Metagenomics
MBDP-102

Jenni Hultman, Antti Karkman, Tatiana Demina
Department of Microbiology
University of Helsinki

Learning goals

- Foundational skills to work with metagenomic data
- Familiarity and practice with bioinformatics tools
- Perspective and confidence to apply these skills in your own work
- Empower you to ask and answer the questions you have of your own data

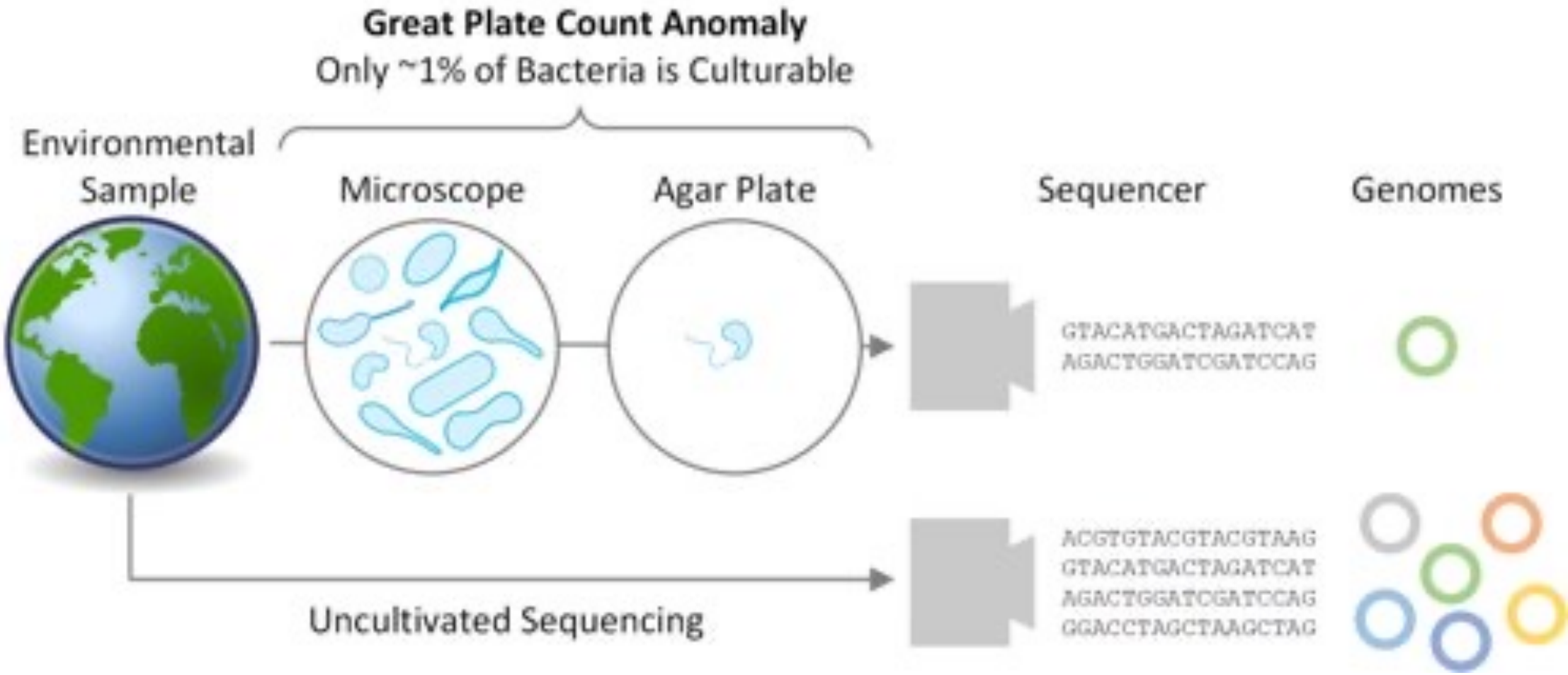
https://aindustryreports.com/2019/07/23/metagenomics-market/

This course

- Hands-on
- Materials available during and after the course
  - Github
- Mix of lectures, tutorials and practice
- Ask questions but if there is an error read the error message first
- Learn from each other as well as instructors
- Schedule available but flexible
- Lunch every day around noon, coffee 14.30

You learn programming by programming

Various authors at CSC


Great Plate Count Anomaly
Only ~1% of Bacteria is Culturable
Environmental Sample
Microscope
Agar Plate
Sequencer
Genomes
GTACATGACTAGATCAT
AGACTGGATCGATCCAG
ACGTGTACGTACGTAAG
GTACATGACTAGATCAT
AGACTGGATCGATCCAG
GGACCTAGCTAAGCTAG
Uncultivated Sequencing

Metagenomics

- Jo Handelsman 1998

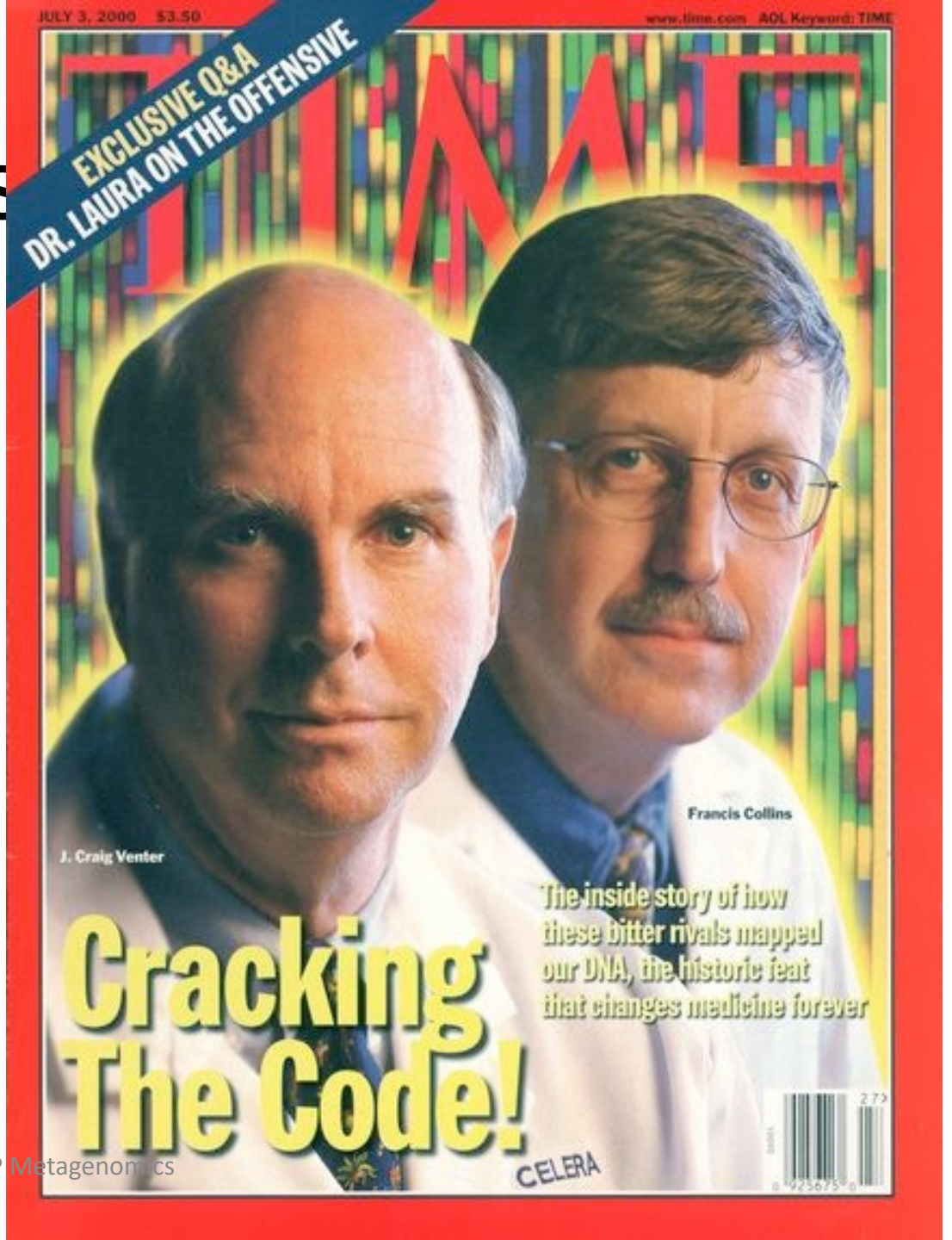
”the application of modern genomics techniques to the study of communities of microbial **organisms directly in their natural environments**, bypassing the need for isolation and lab cultivation of individual species”

H mics

- 19 ide en
- 19 rR
- 20 se dif
- 20 me

un

Sea



RESEARCH ARTICLE

Environmental Genome Shotgun Sequencing of the Sargasso Sea

J. Craig Venter,[1*] Karin Remington,[1] John F. Heidelberg,[3] Aaron L. Halpern,[2] Doug Rusch,[2] Jonathan A. Eisen,[3] Dongying Wu,[3] Ian Paulsen,[3] Karen E. Nelson,[3] William Nelson,[3] Derrick E. Fouts,[3] Samuel Levy,[2] Anthony H. Knap,[6] Michael W. Lomas,[6] Ken Nealson,[5] Owen White,[3] Jeremy Peterson,[3] Jeff Hoffman,[1] Rachel Parsons,[6] Holly Baden-Tillson,[1] Cynthia Pfannkoch,[1] Yu-Hui Rogers,[4] Hamilton O. Smith[1]

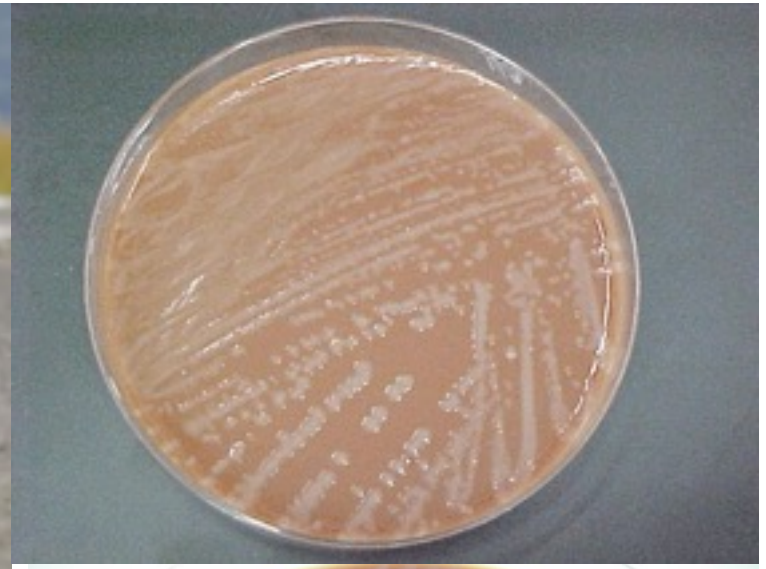
We have applied "whole-genome shotgun sequencing" to microbial populations collected en masse on tangential flow and impact filters from seawater samples collected from the Sargasso Sea near Bermuda. A total of 1.045 billion base pairs of nonredundant sequence was generated, annotated, and analyzed to elucidate the gene content, diversity, and relative abundance of the organisms within these environmental samples. These data are estimated to derive from at least 1800 genomic species based on sequence relatedness, including 148 previously unknown bacterial phylotypes. We have identified over 1.2 million previously unknown genes represented in these samples, including more than 782 new rhodopsin-like photoreceptors. Variation in species present and stoichiometry suggests substantial oceanic microbial diversity.

- 148 previously unknown bacterial phylotypes
- 1.2 million previously unknown genes
  - 782 new rhodopsin like genes

Metagenomes and microbes

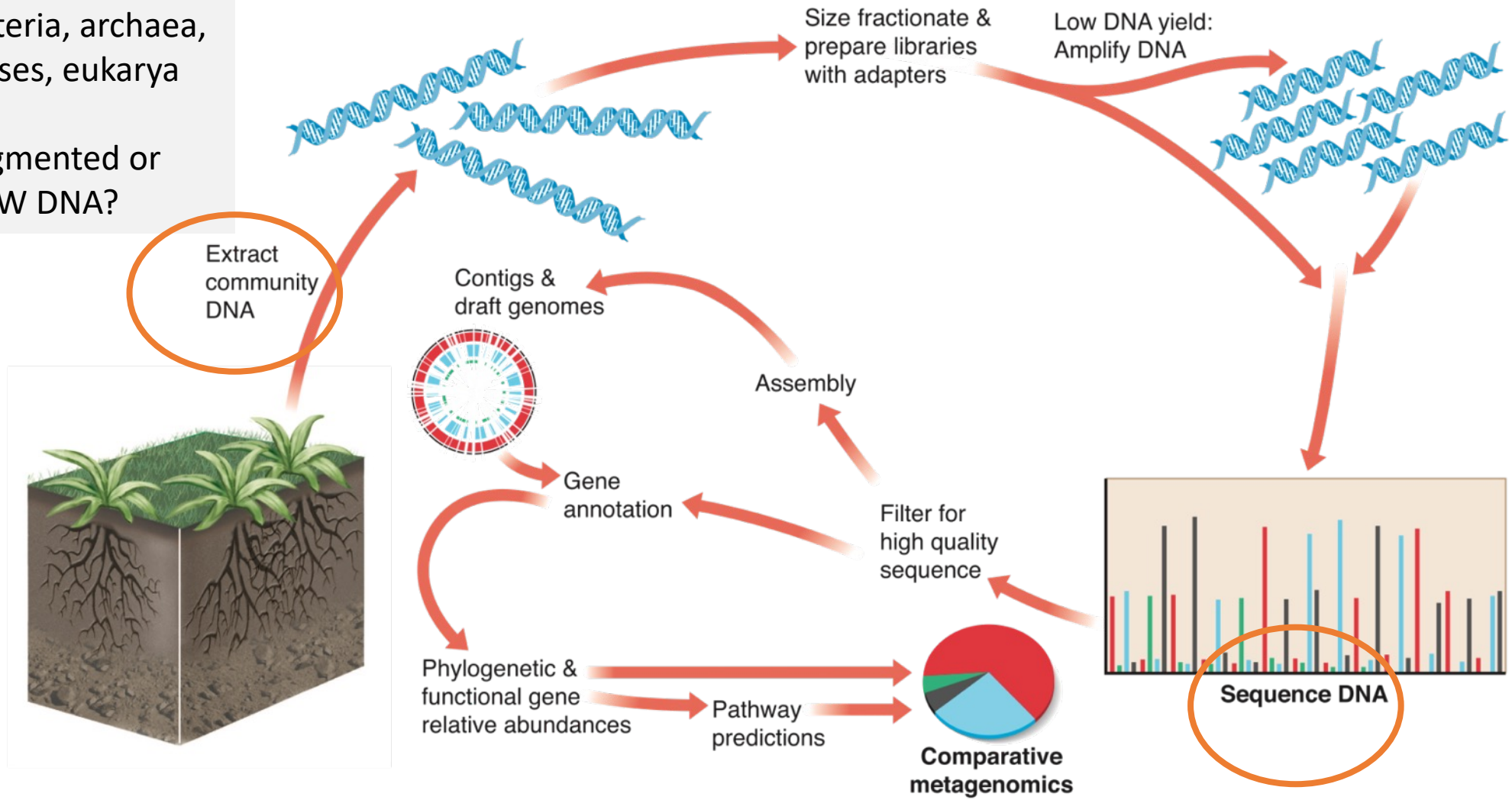
- Often metagenome is mostly microbial (Bacteria, Archaea, Fungi, some protist)
- Why metagenomes are high in microbes?
- Microbes are hard to study: small, diversity and numbers are high
- Phenotype

Metagenomes and microbes



DNA from all bacteria, archaea, viruses, eukarya

Fragmented or HMW DNA?



Soil metagenomics methodology.

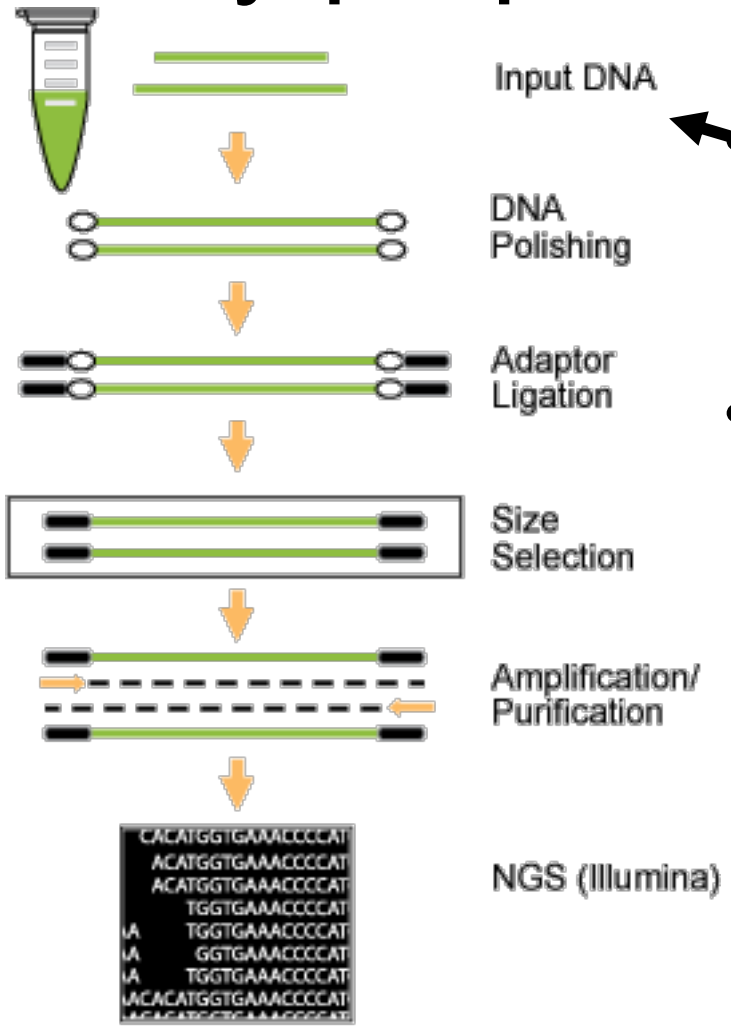
Illumina, PacBio, Oxford Nanopore

Paired end, single read, size of library

Sequencing depth, replicates

Jansson 2011: Towards tera-terra. Microbe June 2011

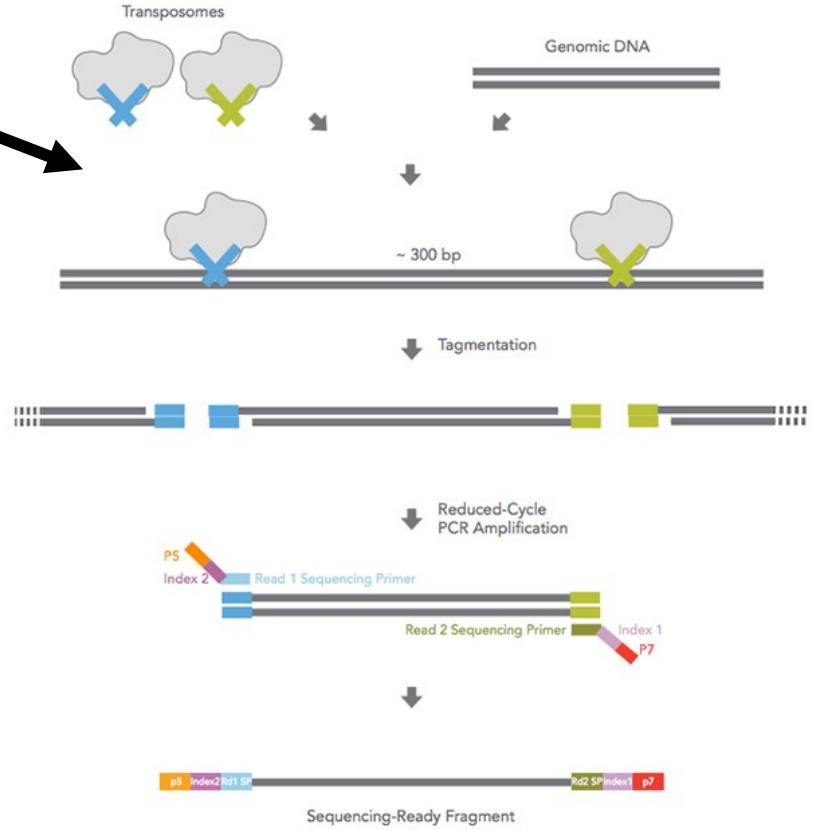
Library preparation



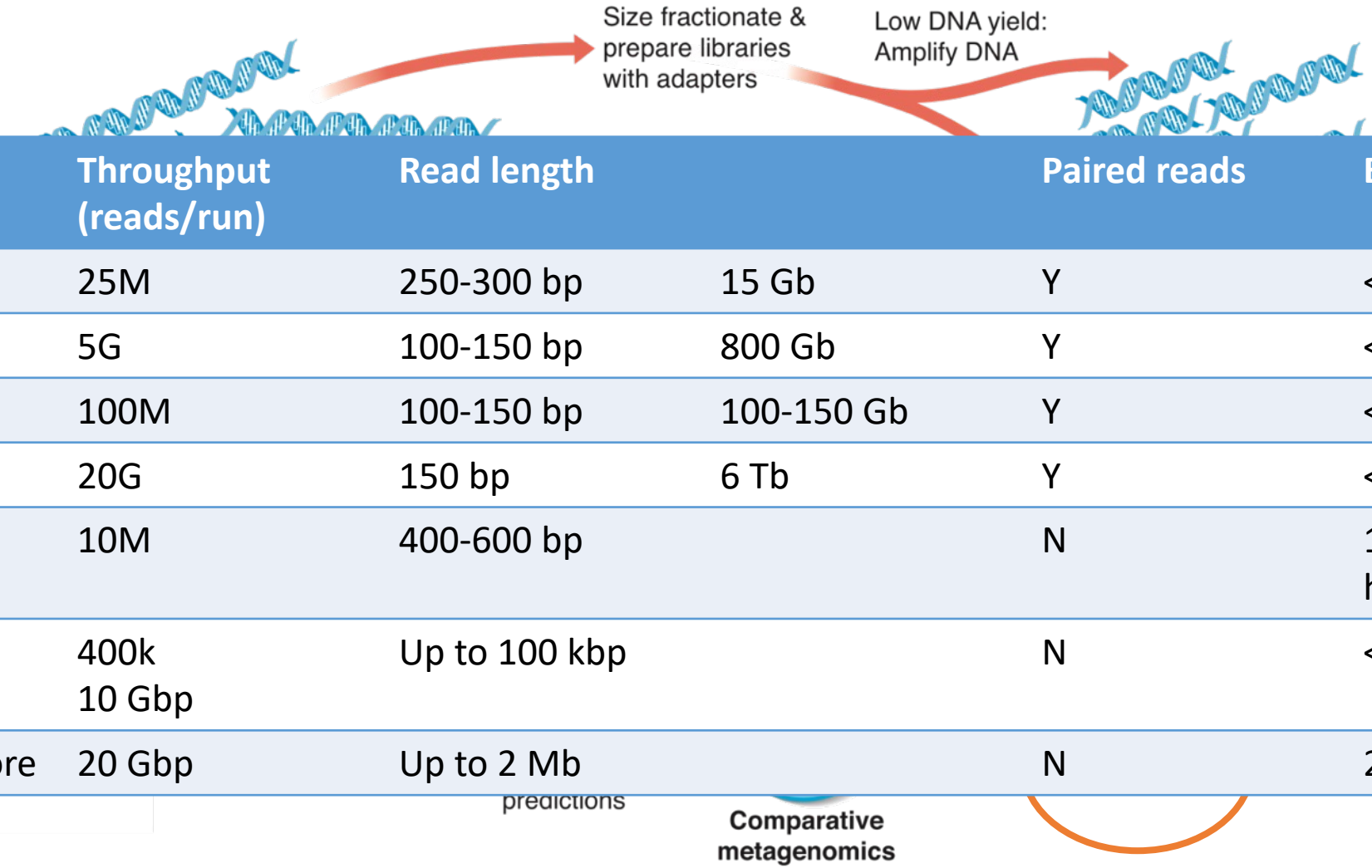
DNA shearing and adapter ligation

- DNA quantity

- **Nextera transposases**
  - Inhibitors (in soil)
  - 1-10 ng of DNA
  - Transposase:DNA ratio



DNA from all
bacteria, archaea
virus

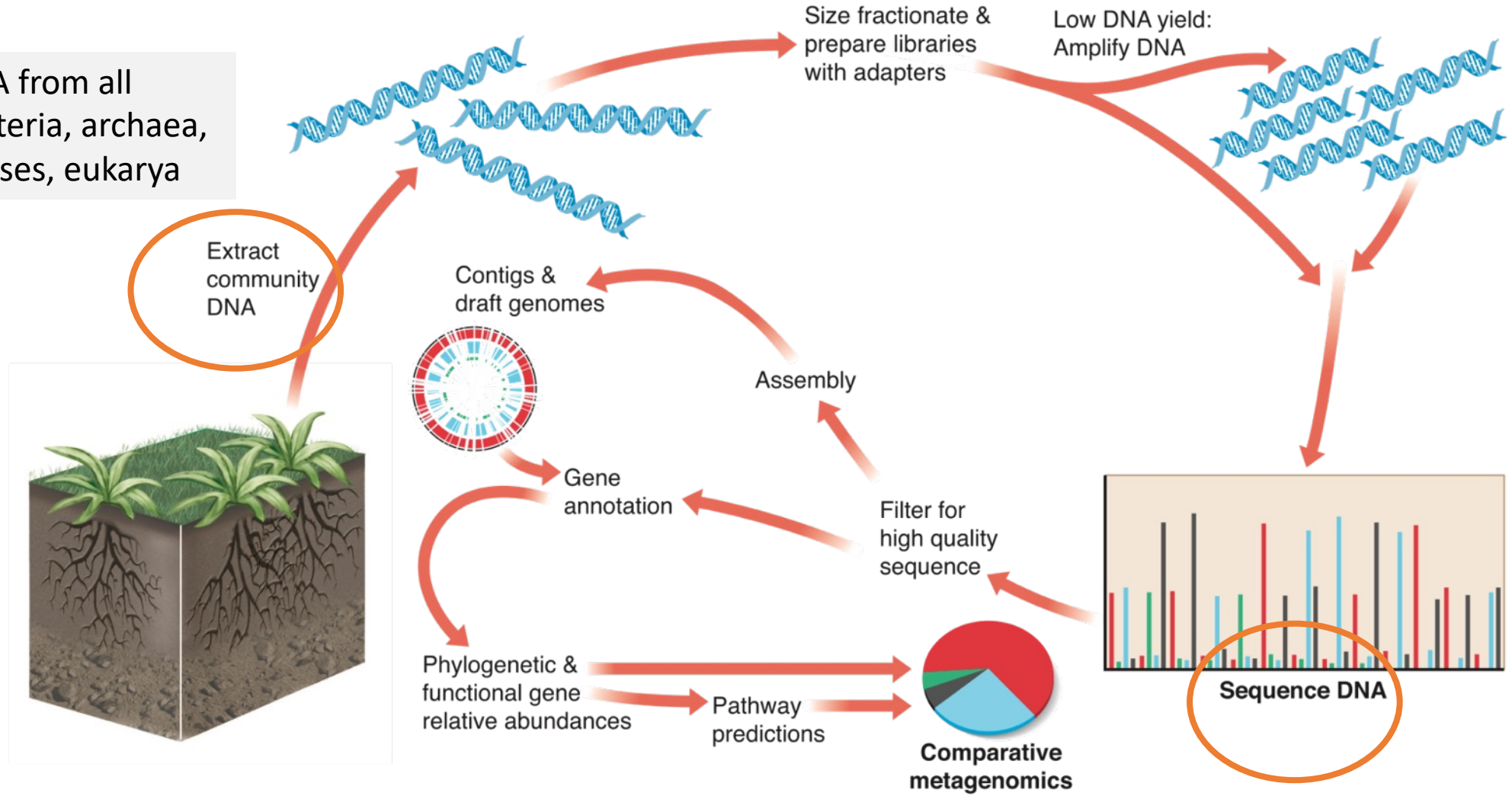


Soil metagenomics methodology.

Technology	Throughput (reads/run)	Read length		Paired reads	Errors?
MiSeq	25M	250-300 bp	15 Gb	Y	<1%
HiSeq	5G	100-150 bp	800 Gb	Y	<1%
NextSeq	100M	100-150 bp	100-150 Gb	Y	<1%
NovaSeq	20G	150 bp	6 Tb	Y	<1%
IonTorrent	10M	400-600 bp		N	1-2% (indels, homopolymers)
Pacific Biosciences	400k 10 Gbp	Up to 100 kbp		N	<1%
Oxford Nanopore	20 Gbp	Up to 2 Mb		N	2-13%

Jansson 2011: Towards tera-terra. Microbe June 2011

DNA from all bacteria, archaea, viruses, eukarya



Illumina, PacBio, Oxford Nanopore, Ion Torrent

Paired end, single read, size of library

Sequencing depth, replicates

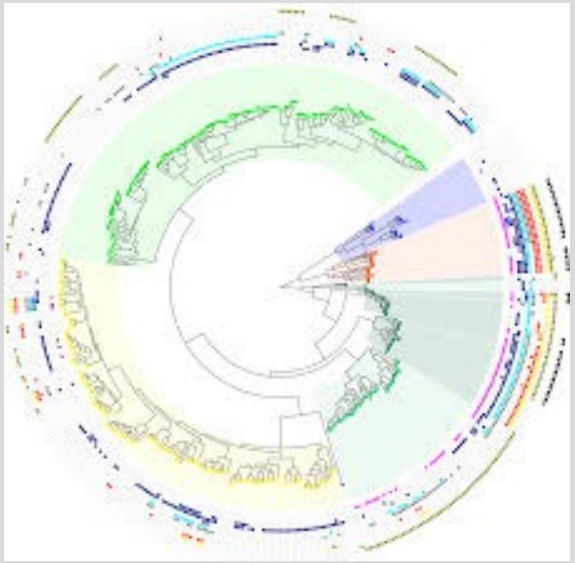
Soil metagenomics methodology.

Jansson 2011: Towards tera-terra. Microbe June 2011

With metagenomics

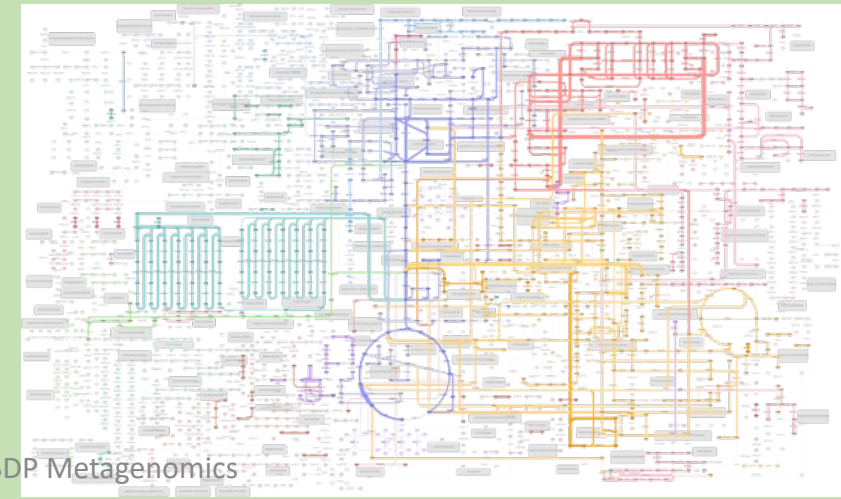
Microbial community composition – who is there

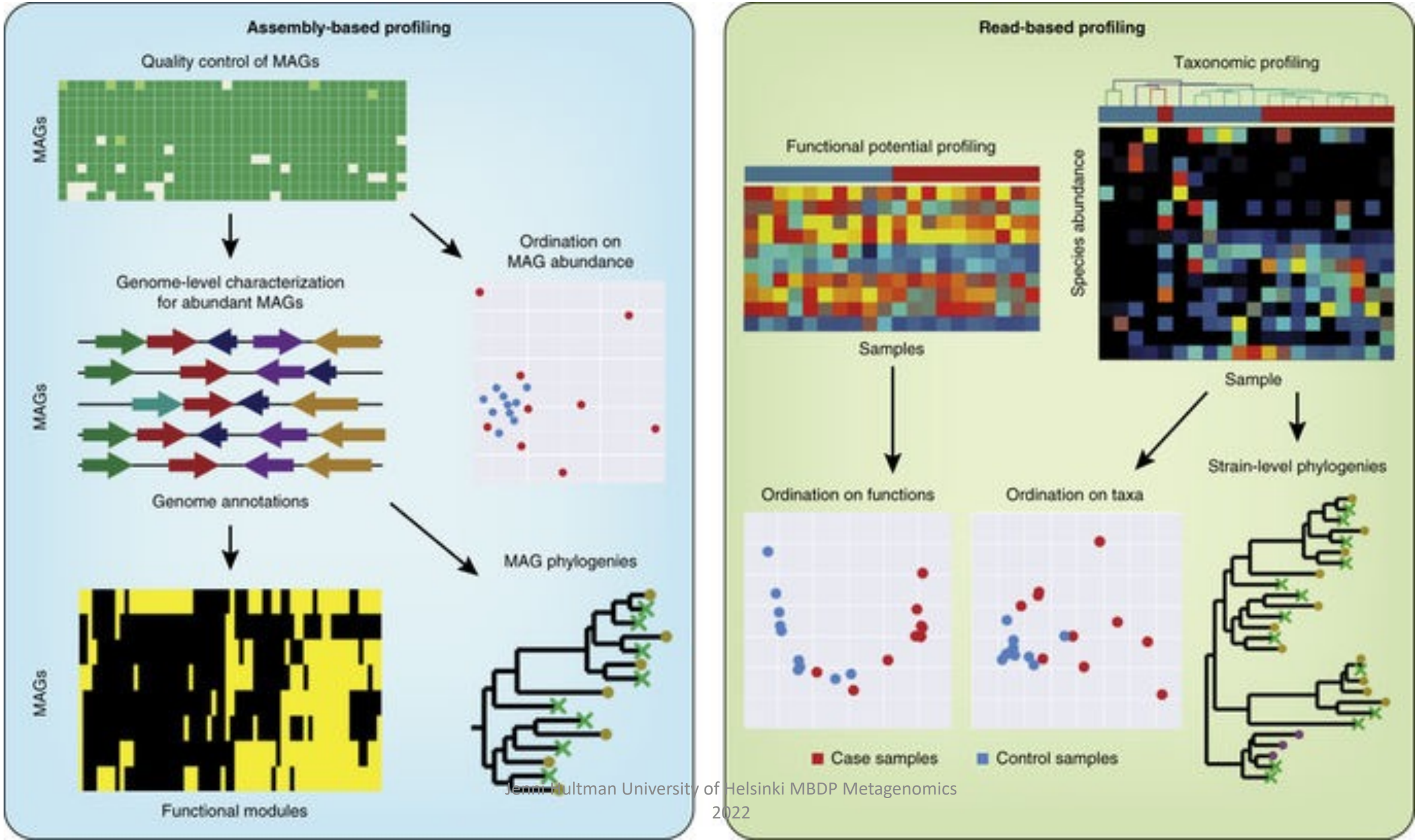
16S/18S, ITS



Microbial community function – what can they do

Protein coding sequences




Assembly-based profiling
Quality control of MAGs
MAGs
Genome-level characterization for abundant MAGs
MAGs
Genome annotations
MAGs
Functional modules
Ordination on MAG abundance
MAG phylogenies
Read-based profiling
Taxonomic profiling
Functional potential profiling
Species abundance
Samples
Sample
Strain-level phylogenies
Ordination on functions
Ordination on taxa
Case samples
Control samples

Dataset for this course

- 23 samples, we have 3 of them
- 2 sequenced with Illumina only, 1 with Nanopore and Illumina
- “uncover abundant undescribed lineages belonging to important functional groups”

nature COMMUNICATIONS

ARTICLE

Check for updates

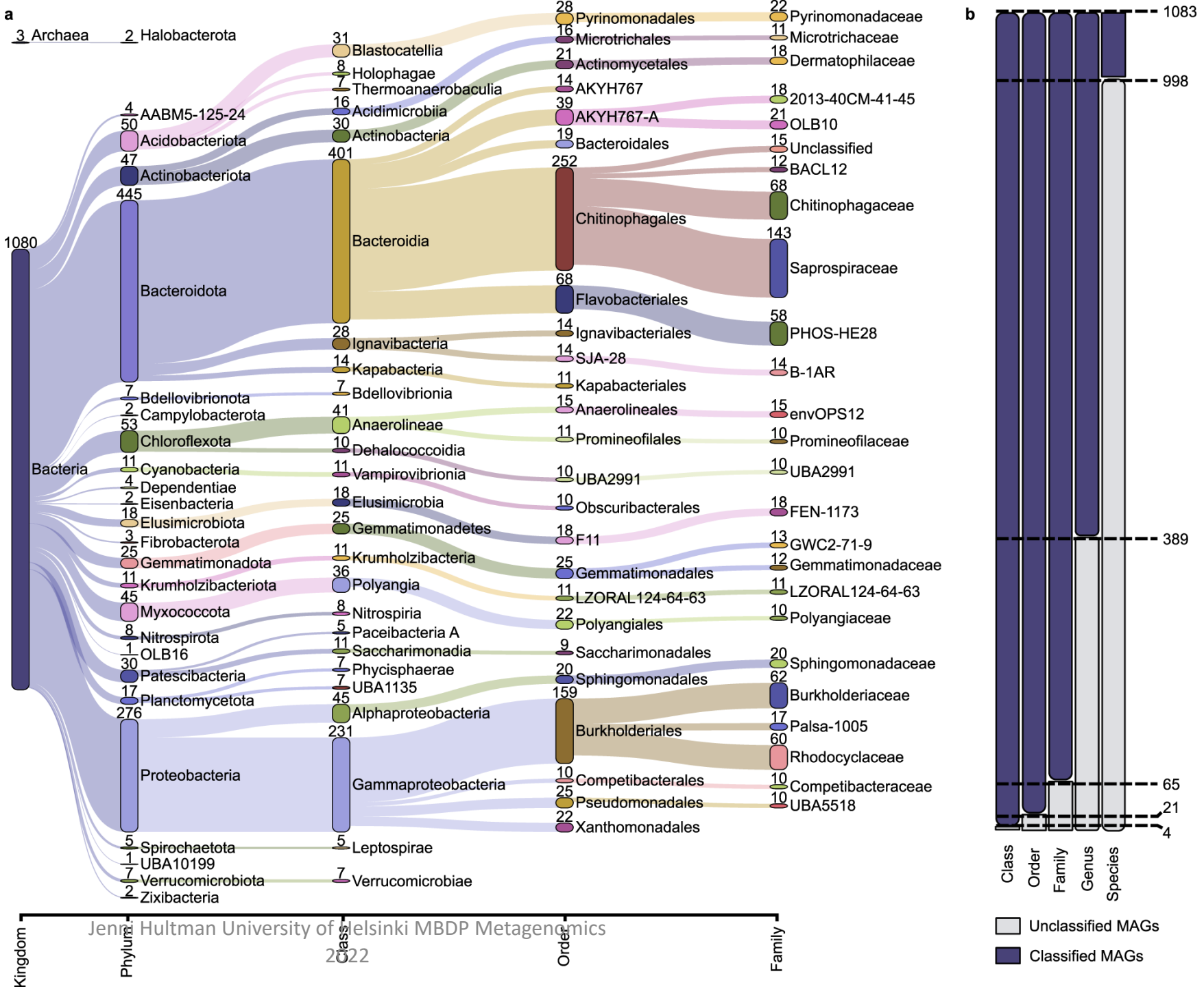
https://doi.org/10.1038/s41467-021-22203-2 OPEN

Connecting structure to function with the recovery of over 1000 high-quality metagenome-assembled genomes from activated sludge using long-read sequencing

Caitlin M. Singleton[1], Francesca Petriglieri[1], Jannie M. Kristensen[1], Rasmus H. Kirkegaard[1], Thomas Y. Michaelsen[1], Martin H. Andersen[1], Zivile Kondrotaite[1], Søren M. Karst[1], Morten S. Dueholm[1], Per H. Nielsen[1✉] & Mads Albertsen[1✉]

- Methodology for the high-throughput production of HQ MAGs and its application to investigate complex microbial communities, focusing on the AS system

- HQ MAGs with full-length 16S rRNA genes



QC and trimming

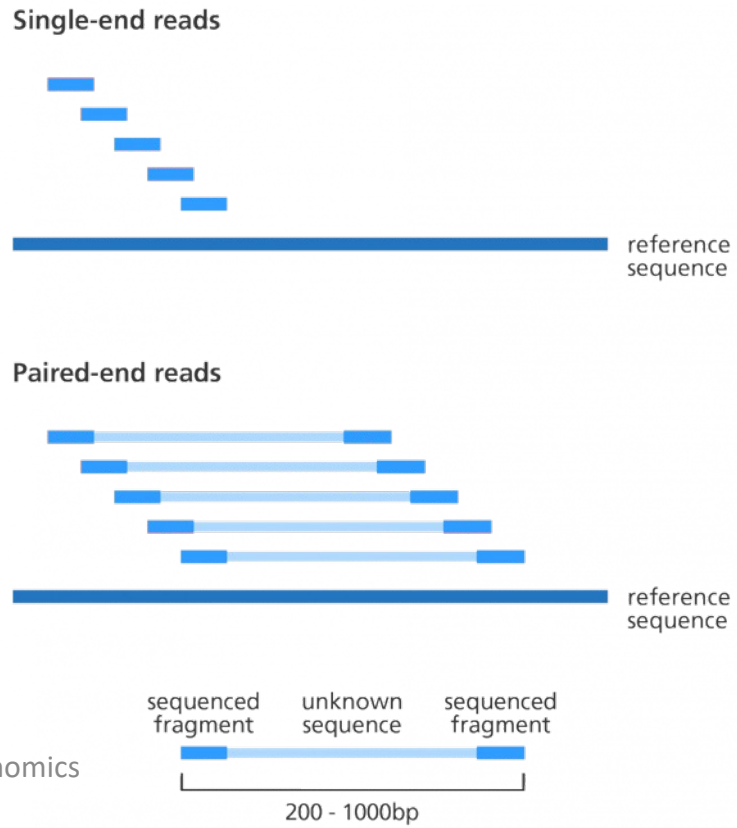
Raw data Illumina

- We have paired end data on the sequenced metagenomes

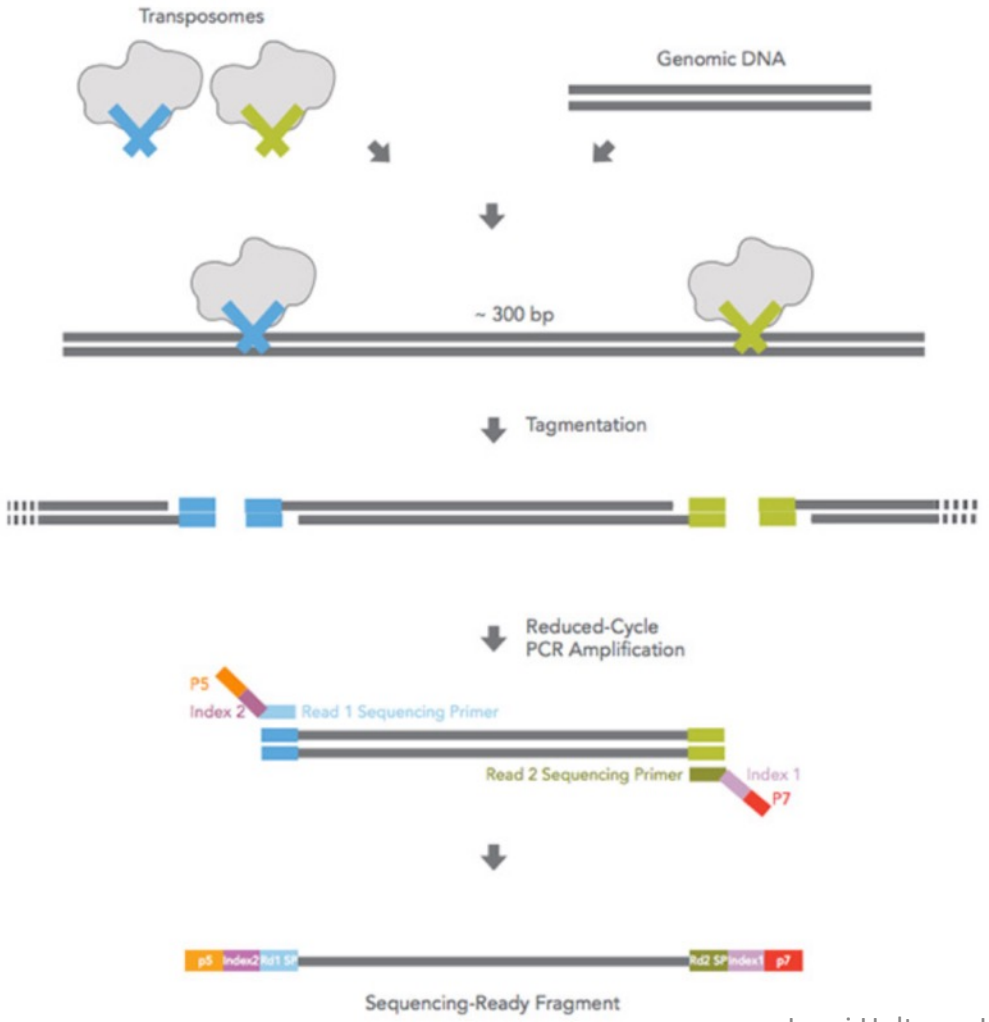
  A004_07004-B_TTGCATGT_GACTCGCA_run20171107N_S4_R1_001.fastq

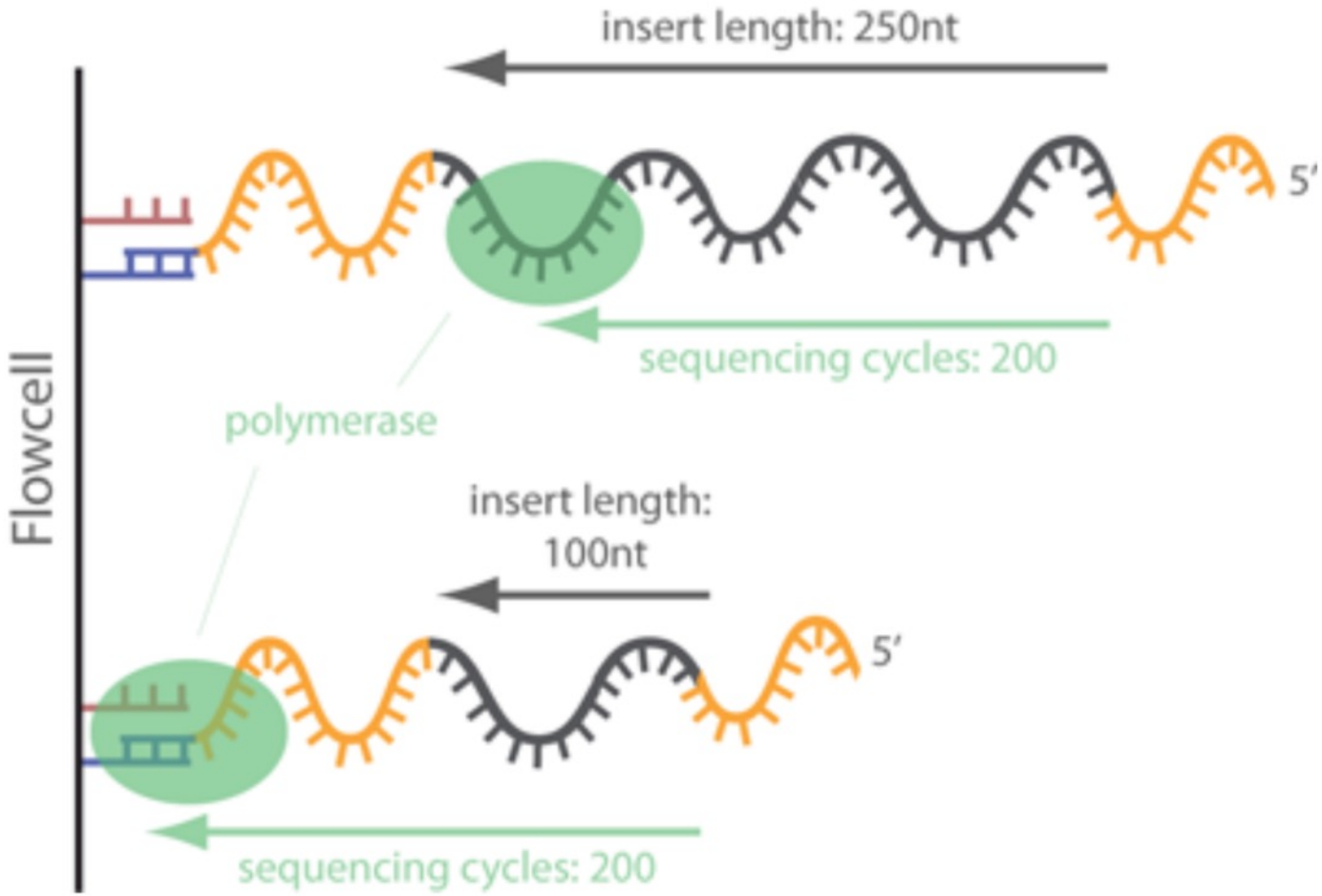
  A004_07004-B_TTGCATGT_GACTCGCA_run20171107N_S4_R2_001.fastq

- 150 + 150 bp
- Can contain
  - Sequence adapters
  - low quality sequence (usually in the end)
    - Occurrence: substitutions > indels
    - Quality scores: substitutions < indels
    - Overall quality: R1 > R2; beginning > end
- Need to check quality and trim the reads



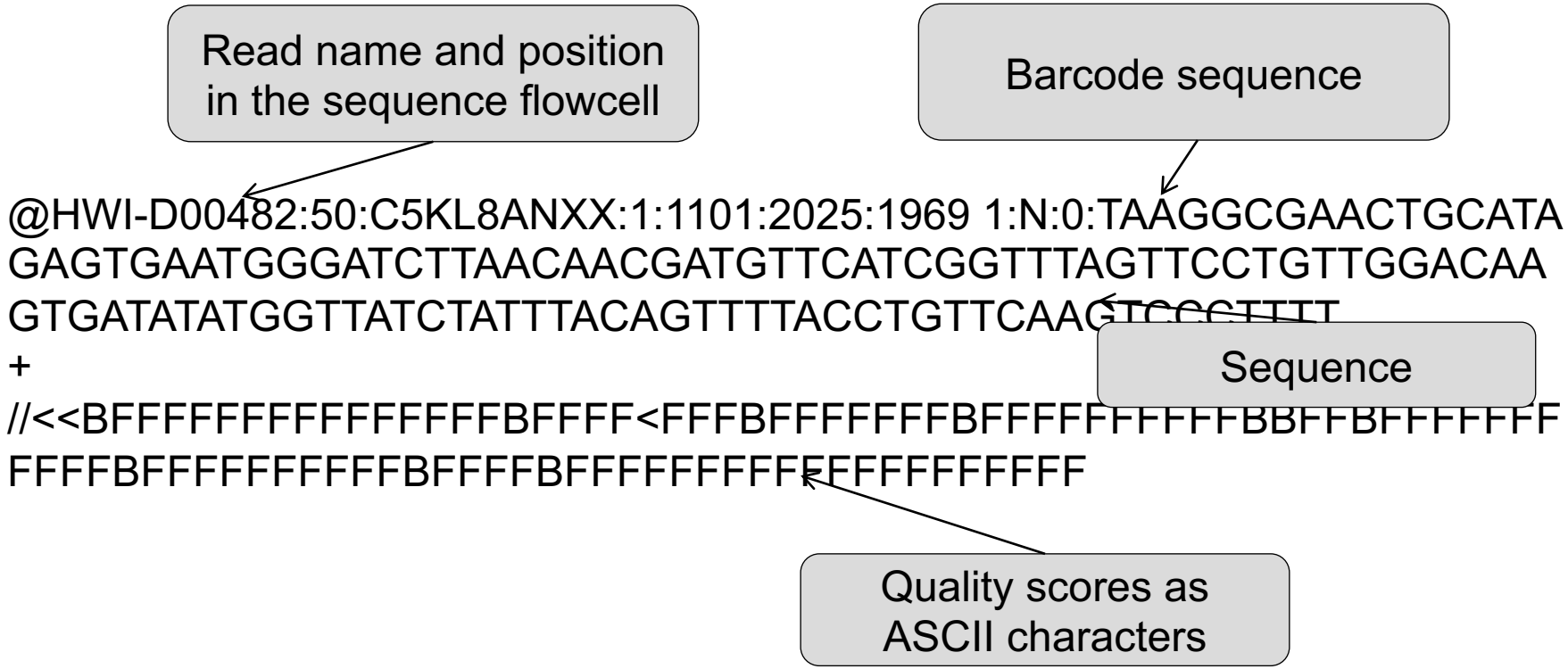
Adapter contamination





Fastq

- Sequence data is commonly delivered in FASTQ format. No chromatograms!



Quality scores

- measure of the quality of the identification of the bases generated by sequencer
- Phred-score

Phred Quality Score	Probability of incorrect base call	Base call accuracy	ASCHII
10	1 in 10	90%	+
20	1 in 100	99%	5
30	1 in 1000	99.9%	?
40	1 in 10000	99.99%	I

- Phred score above 20-25 considered as acceptable
  - 1 mistake in 100

@HWI-D00482:50:C5KL8ANXX:1:1101:2025:1969 1:N:0:TAAGGCGAACTGCATA
GAGTGAATGGGATCTTAACAACGATGTTCATCGGTTTAGTTCCTGTTGGACAAGTGATATATGGTTATCT
ATTTACAGTTTTACCTGTTCAAGTCCCTTTT
+
//<<BFFFFFFFFFFFFFFFBFFFF<FFFBFFFFFFFBFFFFFFFFFFBBFFBFFFFFFFFFFFBFFFFFFFFFFBFFFF
BFFFFFFFFFFFFFFFFFFF

ASCII_BASE=33 Illumina, Ion Torrent, PacBio and Sanger

Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII	Q	P_error	ASCII
0	1.00000	33 !	11	0.07943	44 ,	22	0.00631	55 7	33	0.00050	66 B
1	0.79433	34 "	12	0.06310	45 -	23	0.00501	56 8	34	0.00040	67 C
2	0.63096	35 #	13	0.05012	46 .	24	0.00398	57 9	35	0.00032	68 D
3	0.50119	36 $	14	0.03981	47 /	25	0.00316	58 :	36	0.00025	69 E
4	0.39811	37 %	15	0.03162	48 0	26	0.00251	59 ;	37	0.00020	70 F
5	0.31623	38 &	16	0.02512	49 1	27	0.00200	60 <	38	0.00016	71 G
6	0.25119	39 '	17	0.01995	50 2	28	0.00158	61 =	39	0.00013	72 H
7	0.19953	40 (	18	0.01585	51 3	29	0.00126	62 >	40	0.00010	73 I
8	0.15849	41)	19	0.01259	52 4	30	0.00100	63 ?	41	0.00008	74 J
9	0.12589	42 *	20	0.01000	53 5	31	0.00079	64 @	42	0.00006	75 K
10	0.10000	43 +	21	0.00794	54 6	32	0.00063	65 A			

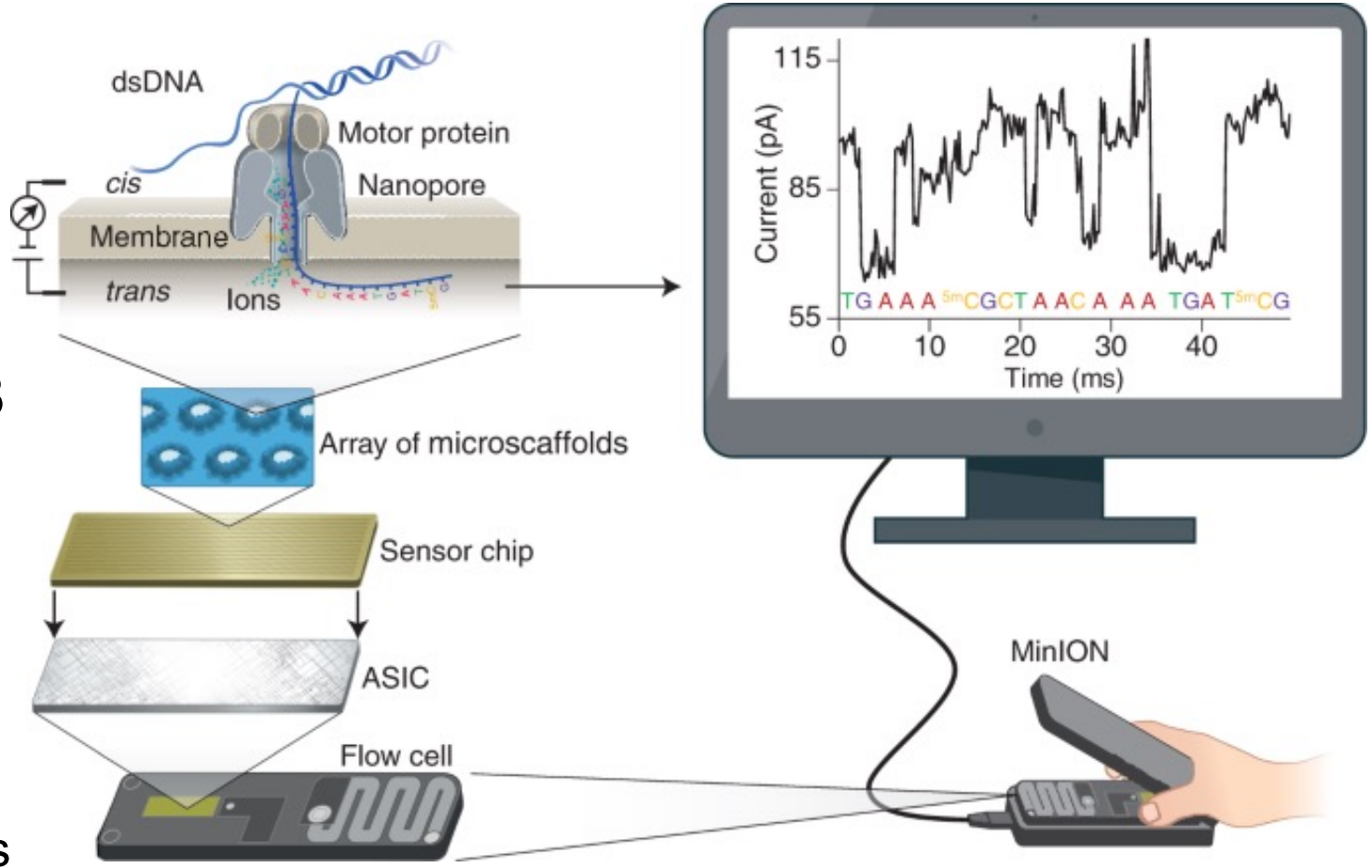
Quality filtering

- Removal of low-quality regions and adapters

- Several programs available, we will use **cutadapt** http://cutadapt.readthedocs.io/en/stable/

Nanopore data

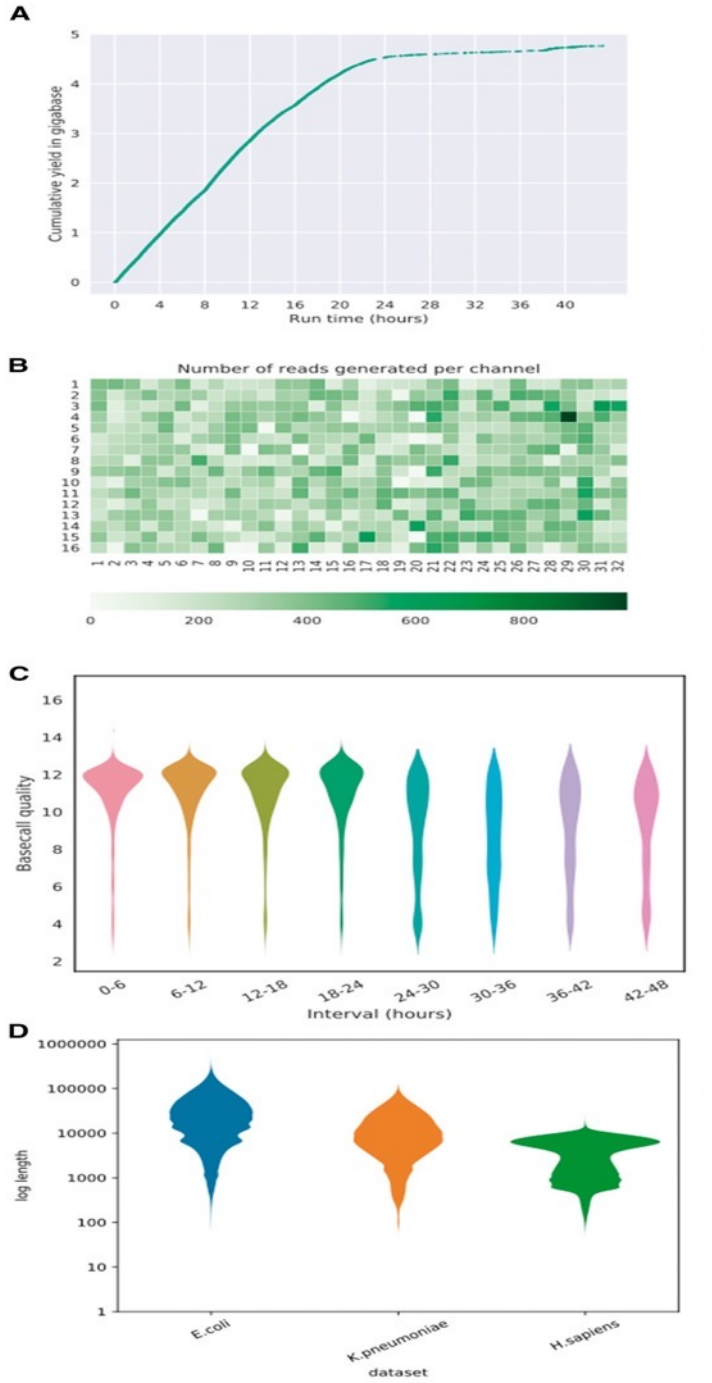
- Quality issues of Nanopore: substitutions
- A MinION flow cell contains 512 channels with 4 nanopores in each channel, for a total of 2,048 nanopores used to sequence DNA or RNA.
  - As nucleotides pass through the nanopore, a characteristic current change is measured and is used to determine the corresponding nucleotide type at ~450 bases per s



https://www.nature.com/articles/s41587-021-01108-x

QC & filtering: NanoPlot, nanoQC, Nanofilt

- Nanoplot: (**A**) Cumulative yield plot (**B**) Flow cell activity heatmap showing number of reads per channel. (**C**) Violin plots comparing base call quality over time. (**D**) NanoComp plot comparing log transformed read lengths of the *E.coli* dataset with a *K.pneumoniae* and human dataset.
- NanoQC
- Nanofilt: Filtering and trimming of long read sequencing data.



https://doi.org/10.1093/bioinformatics/bty149

Garbage in – garbage out

Metagenome assembly

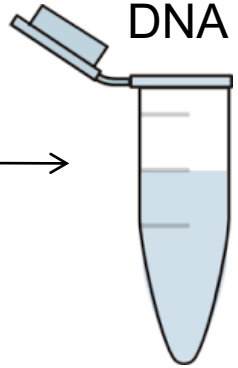
Reconstruct the original genome
from the sequence reads

Jenni Hultman University of Helsinki MBDP
Metagenomics 2022

De novo genome assembly

- Putting the sequence reads together
- In an ideal world:

bacteria



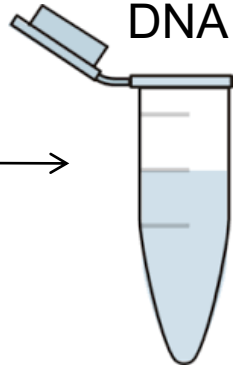
Sequencer

Genome sequence of the chromosome

De novo metagenome assembly

- Putting the sequence reads together
- In an ideal world:

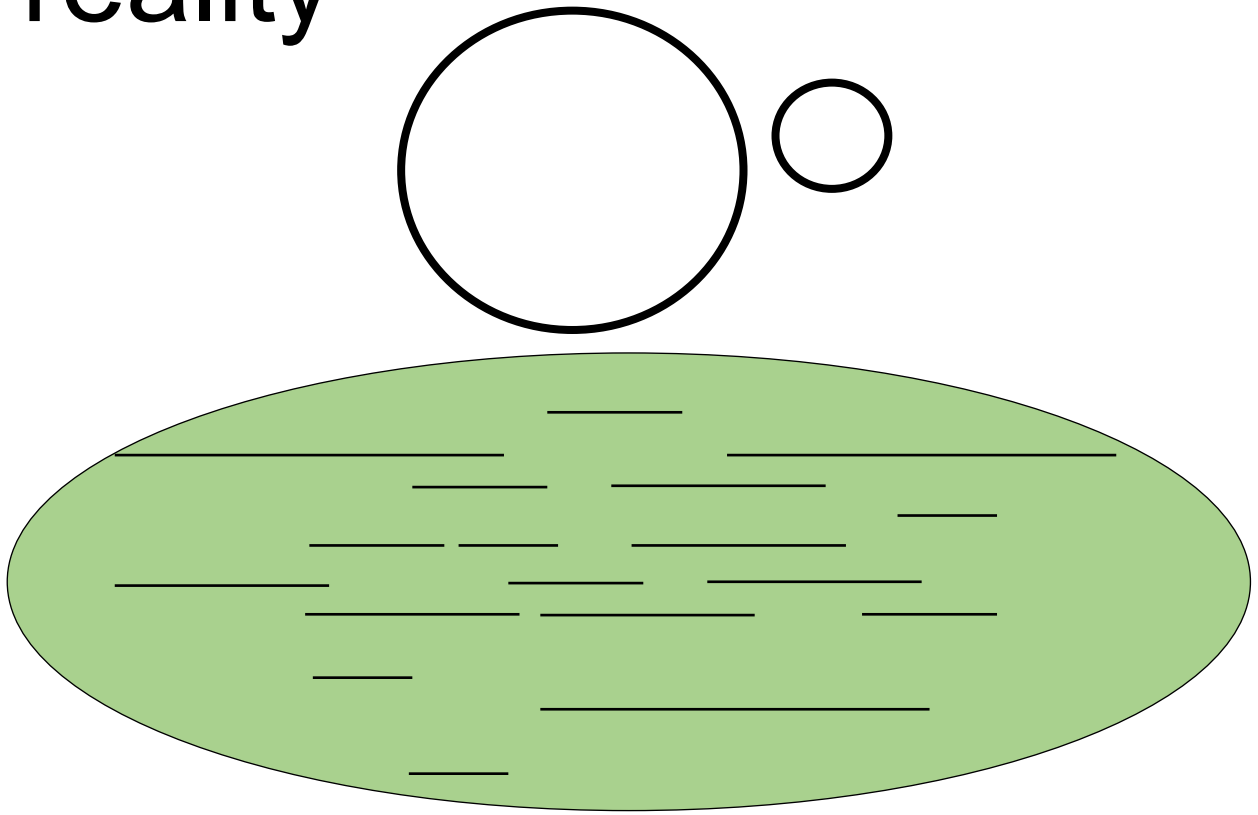
community



Sequencer

Genome sequence of the community

In reality

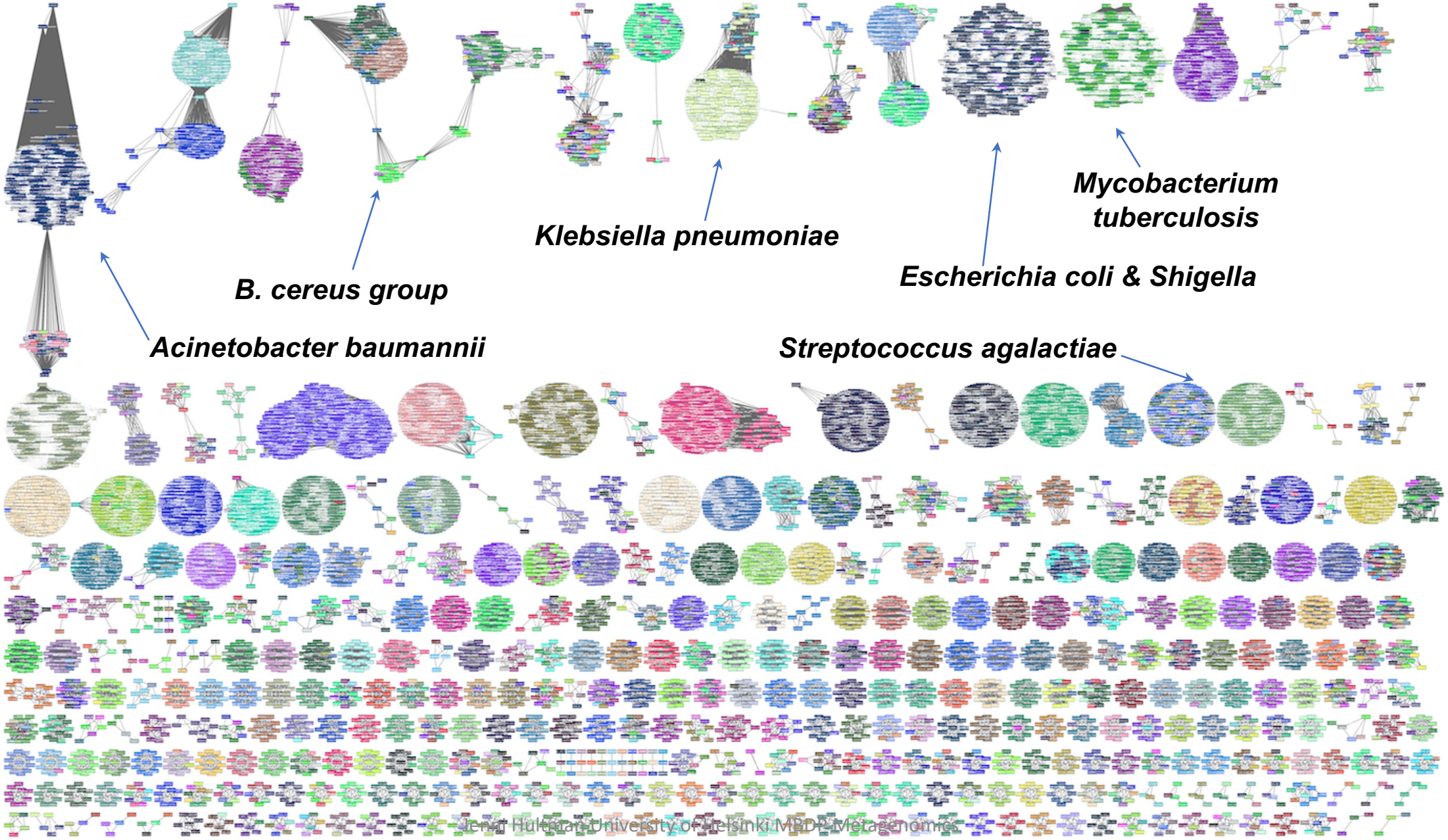


Genome

Fragmentation to short pieces

File

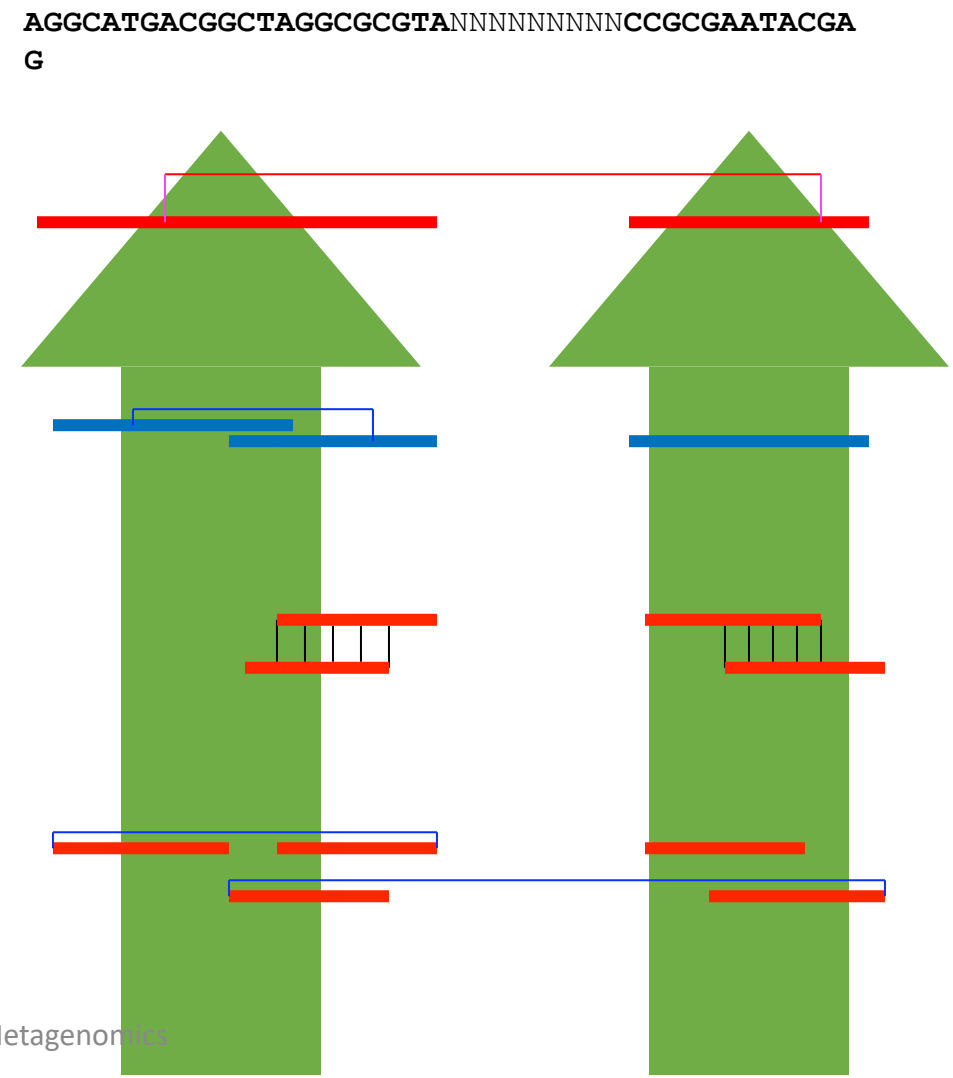
Fastq-file with sequences of the fragments



Slide from Todd Treangen

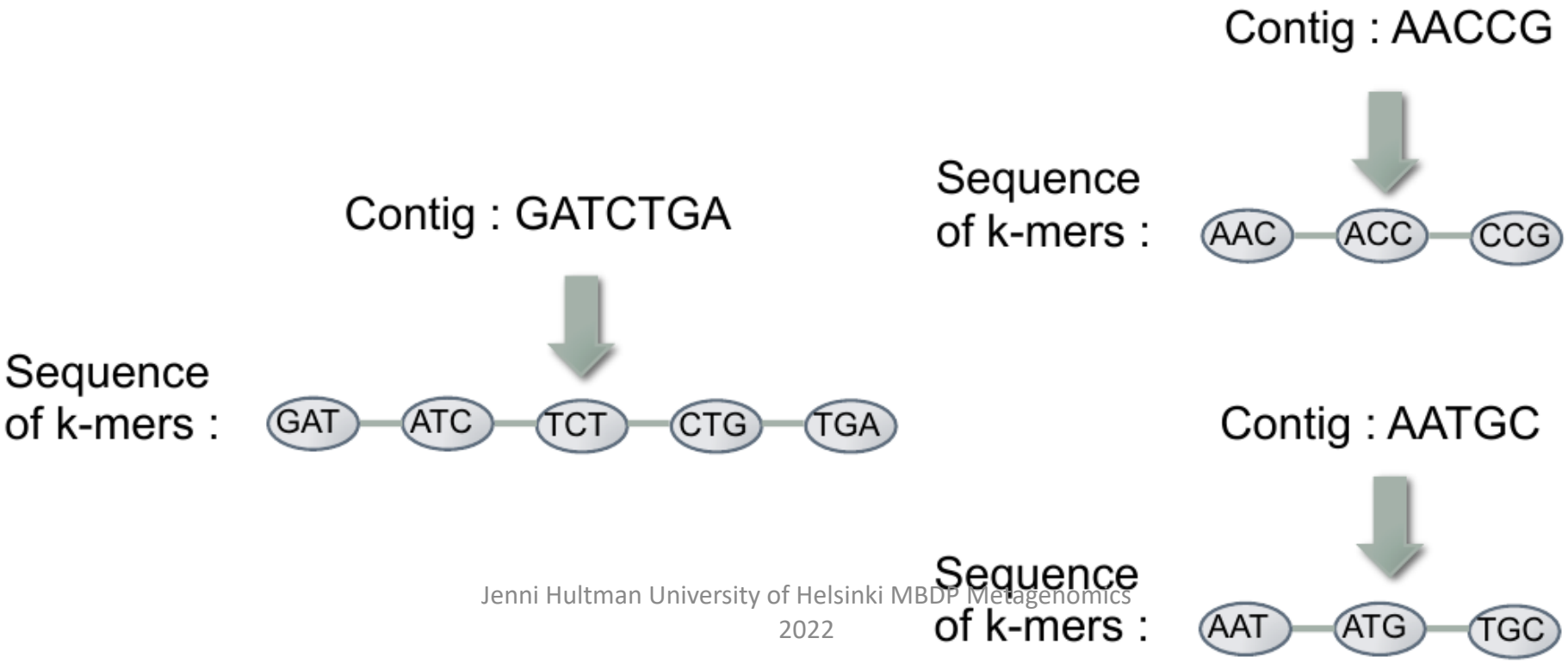
Glossary

- **Consensus**
  - Multiple alignment of sequences
- **Scaffold**=contigs + gaps
  - group of contigs that can be ordered and oriented with respect to each other (usually with the help of mate--pair data)
- **Contigs**
  - contiguous segment of DNA reconstructed (unambiguously) from a set of reads
- **Overlaps**
  - Shared sequences between the suffix of one read and the prefix of another
- **Reads**
  - segment of DNA "read" by a sequencing instrument
- **Mate-pairs, paired ends**
  - pair of reads whose distance from each other within the genome is approximately known

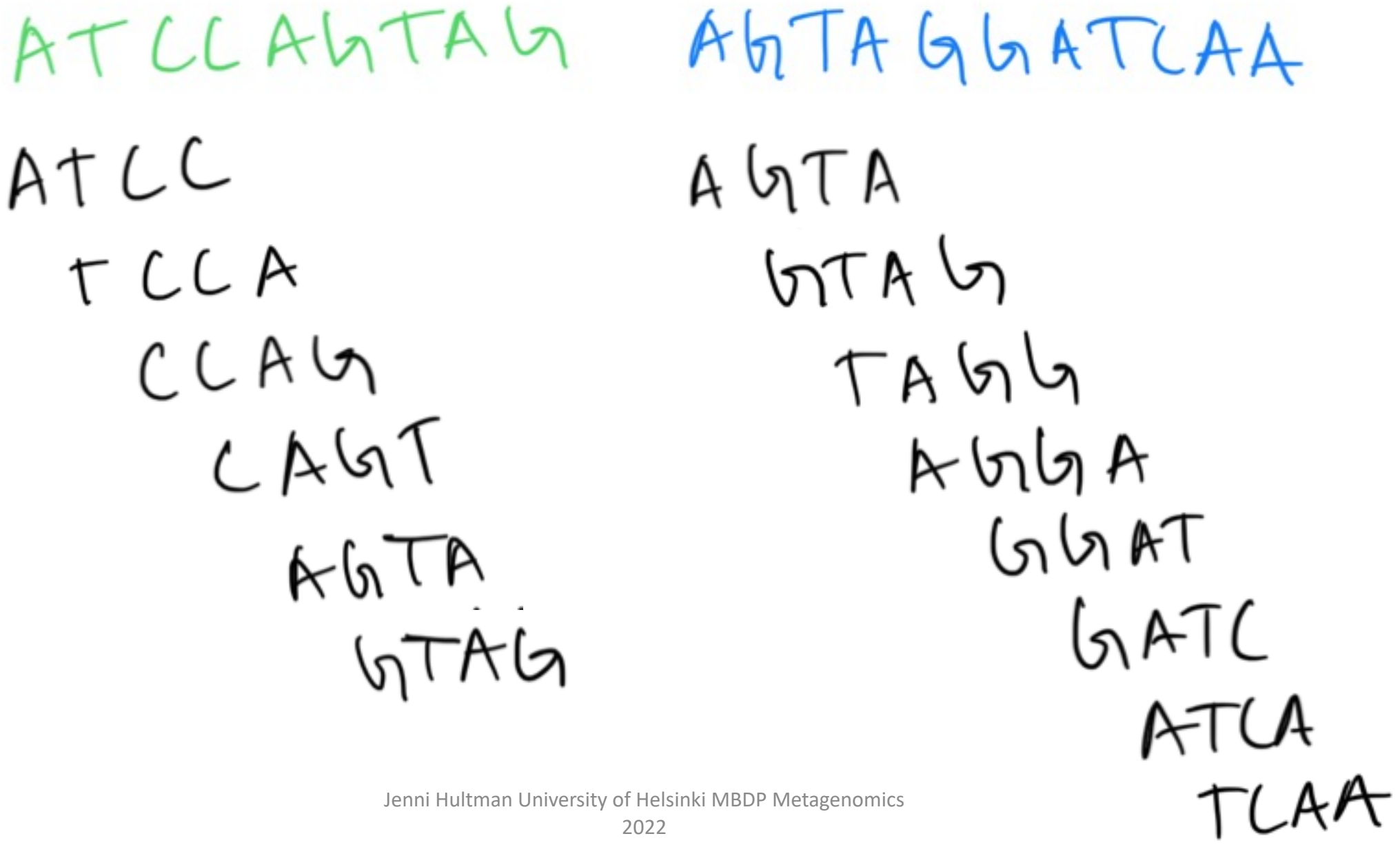


de Bruijin graph for short reads

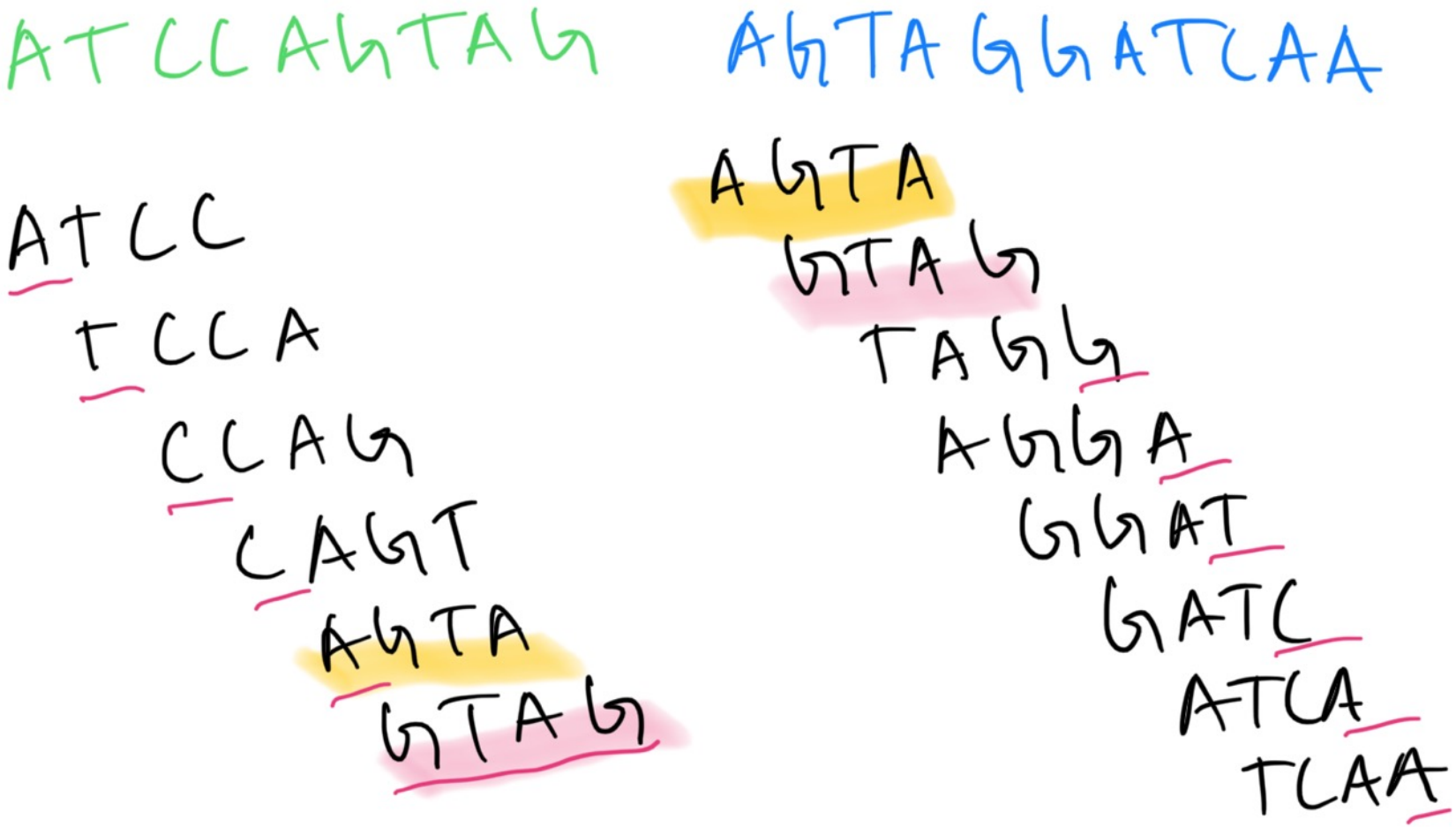
- 1,000,000 sequences cannot be compared with with each read (10^6*10^6 comparisons)
- Use of k-mers
  - Fragments of k-length



How k-mers work in assembly kmer=4




ATCCAGTAG
AGTAGGATCAA
ATCC
TCCA
CCAG
CAGT
AGTA
GTAG
AGTA
GTAG
TAGG
AGGA
GGAT
GATC
ATCA
TCAA

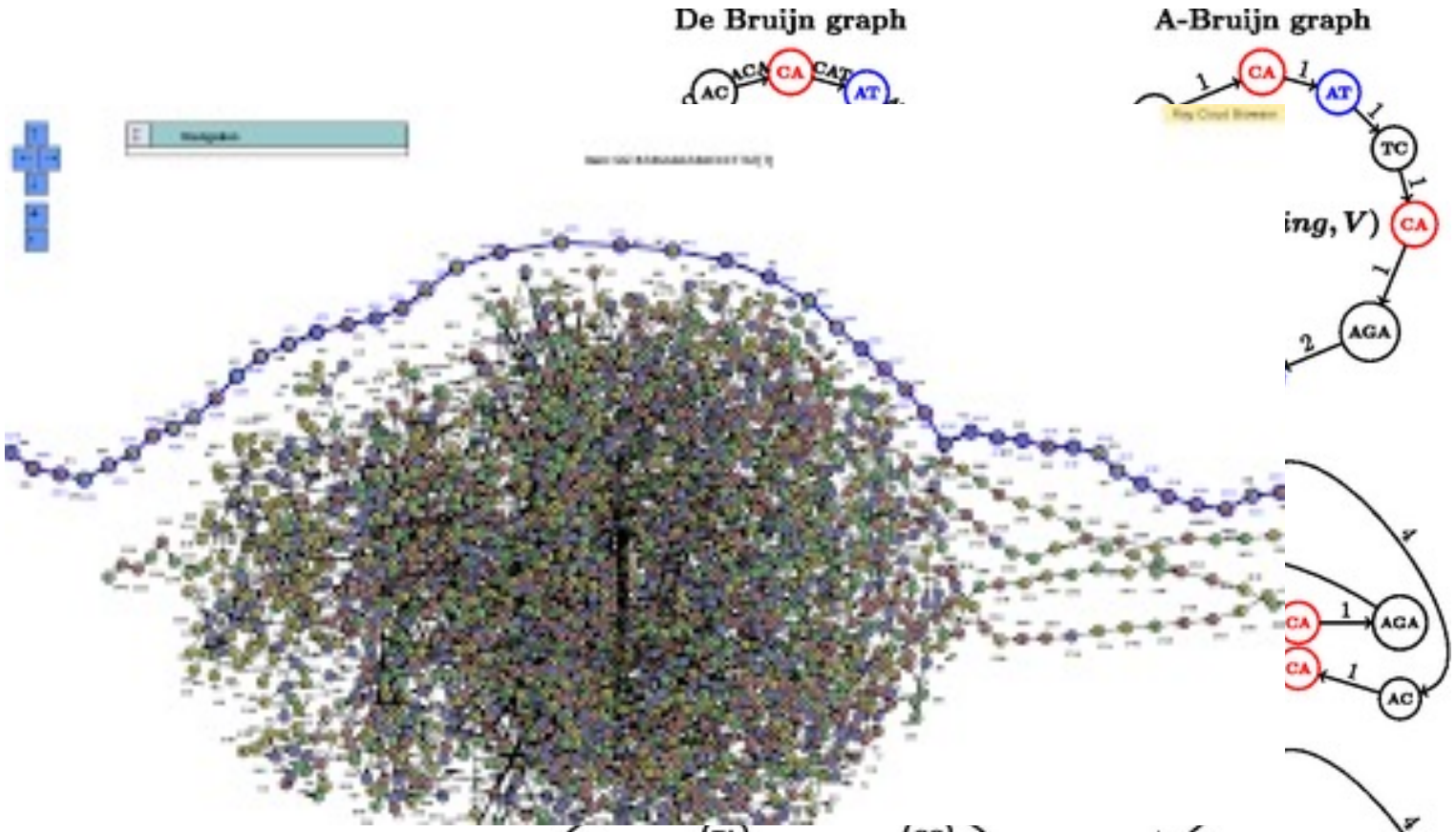

ATCCAGTAG
AGTAGGATCAA
ATCC
TCCA
CCAG
CAGT
AGTA
GTAG
AGTA
GTAG
TAGG
AGGA
GGAT
GATC
ATCA
TCAA
ATCCAGTAGGATCAA

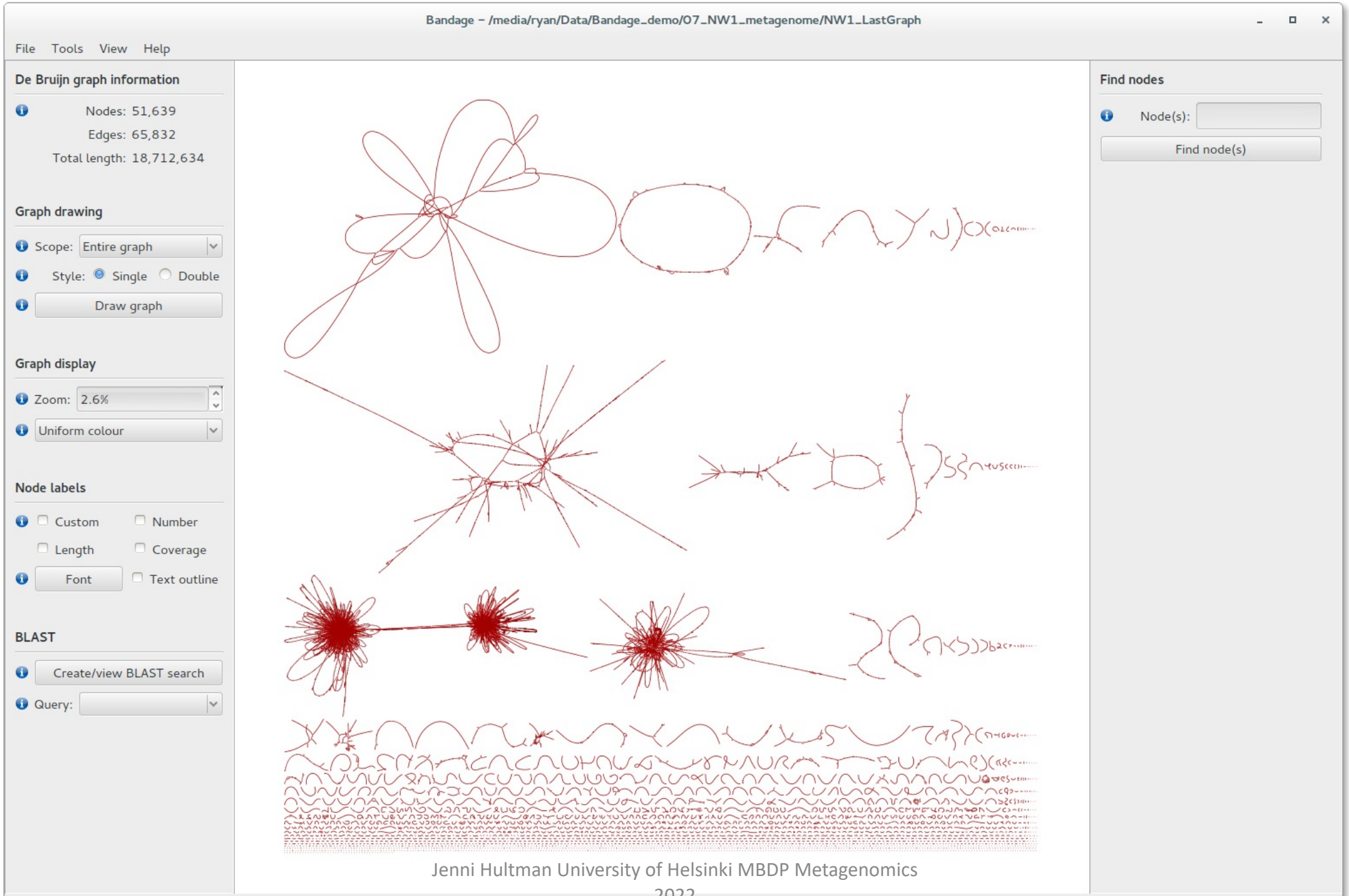
Size of k-mer has huge effect!

Too small -> misassembly (anything can assemble)

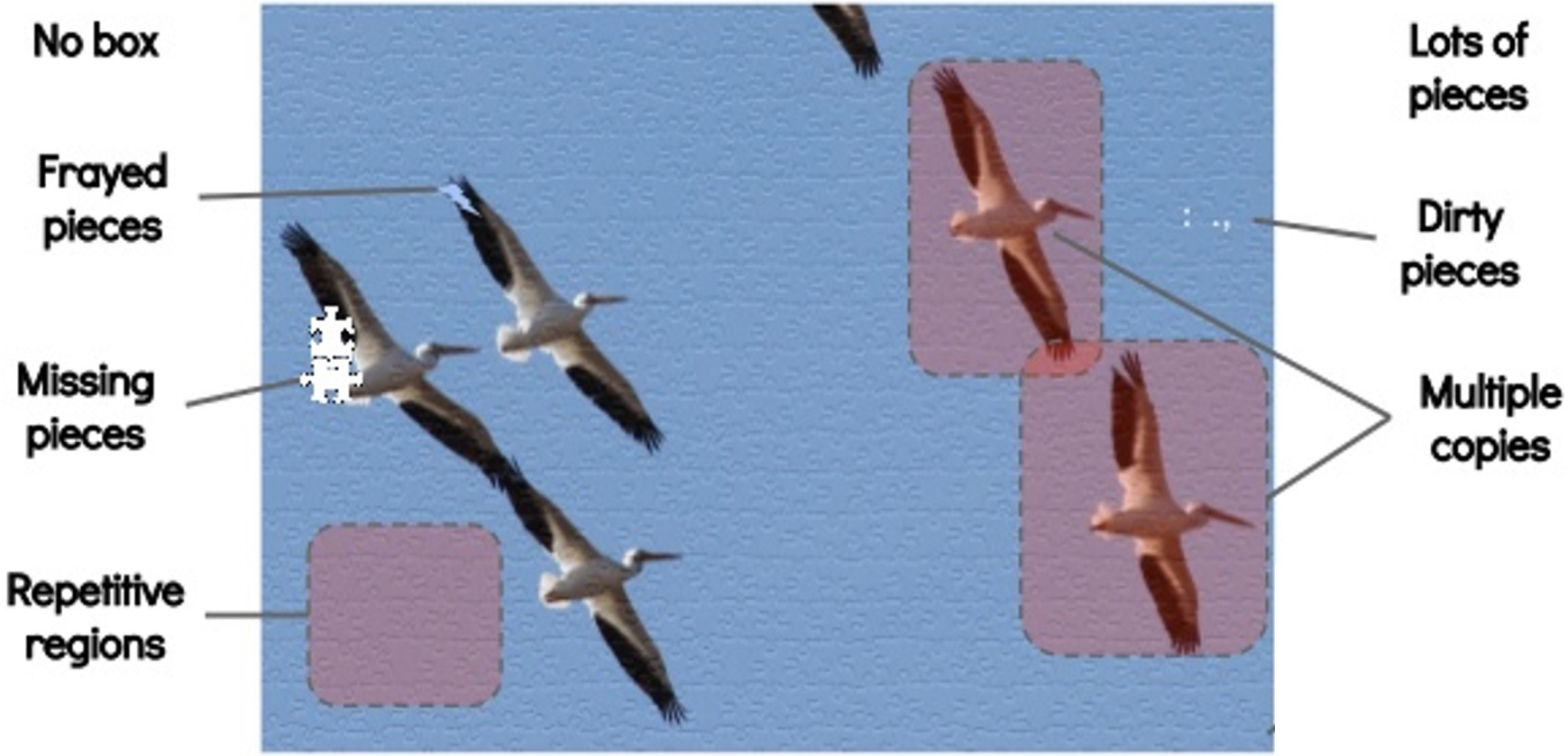
Too long -> no assembly/misassembly

In real life with 10^6 sequences




Bandage – /media/ryan/Data/Bandage_demo/07_NW1_metagenome/NW1_LastGraph
File Tools View Help
De Bruijn graph information
Nodes: 51,639
Edges: 65,832
Total length: 18,712,634
Graph drawing
Scope: Entire graph
Style: Single Double
Draw graph
Graph display
Zoom: 2.6%
Uniform colour
Node labels
Custom Number
Length Coverage
Font Text outline
BLAST
Create/view BLAST search
Query:
Find nodes
Node(s):
Find node(s)

What makes a puzzle hard?



Slide taken from Torsten Seemann

What makes **metagenome** assembly tricky?

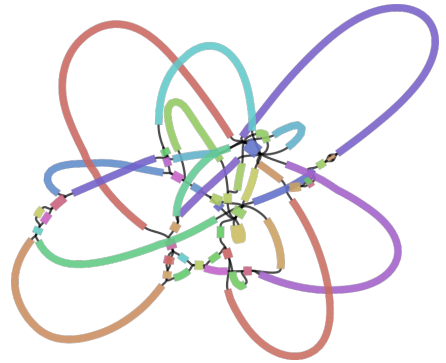
- Many pieces (computational)
- Errors in sequence (which is correct?)
- Missing fragments
- Repetitive fragments (tandem, interspersed)
- Multiple copies (rRNA gene as an example)
- Circular genome: no starting point
- Conserved genes
- Community composes of strains, species, genera with high similarity
- Diverse environments, how much data is needed?

Metagenomic assembly

- Only fraction of reads assembles
  - Does not represent the whole community
- Can and will contain errors
- Testing different assemblers
- More or less sequences
- Co-assembly in some cases

- We have long reads! They help

Assembling long reads with Flye

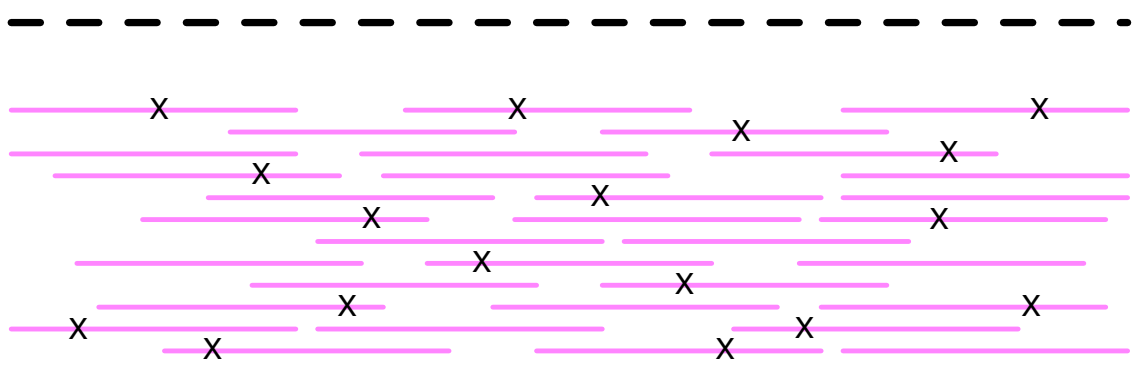


- **Flye** is a de novo assembler for single-molecule sequencing reads, such as those produced by PacBio and Oxford Nanopore Technologies.
  - special mode for metagenome assembly.
- repeat graph as the core data structure. Compared to de Bruijn graphs (which require exact k-mer matches), repeat graphs are built using approximate sequence matches, and can tolerate the higher noise of SMS reads.

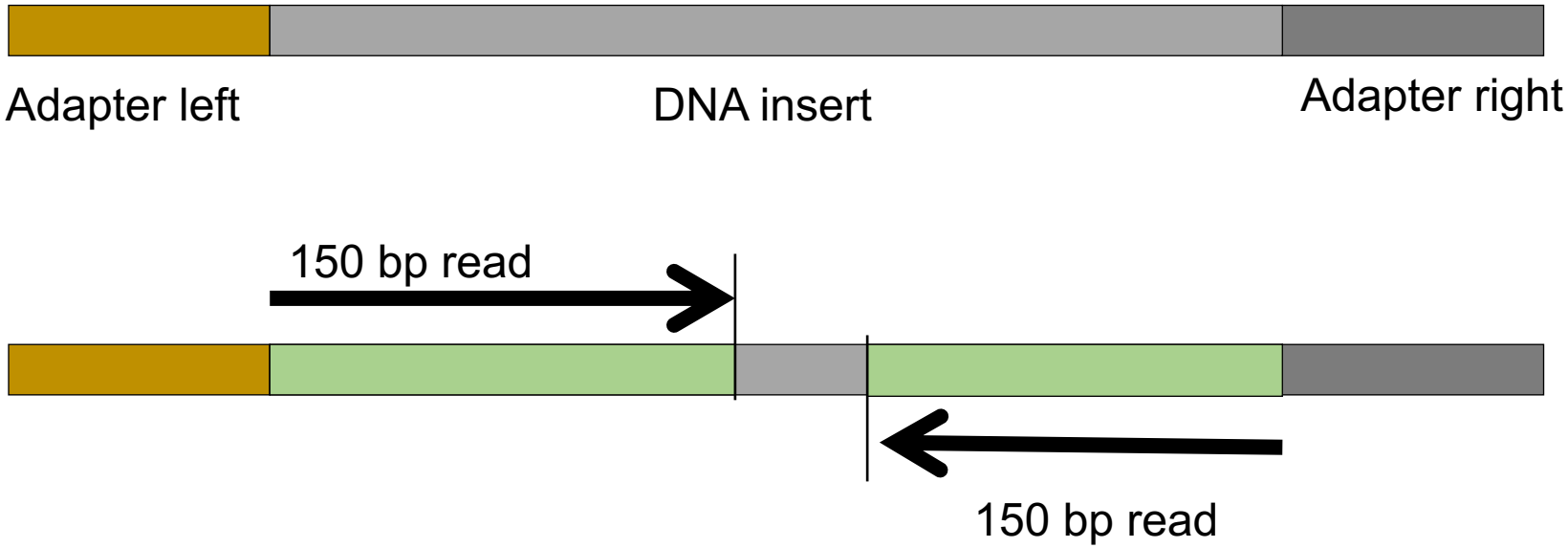
Coverage

- Coverage describes the average number of reads that align to, or "cover," known reference base

- Average coverage



Paired-end sequence

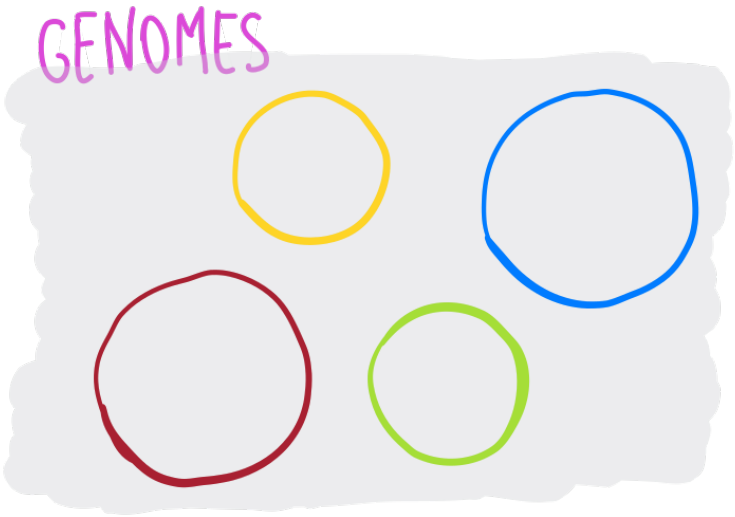


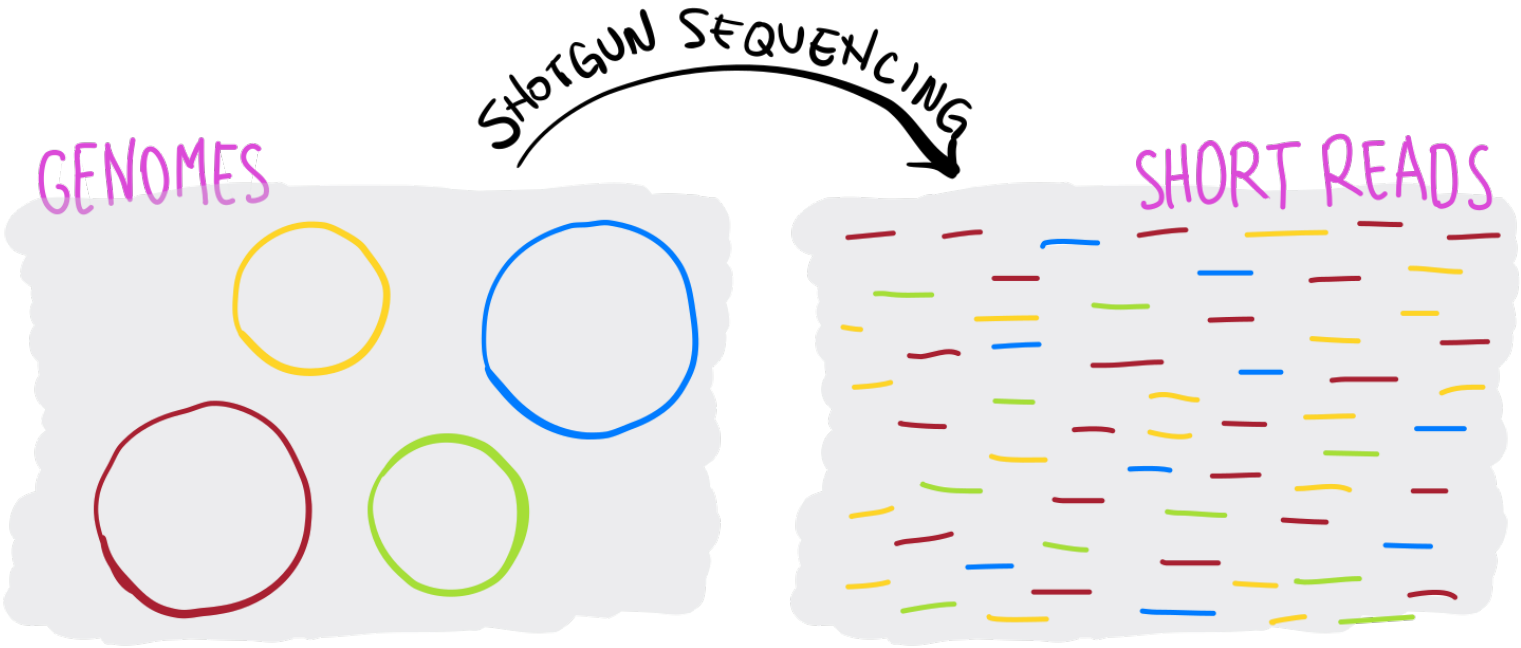
- Helps in assembly, tell to assembler to aid correct assembly
- Also insert size needed

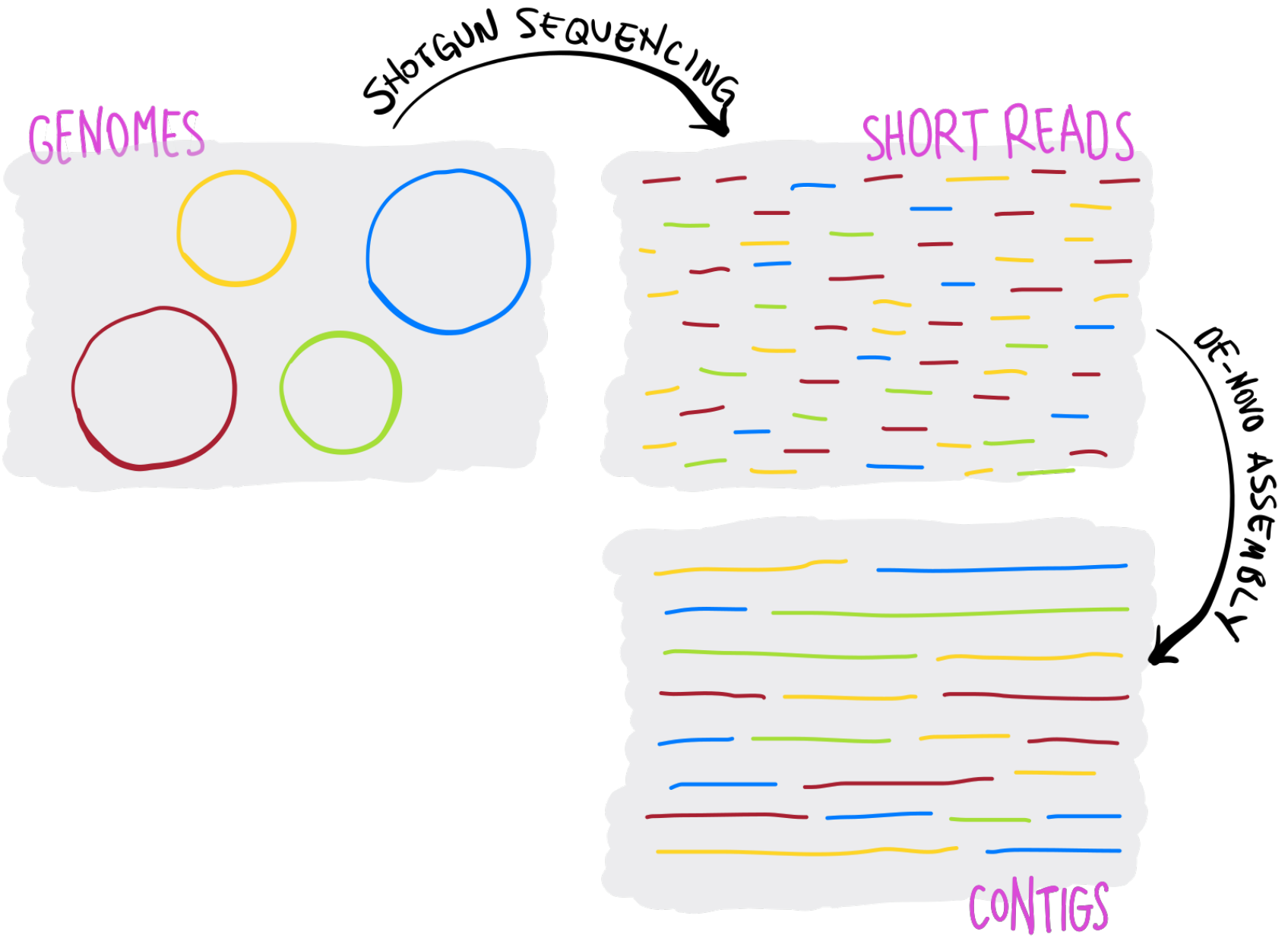
Reconstructing genomes from metagenomes

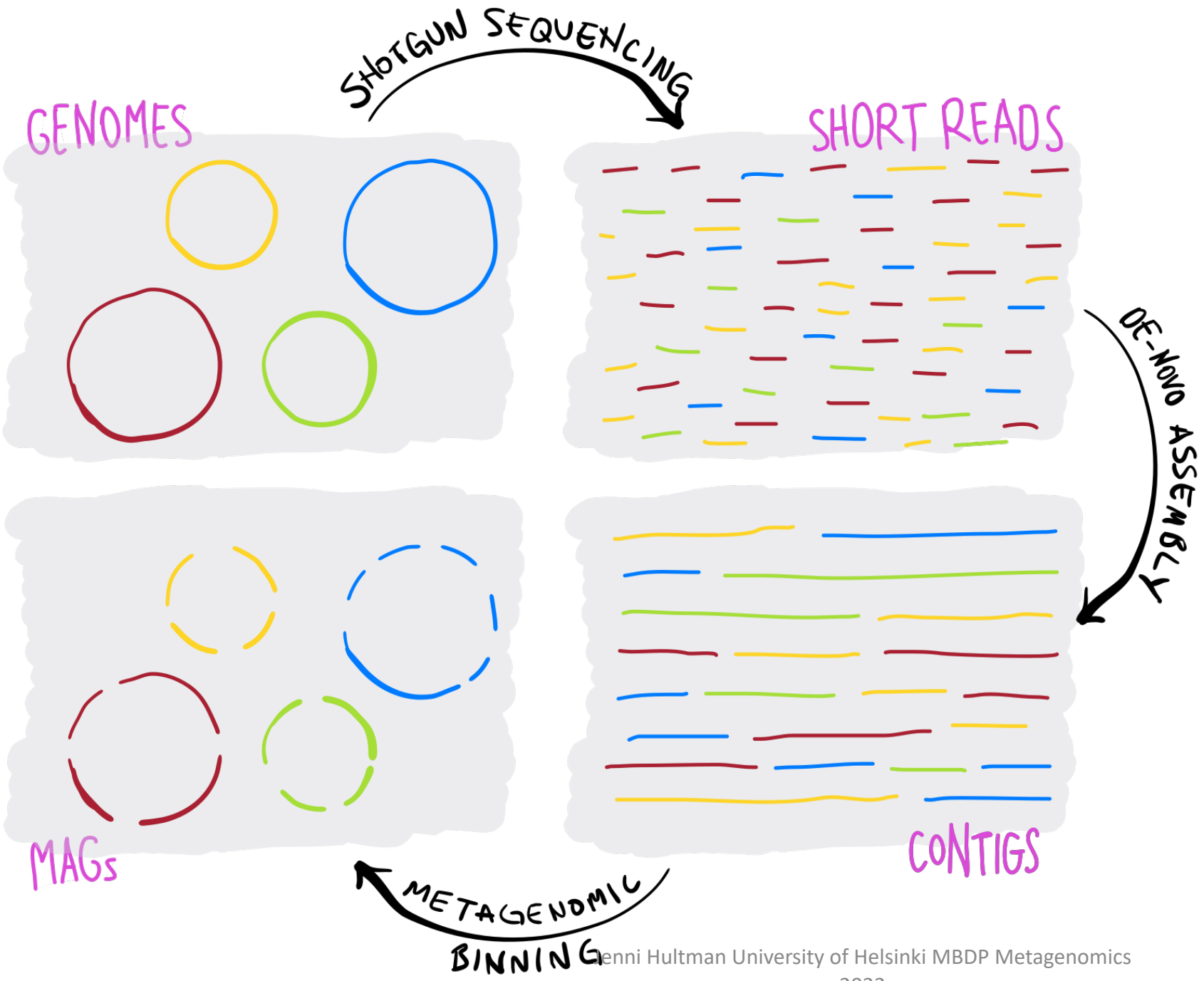
Jenni Hultman University of Helsinki MBDP Metagenomics 2022

http://merenlab.org/momics


GENOMES


SHOTGUN SEQUENCING
GENOMES
SHORT READS

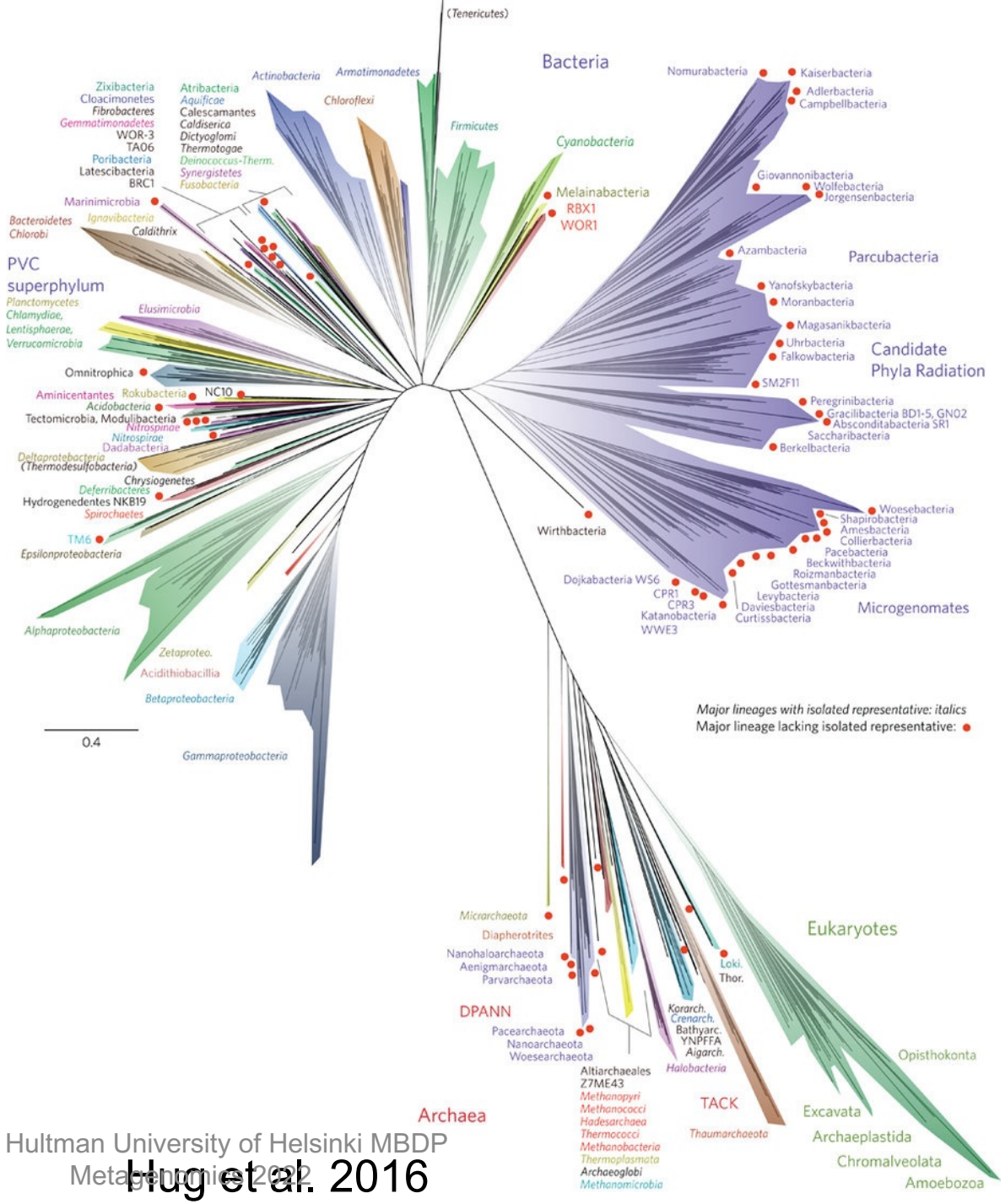

SHOTGUN SEQUENCING
GENOMES
SHORT READS
DE-NOVO ASSEMBLY
CONTIGS



Same community without cultivation of the members

Omics

- Metagenomics
- Metatranscriptomics
- Metaproteomics
- Meta-metabolomics



Hug et al. 2016

Omics

- Metagenomic
- Metatranscrip
- Metaproteom
- Meta-metabo

All omics need good reference genomes/databases

Hug et al. 2016

Overview of generic* metagenomics workflow

*This is generic; specific workflows can vary on the order of steps here and how they are done.

When working with your own data you should never follow any pipeline blindly. There can be critical differences based on your data.

sequencing facility → fastq files → demultiplex (split samples by barodes) → quality filter/trim (remove adapters/**primers**) → fasta files

might be done by sequencing facility

fastqc/multiqc

@HISEQ2500:282:1:1101:1220:1944 1
ATCGGATCG...
+
<G.<G<AGGII...

Some tools:
- sabre
- fastx_demux (usearch/vsearch)
- idemp
- fastx barcode splitter (fastx-toolkit)

Some tools:
- trimmomatic
- bbduk.sh (bbtools suite of tools)

>HISEQ2500:282:1:1101:1220:1944 1
ATCGGATCG...

no-assembly path

read-based analysis

Some tools:
- TIPP/SEPP
- metaphlan2
- humann2
- sourmash
- kraken

assembly path

consider testing assemblies with and w/o

digital normalization

Some tools:
- bbnorm
- diginorm

Count Table

	Sample_A	Sample_B	...
obj_1	0	428	...
obj_2	306	323	...
obj_3	217	1	...
...	...	...	...

MetaQUAST is a great tool for comparing assemblies

Analysis

Some tools:
- phyloseq
- Breakaway
- DivNet
- CORNCOB
- SpiecEasi
- MaAsLin
- DESeq2

(co)-assembly

map individual sample reads to (co)-assembly

Generate coverage information (mapping)

Some tools:
- bowtie2
- bwa

Some assemblers and tools:
- Megahit (assembler)
- SPAdes (assembler)
- idba-ud (assembler)
- MetAMOS (assembler and analysis pipeline)
- MetaCompass (reference-guided)
- MetagenomeScope (visualize assembly graphs)

Gene calling Functional/taxonomic profiling

Some tools:
- prodigal (identifies open reading frames)
- prokka (runs prodigal and performs annotations)
- GHOSTKOALA (web-hosted KEGG annotations)
- BLAST (protein nr db/refseq/COGs)

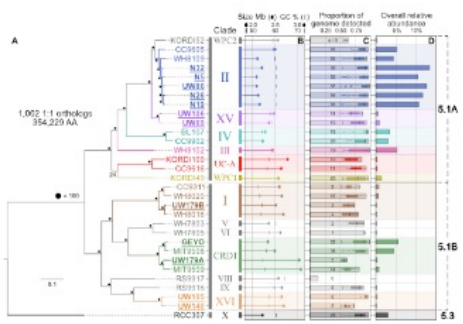
Recovering genomes from metagenomes → **A note on MAGs:** MAGs (metagenome-assembled genomes) are not the same thing as isolate genomes. They are composite representative genomes of closely related genomic lineages.

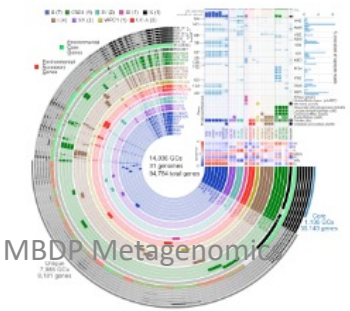
Some tools:
- anvi'o (interactive manual curation of bins; and much more)
- CONCOCT (kmer-based and coverage-based binning; also incorporated in anvi'o)
- COCACOLA (kmer-based, coverage-based, and incorporates paired-read linkage of contigs)
- MetaBAT2 (kmer-based and coverage-based binning tool)
- BinSanity (primarily coverage-based, optional second round kmer-based binning tool)
- checkm (genome-level taxonomy; and much more)
- DASTool (a tool for evaluating bins recovered by different methods)
- DESMAN (tool aimed at resolving strains)

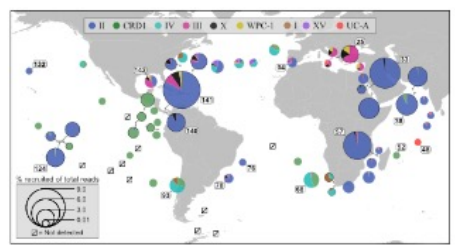
Some common genomics stuff

Phylogenomics
Comparative genomics
Pangenomics
Env. distributions

Some tools:
- anvi'o (integrated HMMs for common single-copy gene sets; integrated pangenomic workflow for identifying orthologs via OrthoMCL)
- PanOCT (identifies orthologs utilizing synteny information)
- StrainPhlAn/PanPhlAn (tools for strain-level analyses)
- MUSCLE (alignment software)
- FastTree (very fast pseudo-maximum likelihood tree builder)
- RAxML (maximum likelihood tree builder)
- Mauve (whole-genome alignment)







astrobiomike.github.io

CSC greetings

- You all have CSC account with 1000 billing units
  - But not project where to do more intensive computing
  - You can run out of billing units
    - **saldo**, should not be negative
    - Check from www.mycsc.fi
- For this course we have a project MBDP_metagenomics_2022, Jenni will add you
- Make sure that when you work with **real data** you have **a PI who has a project** with enough billing units and you are member of that project