

Stellenbosch

UNIVERSITY
IYUNIVESITHI
UNIVERSITEIT



www.ucd.ie/cfs

@CfsUcd

Advanced Applications of Next Generation Sequencing in Food Safety

Dr Guerrino Macori, BSc, MSc, PhD

Assistant Professor

School of Biology and Environmental Science

University College Dublin, Ireland

UCD Centre for Food Safety

guerrino.macori@ucd.ie

@guerrinomacori



DAY 1

Monday 22nd January 2024

09:00 Stellenbosch welcome UCD to laboratory facilities

12:30 Lunch

SESSION 1: INTRODUCTION TO NEXT GENERATION SEQUENCING (NGS)
technologies, applications, and an introduction to bioinformatics

13:30 Professor Pieter A. Gouws, Stellenbosch University, South Africa

- Welcome, opening remarks and introduction to the program

14:30 Dr Guerrino Macori, University College Dublin, Ireland

- Overview of Next Generation Sequencing Technologies -what do we need to know

15:00 Professor Séamus Fanning, University College Dublin, Ireland

- Next Generation Sequencing Technologies in the context of risk assessment & food safety|

15:30 Dr Guerrino Macori, University College Dublin, Ireland

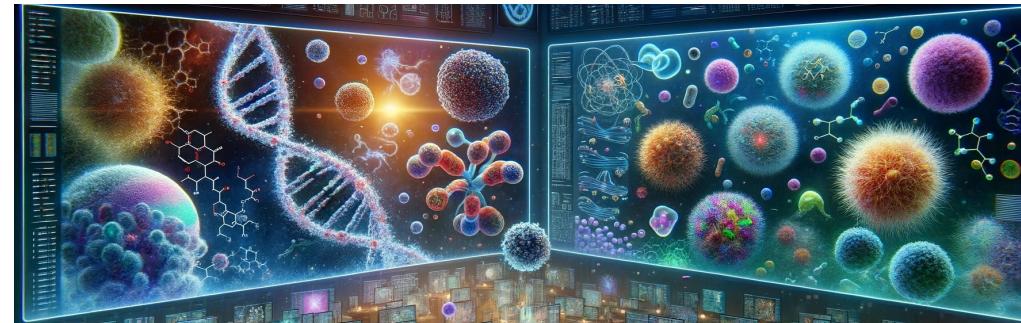
- Introduction to Bioinformatics and its applications

16:30 The industry perspective –

Karin Carstensen (Woolworths)

Dr Tynan Marais (WhiteSci)



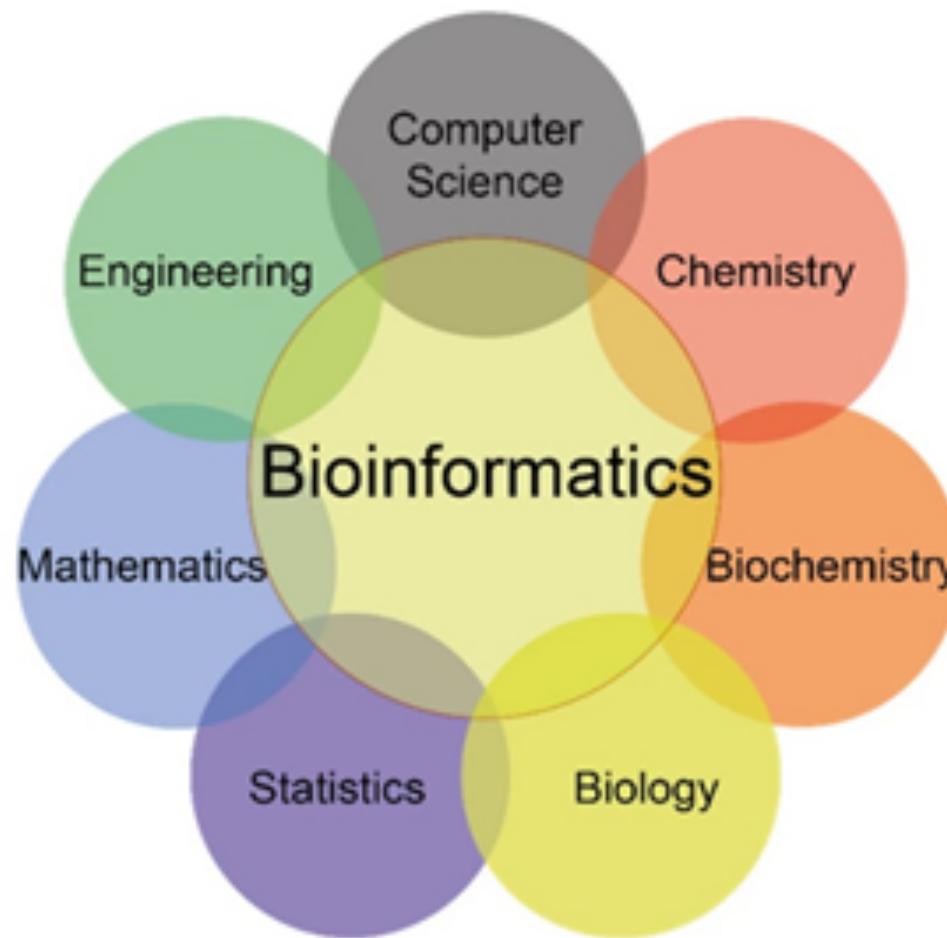


Overview of NGS Technologies - what do we need to know



Background & History
Next Generation Sequencing
Overview of NGS Technologies
Data Generation and Applications

What is Bioinformatics?



Bioinformatics Defined

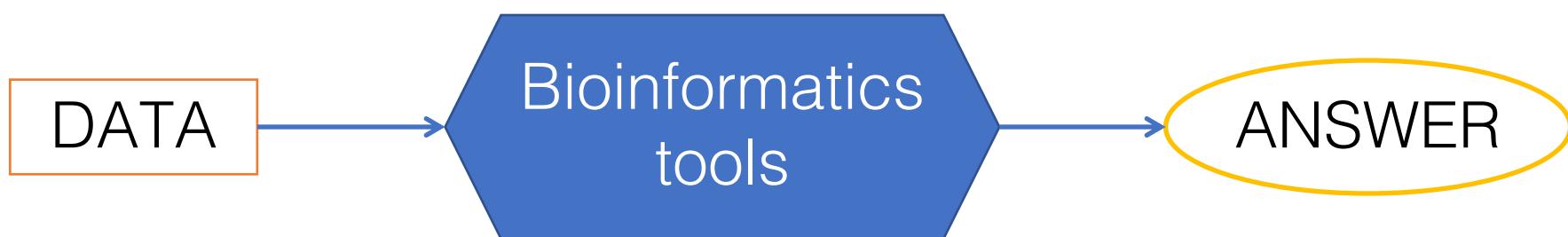
(Oxford English Dictionary):

The branch of science concerned with information and information flow in biological systems, esp. the use of computational methods in genetics and genomics

Background

Flawed view of bioinformatics

Most biologists believe bioinformatics works the following way

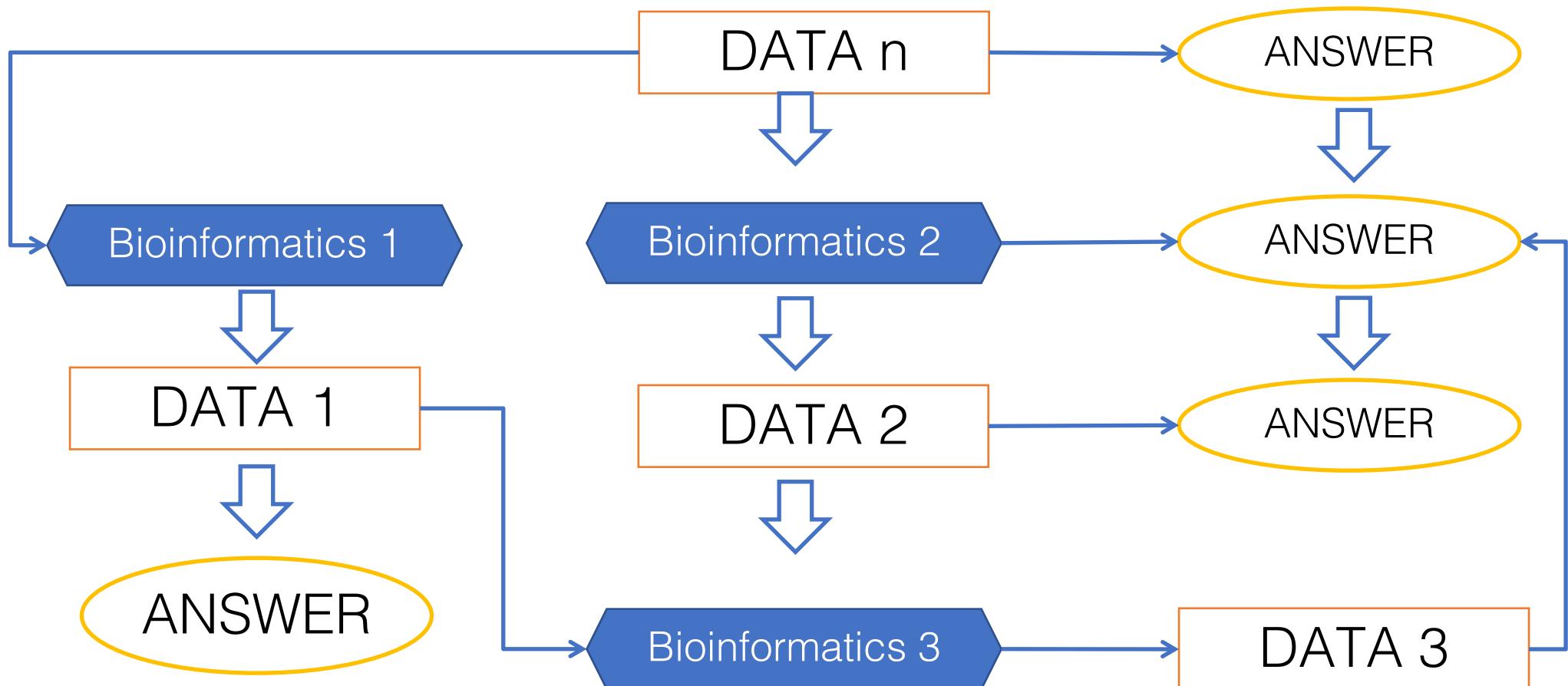


Misconception: scientist need to find the right tool that gives them the answer

This world view is surprisingly pervasive

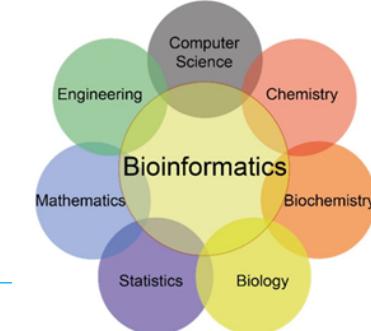
Background

Bioinformatics in reality





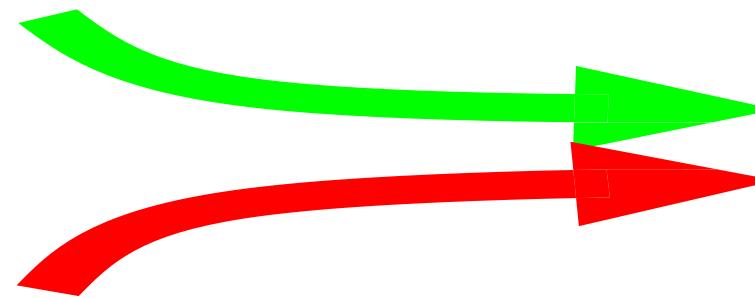
What is Bioinformatics?



The field of science in which biology, computer science and information technology merge into a single discipline

Biologists

collect molecular data: DNA & Protein sequences, gene expression, epigenetics etc.



Bioinformaticians

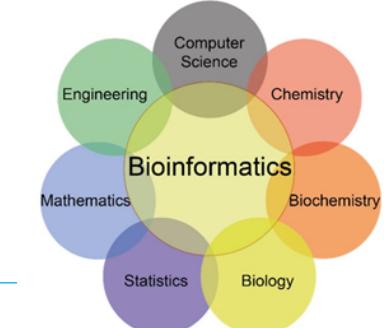
Study biological questions by analysing molecular data

Computer scientists

(+Mathematicians, Statisticians, etc.)

Develop tools, pieces of software, algorithms to store and analyze the data.

What is Bioinformatics?

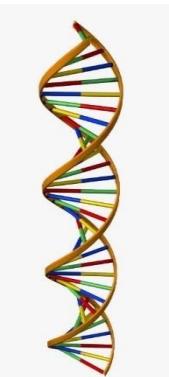


- Bioinformatics is the application of computer science and technology to biology and medicine.
- Bioinformatics makes it possible to analyse large quantities of biological data and can be used to:
 - Search Biological databases
 - Compare sequences
 - Draw Molecular structures

Applications

Organise, Process, and Make Sense of Complex Biological Data Sets

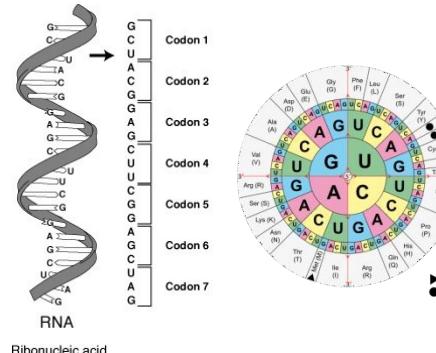
DNA



Bioinformatics Tools:

DNA Sequencing Identify Mutations in DNA
Describe complex populations

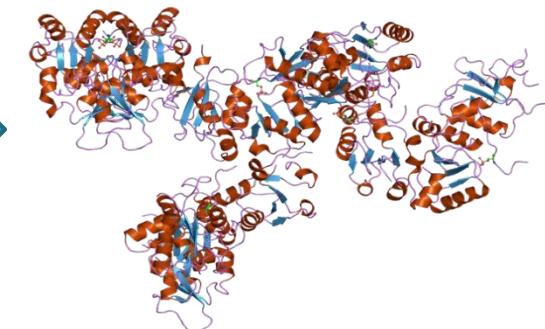
RNA



Bioinformatics Tools:

RNA Sequencing Identify tissue specific gene expression.

Protein



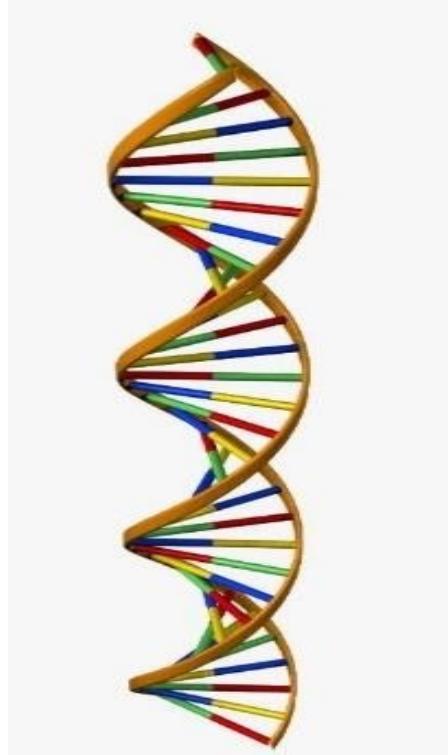
Bioinformatics Tools:

Protein 3D Structure visualization.

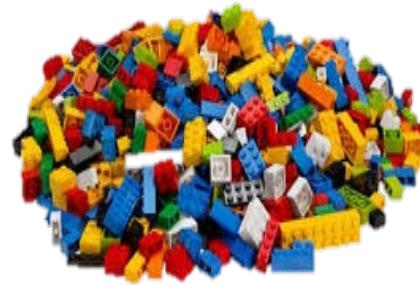
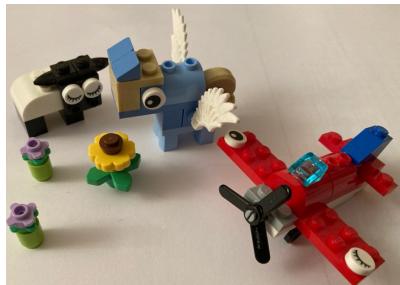
Question: What kinds of scientific questions can we answer with bioinformatics tools?

Applications

Organise, Process, and Make Sense of Complex Biological Data Sets

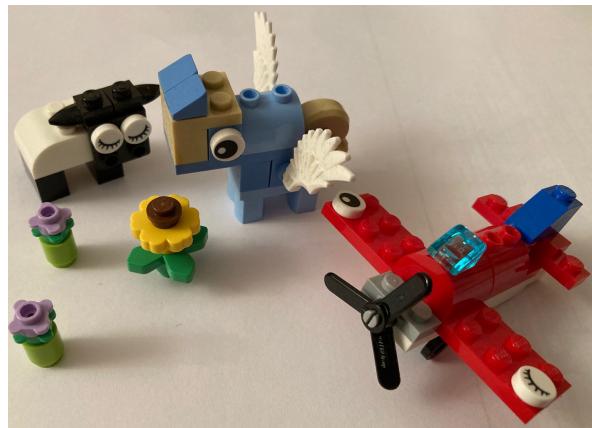


- Whole Genome Sequencing
- Metagenomics analysis



- Microbial whole-genome sequencing (WGS)
 - Sequencing entire genomes
 - Discover new variants / detect low-frequency *quasispecies* in culture or clinical samples
 - Sequence multiple organisms in parallel
- Metagenomes
 - Describe diversity
 - Taxonomic identification
 - WGS from Metagenomics: Metagenome-Assembled Genome (MAG)
 - Functional analysis – describing specific Virulence factors

Applications



WGS from Metagenomics

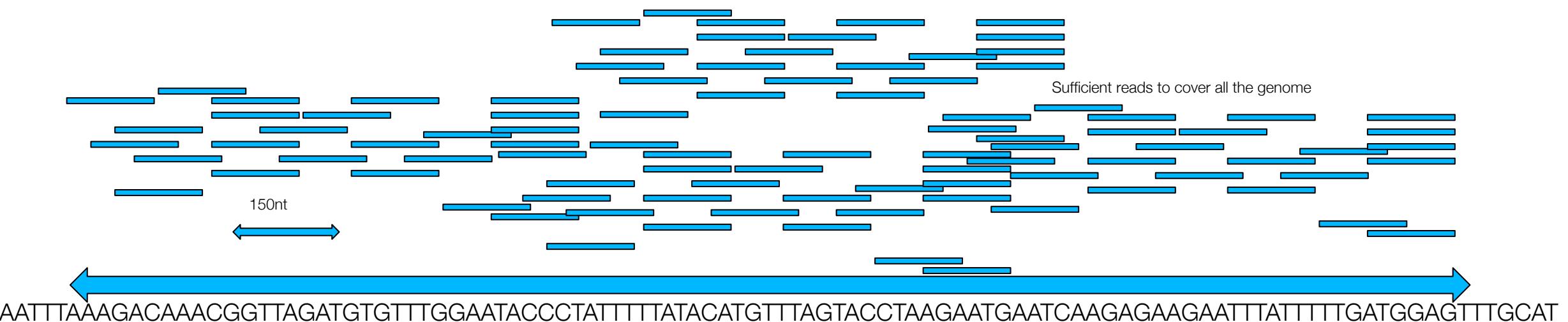
Metagenome-Assembled Genome (MAG) is a single-taxon assembly based on one or more binned metagenomes that has been asserted to be a close representation to an actual individual genome
(that could match an already existing isolate or represent a novel isolate).

Applications

Reads length among sequencing technologies and platforms

- Sequence depth

Technology	Maximum read length (L)
Illumina	600 bp
SOLiD	75 bp
IonTorrent	400 bp
454	~700 bp (up to 1kb)
Nanopore	5.4-10 kb
PacBio	~15kb (up to 40 kb)

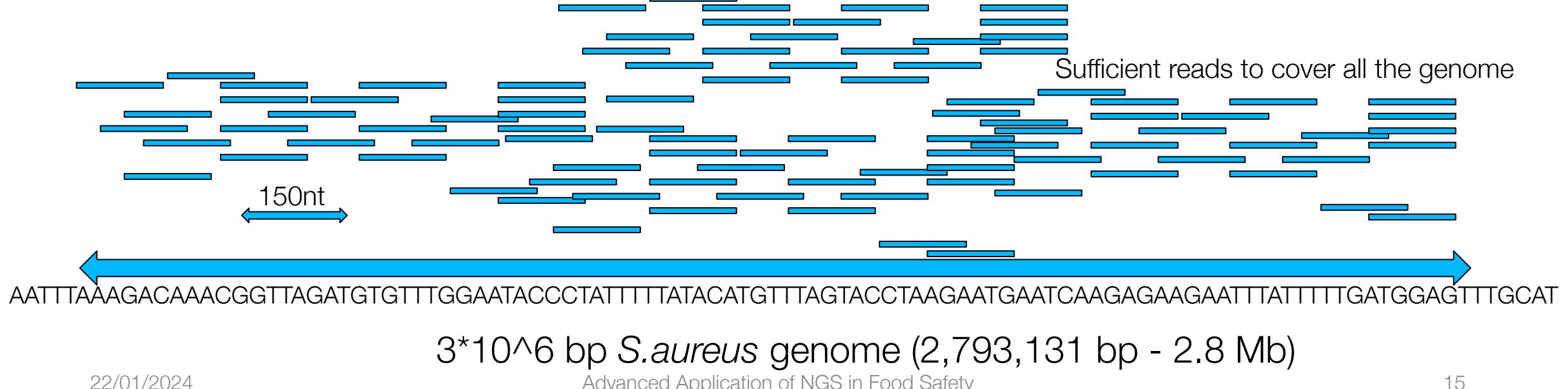


3×10^6 bp *S. aureus* genome (2,793,131 bp - 2.8 Mb)

- Sequence depth

C = coverage (100x)

L = read length (MiSeq = 600 bp)



- Sequence depth

Exercise:

Project - Whole Genome Sequencing of bacteria responsible of FBO ($G=6 \times 10^6$)
Calculate the number of reads for the estimating coverage ($C=100$) for MiSeq
($L=600$)

The Lander/Waterman equation for computing coverage

Coverage Equation

$$C = LN / G$$

C = coverage (100x)

G = genome length = 6×10^6 bp

L = read length (MiSeq = 600)

N = number of reads (X)

$$100 = \frac{600N}{2,793,131}$$

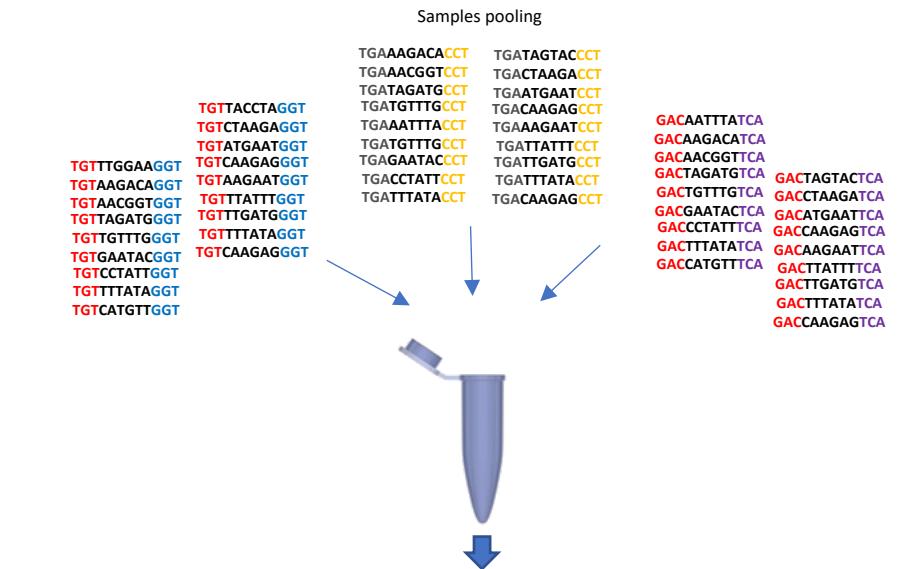
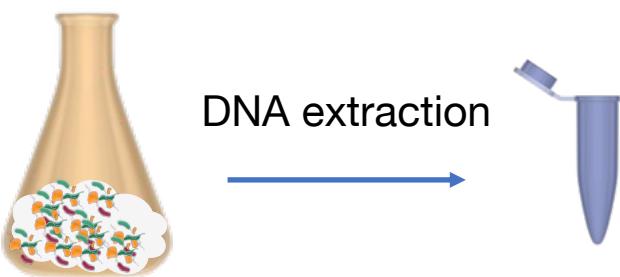
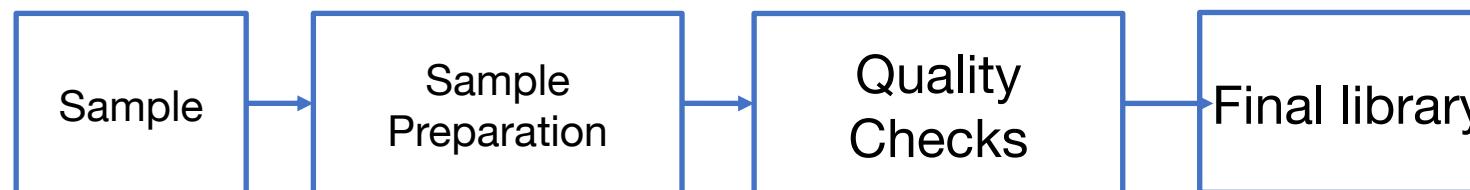
$$N = \frac{100 * 2,793,131}{600}$$

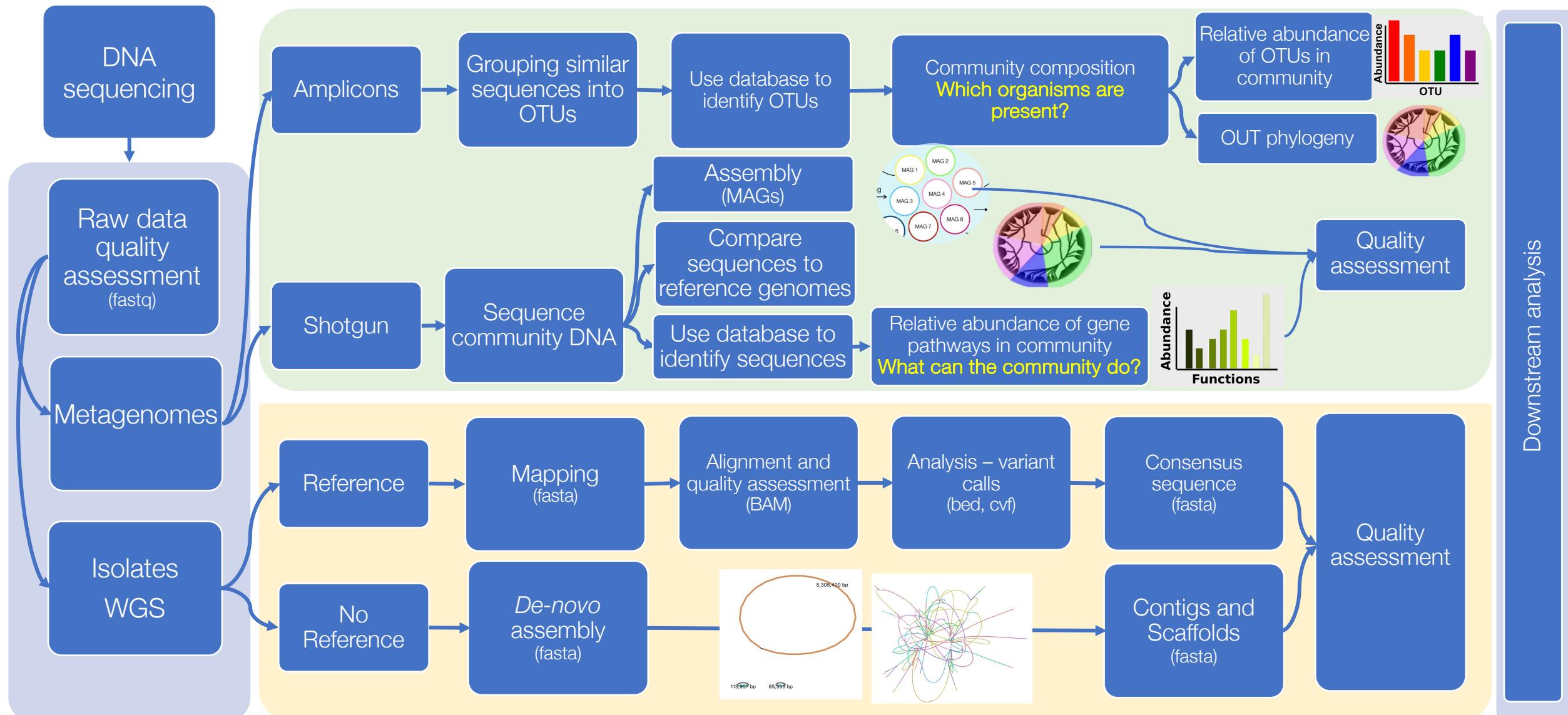


465,522

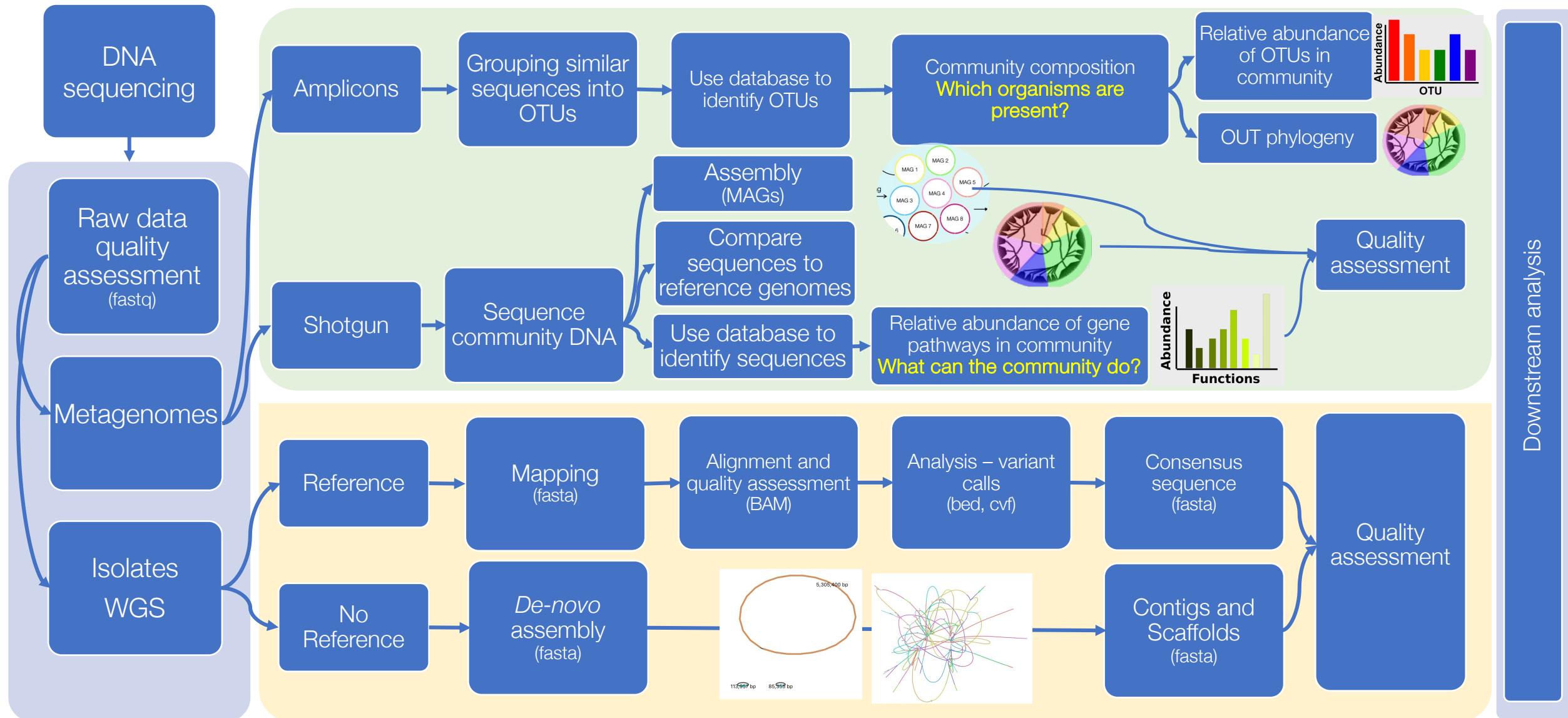
AATTAAAGACAAACGGTTAGATGTGTTGGAATACCCTATTTTATACATGTTAGTACCTAAGAACATGAAGAGAAGAATTATTTGATGGAGTTGCAT

Applications

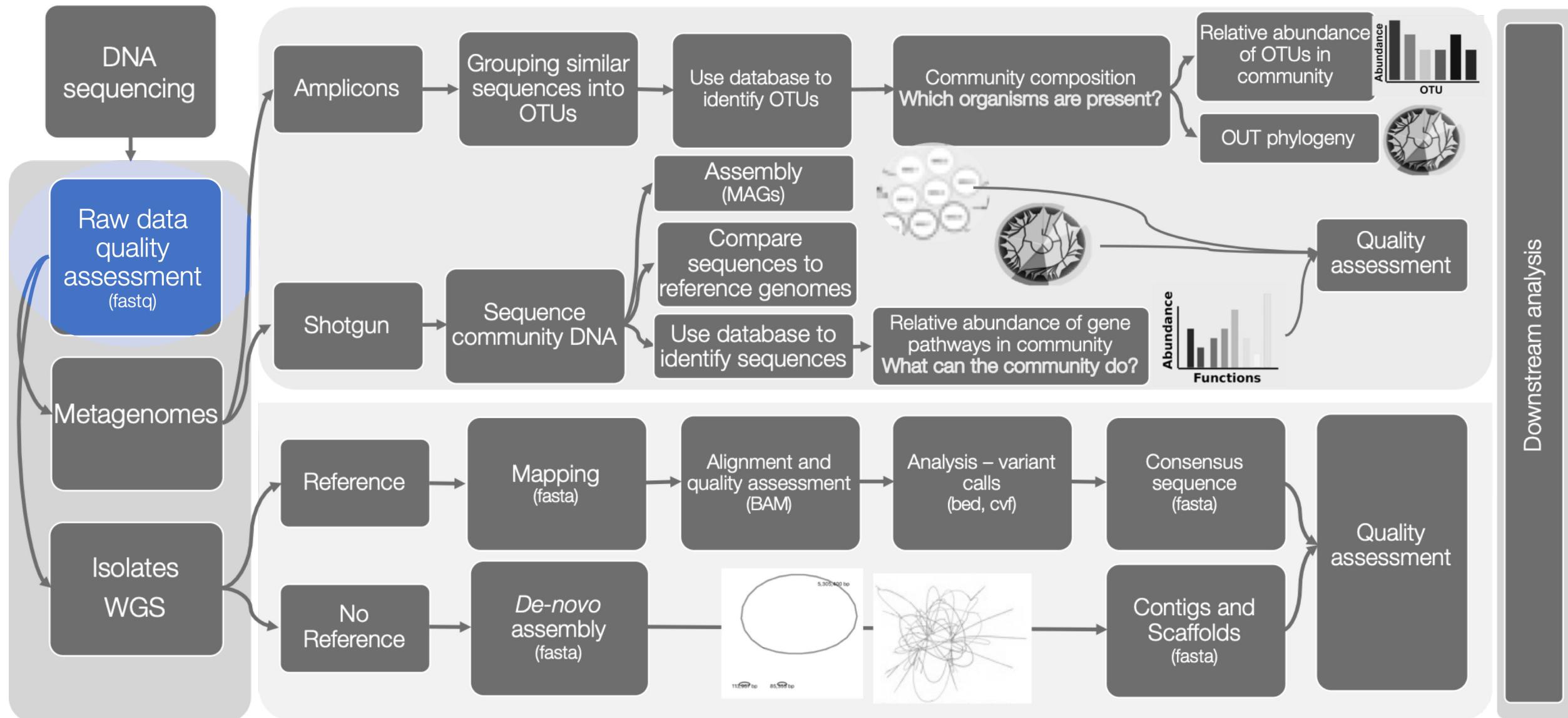




Applications - Food Safety Bioinformatics



Applications - Food Safety Bioinformatics



Applications - Food Safety Bioinformatics

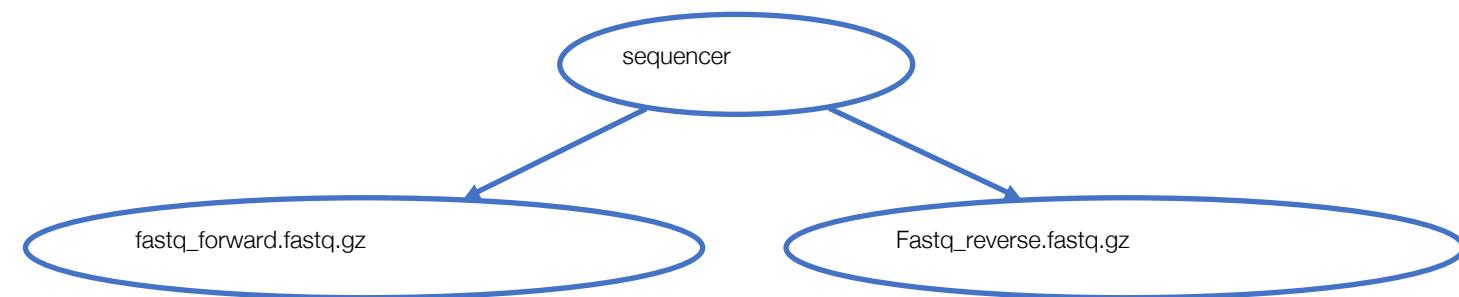
Text-based format for storing both a DNA sequence and its corresponding quality scores

Typically you will get your NGS reads in FASTQ format

- Oxford Nanopore – FAST5 format
- PacBio – HDF5 format

Four lines per sequence (read)

- @ - The UNIQUE sequence name –
- The nucleotide sequence {A, C, G, T, N} – on 1 line
- + - The quality line break
- The quality scores {ASCII characters} – on 1 line



The quality line is always the same length as the sequence line

- The 1st quality score corresponds to the quality of the 1st sequence, and so on...

Multiple FASTQ sequence per file – millions

Sequences can be different length

```
@M05020:90:000000000-CLDT3:1:1101:8838:1072 1:N:0:NTTACTCG+TATAGCCN  
CCTGCGCGGTGCGACTACGACCTGGCGCTGCCGATCATGGCTGGCACTGGCGCGGGN  
+  
CCCCCDGGGGGFEG7FGGGGGGGGGGGFFGC@GGGGGGEGGGGGFGGCGFFF8FCFGGDF#:A
```

-> Sequence name
-> Sequence bases

-> Quality scores

Quality assessment raw data

Fastq

- Text-based format for storing both a DNA sequence and its corresponding quality scores

```
@M05020:90:000000000-CLDT3:1:1101:8838:1072 1:N:0:NTTACTCG+TATAGCCN  
CCTGCGCGGTGCGACTACGACCTGGCGCTGCCGATCATGGCTGGCACTGGCGCGGGN  
+  
CCCCCDGGGGGFEG7FGGGGGGGGGFFGC@GGGGGGEGGGGFGGCGFFF8FCFGGDF#:A
```

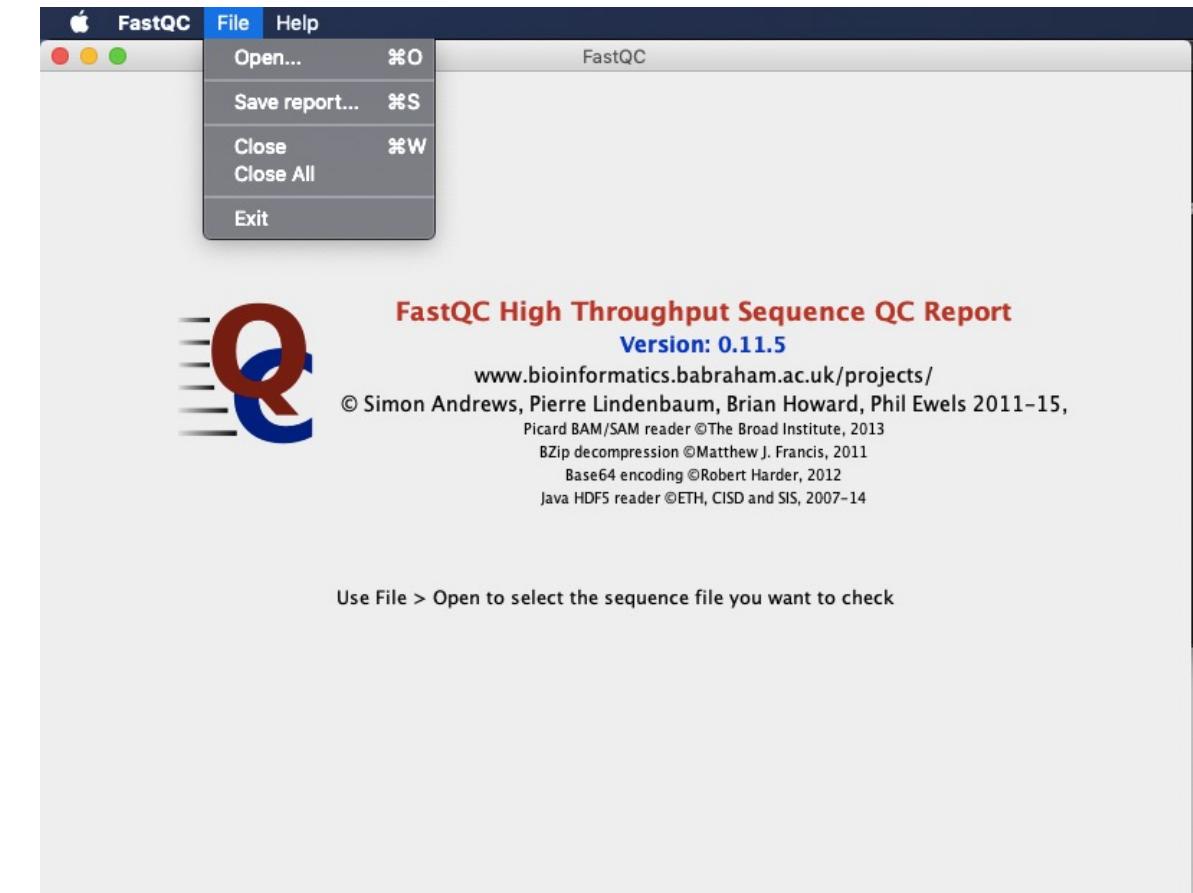
-> Sequence name
-> Sequence bases
-> Quality line break
-> Quality scores

@M05020:90:000000000-CLDT3:1:1101:8838:1072 1:N:0:NTTACTCG+TATAGCCN the unique instrument name
@M05020:90:000000000-CLDT3:1:1101:8838:1072 1:N:0:NTTACTCG+TATAGCCN the run id
@M05020:90:000000000-CLDT3:1:1101:8838:1072 1:N:0:NTTACTCG+TATAGCCN the flowcell id
@M05020:90:000000000-CLDT3:1:1101:8838:1072 1:N:0:NTTACTCG+TATAGCCN flowcell lane
@M05020:90:000000000-CLDT3:1:1101:8838:1072 1:N:0:NTTACTCG+TATAGCCN tile number within the flowcell lane
@M05020:90:000000000-CLDT3:1:1101:8838:1072 1:N:0:NTTACTCG+TATAGCCN 'x'-coordinate of the cluster within the tile
@M05020:90:000000000-CLDT3:1:1101:8838:1072 1:N:0:NTTACTCG+TATAGCCN 'y'-coordinate of the cluster within the tile
@M05020:90:000000000-CLDT3:1:1101:8838:1072 1:N:0:NTTACTCG+TATAGCCN the member of a pair
@M05020:90:000000000-CLDT3:1:1101:8838:1072 1:N:0:NTTACTCG+TATAGCCN Yes or Not if the read fails
@M05020:90:000000000-CLDT3:1:1101:8838:1072 1:N:0:NTTACTCG+TATAGCCN 0 the sequence NOT identified as control (PhiX) >0 sis a control index sequence

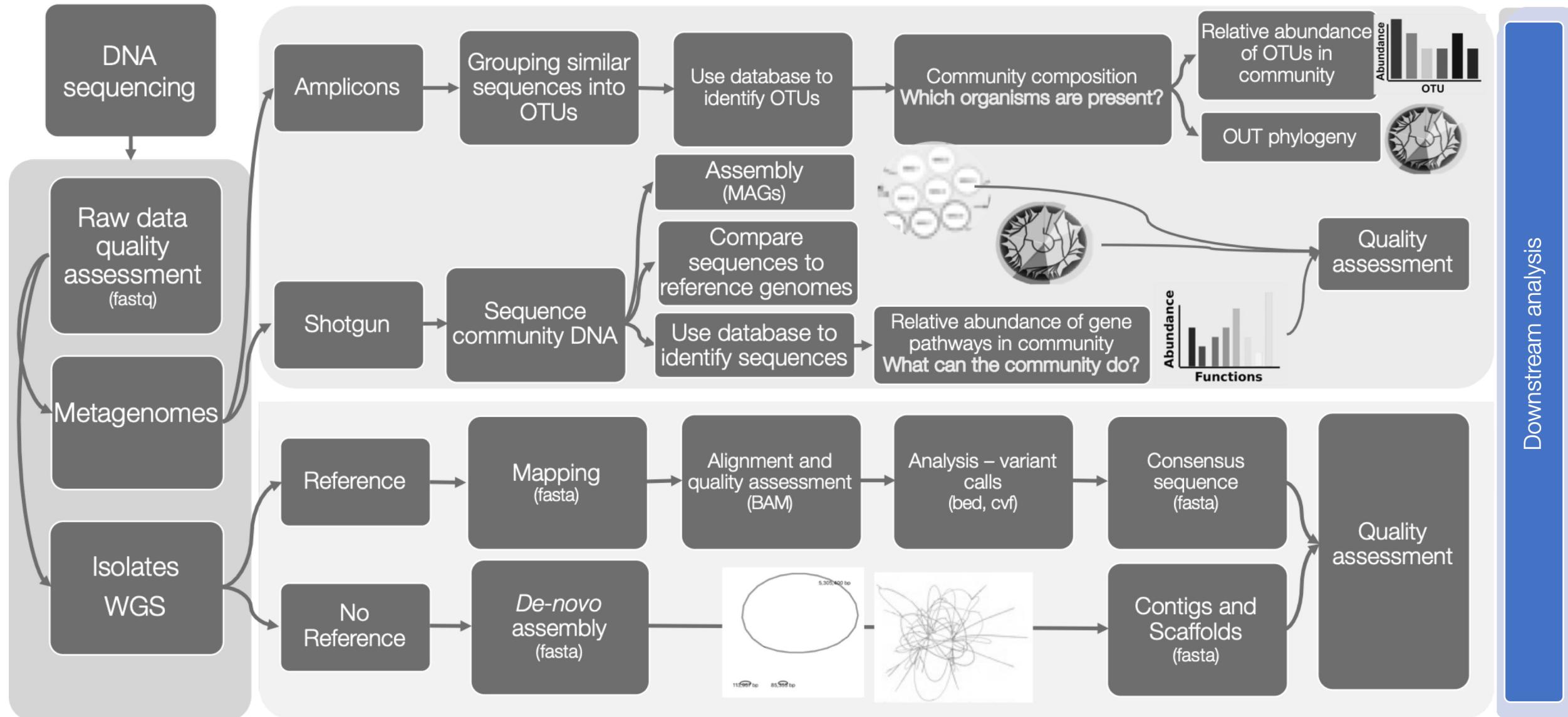
Quality assessment raw data

Fastq

```
guerrinomacori — ~ — java -Xms200m -Xmx250m -Xdock:name=FastQC -Xd...
Last login: Tue Oct 27 20:44:07 on ttys000
guerrinomacori@dhcp-892b0edb ~
[$ fastqc
```



Food Safety Bioinformatics



Annotation

Microbial genome annotation often consists of running an automatic annotation pipeline followed by manual curation of the results

Homology methods to transfer information from a closely related reference genome to the new sequence.

Automatic pipelines with errors and then manual curation step to catch and remove these.

Whole genome annotation

Annotation

Tools for bacterial whole genome annotation

[Prokka](#) – Standalone command line tool, takes just a few minutes per genome. This is the best way to get good quality annotation in a flash, which is particularly useful if you have loads of genomes or need to annotate a pangenome or metagenome. Note however that the quality of functional information is not as good as RAST, and you will need several extra steps if you want to do functional profiling and pathway analysis of your genome(s)... which is in-built in RAST.

[PGAP](#): NCBI has developed an automatic prokaryotic genome annotation pipeline that combines ab initio gene prediction algorithms with homology based methods. The first version of NCBI Prokaryotic Genome Automatic Annotation Pipeline (PGAAP; [see Pubmed Article](#)) developed in 2005 has been replaced with an upgraded version that is capable of processing a larger data volume. NCBI's annotation pipeline depends on several internal databases and is not currently available for download or use outside of the NCBI environment.

[BEACON](#) (automated tool for Bacterial GEnome Annotation ComparisON), a fast tool for an automated and a systematic comparison of different annotations of single genomes. The extended annotation assigns putative functions to many genes with unknown functions. BEACON is available under GNU General Public License version 3.0 and is accessible at: <http://www.cbrc.kaust.edu.sa/BEACON/>.

[BlastKOALA](#): Assigns K numbers to the user's sequence data by BLAST searches, respectively, against a nonredundant set of KEGG GENES. KOALA (KEGG Orthology And Links Annotation) is KEGG's internal annotation tool for K number assignment of KEGG GENES using SSEARCH computation. Annotate Sequence in KEGG Mapper and Pathogen Checker in KEGG Pathogen are special interfaces to this server and can be executed in an interactive mode. BlastKOALA is suitable for annotating fully sequenced genomes.

[PAGIT](#): Provides a toolkit for improving the quality of genome assemblies created via an assembly software. PAGIT compiled four tools: (i) ABACAS which classifies and orientates contigs and estimates the sizes of gaps between them; (ii) IMAGE uses paired-end reads to extend contigs and close gaps within the scaffolds; (iii) ICORN for identifying and correcting small errors in consensus sequences and; (iv) RATT for help annotation. The software was mainly created to analyze parasite genomes of up to about 300 Mb.

[MAKER](#): A portable and easily configurable genome annotation pipeline. MAKER allows smaller eukaryotic and prokaryotic genome projects to independently annotate their genomes and to create genome databases. It identifies repeats, aligns ESTs and proteins to a genome, produces ab-initio gene predictions and automatically synthesizes these data into gene annotations having evidence-based quality values. MAKER's inputs are minimal and its outputs can be directly loaded into a Generic Model Organism Database (GMOD). They can also be viewed in the Apollo genome browser; this feature of MAKER provides an easy means to annotate, view and edit individual contigs and BACs without the overhead of a database. MAKER is available for download and can be tested online via the MAKER Web Annotation Service (MWAS).

[MyPro](#) is a software pipeline for high-quality prokaryotic genome assembly and annotation. It was validated on 18 oral streptococcal strains to produce submission-ready, annotated draft genomes. MyPro installed as a virtual machine and supported by updated databases will enable biologists to perform quality prokaryotic genome assembly and annotation with ease.

Annotation

Downstream analysis



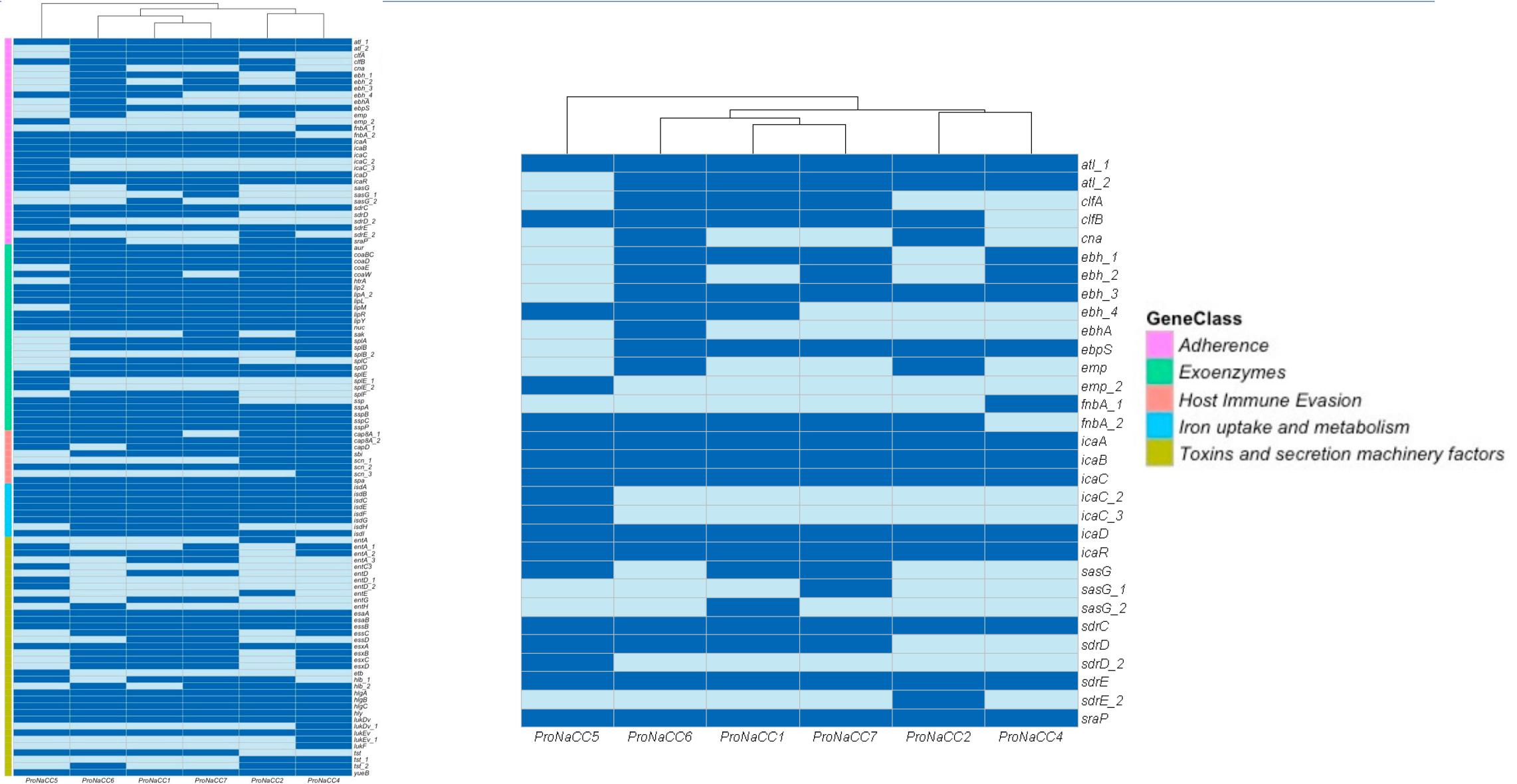
Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014 Jul 15;30(14):2068-9. [PMID:24642063](#)

Prokka is a software tool for the rapid annotation of prokaryotic genomes. A typical 4 Mbp genome can be fully annotated in less than 10 minutes on a quad-core computer, and scales well to 32 core SMP systems. It produces GFF3, GBK and SQN files that are ready for editing in Sequin and ultimately submitted to Genbank/DDJB/ENA.

PROKKA_11092021

locus_tag	ftype	length_bp	gene	EC_number	COG	product
JCJMHEKL_00001	CDS	300				hypothetical protein
JCJMHEKL_00002	CDS	1794				hypothetical protein
JCJMHEKL_00003	CDS	300				hypothetical protein
JCJMHEKL_00004	CDS	1239	tyrS_1	6.1.1.1	COG0162	Tyrosine--tRNA ligase
JCJMHEKL_00005	CDS	1257	nicP_1			Porin-like protein NicP
JCJMHEKL_00006	CDS	1209	bcr_1			Bicyclomycin resistance protein
JCJMHEKL_00007	CDS	423	farR			HTH-type transcriptional regulator FarR
JCJMHEKL_00008	CDS	243				hypothetical protein
JCJMHEKL_00009	CDS	1428	ccoN1_1	1.9.3.1	COG3278	Cbb3-type cytochrome c oxidase subunit CcoN1
JCJMHEKL_00010	CDS	1503	norG_1		COG1167	HTH-type transcriptional regulator NorG
JCJMHEKL_00011	CDS	1431				hypothetical protein
JCJMHEKL_00012	CDS	1674	sir_1	1.8.7.1	COG0155	Sulfite reductase [ferredoxin]
JCJMHEKL_00013	CDS	489				hypothetical protein
JCJMHEKL_00014	CDS	807	hpcH	4.1.2.52	COG3836	4-hydroxy-2-oxo-heptane-1,7-dioate aldolase
JCJMHEKL_00015	CDS	804	hpcG	4.2.1.163		2-oxo-hept-4-ene-1,7-dioate hydratase
JCJMHEKL_00016	CDS	1305	nicT			Putative metabolite transport protein NicT
JCJMHEKL_00017	CDS	393	hpcD_1	5.3.3.10		5-carboxymethyl-2-hydroxymuconate Delta-isomerase
JCJMHEKL_00018	CDS	924	hpcB	1.13.11.15	COG3384	3,4-dihydroxyphenylacetate 2,3-dioxygenase
JCJMHEKL_00019	CDS	1461	betB_1	1.2.1.8	COG1012	NAD/NADP-dependent betaine aldehyde dehydrogenase
JCJMHEKL_00020	CDS	780	hpcE_1		COG0179	Homoprotocatechuate catabolism bifunctional isomerase/decarboxylase
JCJMHEKL_00021	CDS	660	hpcE_2		COG0179	Homoprotocatechuate catabolism bifunctional isomerase/decarboxylase
JCJMHEKL_00022	CDS	912	rhaR_1			HTH-type transcriptional activator RhaR
JCJMHEKL_00023	CDS	807	neo	2.7.1.95		Aminoglycoside 3'-phosphotransferase
JCJMHEKL_00024	CDS	570				hypothetical protein
JCJMHEKL_00025	CDS	2187	bphP	2.7.13.3		Bacteriophytocrome
JCJMHEKL_00026	CDS	588				hypothetical protein
JCJMHEKL_00027	CDS	1386	ppnN	3.2.2.-	COG1611	Pyrimidine/purine nucleotide 5'-monophosphate nucleosidase
JCJMHEKL_00028	CDS	510				hypothetical protein
JCJMHEKL_00029	CDS	1191	sotB_1			Sugar efflux transporter B

Downstream analysis

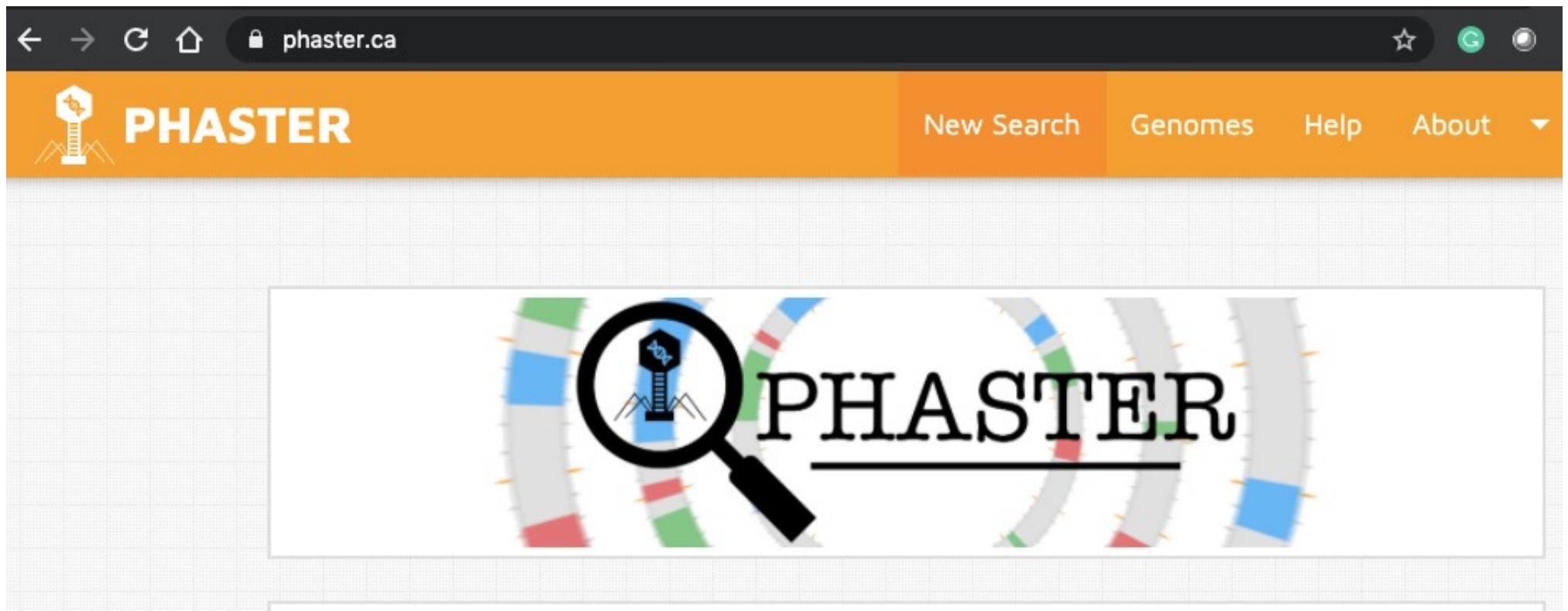


Annotation

In silico Phage Hunting

PHASTER (PHAge Search Tool Enhanced Release)

Downstream analysis



Annotation

In silico Phage Hunting

PHASTER (PHAge Search Tool Enhanced Release)

Downstream analysis

Total: 1 prophage regions have been identified, of which 1 regions are intact, 0 regions are incomplete, and 0 regions are questionable.

Region	Region Length	Completeness	Score	# Total Proteins	Region Position	Most Common Phage	GC %	Details
1	38.2Kb	intact	131	55	156682-194926 ⓘ	PHAGE_Pseudo_JD024_NC_024330(51)	64.35%	Show ⓘ

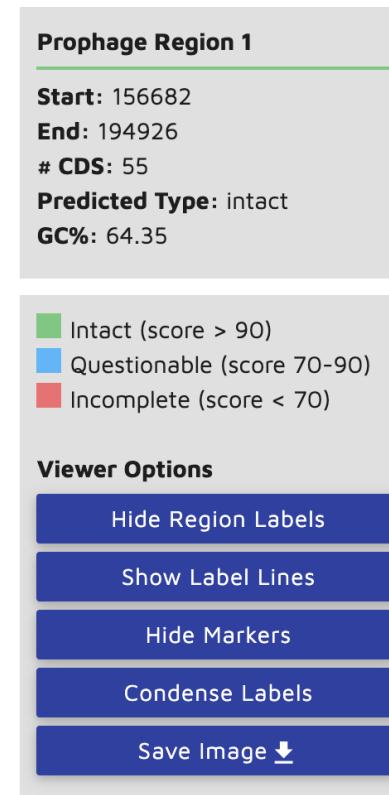
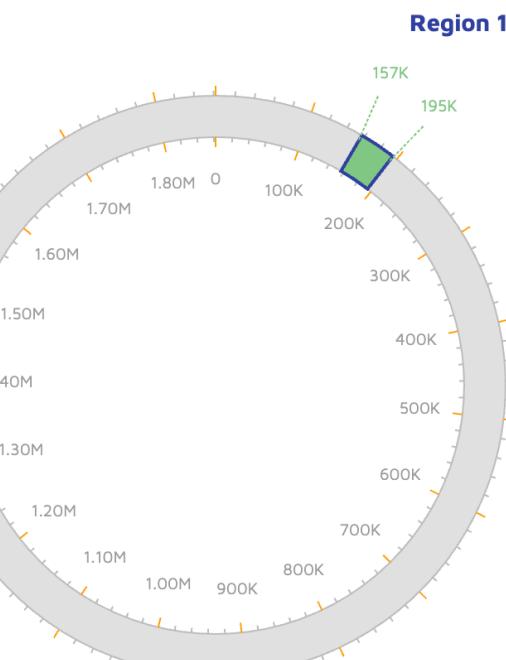
- Intact (score > 90)
- Questionable (score 70-90)
- Incomplete (score < 70)

Annotation

In silico Phage Hunting

Downstream analysis

[">NODE_1_length_1862038_cov_39.509666"]



Region	1
Region Length	38.2Kb
Completeness(score)	intact(131)
Capsid Keyword	transposase,capsid,virion,tail
Region Position	156682-194926
RNA	0
Total Proteins	55
Phage Hit Proteins	55
Hypothetical Proteins	0
Phage + Hypothetical Protein %	100%

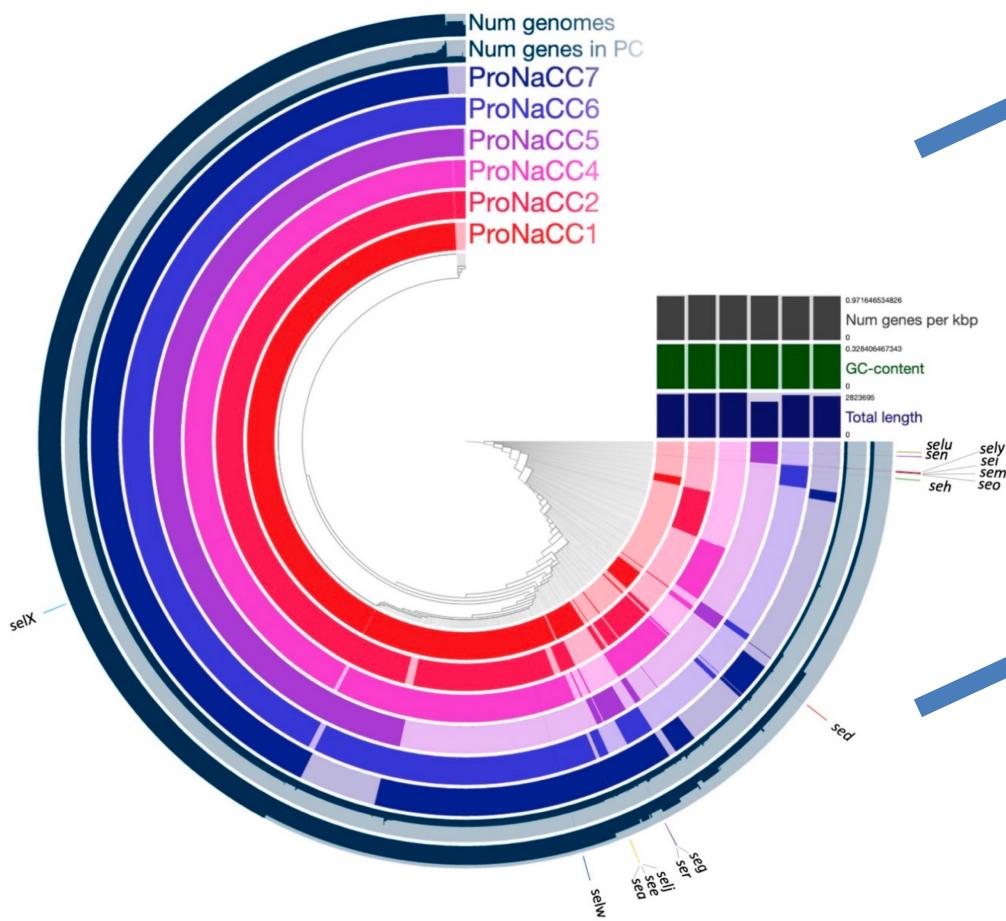
Length: 1862038 bps

Phages: 1

Annotation

In silico Phage Hunting

Downstream analysis



GGTCGTAATATCGCACGTCCATATGTTGGTGAACCAGGAAACTTTAC
ACGTACATCTAATCGACATGACTATGCCTTAAAACCTTTGGTAAAAA
CTGTCTTAGATCATTGAAAGACGGTGGTTATGATGTTATTGCCATC
GGTAAAATTAAATGACATTTATGATGGTGAAGGTGTAACAGAACGGT
TCGTACGAGAGTAACATGGACGGTATGGATCAATTGATGAAAATTG
TTAAGAAAAGATTTACAGGTATTAGCTTCTTAAACTTAGTAGACTTT
GATGCATTATACGGTCATCGTCGTGATAAACCCAGGTTATGCACAAGC
AATTAAAGATTTCGATGATCGCTTCCCAGAACTGTTAGCAACTTAA
AAGAAGACGATTAGTAATTATTACAGCAGACCATGGTAATGACCCG
ACAGGCCAGGTACGGACCACGAGAGAATATATCCCAGTAATTAT
GTACAGTCCGAAATTAAAGGTGGTCATGCACTAGAAAAGTGATACTA
CATTCAAGTTCTATCGGTGCAACTATAGCAGATAATTCAACGTAACA
TTACCAGAGTTGGTAAAAGTTATTAAAGGAATTGAAATAGAATAAA
ATTTAGATATTATAAAAACAGCAGTGAAGTTAACTATAACAATAGTT
TTCTTCACTGCTTTTATTATAATAGAGAAACGTAAGACG

GGTCGTAATATCGCACGTCCATATGTTGGTGAACCAGGAAACTTTAC
ACGTACATCTAATCGACATGACTATGCCTTAAAACCTTTGGTAAAAA
CTGTCTTAGATCATTGAAAGACGGTGGTTATGATGTTATTGCCATC
GGTAAAATTAAATGACATTTATGATGGTGAAGGTGTAACAGAACGGT
TCGTACGAGAGTAACATGGACGGTATGGATCAATTGATGAAAATTG
TTAAGAAAAGATTTACAGGTATTAGCTTCTTAAACTTAGTAGACTTT
GATGCATTATACGGTCATCGTCGTGATAAACCCAGGTTATGCACAAGC
AATTAAAGATTTCGATGATCGCTTCCCAGAACTGTTAGCAACTTAA
AAGAAGACGATTAGTAATTATTACAGCAGACCATGGTAATGACCCG
ACAGGCCAGGTACGGACCACGAGAGAATATATCCCAGTAATTAT
GTACAGTCCGAAATTAAAGGTGGTCATGCACTAGAAAAGTGATACTA
CATTCAAGTTCTATCGGTGCAACTATAGCAGATAATTCAACGTAACA
TTACCAGAGTTGGTAAAAGTTATTAAAGGAATTGAAATAGAATAAA
ATTTAGATATTATAAAAACAGCAGTGAAGTTAACTATAACAATAGTT
TTCTTCACTGCTTTTATTATAATAGAGAAACGTAAGACG

Publicly available antibiotic resistance and virulence databases

Antimicrobial resistance databases	Virulence databases
CARD (https://card.mcmaster.ca/) (23)	VFDB (http://www.mgc.ac.cn/VFs/) (24)
RAC (http://rac.aihi.mq.edu.au/rac/) (25)	PATRIC (https://www.patricbrc.org/) (26, 27)
ResFinder (https://cge.cbs.dtu.dk/services/data.php) (21)	Victors (http://www.phidias.us/victors/) (28)
ARDB (https://ardb.cbcn.umd.edu/) (29)	PHI-BASE (http://www.phi-base.org/) (30-33)
NDARO https://www.ncbi.nlm.nih.gov/pathogens/antimicrobial-resistance/	MvirDB (http://mvirdb.llnl.gov/) (34)

Options bioinformatics analysis for Food Safety

- Command line
- Stand alone web-resources (for example <http://www.genomicepidemiology.org/services/>)
- Online pipelines (for example <https://galaxyproject.org/eu/>)

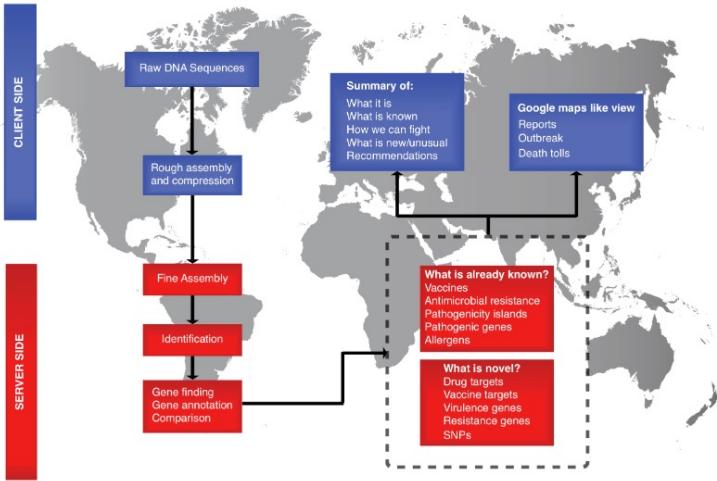
Bioinformatic tools for characterising bacterial pathogens

Live exercise
Using command line/terminal

```
abricate --db resfinder CFS1073_ill.fasta
```

Center for Genomic Epidemiology

Home Services Publications



Welcome to the Center for Genomic Epidemiology

The use of sequencing technologies is currently transforming almost every aspect of biological science. In relation to infectious diseases, the advances are rapidly changing our scientific discoveries, as well as diagnostic and outbreak investigations. The ability to analyze sequencing data and take advantage of the rapid progress, is however, not equally distributed between institutions and countries.

The aim of the Center for Genomic Epidemiology (CGE) is to provide access to bioinformatics resources also for those with limited experience and thereby allow all countries, institutions and individuals to take advantage of the novel sequencing technologies. Doing so, we hope to facilitate more open data sharing around the world and provide more equal opportunities for all.

CGE is entirely non-commercial and operates a number of free online bioinformatics services. Funding is provided as core funding from the Technical University of Denmark (DTU) and from a range of public and private sources as listed below.

If you want to read more about us and our research activities, please visit the [Global Surveillance website](#).

Overview of Services

Phenotyping

[ResFinder](#)

Identification of acquired antibiotic resistance genes.

[ResFinderFG](#)

Identification of functional metagenomic antibiotic resistance determinants.

[LRE-finder](#)

Identification of genes and mutations leading to linezolid resistance.

[KmerResistance](#)

Identification of acquired antibiotic resistance genes using Kmers.

[PathogenFinder](#)

Prediction of a bacteria's pathogenicity towards human hosts.

[VirulenceFinder](#)

Identification of acquired virulence genes.

[Restriction-ModificationFinder](#)

Determination of Restriction-Modification sites (based on REBASE.)

[SPIFinder](#)

SPIFinder identifies Salmonella Pathogenicity Islands.

[ToxFinder](#)

ToxFinder identifies genes involved in mycotoxin synthesis.

Typing

[MLST](#)

Multi Locus Sequence Typing (MLST) from an assembled genome or from a set of reads.

[PlasmidFinder](#)

PlasmidFinder identifies plasmids in total or partial sequenced isolates of bacteria.

[pMLST](#)

Multi Locus Sequence Typing (MLST) from an assembled plasmid or

Phylogeny

[MINTyper](#)

Identification of SNPs with automatic filtering, masking and site validation together with inferred phylogeny based on both long and short sequencing data.

[CSIPhylogeny](#)

CSI Phylogeny calls SNPs, filters the SNPs, does site validation and infers a phylogeny based on the concatenated alignment of the high quality* SNPs.

[NDtree](#)

NDtree constructs phylogenetic trees from Single-End or Pair-End FASTQ files.

[Evergreen](#)

Evergreen generates a forest of constantly updated phylogenetic trees with publicly available whole-genome sequencing data from foodborne, bacterial isolates that were deposited in the short sequencing read archives (NCBI SRA/ENA).

[TreeViewer](#)

Phylogeny Tree Viewer.

Metagenomics

[CCMetagen](#)

CCMetagen: Comprehensive and accurate identification of eukaryotes and prokaryotes in metagenomic data.

PCR-tools

[RUCS](#)

RUCS: Rapid Identification of PCR Primers Pairs for Unique Core Sequences.

Other

[MyKMAfinder](#)

Center for Genomic Epidemiology

Home Services Publications Contact

ResFinder 4.1

Service Instructions Output Article abstract Citations Overview of genes Database history

ResFinder identifies acquired genes and/or finds chromosomal mutations mediating antimicrobial resistance in total or partial DNA sequence of bacteria.

ResFinder and PointFinder software: (2022-08-08)
ResFinder database: (2022-05-24)
PointFinder database: (2022-04-22)

For analysis part of EFSA, go to [ResFinder-EFSA](#)

[Chromosomal point mutations](#)

[Acquired antimicrobial resistance genes](#)

Select species

*Comprehensive point mutation database exists

Select type of your reads

Choose Field Name Size Progress Status

If you get an "Xcode forbidden, Error 403" Make sure the start of the web address is https and not just http. Fix it by clicking here.

<input checked="" type="checkbox"/> Choose Field	Name	Size	Progress	Status
<input type="button" value="Upload"/>	<input type="button" value="Remove"/>			

Live exercise web resources

<https://www.genomicepidemiology.org/services/>

Center for Genomic Epidemiology

[Home](#)[Services](#)[Publications](#)[Contact](#)

ResFinder 4.1

Service [Instructions](#) [Output](#) [Article abstract](#) [Citations](#) [Overview of genes](#) [Database history](#)

ResFinder identifies acquired genes and/or finds chromosomal mutations mediating antimicrobial resistance in total or partial DNA sequence of bacteria.

The database is curated by:
Frank Møller Aarestrup
(click to contact)

ResFinder and PointFinder software: (2022-08-08)
ResFinder database: (2022-05-24)
PointFinder database: (2022-04-22)

For analysis part of EFSA, go to [ResFinder-EFSA](#)

Chromosomal point mutations □

Acquired antimicrobial resistance genes □

Select species
Campylobacter spp.*

*Chromosomal point mutation database exists

Select type of your reads
Assembled Genome/Contigs

If you get an "Access forbidden. Error 403": Make sure the start of the web address is https and not just http. Fix it by clicking [here](#).

Choose File(s)
Name Size Progress Status

Try different services using
the assemblies provided

Live exercise
web resources

<http://www.genomicepidemiology.org/services/>

Galaxy Europe

The homepage of the European Galaxy community

Galaxy is an **open-source** platform for **FAIR** data analysis that enables users to:

- use **tools** from various domains (that can be plugged into **workflows**) through its graphical web interface.
- run code in **interactive environments** (RStudio, Jupyter...) along with other tools or workflows.
- **manage data** by sharing and publishing results, workflows, and visualizations.
- **ensure reproducibility** by capturing the necessary information to repeat and understand data analyses.

The **Galaxy Community** is actively involved in helping the ecosystem improve and sharing scientific discoveries.



Live exercise
Using <https://galaxyproject.org/eu/>

Bioinformatic tools for characterising bacterial pathogens

← → ⌂ 🔍 usegalaxy.org

Galaxy Workflow Visualize Shared Data Help Login or Register 🔔 🎓 📱 G Update

Using 0%

Tools search tools Upload Data

Get Data Send Data Collection Operations

GENERAL TEXT TOOLS

- Text Manipulation
- Filter and Sort
- Join, Subtract and Group
- Datamash

GENOMIC FILE MANIPULATION

- FASTA/FASTQ
- FASTQ Quality Control
- SAM/BAM
- BED
- VCF/BCF
- Nanopore

Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources. You can install your own Galaxy by following the tutorial and choose from thousands of tools from the Tool Shed.

Galaxy in numbers

- 53 talks
- 47 speakers
- 43 posters
- 15 BoFs
- 131 participants
- 52 CoFest Heros
- 31 countries
- 130 Video hours
- 75 hours
- 100 instructors
- Beers

Galaxy version 22.05.1, commit ace1e13da34b67275a757f83c3f766d296bc8a1f

History + 🔍

search datasets

Unnamed history

0 B 0 🔍

This history is empty.
You can load your own data or get data from an external source.

Live exercise

Using <https://galaxyproject.org/eu/>

Summary and Glossary

- **Big data/grid/cloud** - With the increasing volume and heterogeneity of data sets (often referred to as “Big Data”), high performance computing is needed for analysis of the data. Many bioinformatics methods have been adapted to run on clusters of multiple computers (grid computing) and on large remotely located servers (cloud computing). Galaxy and KNIME are two popular software solutions to integrate and distribute larger data analysis tasks to the grid/cloud. Examples of cloud-based bioinformatics applications: HBLAST (BLAST, the most used bioinformatics sequence alignment tool); TPP, a proteomic analysis tool; HIPPIE: promoter analysis provided as Amazon Machine Image; BG: bacterial genome annotation based on Amazon Web Services .
- **Data mining** - Statistical and machine learning techniques to determine trends in typically large data sets. Unsupervised techniques (sample grouping is not explicitly used in the analysis) include: principal component analysis (PCA) and clustering algorithms (e.g. K-means, hierarchical). Supervised techniques (sample grouping is taken into account) include: ANOVA, Mann-Whitney U test, partial least squares analysis (PLS), machine learning (e.g. by support vector machines (SVM) and random forest): a computational model is trained to use properties derived from samples to predict the status of samples.
- **Virtual machines (VM)** - A large computer file (disk image) that consists of an operating system (e.g. Linux), software tools and data. The image can be run on an actual computer using virtual machine software that emulates an actual computer. In other words, a computer in a computer. The advantage of VMs is that they are portable (can be run on many different types of computer hardware), easy to backup and more straight-forward to maintain. Examples of the use of VMs are the generic bioinformatics tools in the NEBC Bio-Linux distribution and the 16s analysis suite Qiime
- **Databases** - Databases are organized collections of biological data. Bioinformatics is only successful if databases with high-quality data are available, together with structured vocabularies that describe the content of the data sets. An updated overview of relevant biological databases can be found here: http://www.oxfordjournals.org/our_journals/nar/database/cap/.

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4793891/>

Summary and Glossary

- **Genome sequencing** - Determining the complete genome sequence of a microbial strain of interest. Next-generation sequencing (NGS) techniques allow for high-throughput and high-quality sequencing results. Especially the combination of different techniques (e.g. Illumina and Pacific Biosystems or PacBio) result in high-quality (circular) genomes.
- **Sequencing data (FASTQ)** - Sequencing data are represented in FASTQ format. These files provide, next to the raw sequence data, additional information regarding the quality of the reads. In this manner, quality control and trimming can be applied.
- **Assembly** - Raw sequence reads of different NGS technologies can be assembled into contigs, long stretches of DNA sequence representing part of the genome. Most of the assembly methods are based on alignment of sequence reads with each other (de novo assembly) or against a reference genome (mapping assembly), thereby generating long DNA sequences from the fragments generated by the sequencing. Some examples of assembly tools are SPAdes, Ray, MIRA and IDBA.
- **Scaffolding** - Organizing the contigs from the assembly to larger, gapped, DNA sequences. Some NGS techniques (e.g. Illumina) allow the synthesis of paired end (PE) or mate pair (MP) libraries; libraries with a fixed insert size that are sequenced at both ends. As reads span a larger DNA fragment, the matched reads pairs can be used to order contigs, even if the sequence in between the contigs has not been assembled. In general, most assembly tools allow for scaffolding, but also dedicated tools exist, such as SSPACE.
- **Gap closure strategies** - After scaffolding, genome sequences will most often contain gaps. Common strategies to fill these gaps are generating new sequencing data using, for example, PacBio's long reads, or predicting the most likely order and orientation of the contigs using bioinformatics tools like Projector2 or Mauve. These tools infer contig order by comparing them to one or more reference sequences.
- **Gene function annotation** - Gene function is typically inferred from similarity in amino acid sequence. Gene functions can be predicted by comparing sequences to databases containing genes with known functions with tools like RAST and Prokka.
- **Orthology** - Genes in different organisms are orthologous when they were the same gene in the last common ancestor. Reconstructing the evolutionary history of genes allows the prediction of functional equivalence (i.e. orthologous genes are likely to have similar functions). Tools are OrthoMCL and Orthogologue.

Summary and Glossary

- **Comparative genomics** - All analyses in which genome sequences or genome content of multiple organisms are compared.
- **Metabolic modelling** - Prediction of growth, and recruitment of metabolic pathways, of microbes by using the genome sequence as an inventory of all possible metabolic reactions. Genome-scale metabolic models can be constructed using automated or comparative genomics analyses. Once constructed, the models can be used to simulate growth by, for example, flux balance analysis (FBA) and to determine the boundaries of fluxes by flux variability analysis (FVA). Tools for modelling are Pysces, the SEED and VANTED.
- **Microbiome analysis** - All microbes present in a particular niche are termed a microbiome. Analysis of microbiomes can be done using different next-generation sequencing-based techniques (see below).
- **16s rRNA sequencing** - 16s amplicon sequencing is the generation of sequence reads from conserved regions of the 16s gene. Amplicon sequencing (e.g. by Illumina) is used to identify the bacterial (and sometimes archaeal) component of microbial communities. Examples of software to infer community composition from sequencing data are Qiime and Mothur. 16s sequencing is a relatively cheap and well-established technique and as such an ideal starting point for characterization of complex cultures for which limited prior knowledge is available.
- **Functional prediction** - 16s sequences derived from a particular ecological niche indicate the taxa present and their relative abundance. From these data, presence of gene functions in those taxa can be performed using, e.g., PICRUSt. It infers the presence of gene functions in given taxa using already sequenced genomes part of the same taxa.
- **Shotgun metagenomics and metatranscriptomics** - Random fragments of the DNA or (enriched) mRNA of a given microbiome are sequenced with next-generation sequencing. Metagenomics and metatranscriptomics techniques are powerful, as they allow circumventing growing microbes while still determining their gene content or gene expression. This provides insight into the molecular functions encoded by the DNA, taxonomic assignment of that DNA fragments or inferring similar information for expressed mRNAs. This method can be used in addition to sequencing of individual isolates from complex cultures. Sequencing of individual isolates, however, has the advantage that comparative genomics can be done and metabolic models can be built more straight-forwardly, provided that the isolates under study are representative of the biodiversity present in the complex culture.

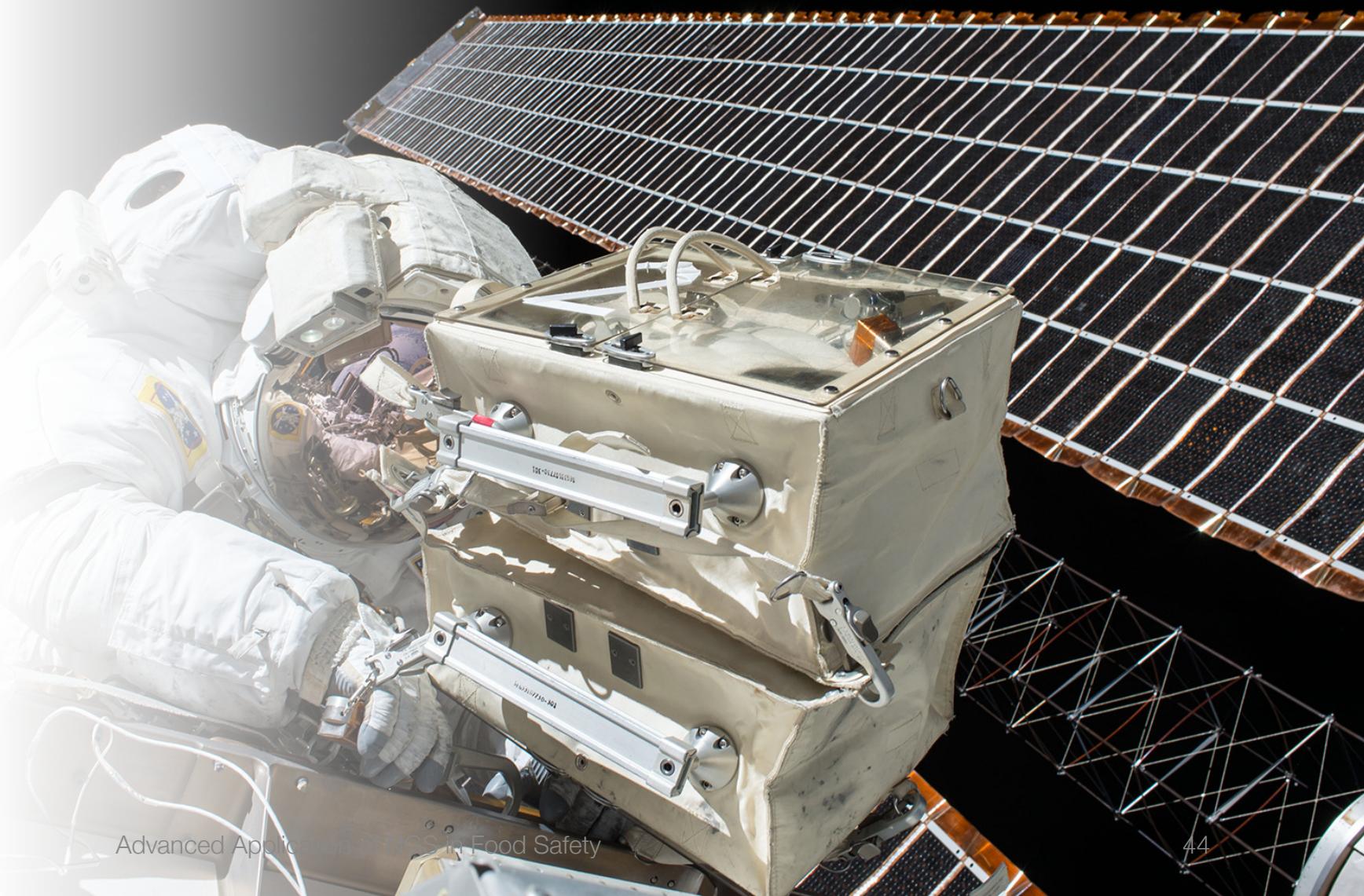
Summary and Glossary

- **Assembly** - Using the sequence overlap, the DNA/RNA-derived sequences can be assembled into larger contigs. Functional annotation of these larger fragments is more straight-forward, but the fraction of reads that can be assembled into contigs depends on both the complexity of the microbiome (many different microbes with varying abundances) as well as the presence of microdiversity (many different microbes with similar genome sequences).
- **Annotation** - Similar to the genome of a single bacterium, the sequences of a metagenome can be functionally and taxonomically annotated by comparing (assembled) sequences or predicted gene products against one or more reference databases with sequences with known functions from known taxonomic origin. Gene context such as operons are, however, primarily missing in shotgun metagenomics reads/contigs. A few tools are: PhymmBL (taxonomic classification using sequence-based models), MG-RAST (functional and taxonomic classification using alignment to reference databases) and MetaPhlAn (taxonomy prediction using taxon-specific marker genes).
- **Strain typing and tracking** - Pinpointing the presence of a particular microbe (strain) in a biological sample. Using MLST markers (multi-locus sequence typing), PCR based on unique DNA fragments or strain-specific markers, the abundance of particular strains can be followed during the course of a fermentation. Potential downside of these techniques is that only known biodiversity can be traced. Therefore, the performance must be evaluated on new strains. New biodiversity can be uncovered, provided that a genomic target is well-designed (e.g. targeting a gene that is single copy with sufficient resolution to distinguish between strains).
- **Predicting phenotypes** - Gene–trait matching: machine learning or statistics methods are used to predict the phenotype of a bacterial strain based on the presence/absence of particular genes, (parts of) metabolic pathways or classifications from experts. Transcriptome–trait matching: gene expression data (based on microarray or RNAseq) instead of gene presence are used. Transcriptome data from multiple strains grown under the same condition or the same strain grown under different conditions can be used.
- **Metabolomics** - The simultaneous measurement of multiple metabolites in biological samples. Metabolomics is a technique that can be applied to describe reaction products of microorganisms in defined media and in food samples. Its data are very suitable to be associated to results from sensory measurements.

Summary and Glossary

- **consensus sequence** - is the calculated order of most frequent residues, either nucleotide or amino acid, found at each position in a sequence alignment. It serves as a simplified representation of the population. It represents the results of multiple sequence alignments in which related sequences are compared to each other and similar sequence motifs are calculated. Such information is important when considering sequence-dependent enzymes such as RNA polymerase.

- In the modern world, it often seems that the age of exploration is over.
- We have been on the Moon, there are research bases on Antarctica, we can look at atoms, we are exploring Mars, working routinely on the ISS
- You may ask yourself are there wonders waiting to be discovered?
- After all, it may seem that all we are left with is to refine and improve on prior observations.



Everyone has the chance to make unique and fascinating discoveries



The screenshot shows the National Library of Medicine's Sequence Read Archive (SRA) page. At the top, the NIH logo and "National Library of Medicine" are displayed, along with a user profile for "g.macori". Below the header is a search bar with "SRA" selected and an "Advanced" link. To the right of the search bar are "Search" and "Help" buttons. A large banner on the left features a blue, glowing, abstract molecular or cellular structure. To its right, the text "SRA - Now available on the cloud" is displayed in bold. Below this, a detailed description reads: "Sequence Read Archive (SRA) data, available through multiple cloud providers and NCBI servers, is the largest publicly available repository of high throughput sequencing data. The archive accepts data from all branches of life as well as metagenomic and environmental surveys. SRA stores raw sequencing data and alignment information to enhance reproducibility and facilitate new discoveries through data analysis." At the bottom of the page, the "GeneLab" logo is visible, followed by the text "Open Science for Life in Space" and a navigation menu with links to Home, About, Data & Tools, Research & Resources, Working Groups, and Help. A search bar with "Keywords" and a magnifying glass icon is also present. The main content area features a large image of a satellite in space against a backdrop of Earth's horizon and a colorful light gradient. Below this image, a welcome message states: "Welcome to NASA GeneLab - the first comprehensive space-related omics database; users can upload, download, share, store, and analyze spaceflight and spaceflight-relevant data from experiments using model organisms." At the bottom, there are six functional buttons: "Data Repository" (Search and upload spaceflight datasets), "Analysis Workflows" (Processing pipelines for sequencing data), "Environmental Data" (Radiation data collected during experiments conducted in space), "Collaborative Workspace" (Share, organize, and store files), "Submit Data" (Have space-relevant data to submit to GeneLab?), and "Visualize Data" (Interact with GeneLab processed data).



Thank you for the attention!



Stellenbosch
UNIVERSITY
IYUNIVESITHI
UNIVERSITEIT



www.ucd.ie/cfs
[@CfsUcd](https://twitter.com/CfsUcd)