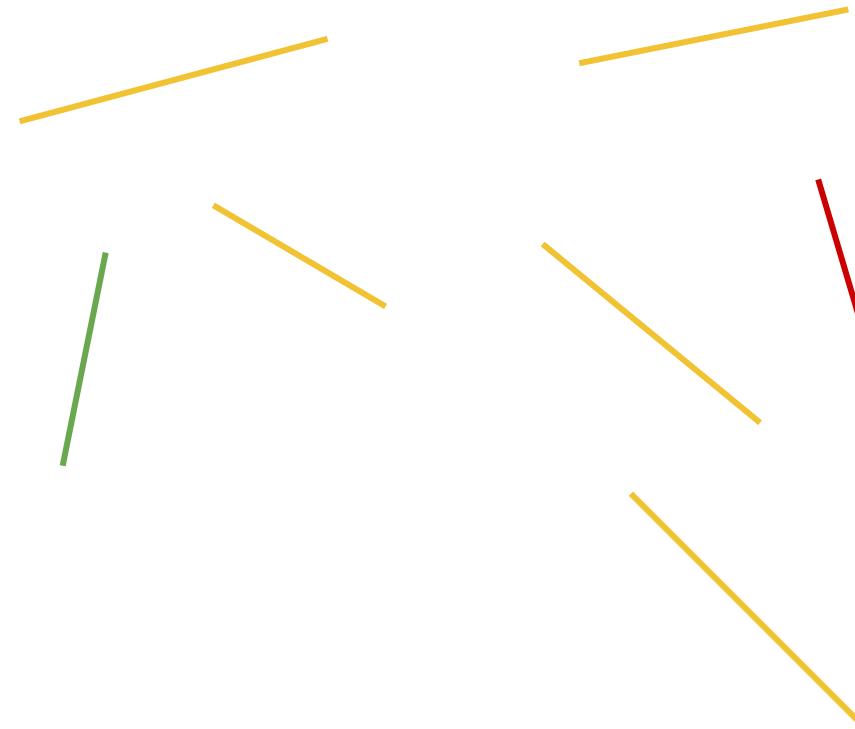


### 3 Metagenomics

- BWA
- Bowtie2



Reference genome



### 3 Metagenomics

NCBI Resources ▾ How To ▾ [Sign in to NCBI](#)

GenBank

GenBank ▾ Submit ▾ Genomes ▾ WGS ▾ Metagenomes ▾ TPA ▾ TSA ▾ INSDC ▾ Other ▾

## GenBank Overview

### What is GenBank?

GenBank® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (*Nucleic Acids Research*, 2013 Jan;41(D1):D36-42). GenBank is part of the [International Nucleotide Sequence Database Collaboration](#), which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data on a daily basis.

A GenBank release occurs every two months and is available from the [ftp](#) site. The [release notes](#) for the current version of GenBank provide detailed information about the release and notifications of upcoming changes to GenBank. Release notes for [previous GenBank releases](#) are also available. GenBank growth statistics for both the traditional GenBank divisions and the WGS division are available from each release. GenBank growth [statistics](#) for both the traditional GenBank divisions and the WGS division are available from each release.

An [annotated sample GenBank record](#) for a *Saccharomyces cerevisiae* gene demonstrates many of the features of the GenBank flat file format.

### GenBank Resources

[GenBank Home](#)

[Submission Types](#)

[Submission Tools](#)

[Search GenBank](#)

[Update GenBank Records](#)

<https://www.ncbi.nlm.nih.gov/genome/>

### 3 Metagenomics

#### Ovis aries (sheep)

Representative genome: [Ovis aries \(assembly Oar\\_v4.0\)](#)

Download sequences in FASTA format for [genome](#), [transcript](#), [protein](#)

Download genome annotation in [GFF](#), [GenBank](#) or [tabular](#) format

BLAST against Ovis aries [genome](#)

All 4 genomes for species:

Browse the [list](#)

Download sequence and annotation from [RefSeq](#) or [GenBank](#)

Display Settings: ▾ Overview

Send to: ▾

[Organism Overview](#) ; [Genome Assembly and Annotation report \[4\]](#) ; [Organelle Annotation Report \[1\]](#)

ID: 83



## Ovis aries (sheep)

sheep

Lineage: [Eukaryota\[2589\]](#); [Metazoa\[856\]](#); [Chordata\[366\]](#); [Craniata\[358\]](#); [Vertebrata\[358\]](#); [Euteleostomi\[352\]](#); [Mammalia\[150\]](#); [Eutheria\[145\]](#); [Laurasiatheria\[69\]](#); [Cetartiodactyla\[31\]](#); [Ruminantia\[18\]](#); [Pecora\[18\]](#); [Bovidae\[11\]](#); [Caprinae\[5\]](#); [Ovis\[2\]](#); [Ovis aries\[1\]](#)

*Ovis aries*, the domestic sheep, is an important livestock animal. It is a wool producing animal that has economic importance throughout the world. The genome sequence will facilitate the identification of genes that may provide explanations for the vast range of mammalian variation and help us to better understand the genetic basis of diseases [More...](#)

### Summary

Sequence data: genome assemblies: 4; sequence reads: 7 (See [Genome Assembly and Annotation report](#))

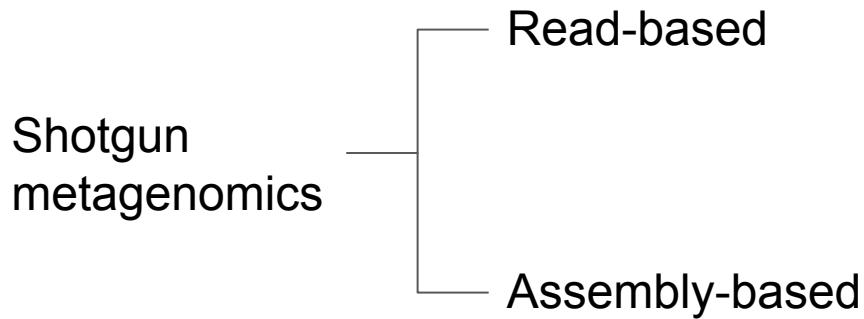
Statistics: median total length (Mb): 2738.01

median protein count: 48308

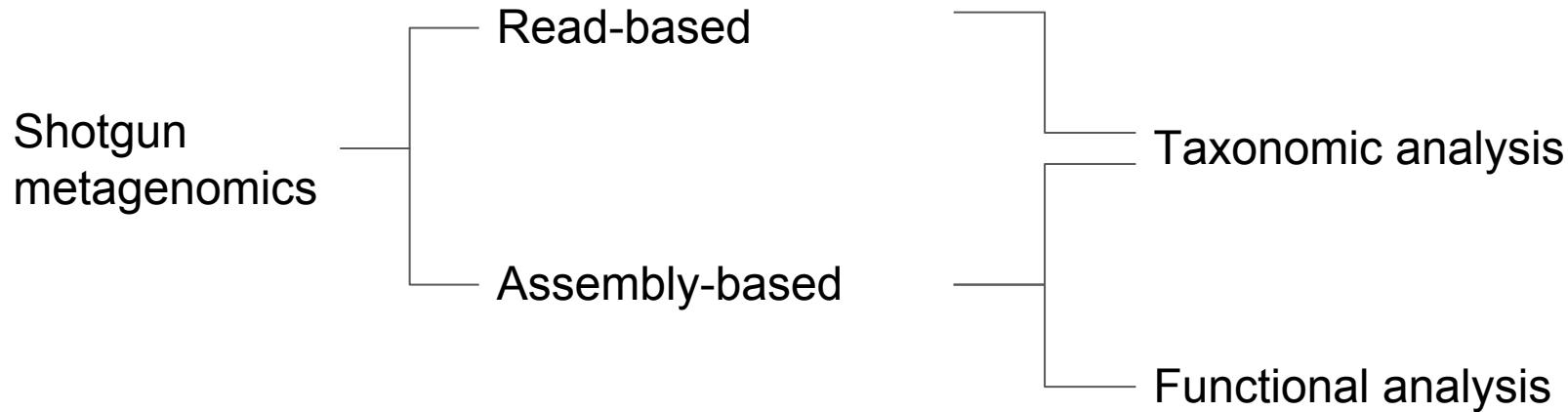
median GC%: 42.1898

NCBI Annotation Release: 102

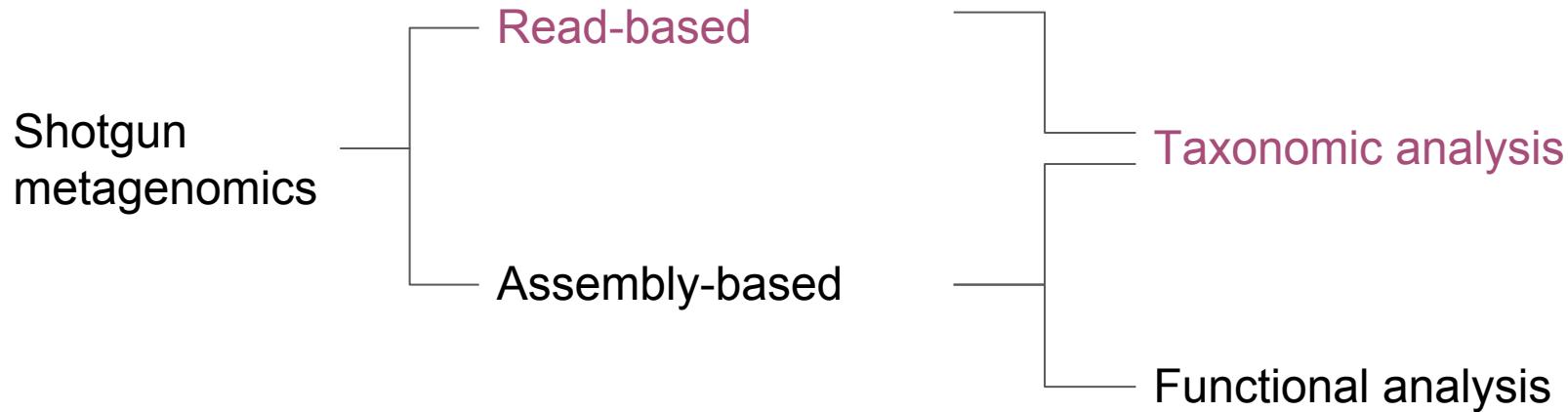
### 3 Metagenomics



### 3 Metagenomics

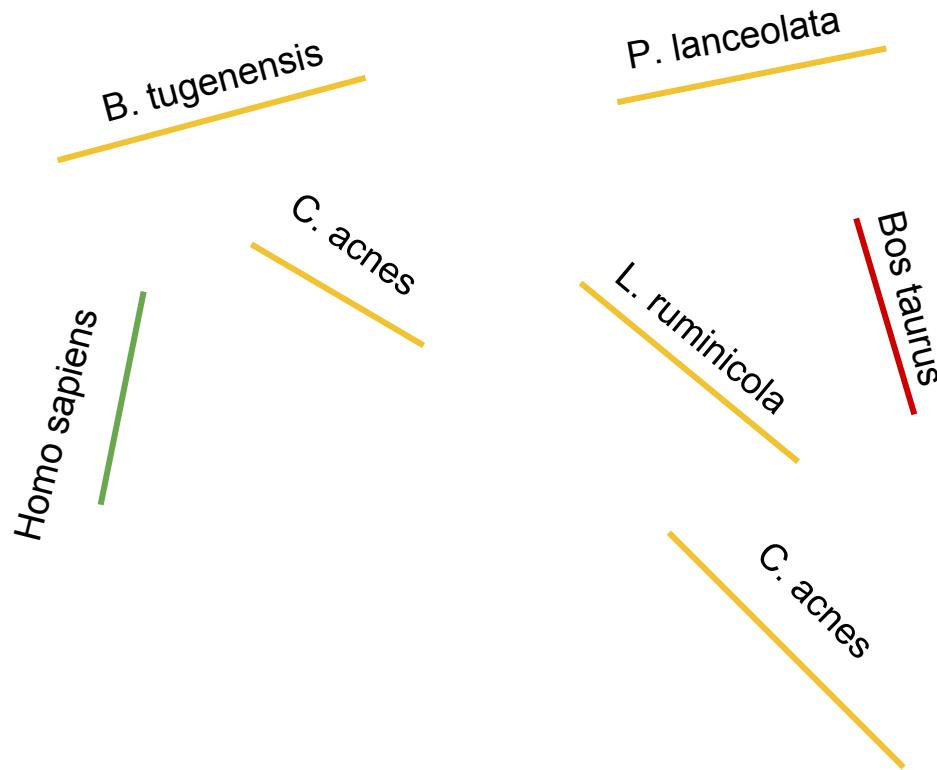


### 3 Metagenomics



### 3 Metagenomics

- Kraken (kmer)
- MGmapper (mapping)
- Centrifuge (FM-index)



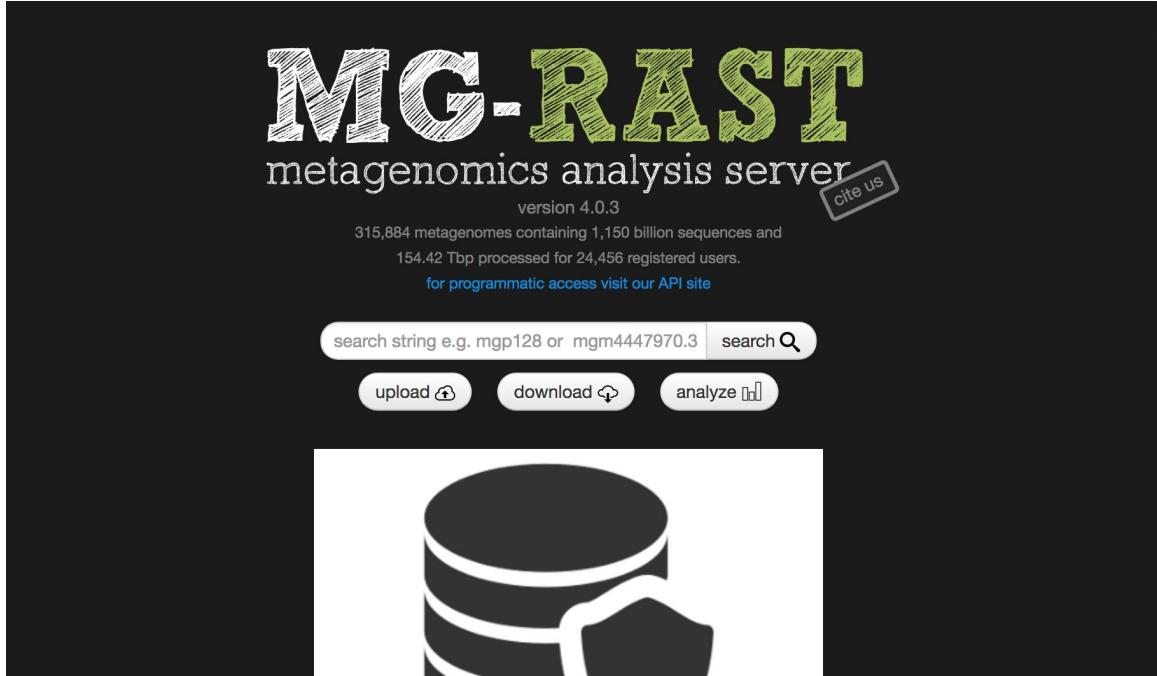
### 3 Metagenomics



[Sign up](#)

[Log in](#)

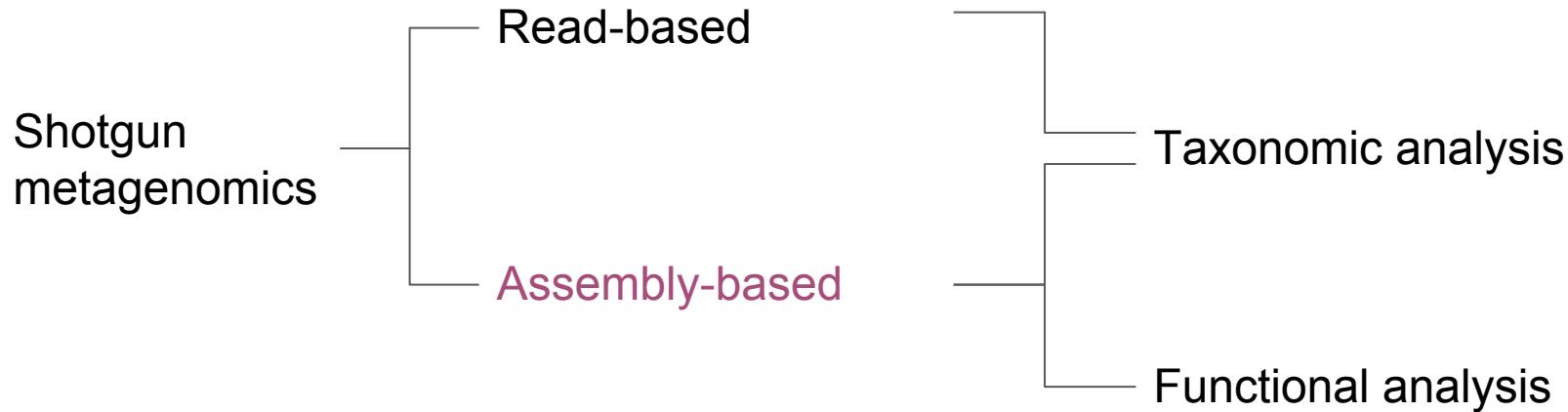
One Codex is a **data platform for applied microbial genomics**, enabling new and valuable applications in clinical diagnostics, food safety, and biosecurity



The image shows the homepage of the MG-RAST metagenomics analysis server. The title "MG-RAST" is prominently displayed in large, stylized green letters, with "metagenomics analysis server" in smaller white text below it. A sub-header "version 4.0.3" is visible. Below the title, statistics are provided: "315,884 metagenomes containing 1,150 billion sequences and 154.42 Tbp processed for 24,456 registered users." A link "for programmatic access visit our API site" is also present. At the bottom, there is a search bar with placeholder text "search string e.g. mgp128 or mgm4447970.3" and a "search" button. Below the search bar are three buttons: "upload" with an upward arrow icon, "download" with a downward arrow icon, and "analyze" with a bar chart icon. A large graphic of a database cylinder is centered at the bottom.

<http://metagenomics.anl.gov/>

### 3 Metagenomics



### 3 Metagenomics

Short sequences > low statistical power for  
taxonomic and functional profiling



### 3 Metagenomics

Short sequences > low statistical power for  
taxonomic and functional profiling

80 bp  
100 bp  
125 bp  
150 bp

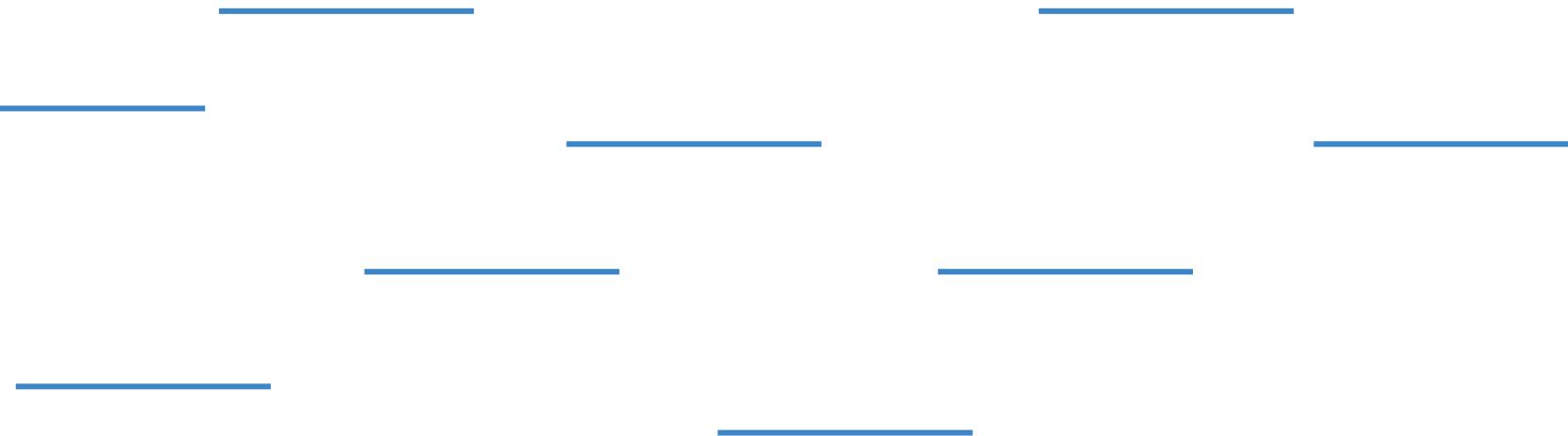


### 3 Metagenomics

- Longer sequences > taxonomic and functional assignments with higher level of confidence
- Coverage information > quantitative measurements



### 3 Metagenomics



Reads

3

## Metagenomics



Universidad  
del País Vasco

Euskal Herriko  
Unibertsitatea



NATURAL HISTORY MUSEUM OF DENMARK  
UNIVERSITY OF COPENHAGEN



# Contigs

### 3 Metagenomics

Single Read



Paired end



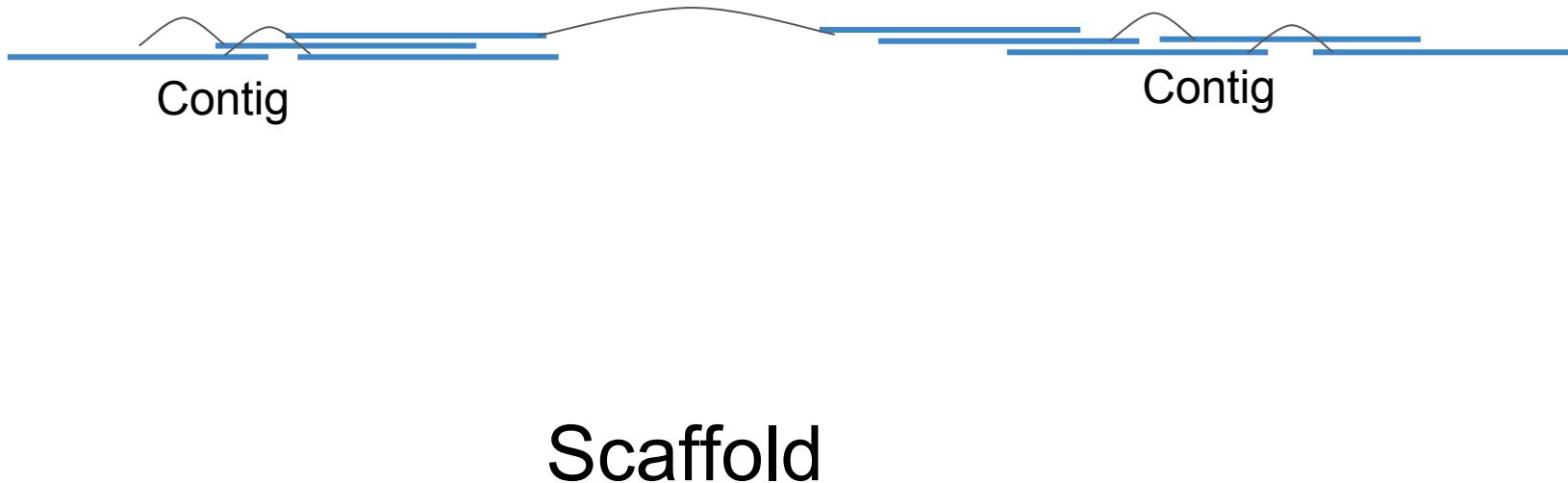
3

## Metagenomics

Reads

3

## Metagenomics



### 3 Metagenomics

## Individual assembly

Each sample is assembled separately

- Less computational resources
- Less bioinformatic steps
- Less efficient for downstream analyses
- Lower predictive capacity
- Worse to compare samples

## Co-assembly

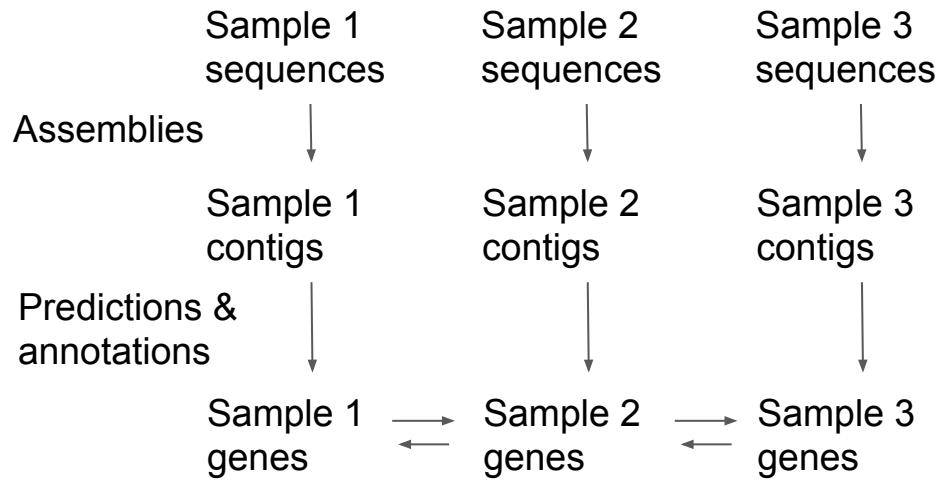
All samples are combined and a single assembly is generated

- High computational demand
- More bioinformatic steps
- More efficient for downstream analyses
- Higher predictive ability
- Better to compare samples

### 3 Metagenomics

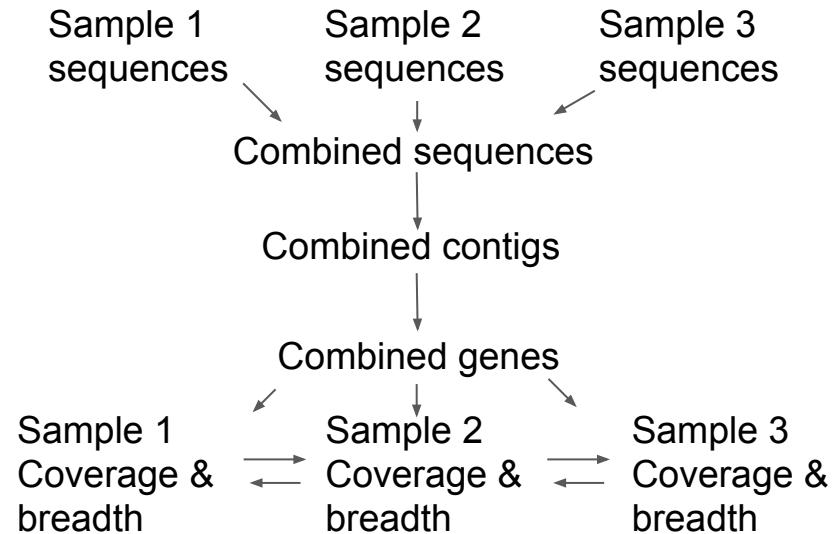
## Individual assembly

Each sample is assembled separately

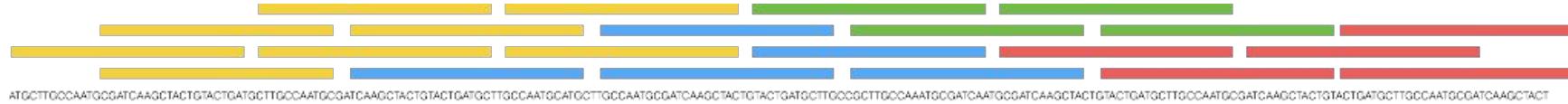


## Co-assembly

All samples are combined and a single assembly is generated

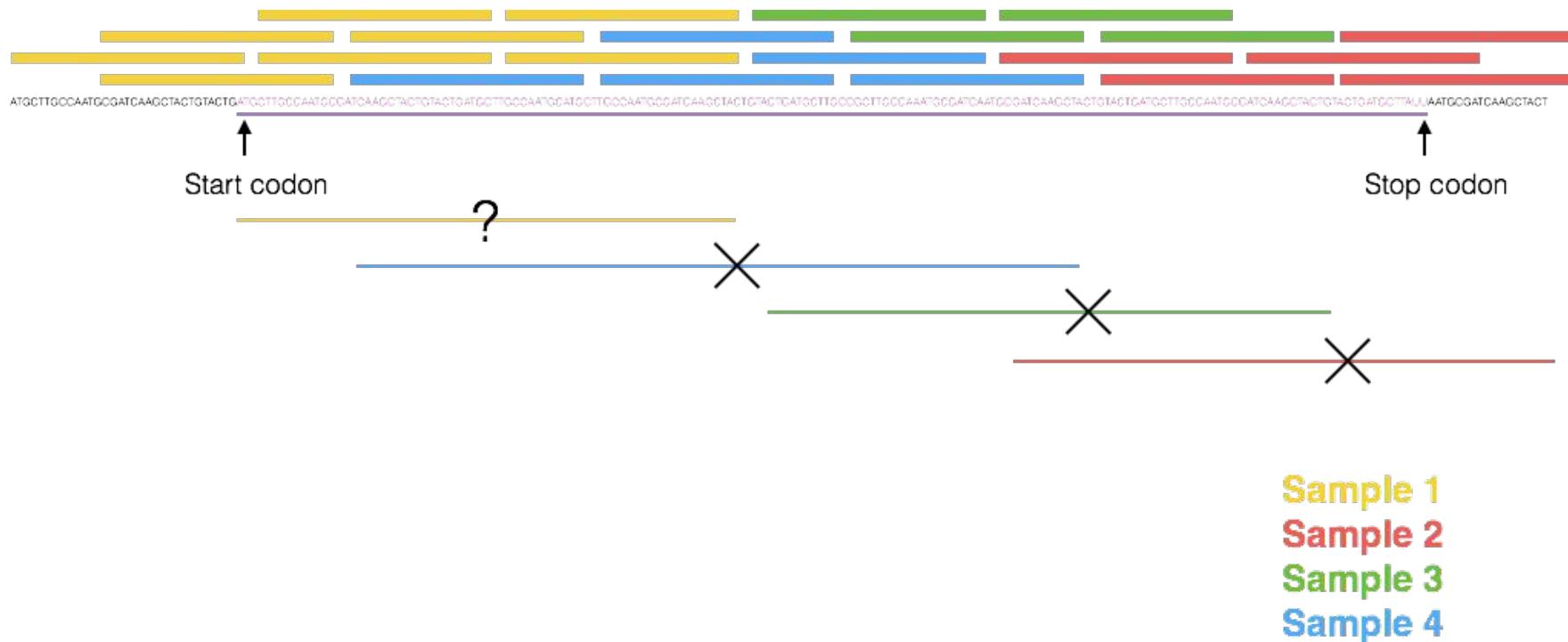


# 3 Metagenomics



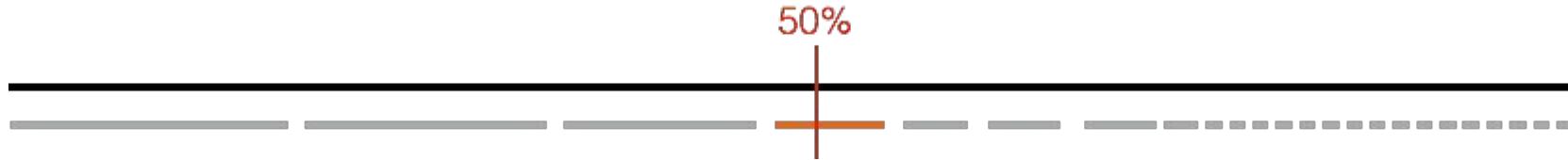
Sample 1  
Sample 2  
Sample 3  
Sample 4

### 3 Metagenomics



### 3 Metagenomics

N50: the shortest sequence length at 50% of the  
(meta)genome



Metagenomics:  $N50 > 1000\text{bp}$

### 3 Metagenomics



Total length: 52,132,606 bp

N50: 1725 bp

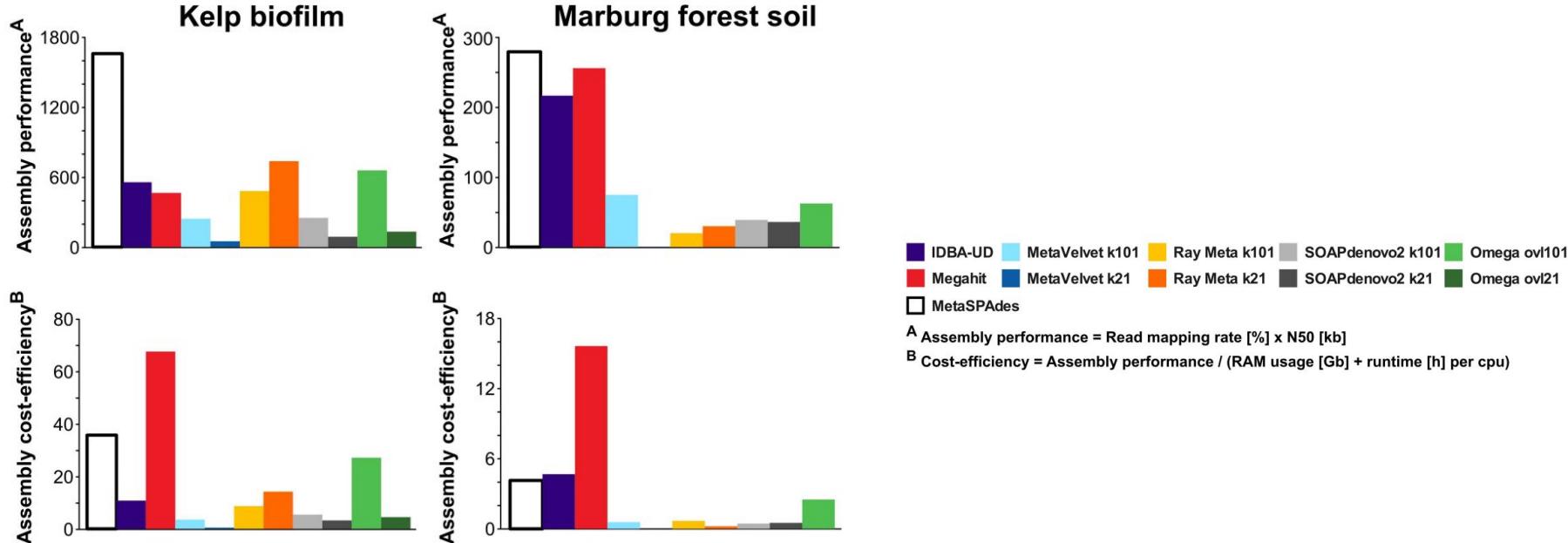
Mean length: 895 bp

Number of contigs: 58,235

### 3 Metagenomics

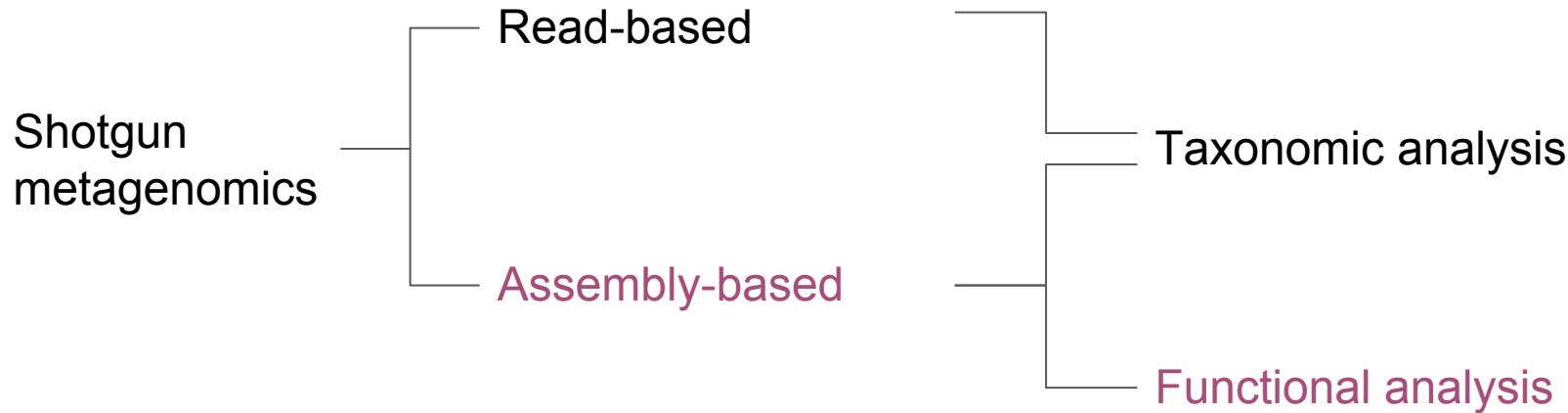


- IDBA-UD
- Megahit
- MetaVelvet
- SOAPdenovo
- Omega
- MetaSPAdes
- ...



Vollmers et al. 2016. Comparing and Evaluating Metagenome Assembly Tools from a Microbiologist's Perspective - Not Only Size Matters! *PLoS ONE* 12(1): e0169662

### 3 Metagenomics



### 3 Metagenomics

GTCCAGACTGCACTATGCGAACGGGCTACCACTAATCGGACTACAGCTAACAGACTGCAACGGG  
CTACCACTAATCGGACTACAGACACAGACTCTAACGCAACGGGCTACCACTAATCGGACTACAGCT  
AACGCAACGGGCTACCACTAATCGGACTACAGCTAACGCATAACACTGACTACGACGAACGGGCT  
ACCACTAATCGGACTACAGCTAACGCAACGGGCTACCACTAATCGGACTACAGCTAACGCAACGG  
GCTACCACTAATCGGACTACAGCTAACGCAACGGGCTACCACTAATCGGACTACAGCTAACGCAAC  
GGGCTACCACTAATCGGACTACAGCTAACGCAACGGGCAACGACTACGACTACCACTAATCGGAC  
TACAGCTAACGCAACGGGCTACCACTAATCGGACTACAGCTAACGCAACGGGCTACCACTAATCG  
GACTACAGCTAACGCAACGGGCTACCACTAATCGGACTACAGCTAACGCAACGGGCTACCACTAAT  
CGGACTACAGCTAACGCAACGGGCTACCACTAATCGGACTACAGCTAAC

### 3 Metagenomics

GTCCAGACTGCACTATGCGAACGGGCTACCACTAATCGGACTACAGCTAACAGACTGCAACGGG  
CTACCACTAATCGGACTACAGACACAGACTCTAACGCAACGGGCTACCACTAATCGGACTACAGCT  
AACGCAACGGGCTACCACTAATCGGACTACAGCTAACGCATAACACTGACTACGACGAACGGGCT  
ACCACTAATCGGACTACAGCTAACGCAACGGGCTACCACTAATCGGACTACAGCTAACGCAACGG  
GCTACCACTAATCGGACTACAGCTAACGCAACGGGCTACCACTAATCGGACTACAGCTAACGCAAC  
GGGCTACCACTAATCGGACTACAGCTAACGCAAACGGGAAACGACTACGACTACCACTAATCGGAC  
TACAGCTAACGCAACGGGCTACCACTAATCGGACTACAGCTAACGCAACGGGCTACCACTAATCG  
GACTACAGCTAACGCAACGGGCTACCACTAATCGGACTACAGCTAACGCAACGGGCTACCACTAAT  
CGGACTACAGCTAACGCAACGGGCTACCACTAATCGGACTACAGCTAAC

Is this a gene?

### 3 Metagenomics

GTCCAGACTGCACTATGCGAACGGGCTACCACTAATCGGACTACAGCTAACAGACTGCAACGGG  
CTACCACTAATCGGACTACAGACACAGACTCTAACGCAACGGGCTACCACTAATCGGACTACAGCT  
AACGCAACGGGCTACCACTAATCGGACTACAGCTAACGCATAACACTGACTACGACGAACGGGCT  
ACCACTAATCGGACTACAGCTAACGCAACGGGCTACCACTAATCGGACTACAGCTAACGCAACGG  
GCTACCACTAATCGGACTACAGCTAACGCAACGGGCTACCACTAATCGGACTACAGCTAACGCAAC  
GGGCTACCACTAATCGGACTACAGCTAACGCAACGGGCAACGACTACGACTACCACTAATCGGAC  
TACAGCTAACGCAACGGGCTACCACTAATCGGACTACA **GCTAACGCAACGGGCTACCACTAATCG**  
**GACTACAGCTAACGCAACGGGCTACCACTAATCGGACTACAGCTAACGCAACGGGCTACCACTAAT**  
**CGGACTACAGCTAACGCAACGGGCTACCACTAATCGGACTACAGCTAAC**

Is this a gene?

### 3 Metagenomics

#### 1) REFERENCE-BASED GENE PREDICTION

#### 2) AB INITIO GENE PREDICTION

**Open reading frame (ORF):** part of a sequence that has the potential to be translated

Start codon (ATG) - Sequence - Stop codon (TAA, TAG, TGA)

Reliability depends on GC content of genomes

High GC content / low AT content -> low number of stop codons

Low GC content / high AT content -> high number of stop codons -> false positives

Different approaches to estimate the reliability of ORFs

### 3 Metagenomics



#### Prokaryotes

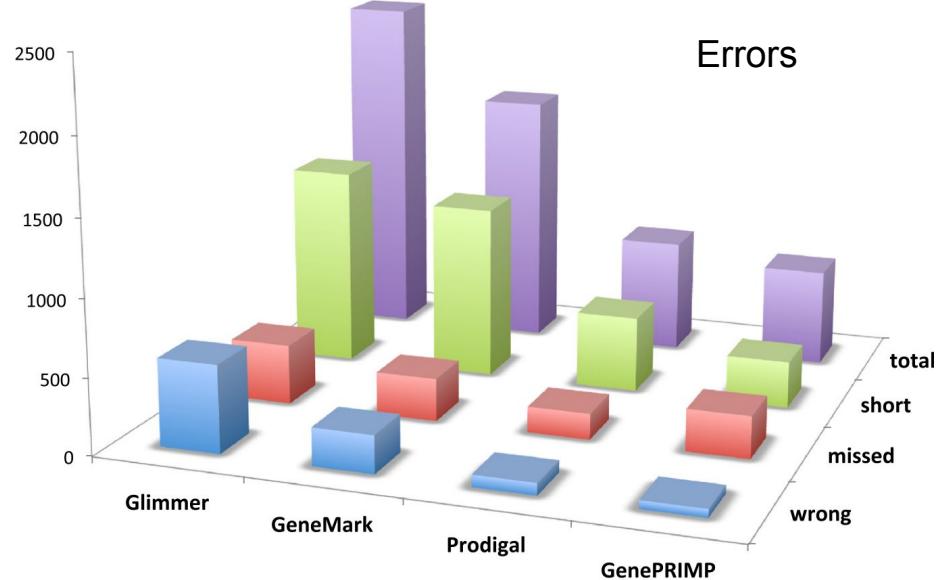
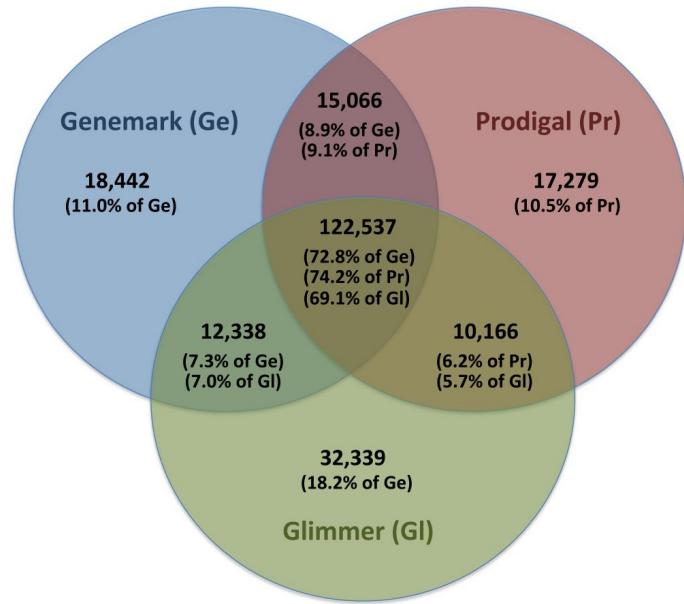
- RAST
- Glimmer
- Genemark
- GenePRIMP
- Prodigal

#### Eukaryotes

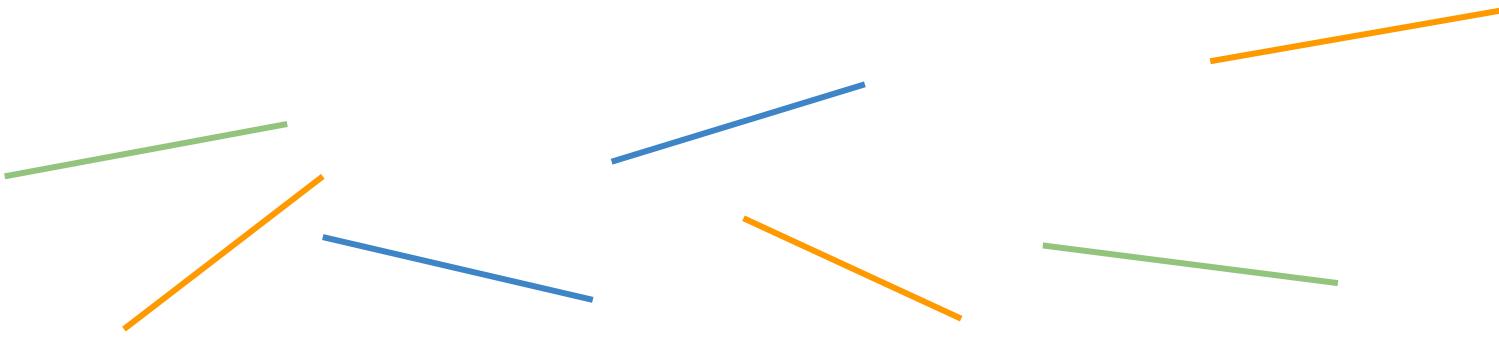
- Augustus
- ...

## 3

# Metagenomics

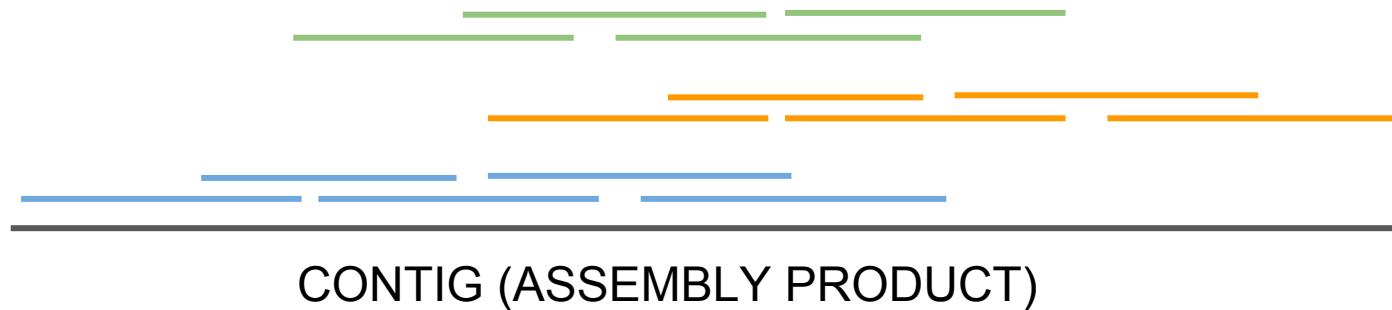


Tripp et al. 2015. Toward a standard in structural genome annotation for prokaryotes. *Standards in Genomic Sciences* 10: 45.



---

CONTIG (ASSEMBLY PRODUCT)





**SAMPLE 1:** 100% breadth, 1.5x depth/coverage

**SAMPLE 2:** 0% breadth, 0x depth/coverage

**SAMPLE 3:** 40% breadth, 0.5x depth/coverage

**SAMPLE 1:** 0% breadth, 0x depth/coverage

**SAMPLE 2:** 100% breadth, 1.8x depth/coverage

**SAMPLE 3:** 20% breadth, 0.1x depth/coverage

### 3 Metagenomics

#### Sample 1

ATGCGATGCCGATGCATGCCCGCGATTAGCTAGGGCTAGCTTACGGCATAT  
01111111222222222222222211111111111111111000000000

#### Sample 2

ATGCGATGCCGATGCATGCCCGCGATTAGCTAGGGCTAGCTTACGGCATAT  
011111111111111111112333333333333333332100000000

### 3 Metagenomics

#### Sample 1

ATGCGATGCCGATGCATGCCCGCGATTAGCTAGGCTAGCTTACGGCATAT  
011111122222222222222222211111111111111111000000000

Coverage: 1.2  
Breadth: 85%

#### Sample 2

ATGCGATGCCGATGCATGCCCGCGATTAGCTAGGCTAGCTTACGGCATAT  
011111111111111111112333333333333333332100000000

Coverage: 1.6  
Breadth: 87%

### 3 Metagenomics

#### Query sequence

```
>MYSEQUENCE
ATGCGAACGGGCTACCACT
```

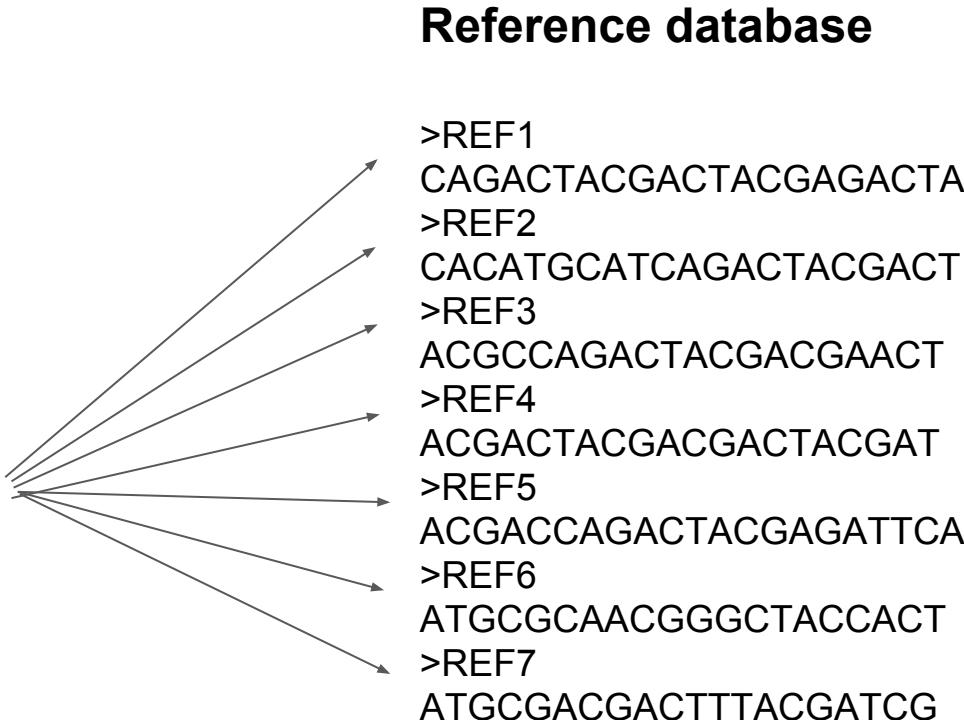
#### Reference database

```
>REF1
CAGACTACGACTACGAGACTA
>REF2
CACATGCATCAGACTACGACT
>REF3
ACGCCAGACTACGACGAACT
>REF4
ACGACTACGACGACTACGAT
>REF5
ACGACCAGACTACGAGATTCA
>REF6
ATGCGAACGGGCTACCACT
>REF7
ATGCGACGACTTACGATCG
```

### 3 Metagenomics

#### Query sequence

```
>MYSEQUENCE  
ATGCGAACGGGCTACCACT
```



### 3 Metagenomics



## Assumption:

Sequence similarity implied taxonomic and functional similarity

Nucleotides

Aminoacids

### 3 Metagenomics

## Reference database

```
>REF1
CAGACTACGACTACGAGACTA
>REF2
CACATGCATCAGACTACGACT
>REF3
ACGCCAGACTACGACGAACT
>REF4
ACGACTACGACGACTACGAT
>REF5
ACGACCAGACTACGAGATTCA
>REF6
ATGCGCAACGGGCTACCACT
>REF7
ATGCGACGACTTACGATCG
```

Each sequence linked to metadata

Usually taxonomic information

- *Prevotella ruminicola*

Sometimes functional information

- Beta-dehydrogenase

Different databases

### 3 Metagenomics

## Function

- Uniprot
- Pfam
- CAZy
- COG
- KEGG
- ...

## Taxonomy

- NCBI Taxonomy
- EMBL-EBI Taxonomy
- BOLD systems
- ...

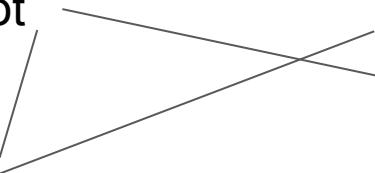
### 3 Metagenomics

## Function

- Uniprot
- Pfam
- CAZy
- COG
- KEGG
- ...

## Taxonomy

- NCBI Taxonomy
- EMBL-EBI Taxonomy
- BOLD systems
- ...



*Often interconnected*

### 3 Metagenomics



- 1) Database
- 2) Aligner

### 3 Metagenomics



- 1) Database
- 2) Aligner

**Database - Aligner**  
NCBI-NT - BLAST  
BOLD - BOLD IDS  
SILVA - SINA

## 3

# Metagenomics

NCBI Resources How To

GenBank Nucleotide Search

GenBank Submit Genomes WGS Metagenomes TPA TSA INSDC Other

## GenBank Overview

### What is GenBank?

GenBank® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences (*Nucleic Acids Research*, 2013 Jan 41(D1):D36-42). GenBank is part of the International Nucleotide Sequence Database Collaboration, which comprises the DNA DataBank of Japan (DDBJ), the European Nucleotide Archive (ENA), and GenBank at NCBI. These three organizations exchange data on a daily basis.

A GenBank release occurs every two months and is available from the [ftp site](#). The [release notes](#) for the current version of GenBank provide detailed information about the release and notifications of upcoming changes to GenBank. Release notes for [previous GenBank releases](#) are also available. GenBank growth statistics for both the traditional GenBank divisions and the WGS division are available from each release. GenBank growth [statistics](#) for both the traditional GenBank divisions and the WGS division are available from each release.

An [annotated sample GenBank record](#) for a *Saccharomyces cerevisiae* gene demonstrates many of the features of the GenBank flat file format.

### Access to GenBank

There are several ways to search and retrieve data from GenBank.

- Search GenBank for sequence identifiers and annotations with [Entrez Nucleotide](#), which is divided into three divisions: [CoreNucleotide](#) (the main collection), [dbEST](#) (Expressed Sequence Tags), and [dbGSS](#) (Genome Survey Sequences).
- Search and align GenBank sequences to a query sequence using [BLAST](#) (Basic Local Alignment Search Tool). BLAST searches CoreNucleotide, dbEST, and dbGSS independently; see [BLAST info](#) for more information about the numerous BLAST databases.
- Search, link, and download sequences programmatically using [NCBI e-utilities](#).
- The ASN.1 and flatfile formats are available at NCBI's anonymous FTP server: <ftp://ftp.ncbi.nlm.nih.gov/ncbi-asn1> and <ftp://ftp.ncbi.nlm.nih.gov/genbank>.

<https://www.ncbi.nlm.nih.gov/genbank/>



NIH U.S. National Library of Medicine NCBI Sign in to NCBI

BLAST® Home Recent Results Saved Strategies Help

**Basic Local Alignment Search Tool**

**May 16 webinar**  
Improved BLAST+ and BLAST databases: now with taxonomic information.  
Mon, 07 May 2018 15:00:00 EST [More BLAST news...](#)

**NEWS**

**BLAST** finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

**Web BLAST**

**Nucleotide BLAST** nucleotide ▶ nucleotide

**blastx** translated nucleotide ▶ protein

**tblastn** protein ▶ translated nucleotide

**Protein BLAST** protein ▶ protein

**BLAST Genomes**

Enter organism common name, scientific name, or tax id  Search

Human Mouse Rat Microbes

<https://blast.ncbi.nlm.nih.gov/Blast.cgi>

### 3 Metagenomics



## Taxonomy

Phylum > Class > Order > Family > Genus > Species

### 3 Metagenomics



## Taxonomy

Phylum > Class > Order > Family > Genus > Species

## KEGG

Domain > Pathway > KO

## EggNog

Class > Family

<http://www.kegg.jp/>



KEGG  [Search](#) [Help](#)  
[» Japanese](#)

## KEGG Home

[Release notes](#)  
[Current statistics](#)  
[Plea from KEGG](#)

## KEGG Database

[KEGG overview](#)  
[Searching KEGG](#)  
[KEGG mapping](#)  
[Color codes](#)

## KEGG Objects

[Pathway maps](#)  
[Brite hierarchies](#)  
[KEGG DB links](#)

## KEGG: Kyoto Encyclopedia of Genes and Genomes

KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies. See [Release notes](#) (May 10, 2018) for new and updated features.

**Announcement:** [HTTPS at www.kegg.jp](https://www.kegg.jp) *New!*

### Main entry point to the KEGG web service

**KEGG2**

[KEGG Table of Contents](#) [\[Update notes | Release history\]](#)

### Data-oriented entry points

### 3 Metagenomics

EggNOG 4.5.1

Search protein or OG 

Navigation

- Home
- Sequence search
- eggNOG-mapper New  
(genome-wide functional annotation)
- Downloads
- API
- Methods
- Viral OGs

Discover EggNOG 4.5.1

A database of orthologous groups and functional annotation



Organisms | Viruses | Orthologous Groups | Trees & Algs.

2,031 | 352 | 190k | 1.9M

Search

News 

Contact 

Tweets by @eggnogdb 

 EggNOG database @eggnogdb nice use of @eggnogdb and eggnoG-mapper for functional annotation and orthology prediction in single-cell transcriptomics of non-model species.

May 8, 2018

  EggNOG database Retweeted  Iñaki Huertas Casares

<http://eggnogdb.embl.de>

### 3 Metagenomics

- Many contigs belong to the same organism
- How can we know which contigs belong to the same organism?

### 3 Metagenomics

- Many contigs belong to the same organism
- How can we know which contigs belong to the same organism?
- What if the organism's information is not in the reference databases?

### 3 Metagenomics

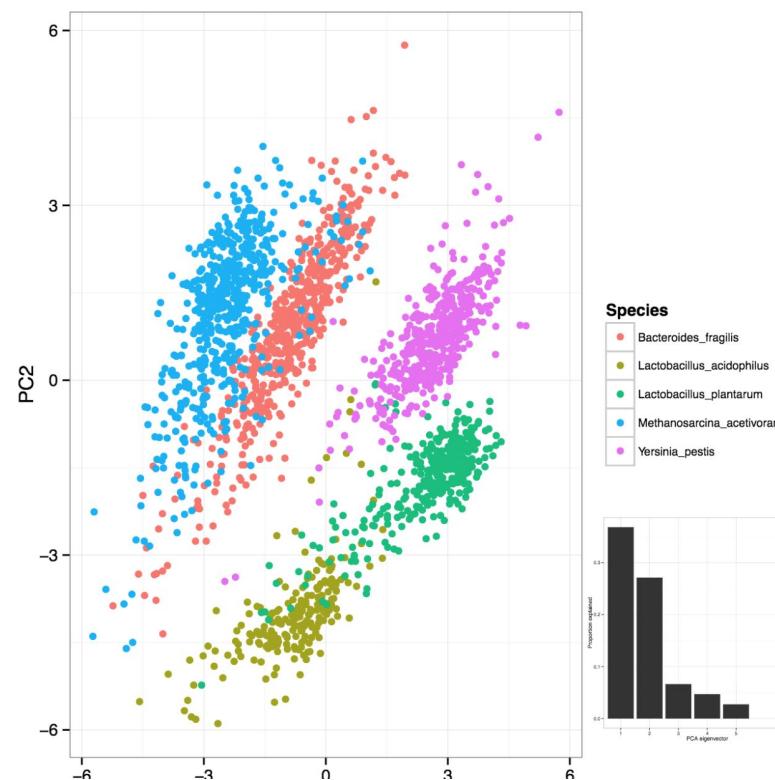
- Many contigs belong to the same organism
- How can we know which contigs belong to the same organism?
- What if the organism's information is not in the reference databases?

Genomic similarity

Coverage ratios

### 3 Metagenomics

Different species have characteristic signatures of overlapping kmer frequencies  
Sandberg et al. 2001 Genome Res



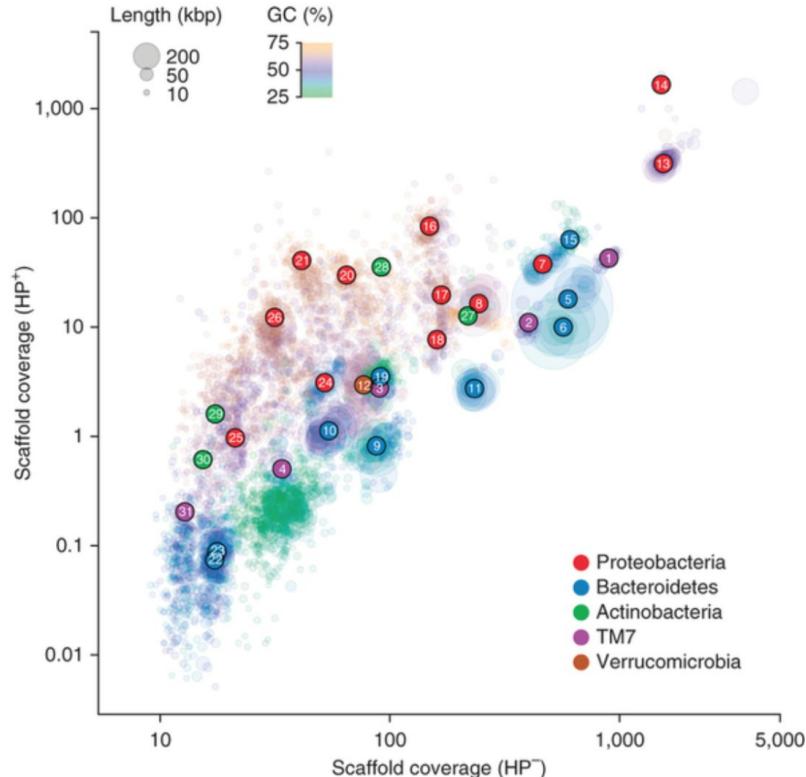
### 3 Metagenomics

Different species have characteristic signatures of GC%

Albertsen et al. 2014 Nature Biotech.

Contigs belonging to the same species show similar coverage ratios

Albertsen et al. 2014 Nature Biotech.



### 3 Metagenomics



- CONCOCT (Alneberg et al. 2014, Nature Methods)
- MetaBAT (Kang et al. 2015, PeerJ)
- MaxBin2 (Wu et al. 2015, Bioinf.)