CHARLES UNIVERSITY

Bioinformatics and Microbiome Analysis MB140P94

Genomics, shotgun data and (meta)genome assembly

Tomáš Větrovský, Iñaki Odriozola and Petr Baldrian
Laboratory of Environmental Microbiology
Institute of Microbiology of the CAS

What is genomics?

Genomics

- interdisciplinary field of science focusing on structure, function and evolution of **genomes**

X

Genetics

- study of individual **genes** and their roles in inheritance

Genomics aims:
- collective characterization and quantification of genes, which direct the production of proteins (with the assistance of enzymes and messenger molecules)

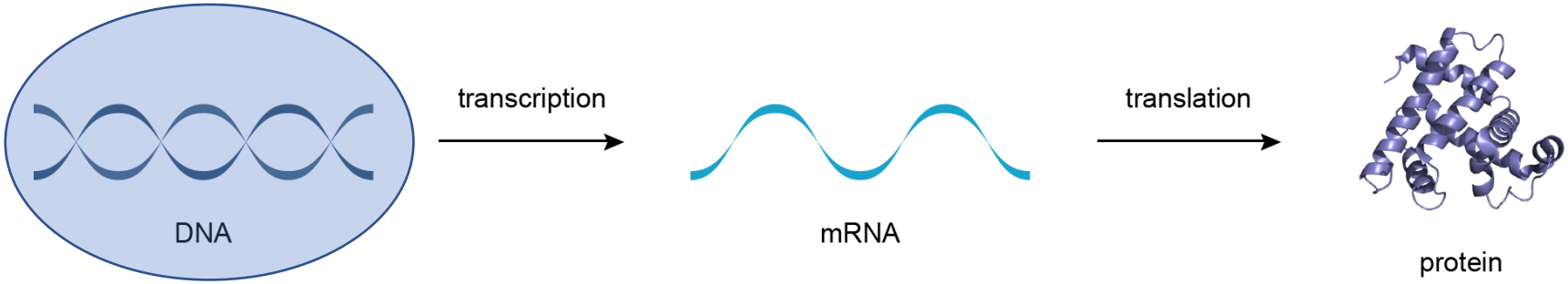
Genomics involves:
- mapping, and editing of genomes
- sequencing and analysis of genomes
- bioinformatics

Genomic has several research areas:
- **metagenomics**
- **structural genomics**
- **functional genomics**
- **comparative genomics**

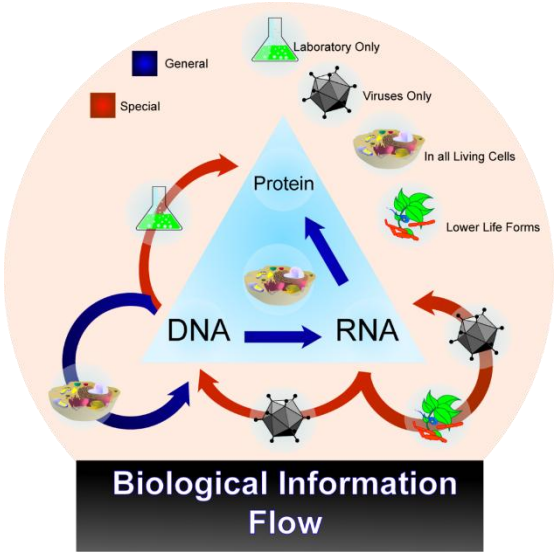
Central dogma of molecular biology

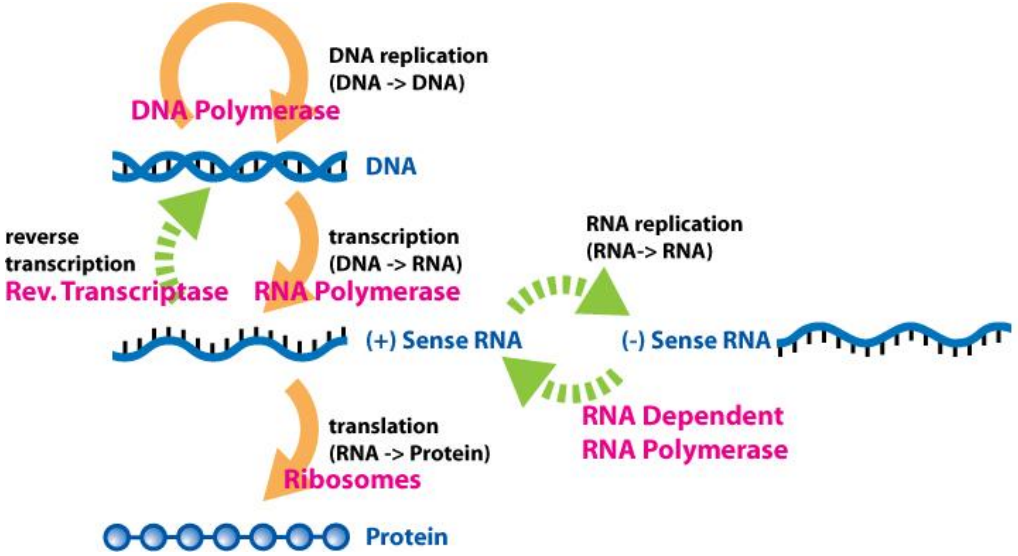
explanation of the flow of genetic information within a biological system

General (Watson 1965)



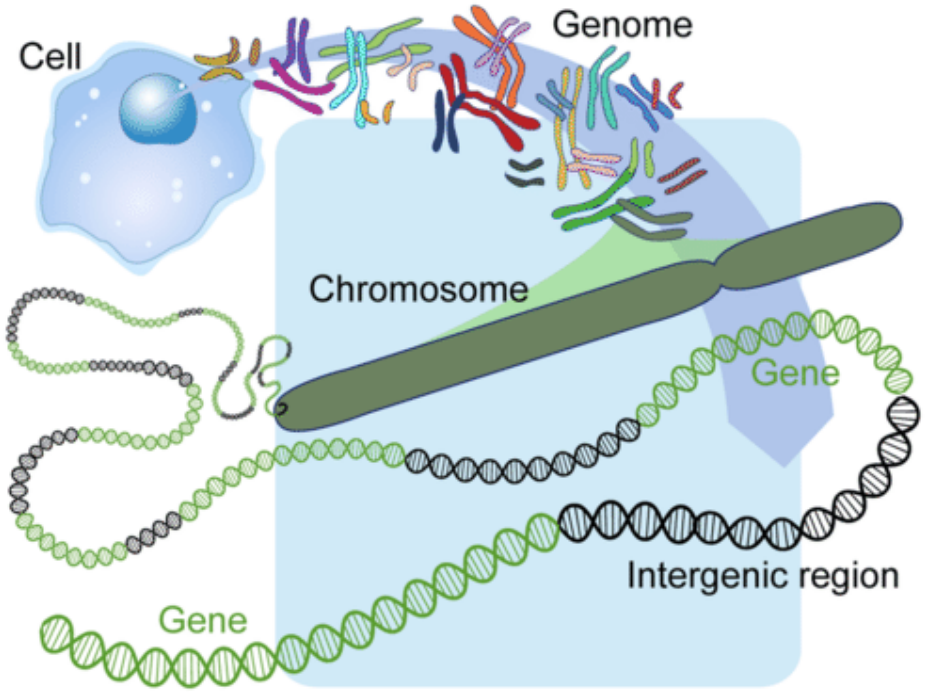
General + Special

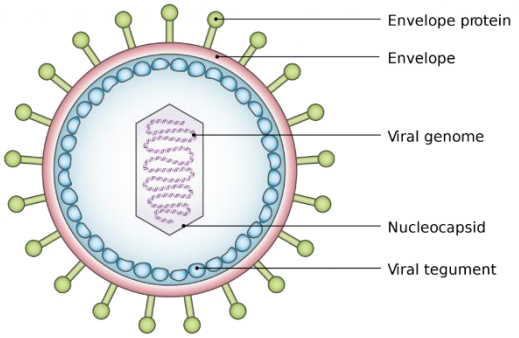


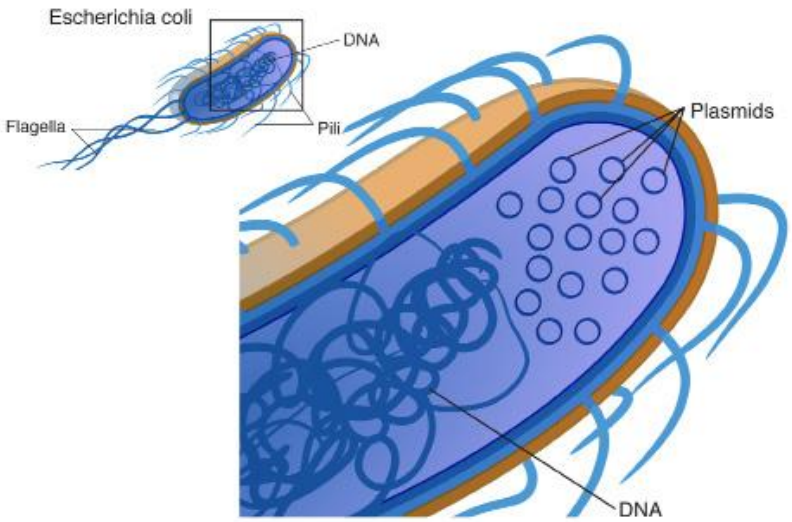


What is genome?

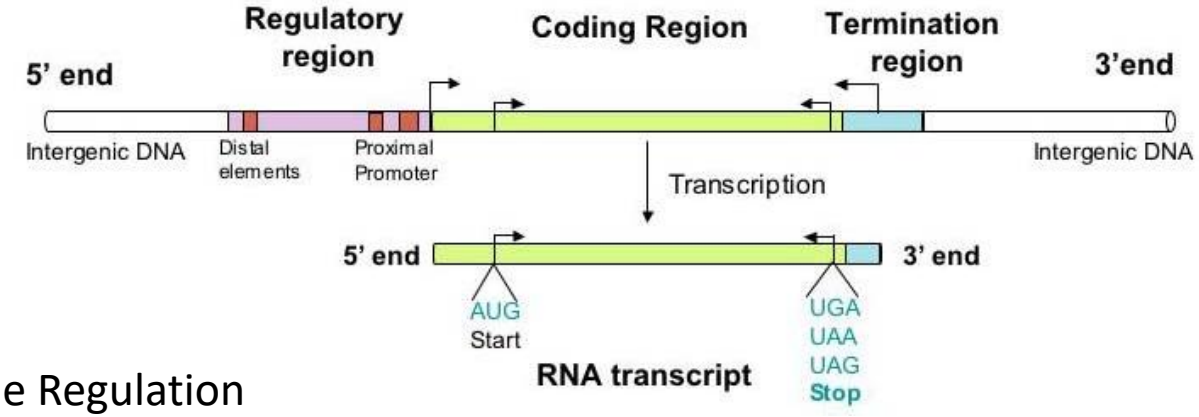
The genome is the genetic material of an organism. It consists of DNA (or RNA in RNA viruses). The genome includes both the **genes** (the coding regions) and the **noncoding DNA**, as well as the **other genetic material** (e.g. mitochondria or chloroplasts).



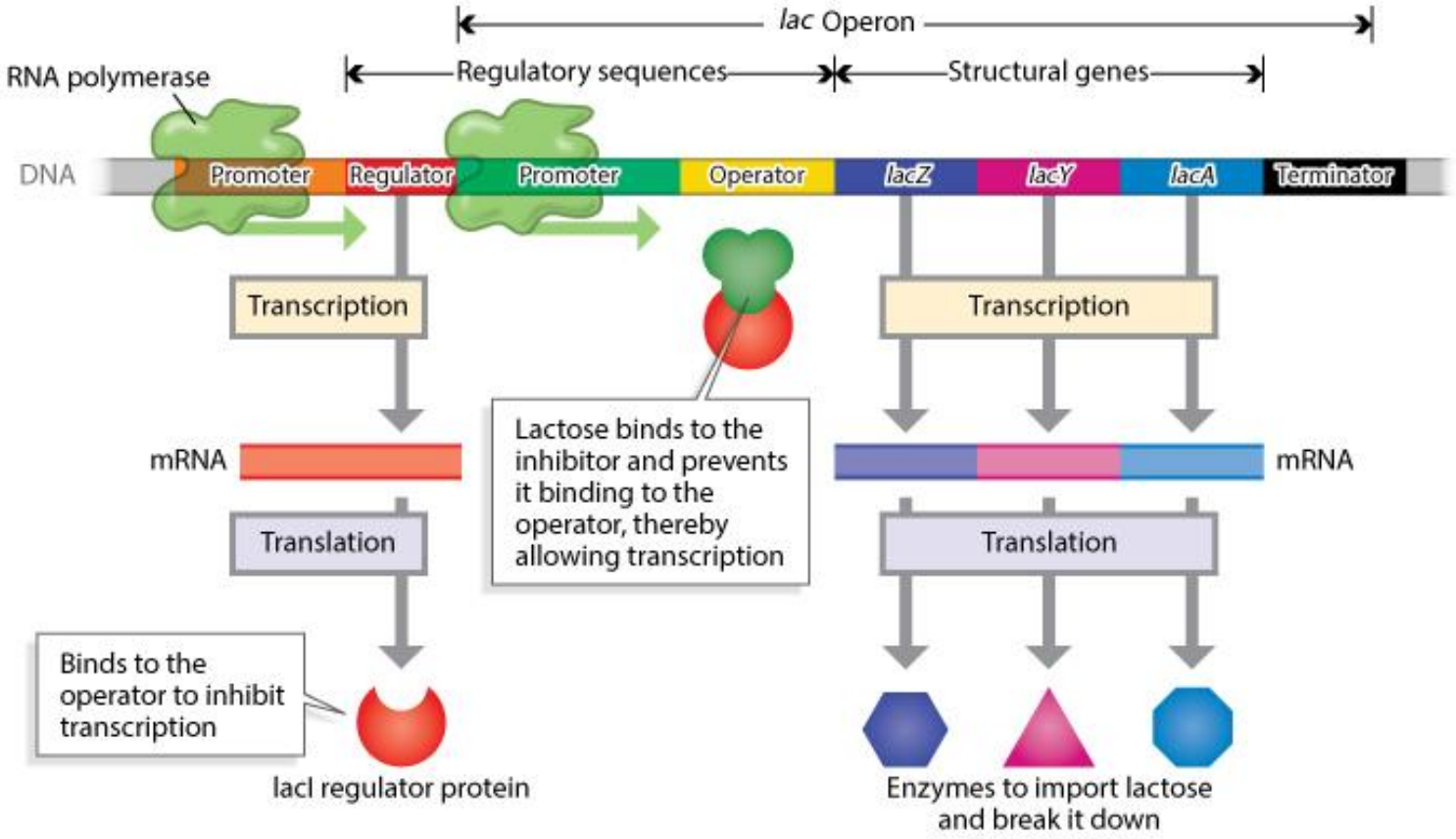




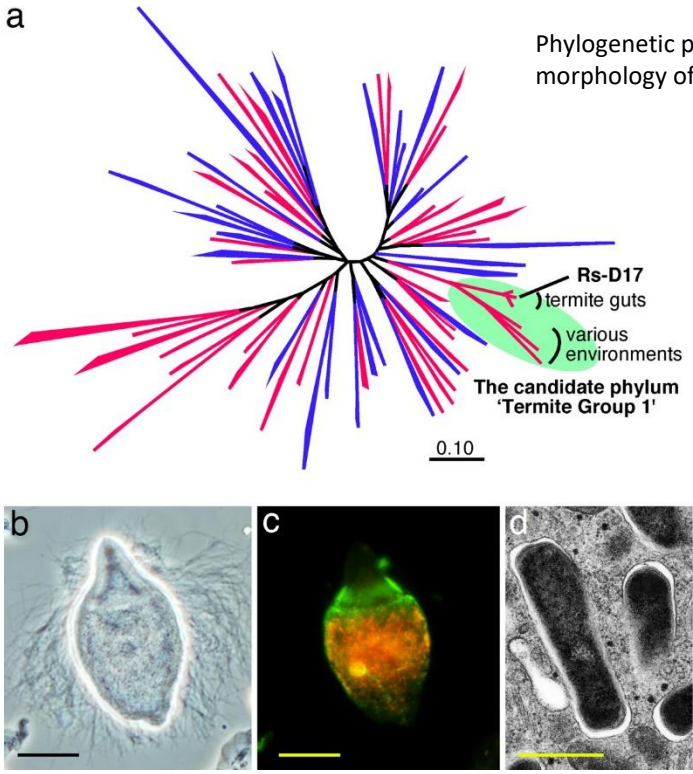
Prokaryotic gene structure



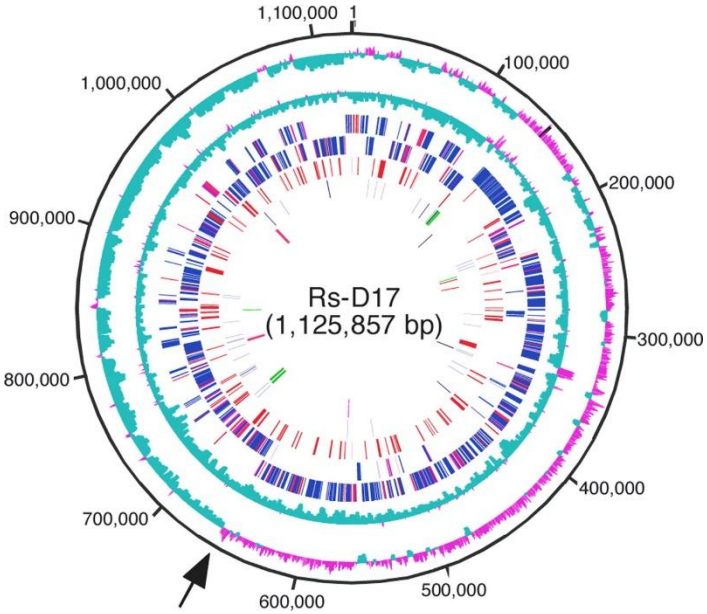
Operons and Prokaryotic Gene Regulation



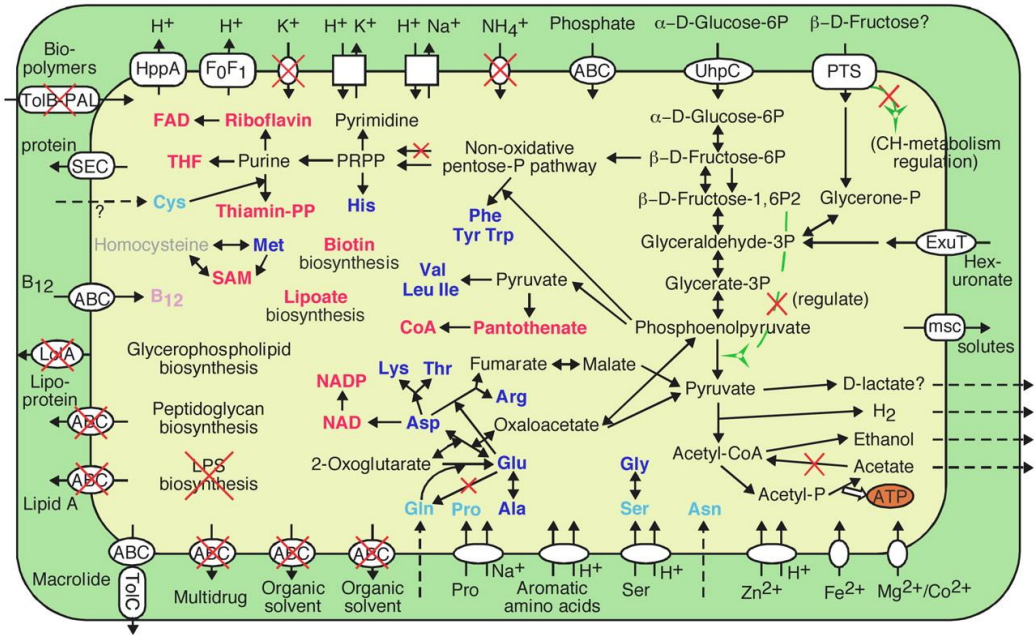
Information encoded in genome of the uncultured Termite Group 1 bacteria in a single host protist cell



Phylogenetic position, localization and morphology of phylotype Rs-D17.



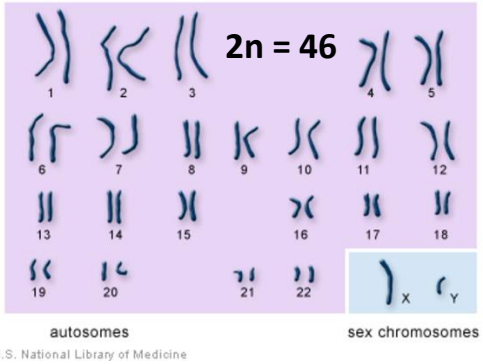
Circular representation of the Rs-D17 chromosome.

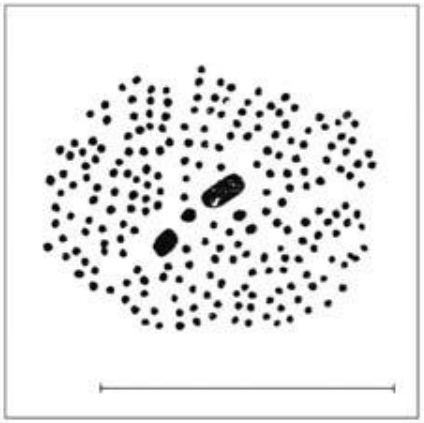


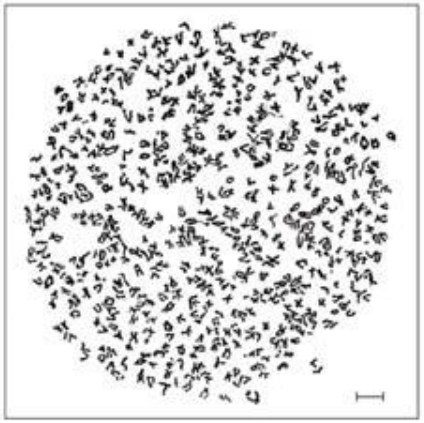
Predicted metabolic pathways of phylotype Rs-D17. Periplasm (green) and cytoplasm (yellow) are shown bounded by outer and inner membranes, respectively. Blue, synthesized amino acids; red, cofactors. Compounds that must be imported are shown in pale colors. Nonfunctional pathways and transporters comprising pseudogenes are marked with red X's. Arrows with broken lines indicate diffusion or transport via some unidentified apparatus. ABC, ATP-binding cassette type transporter; msc, mechanosensitive channel; FAD, flavin adenine dinucleotide; SAM, S-adenosylmethionine; THF, tetrahydrofolate.

Eukaryotic genomes

- linear DNA chromosomes (in the nucleus)
- number of chromosomes varies widely
- exon-intron organization of protein coding genes
- repetitive DNA.
- + mitochondrial and chloroplast genomes







Chromosome number diversity found in animals and plants. (a) The bulldog ant, Myrmecia pilosula, 2n = 4 (from Crosland & Crozier, 1986) (b) The lycaenid butterfly, Lysandra atlantica, 2n = ca. 440 (from de Lesse, 1970) (c) Brachycome dichromosomatica, 2n = 4) (d) Adder's tongue fern Ophioglossum reticulatum 2n = ca. 1440 (from Ninan, 1958). Scale bar = 10 µm.

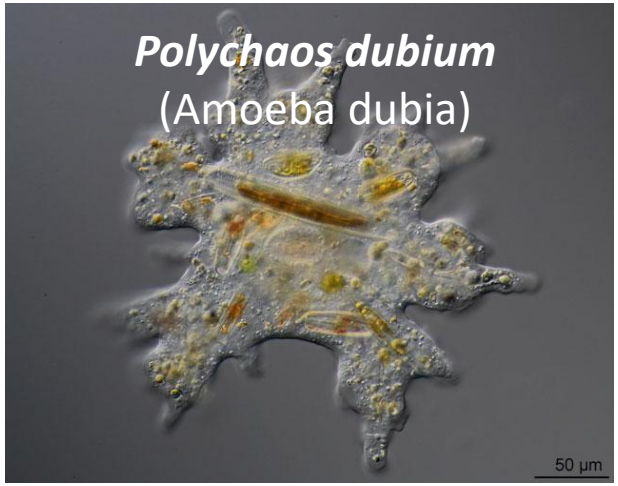
The largest genomes

Rank	Organism	Approximate genome size	Note
1	*Tmesipteris oblanceolata*	**160 Gbp**	largest verified genome (plant)
2	*Paris japonica*	149 Gbp	former record holder (plant)
3	*Protopterus aethiopicus*	130 Gbp	largest animal genome
–	Polychaos dubium	~670 Gbp (unverified)	historical, methodologically dubious figure

Tmesipteris oblanceolata
50x larger than the human genome (which has ~3.2 Gbp)

Nevertheless, it's an extreme example of genome gigantism, typically caused by:

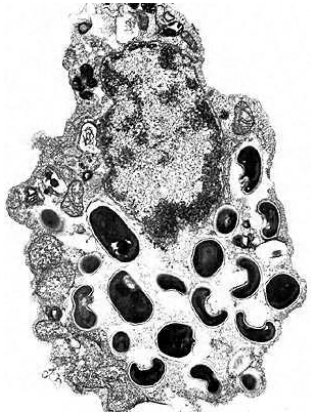
- massive DNA duplications
- accumulation of mobile elements,
- low selective pressure to reduce genome size in some unicellular eukaryotes



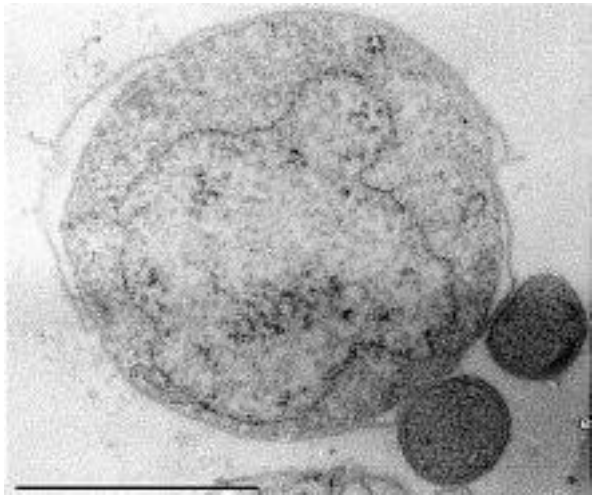
The smallest genomes

Rank	Organism	Approximate genome size	Note
1	“Candidatus Nasuia deltocephalinicola”	112 kbp	smallest confirmed genome of any cellular organism (obligate bacterial endosymbiont of leafhoppers)
2	Nanoarchaeum equitans	490 kbp	smallest archaeal genome; hyper-thermophilic parasite of Ignicoccus
3	Mycoplasma mycoides JCVI-syn3.0	531 kbp	lab-built minimal cell with the smallest genome of any self-replicating organism grown in culture (synthetic) J. Craig Venter Institute
4	“Candidatus Pelagibacter communis”	1.3 Mbp	smallest genome of any free-living organism (marine SAR11 bacterium)
5	Encephalitozoon intestinalis	2.3 Mbp	smallest known eukaryotic nuclear genome (microsporidian parasite)

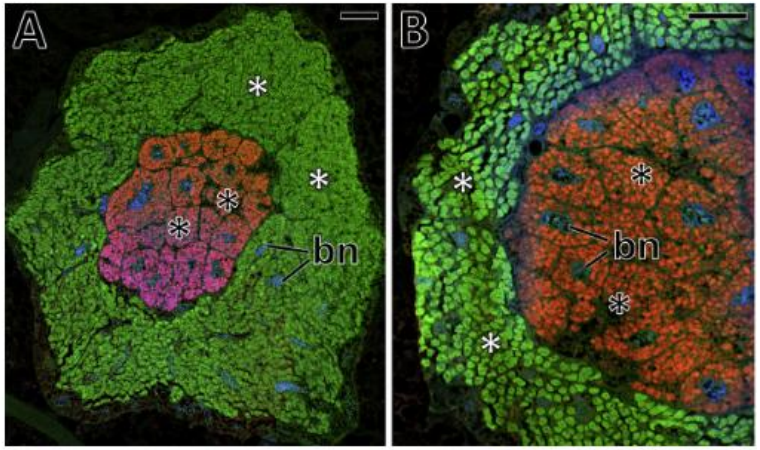
- Viruses and viroids can be much smaller (down to a few hundred nucleotides – 220-250 bp) - but they’re not cellular organisms
- Endosymbionts and parasites can shed many genes they no longer need, which is why they dominate the extreme low end



Encephalitozoon intestinalis



Two *Nanoarchaeum equitans* cells (and its larger host *Ignicoccus*)



What is gene?

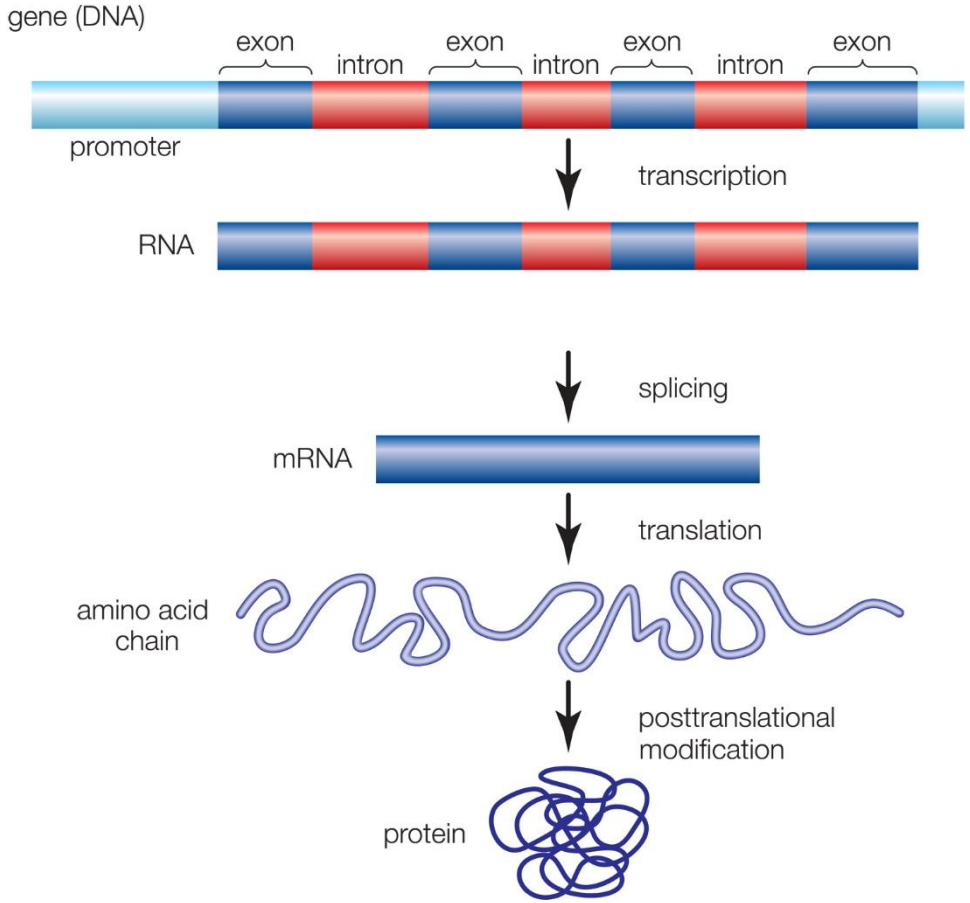
gene is a sequence of DNA or RNA which codes for a **molecule that has a function**

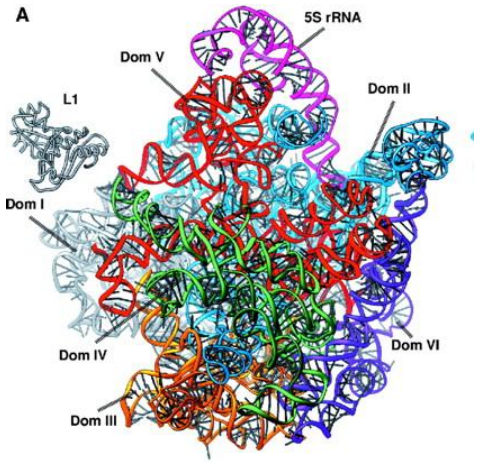
gene expression

DNA is copied into RNA (**transcription**)

RNA can be directly functional **ribozymes** (**ribo**nucleic acid en**zyme**s)

…or be the intermediate template for a protein (**translation**)

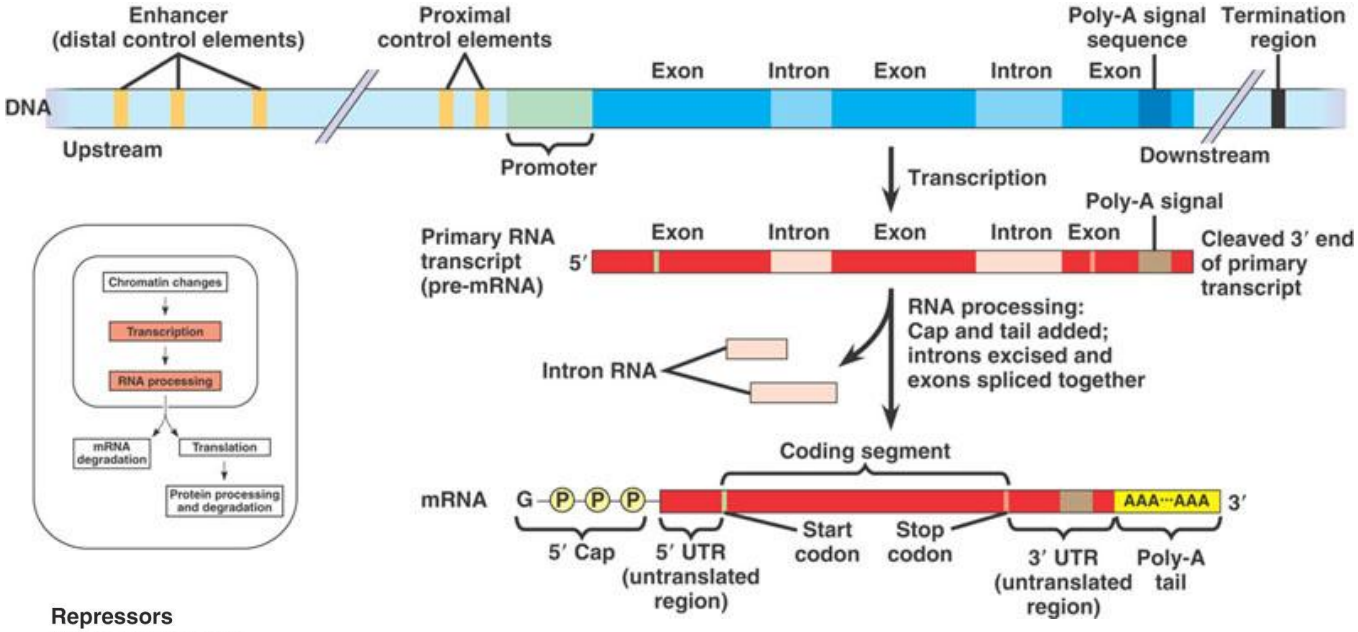




A ribosome is a biological machine that utilizes a ribozyme to translate RNA into proteins

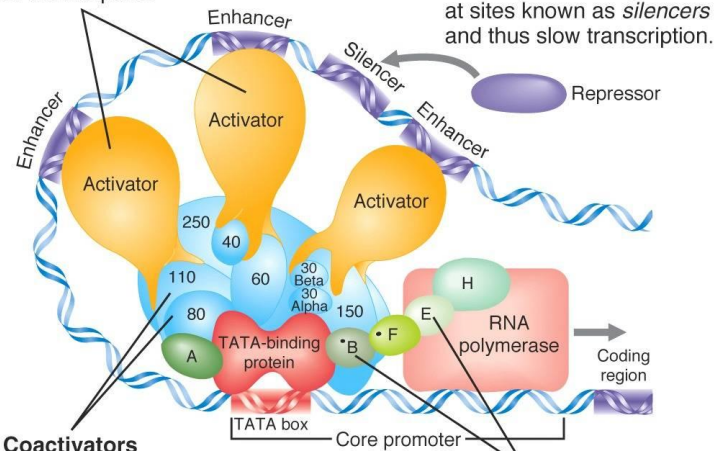
Eukaryotic gene structure

exon-intron organization of protein coding genes



Activators
These proteins bind to genes at sites known as *enhancers* and speed the rate of transcription.

Repressors
These proteins bind to selected sets of genes at sites known as *silencers* and thus slow transcription.

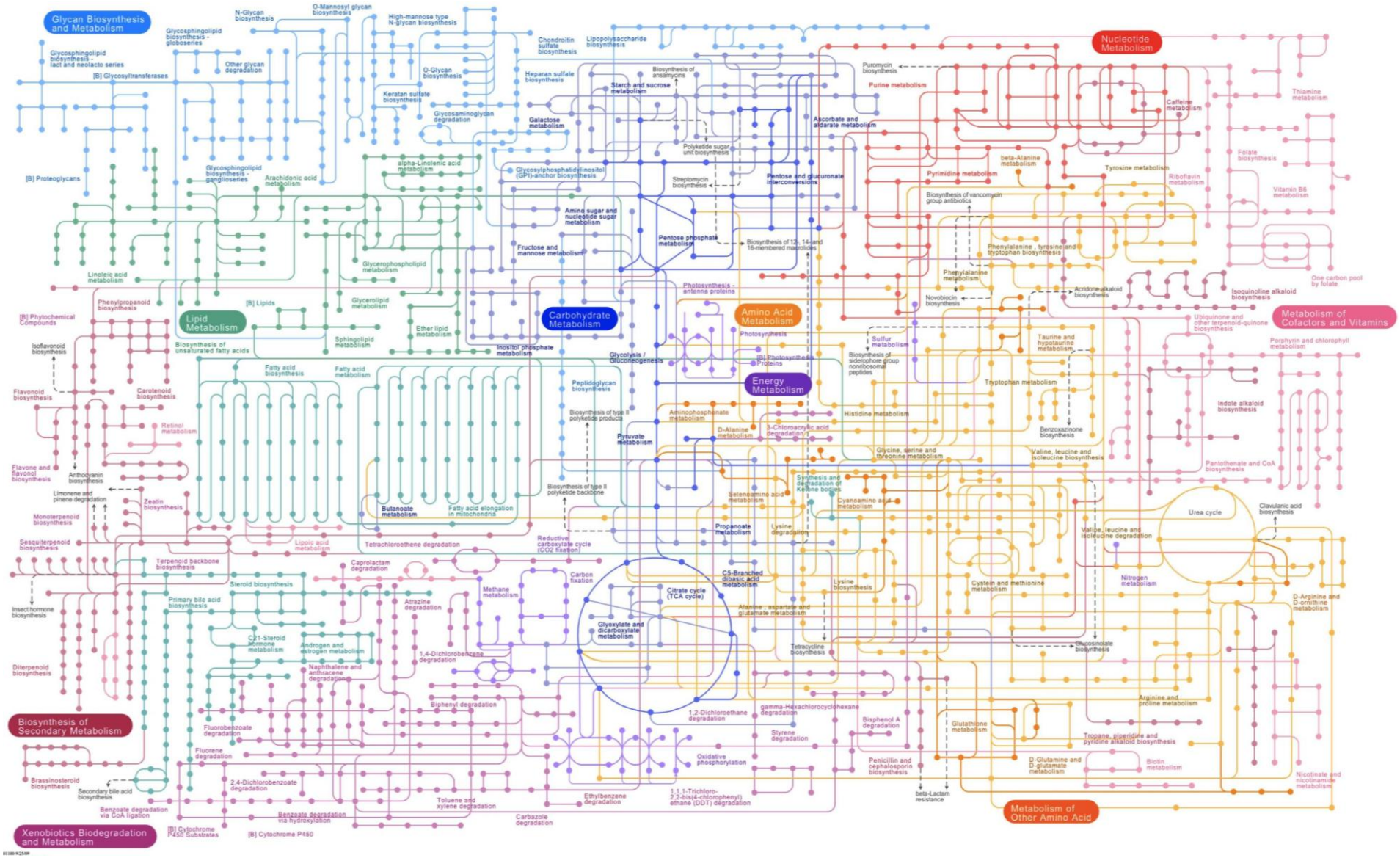


Coactivators
These "adapter" molecules integrate signals from activators and perhaps repressors.

Basal transcription factors
In response to injunctions from activators, these factors position RNA polymerase at the start of

Much complex regulation of expression…

Information encoded in eukaryotic genome…

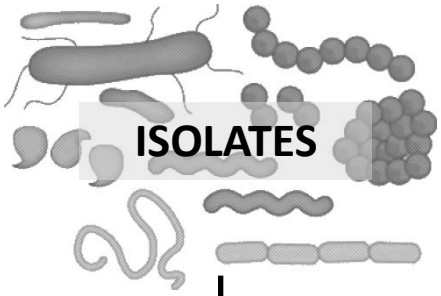


METAGENOMIC APPROACHES

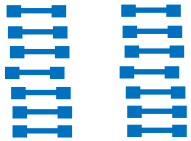
ENVIRONMENTAL SAMPLE

EXTRACTED DNA

ISOLATES



Targeted Gene Sequencing



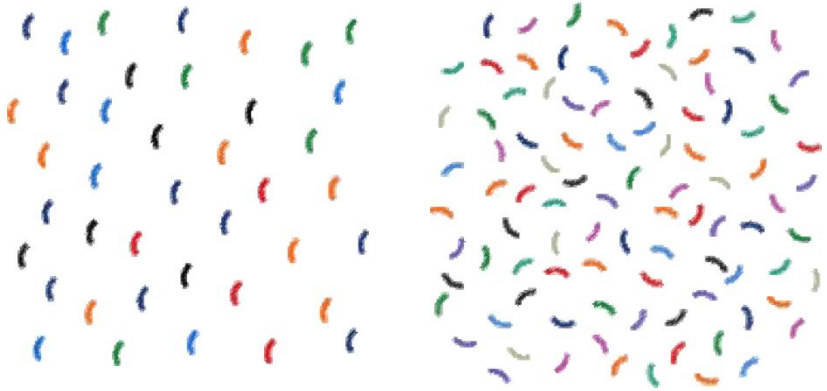
known genes
identity, diversity, distribution, dynamics

Shotgun Metagenome Sequencing

"all" genes
functional potential metabolic reconstruction

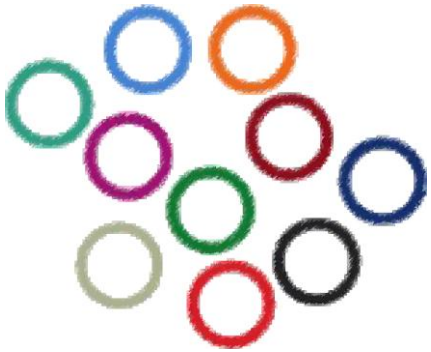
Single Cell Genome Sequencing

genomes
linking phylogeny & function, genome diversification



Gene-centric

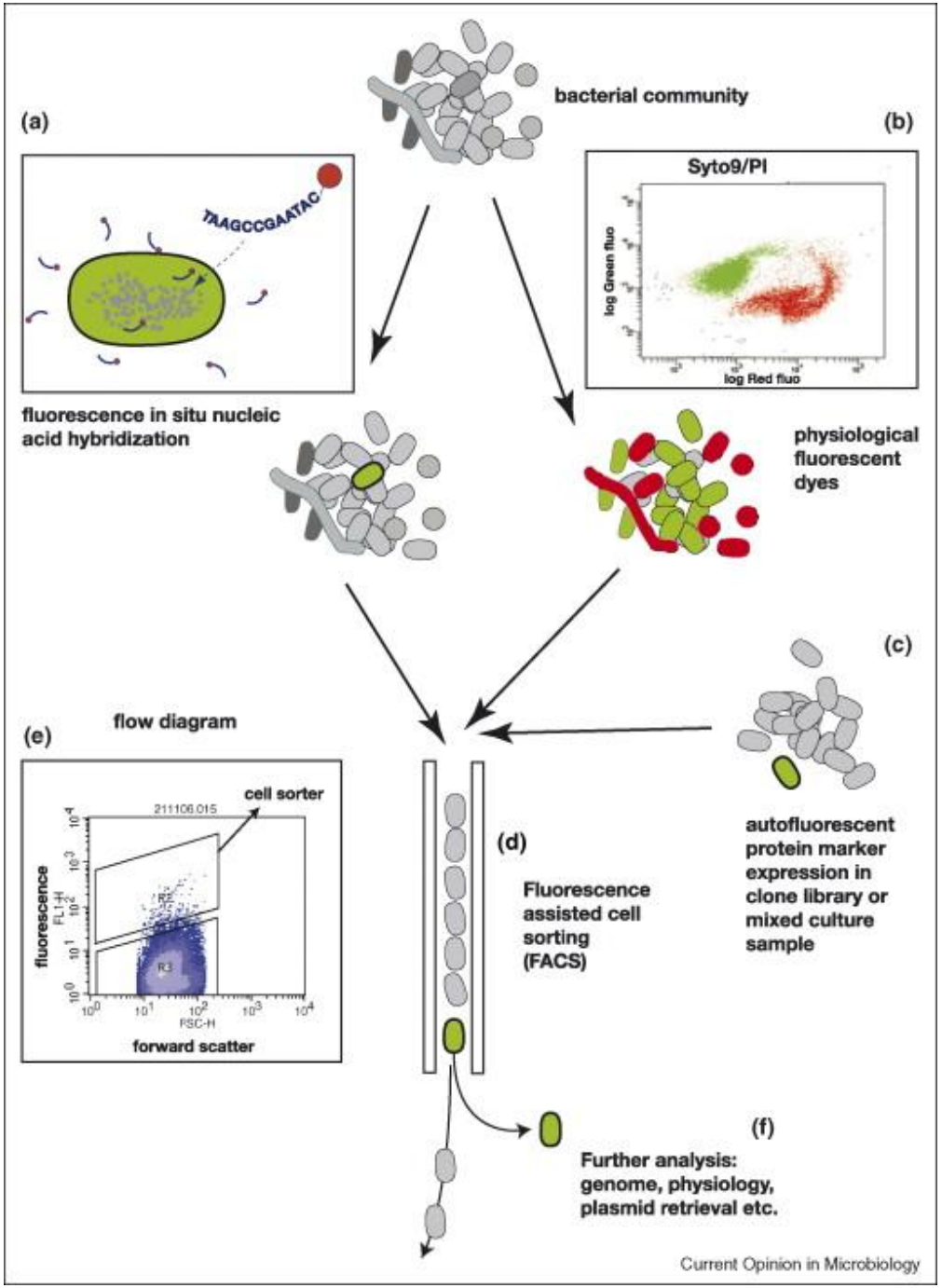
binning to *metagenome* assembled *genomes* (*MAGs*)

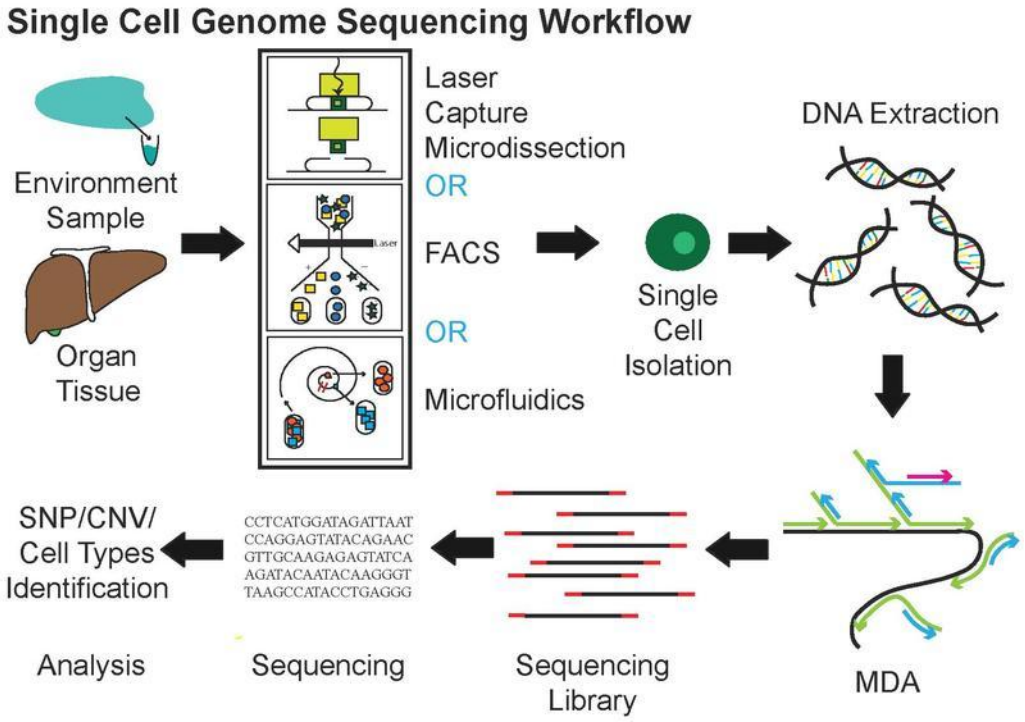


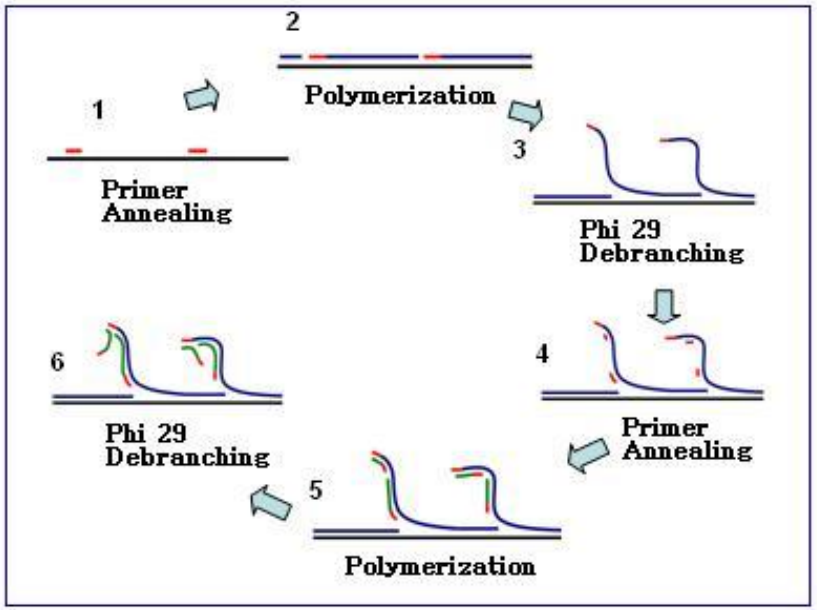
Genome-centric

DNA sequencing of uncultured microbes from single cells

Schematic illustration of the different strategies for using **flow cytometry** and fluorescence assisted cell sorting in environmental microbiology
(a–f) **fluorescence in situ hybridization to target specific cells** in the community and subsequent sorting or characterization (a).
(b) Combinations of physiological **stains to differentiate active from non-active populations**; here, a diagram using SYTO9/propidium iodide co-staining on *Pseudomonas fluorescens*
(c) Illustration of the use of expressed autofluorescent proteins under control of specific promoters to identify and sort cells in, for example clone libraries or for screening purposes.
(e) Illustration of a flow diagram of a population of Escherichia coli with a few mutant cells expressing green fluorescent protein from a modified regulator protein







Multiple displacement amplification (MDA) is a non-PCR based DNA amplification technique. This method can rapidly amplify minute amounts of DNA samples to a reasonable quantity for genomic analysis. The reaction starts by **annealing random hexamer primers** to the template: DNA synthesis is carried out by a **high fidelity enzyme, preferentially Φ29 DNA polymerase**, at a constant temperature. Compared with conventional PCR amplification techniques, MDA generates larger sized products with a lower error frequency. This method has been actively used in whole genome amplification (WGA) and is a promising method for application to single cell genome sequencing and sequencing-based genetic studies.

Shotgun metagenome sequencing pipeline workflow

DNA sequencing

samples properties

input data
(paired-end FASTQ files)

remove residual adapters adapters and
remove N from start and end positions

quality filtering

remove short sequences

digital normalisation

assembly

mapping reads
to contigs

MG-RAST
metagenomics analysis server

annotation of
contigs

binning contigs
to genomes

normalize mapping table

metadata

Results

processing the results with R or another statistical software

1. Adapter trim & quality filter

A. remove residual adapters adapters and remove N from start and end positions - trimmomatic

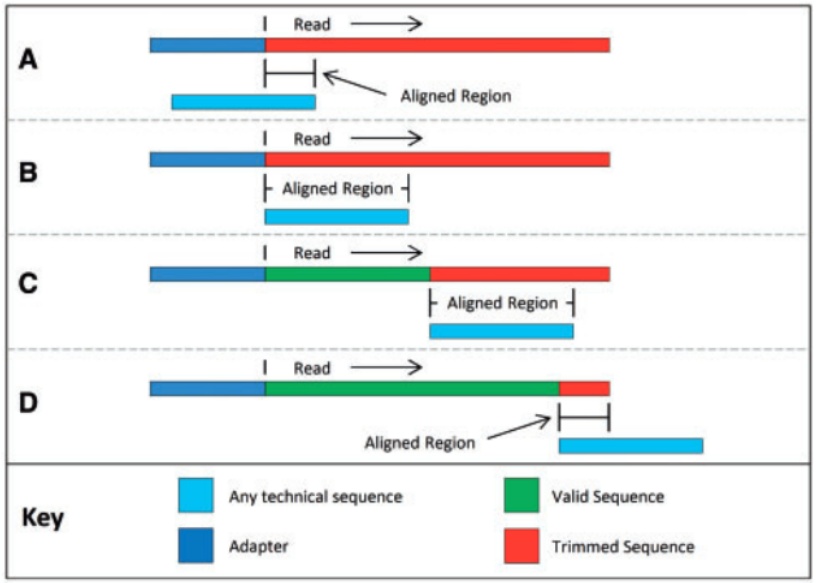
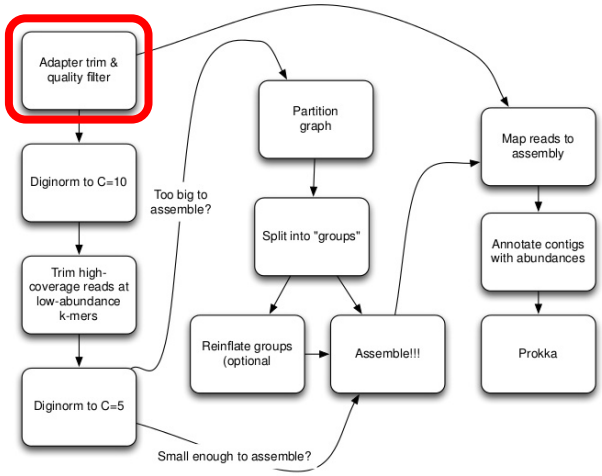


Fig. 1. Putative sequence alignments as tested in simple mode. The alignment process begins with a partial overlap at the 5′ end of the read (**A**), increasing to a full-length 5′ overlap (**B**), followed by full overlaps at all positions (**C**) and finishes with a partial overlap at the 3′ end of the read (**D**). Note that the upstream 'adapter' sequence is for illustration only and is not part of the read or the aligned region



BIOINFORMATICS ORIGINAL PAPER
Vol. 30 no. 15 2014, pages 2114–2120
doi:10.1093/bioinformatics/btu170
Genome analysis
Advance Access publication April 1, 2014

Trimmomatic: a flexible trimmer for Illumina sequence data

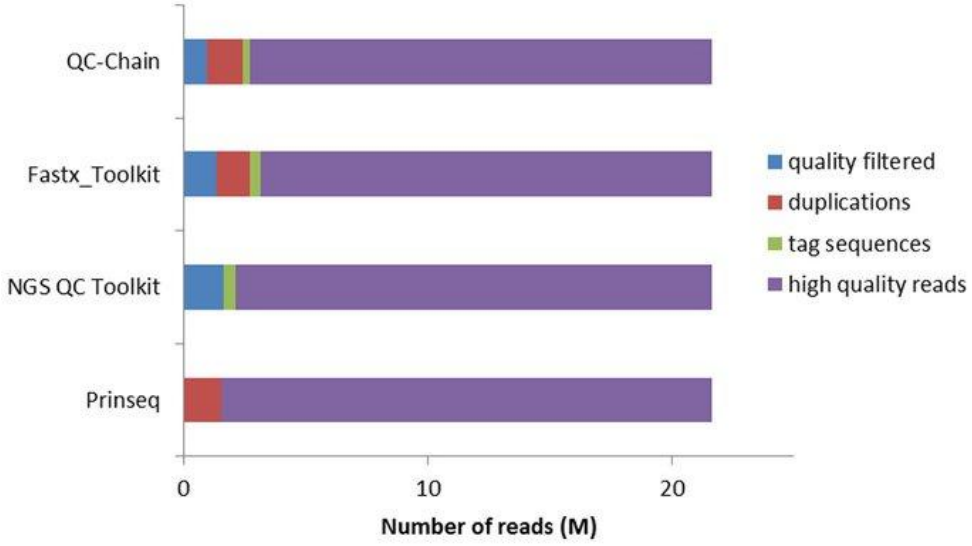
Anthony M. Bolger[1,2], Marc Lohse[1] and Bjoern Usadel[2,3,*]
[1]Department Metabolic Networks, Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, 14476 Golm,[2]Institut für Biologie I, RWTH Aachen, Worringer Weg 3, 52074 Aachen and [3]Institute of Bio- and Geosciences: Plant Sciences, Forschungszentrum Jülich, Leo-Brandt-Straße, 52425 Jülich, Germany
Associate Editor: Inanc Birol

Specific parameters:
LEADING: Cut bases off the start of a read, if below a threshold quality
TRAILING: Cut bases off the end of a read, if below a threshold quality

1. Adapter trim & quality filter

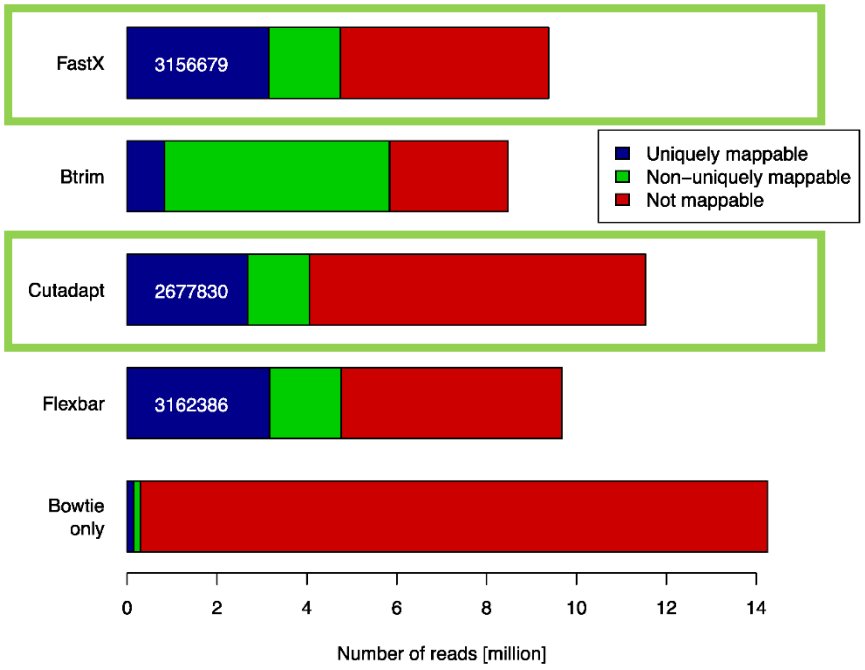
B. quality filtering - fastx

Comparison of read-quality assessment



Three general read-quality trimming procedures, including quality filtration, duplication removal and tag sequences filtration were selected to assess the effects of the tested QC tools.

RNA-seq data – mapping to *C. elegans* genome



Number of returned reads as stratified by subsequent read mapping with Bowtie. Mapping results of untreated reads (Bowtie only) are shown in the bottom most row (control case). The respective adapter removal tools did not return all reads, as some did not pass the respective output filters.

1. Adapter trim & quality filter

C. remove short sequences - Biopieces

Example of command:

```
read_fastq -i ${file} -e base_33 | grab -e 'SEQ_LEN >= 50' | write_fastq -x -o ${file}.cut
```

The Biopieces are a collection of bioinformatics tools that can be pieced together in a very easy and flexible manner to perform both simple and complex tasks. The Biopieces work on a data stream in such a way that the data stream can be passed through several different Biopieces, each performing one specific task: modifying or adding records to the data stream, creating plots, or uploading data to databases and web services. The Biopieces are executed in a command line environment where the data stream is initialized by specific Biopieces which read data from files, databases, or web services, and output records to the data stream that is passed to downstream Biopieces until the data stream is terminated at the end of the analysis as outlined below:

```
read_data | calculate_something | write_results
```

Shotgun metagenome sequencing pipeline workflow

DNA sequencing

samples properties

input data
(paired-end FASTQ files)

remove residual adapters adapters and
remove N from start and end positions

quality filtering

remove short sequences

digital normalisation ?!

assembly

mapping reads
to contigs

MG-RAST
metagenomics analysis server

annotation of
contigs

binning contigs
to genomes

normalize mapping table

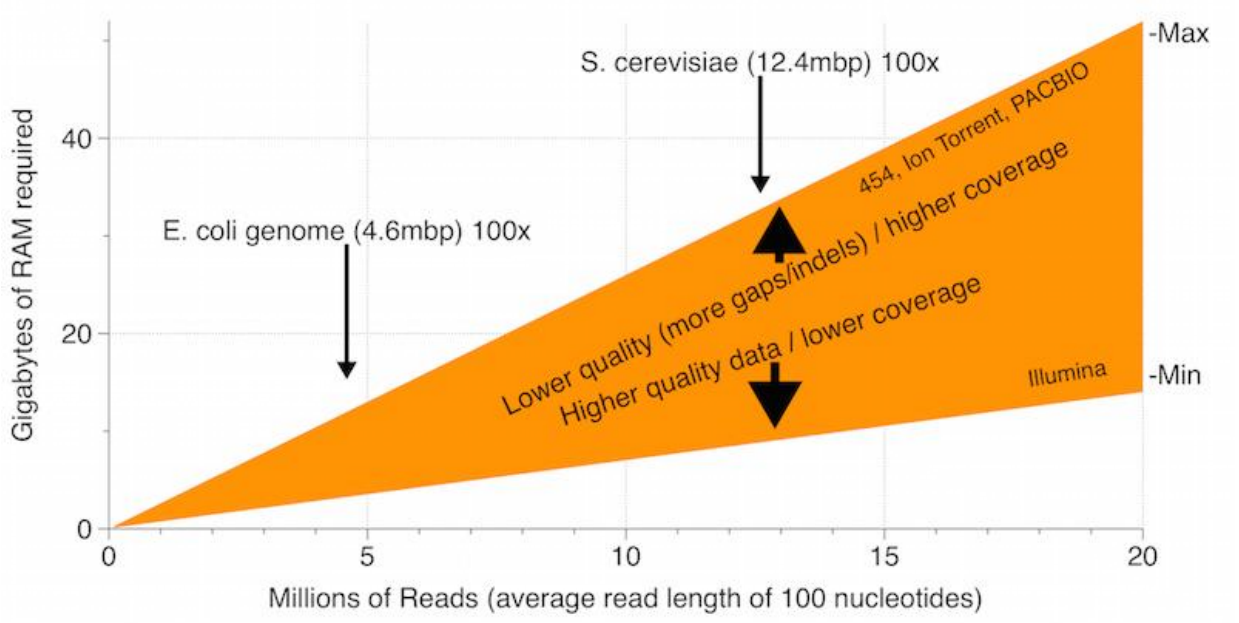
metadata

Results

processing the results with R or another statistical software

Clean data but too big for assembly

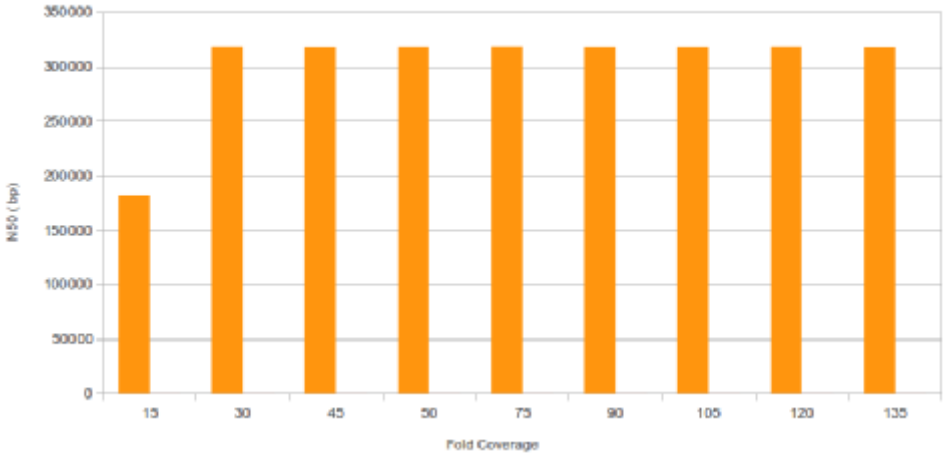
General rules of thumb for RAM requirements:
Assembling data with higher coverage depth will require more RAM – aim for coverage between 50 and 100x
Assembling lower quality data, with more miscalls, indels and gaps, will require more RAM
Doubling the size of your dataset (total nucleotides) will roughly double RAM requirements
For illumina data roughly 1 GB of RAM will be required to assemble a data set of 1 million reads (with an average read length of 100 nucleotides)



https://support.geneious.com/hc/en-us/articles/227534968-What-are-the-hardware-requirements-for-assembly-of-NGS-data-using-the-Geneious-de-novo-Assembler-

Does higher coverage means better assembly?

For bacterium with a small genome size (2.36Mb) then there is **no advantage is sequencing to a higher depth than 27x coverage**. As the number of contigs >1kb in length does not decrease and the number of complete genes and % of genome that is mapped does not continue increasing with greater coverage.



Increasing the amount of sequene will produce a better assembly, but only upto a certain point.

N50 -The length for which all contigs of that length or longer contains at least half of the total of the lengths of the contigs

Need to resolve errors (Illumina 1-2% error rate)
The more coverage there is, the more errors there are.

GCGTCAGGTAG**C**AGACCACCGCCATGGCGACGATG

GCGTCAGGTAGGAGACCACCG**T**CATGGCGACGATG

GCGT**T**AGGTAGGAGACCACCGCCATGGCGACGATG

GCGTCAGGTAGGAGACC**G**CCGCCATGGCGACGATG

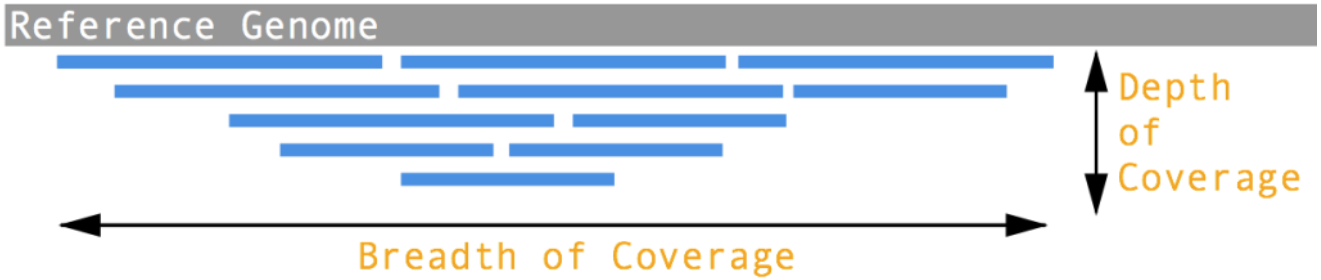
Memory usage ≈ “real” variation + number of errors
Number of errors ≈ size of dataset

https://blogs.warwick.ac.uk/microbialunderground/entry/what_coverage_is/

Digital normalisation

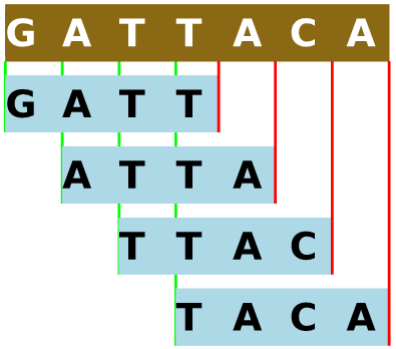
Removing redundant reads -> genome coverage information

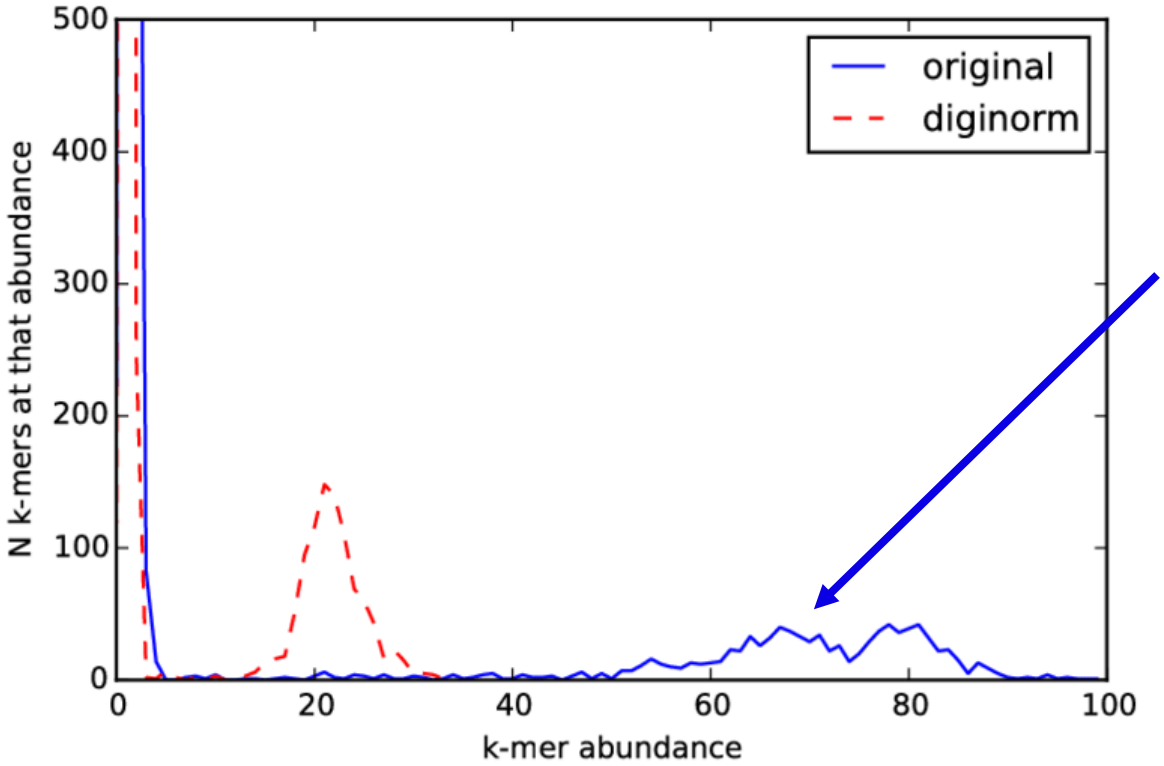
- mapping reads to assembly (traditional approach - chicken egg-problem)



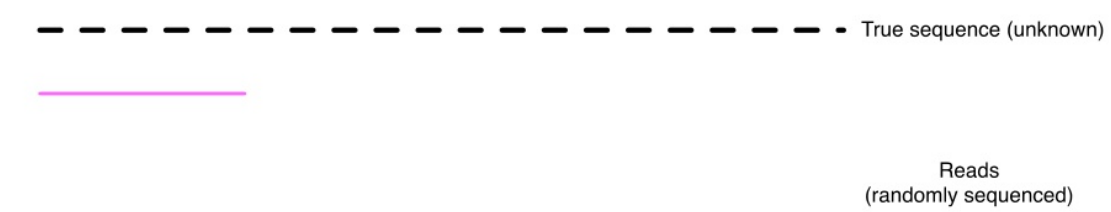
k-mer distribution within the reads
median k-mer abundance

k-mer typically refers to all the possible substrings of length k that are contained in a string





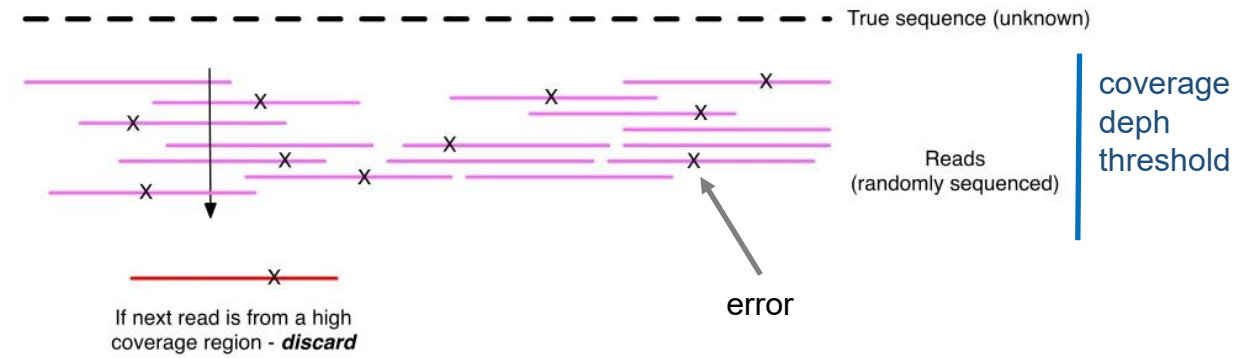
Digital normalisation



median k-mer abundance
is not changed for 1 substitution error when:

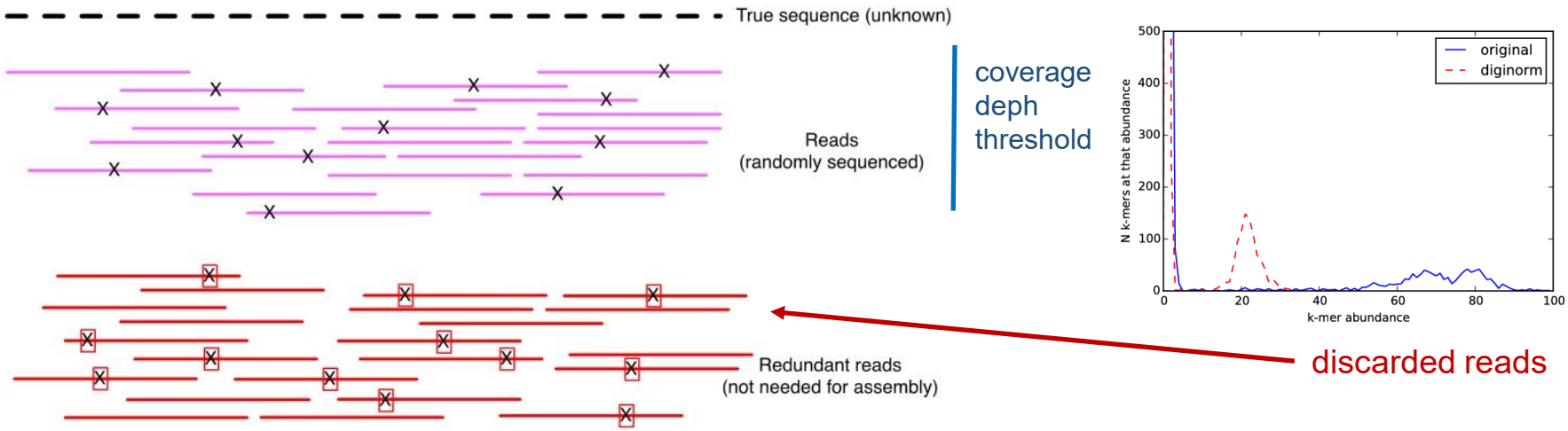
sequence length > 3k-1

(in agreement with Illumina error rate)




```
for read in dataset:
  if estimated_coverage(read) < C:
     accept(read)
  else:
     discard(read)
```

It discards redundant reads and their errors…



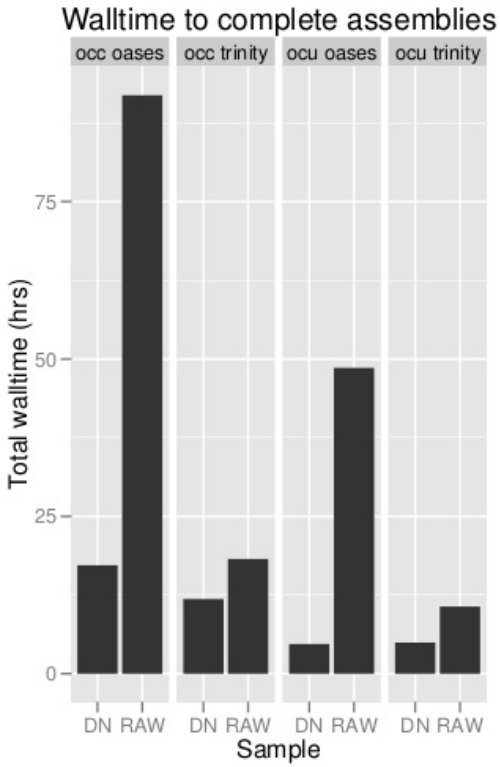
Brown et al. (2012) A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data. arXiv:1203.4802

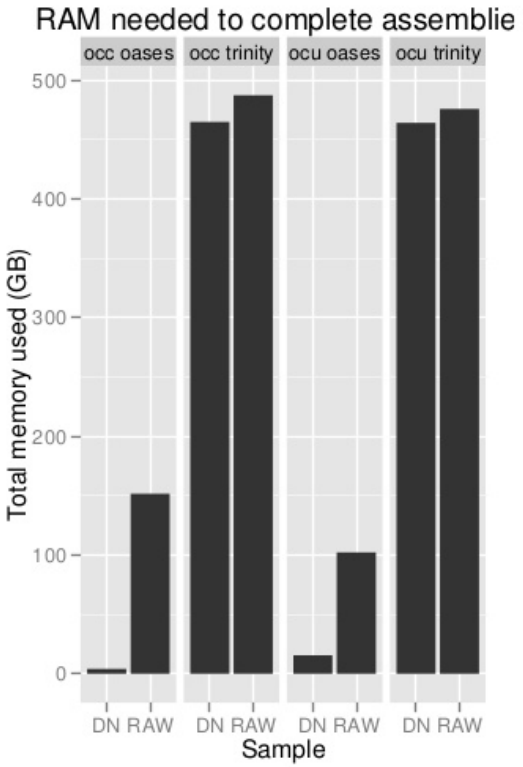
Speed up/efficiency?

Reducing RAM requirements (in some cases) is the only solution how to finish the assembly

Our metagenome - 200 GB of raw data
Our cluster RAM capacity - 256 GB

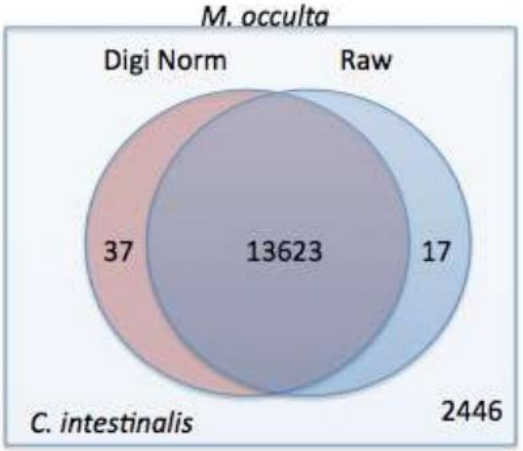
After DN 200 GB of raw data were reduced to 30 GB…allowing assembly with Megahit assembler…



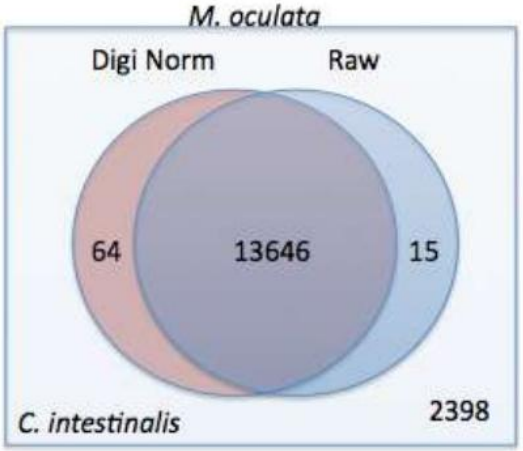


Diginorm increases sensitivity (very slightly :)

Evaluation by homology against a reference genome.



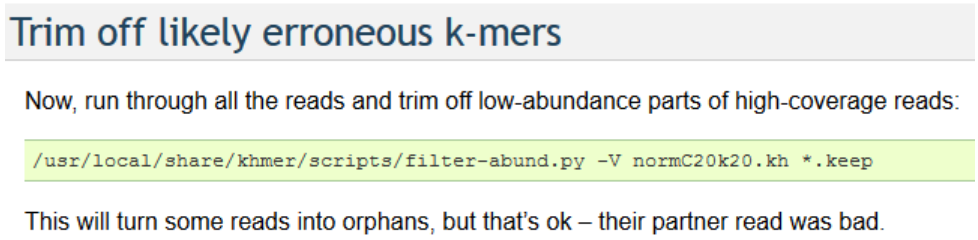
37 extra from diginorm, vs 17 lost;



64 extra from diginorm, vs 15 lost;

Discussion and Criticism…

Removing also some real k-mers… (single pass vs three pass?)



Trim off likely erroneous k-mers

Now, run through all the reads and trim off low-abundance parts of high-coverage reads:

```
/usr/local/share/khmer/scripts/filter-abund.py -V normC20k20.kh *.keep
```

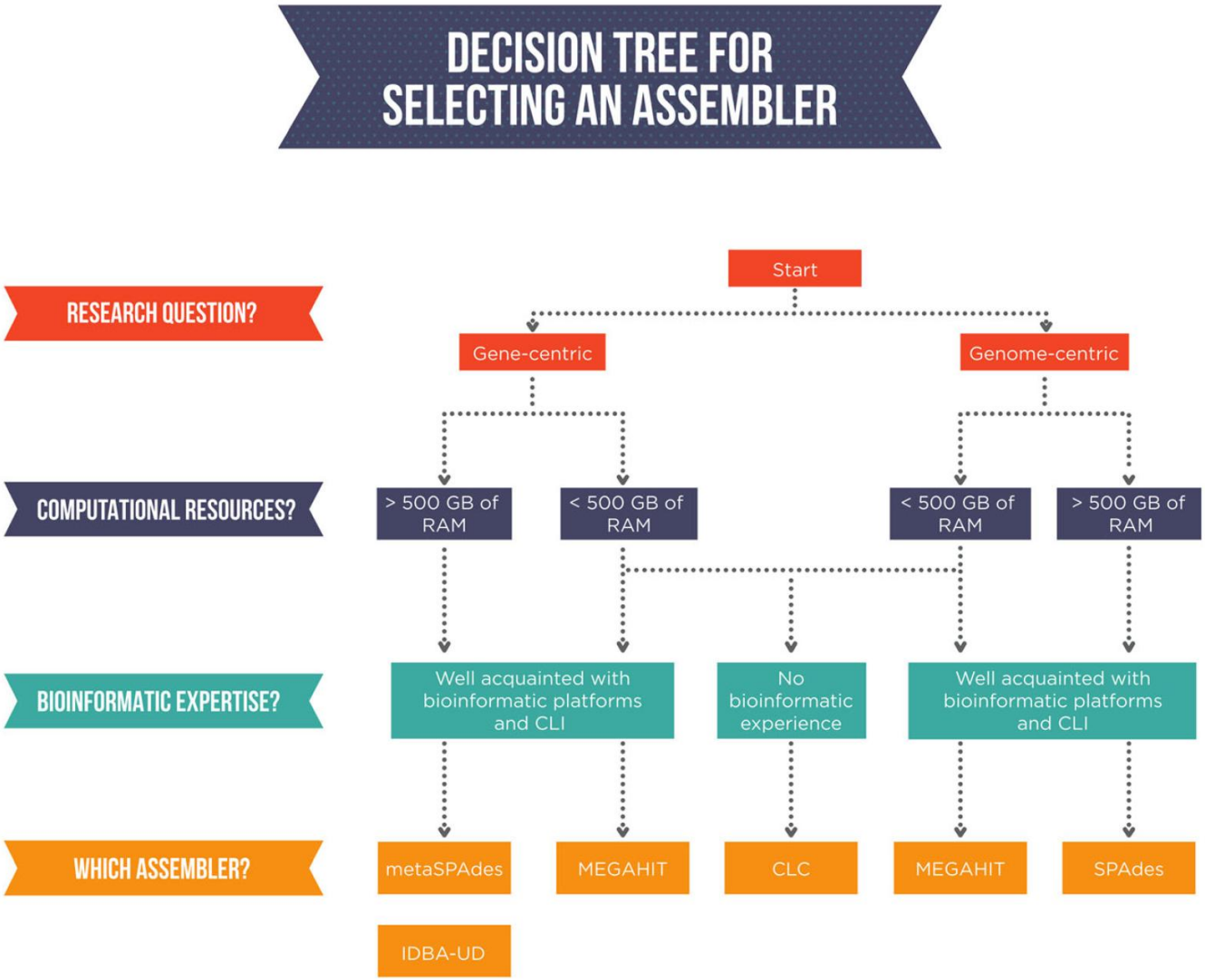
This will turn some reads into orphans, but that's ok – their partner read was bad.

Breaks assembly graph at high-copy repeats…

Working well with some assemblers based on De Bruijn graphs
Velvet (single genomes)
Trinity (meta-transcriptomes)
Megahit (meta-genomes)
x
Problems with assemblers using coverage information from raw data…

Metagenome assembly

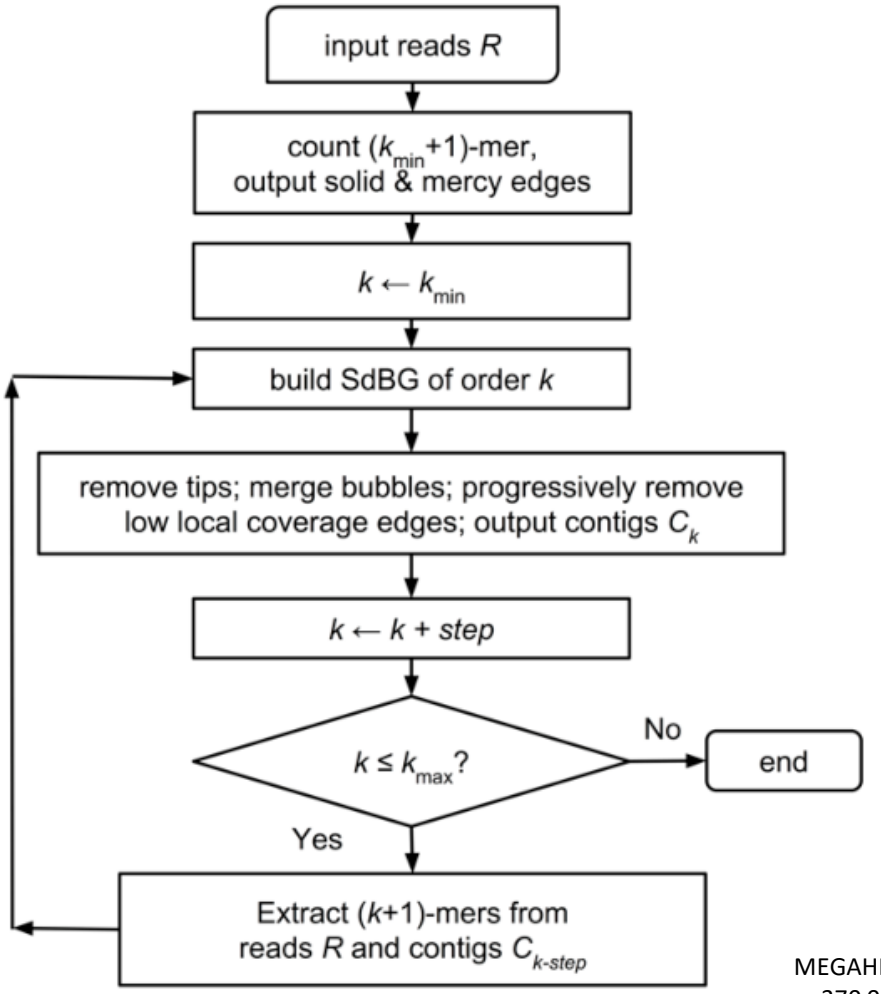
Proposed workflow to select a metagenome assembler based on the research question, the computational resources available and the bioinformatic expertise of the researcher.

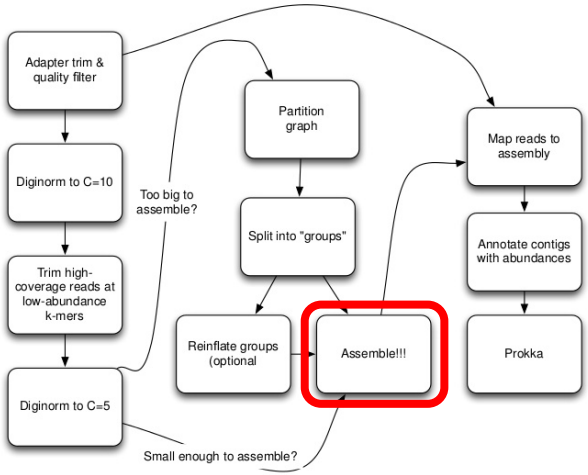


Van der Walt et al. (2017) Assembling metagenomes, one community at a time. BMC Genomics 18:521

3. Assembly - Megahit

The workflow of MEGAHIT



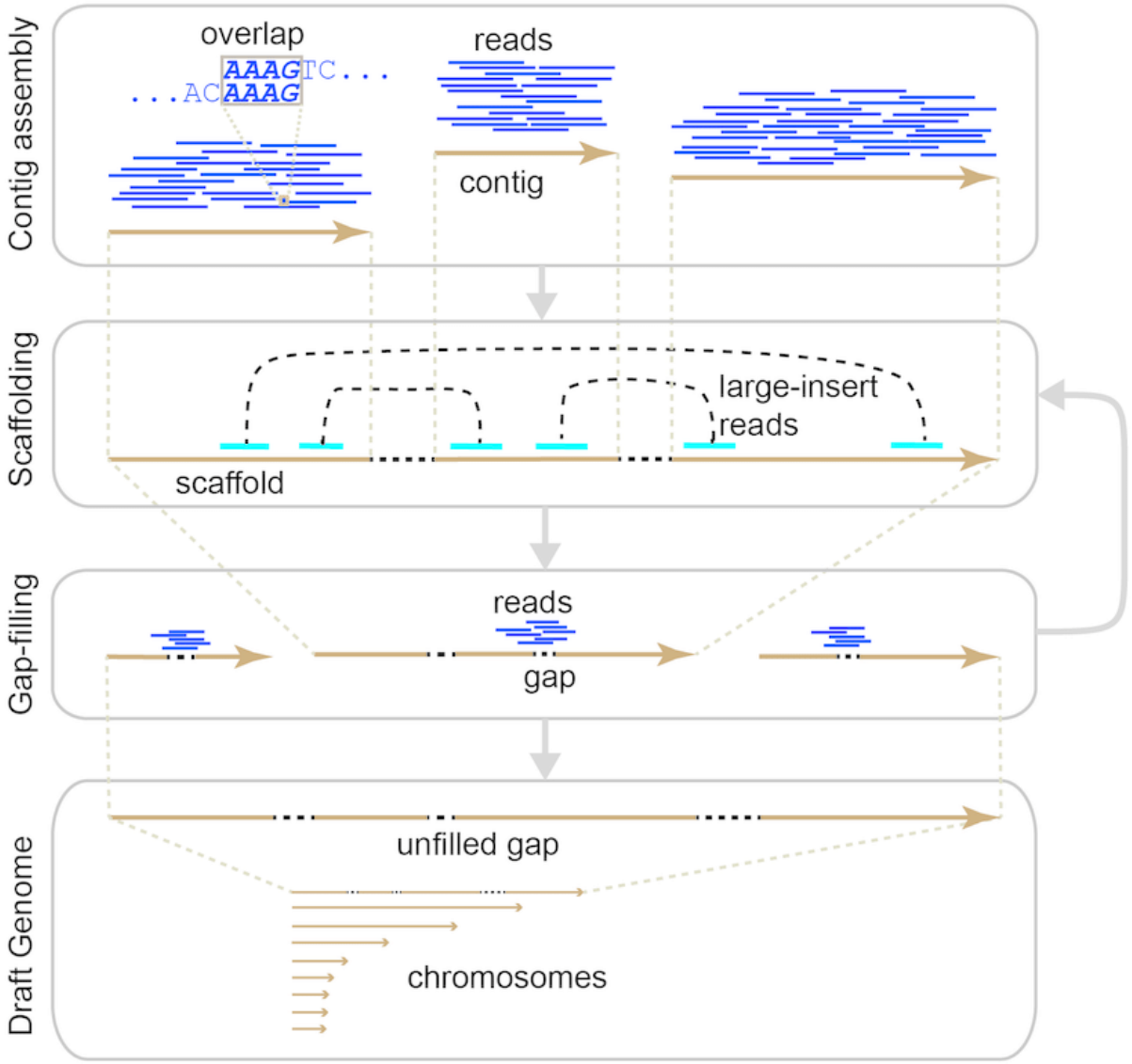


Summary statistics for MEGAHIT, Howe *et al.* and Minia

	MEGAHIT	Howe et al.	Minia
Wall time (h)	44.1	>488	331.4
Peak memory (GB)	345	287	29
Total size (Mbp)	4902	1503	1490
Average length (bp)	633	485	505
N50 (bp)	657	471	488
Longest (bp)	184 210	9397	32 679
# of contigs	7 749 211	3 096 464	2 951 575
# of contigs ≥ 1kbp	841 257	129 513	158 402

MEGAHIT utilizes all 24 CPU threads with options '--k-min 27 --k-max 87 --k-step 10 -m 370 000 000 000'. The wall time for CPU version of MEGAHIT is 99.4 h. Minia does not support multi-threads; it was run with k = 31 and min_abundance = 2. The time and memory of Howe et al. were excerpted from the paper; the time accounts for digital normalization and partitioning only.

Genome assembly



General workflow of the *de novo* assembly of a whole genome

By overlapping reads, contigs are assembled from short reads before scaffolding by large-insert reads, and the remaining gaps are filled.

The scaffolding and gap-filling steps can be iteratively performed until no contigs are scaffolded or no additional gaps are resolved before completion.

Through this procedure, a draft genome consisting of chromosomes is built. Some unfilled gaps may remain in the draft genome.

Sohn & Nam (2016) The present and future of de novo whole-genome assembly. Briefings in Bioinformatics 19(1):bbw096

How to obtain genomes of organisms?

Genome sequencing --> Genome assembly --> Genome annotation

DNA copies of the genome

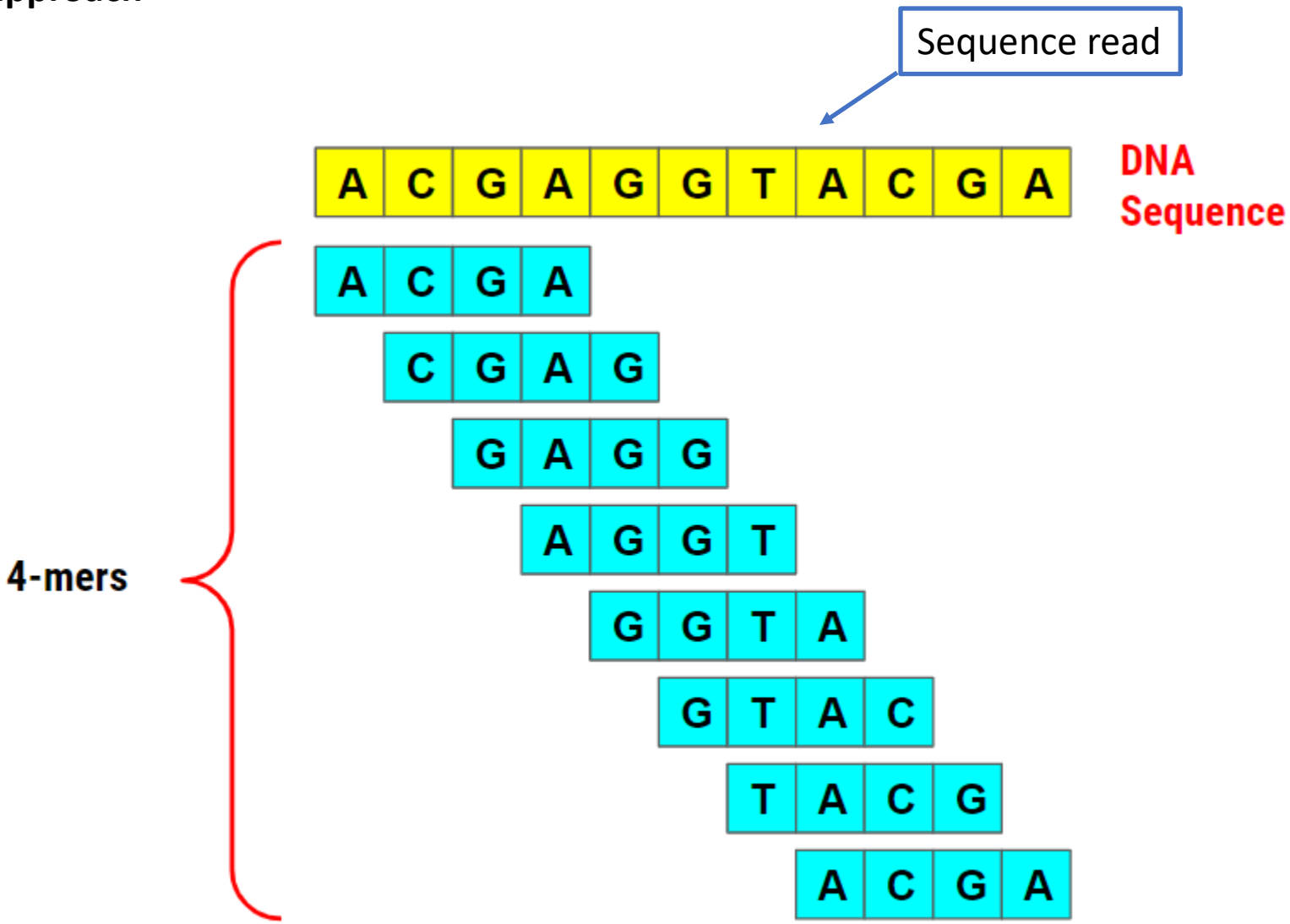
Sequence Reads

Assembled genome

typically for Illumina type short read sequencing, reads of length 150 - 300 bp are produced.

Genome assembly

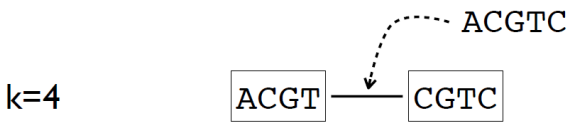
k-mer approach

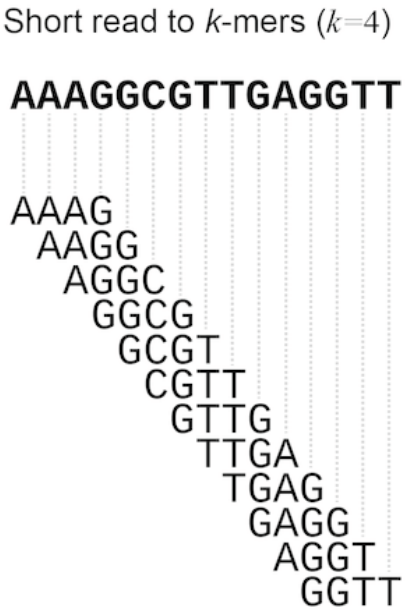


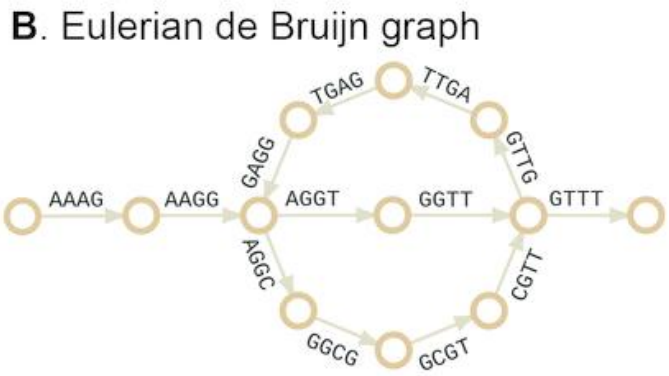
Genome assembly - de Bruijn graph

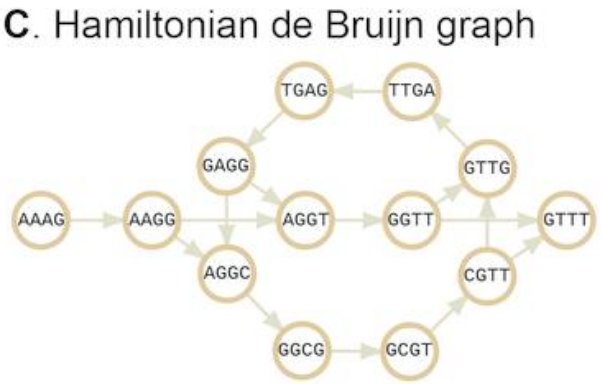
In the de Bruijn graph approach, short reads are split into short k-mers before the de Bruijn graphs are built. In **graph theory**, an *n*-dimensional **De Bruijn graph** of *m* symbols is a **directed graph** representing overlaps between sequences of symbols.

- de Bruijn graph
  - k-dimensional graph over four symbols {A, C, G, T}
  - vertex: k-mer -- a string of k nucleotides
  - edge: (k+1)-mer









In the Hamiltonian approach, the k-mers (or sequences) are the nodes, whereas they are the edges in the Eulerian approach. The k-mers are connected to neighbors by overlapping prefix and suffix (k-1)-mers.

Sohn & Nam (2016) The present and future of de novo whole-genome assembly. Briefings in Bioinformatics 19(1):bbw096

Genome assembly - de Bruijn graph

simple example

reads:

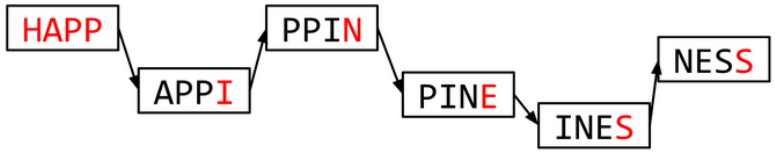
HAPPI PINE INESS APPIN

k = 4 k-mers:

HAPP APPI

PINE PPIN

INES NESS



HAPPINESS

lower k

- more connections
- less chance of resolving small repeats
- higher k-mer coverage

higher k

- less connections
- more chance of resolving small repeats
- lower k-mer coverage

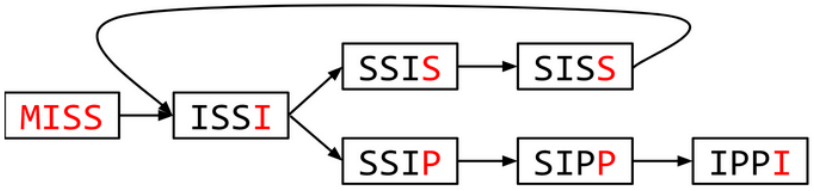
problem of repeats

reads:

MISSIS SSISSI SSIPPI

All 4-mers:

MISS ISSI SSIS SISS SSIP SIPP IPPI



MISSISSIPPI or MISSISSISSISSIPPI or ...

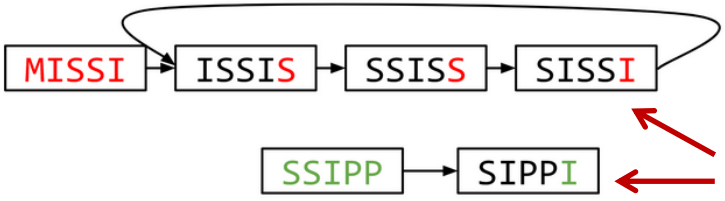
higher k-mer length

reads:

MISSIS SSISSI SSIPPI

This time k = 5 k-mers:

MISSI ISSIS SSISS SISSI SSIPP SIPPI



No connection between these two nodes!

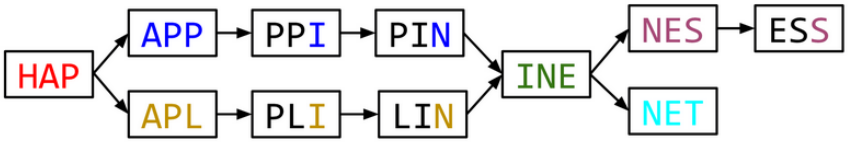
MISSISSIS SSIPPI

Genome assembly - de Bruijn graph

HAPPI INESS APLIN PINET

k = 3 k-mers:

HAP APP PPI INE NES ESS APL PLI LIN PIN NET



6 contigs: HAP APPIN APLIN INE NESS NET

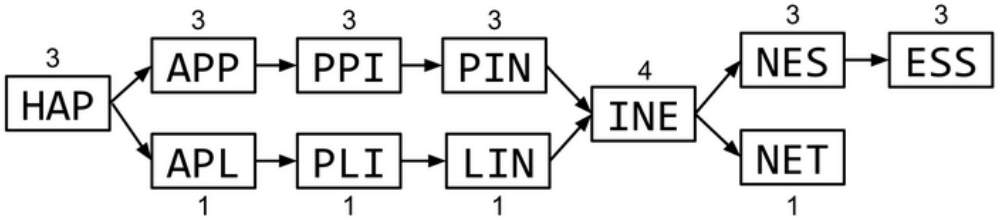
read errors

- randon distribution of errors
- we can count the frequency of each k-mer
- use frequency data to clean de Bruijn graph

HAPPI INESS APLIN PINET
HAPPI INESS
HAPPI INESS

k = 3 k-mers:

HAPx3 APPx3 PPIx3 INEx4 NESx3 ESSx3 APLx1 PLIx1
LINx1 PINx1 NETx1

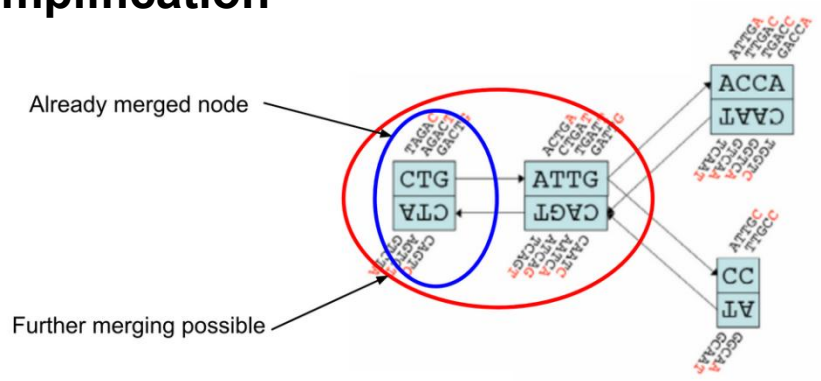


Coverage cutoff

Genome assembly - de Bruijn graph simplification

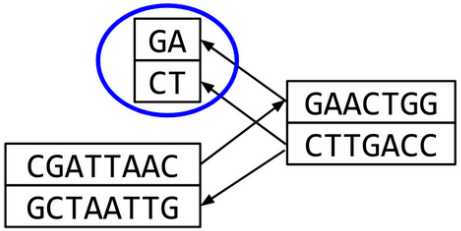
Chain merging

- when there are two connected nodes without a divergence, merg the two nodes



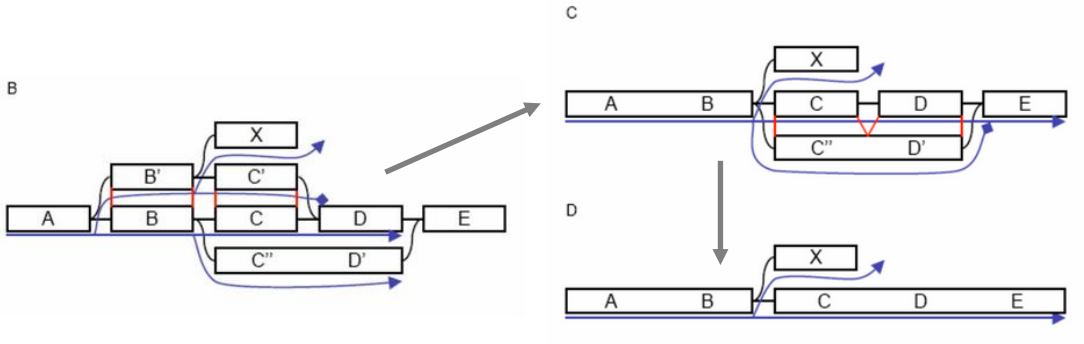
Tip clipping

- clip tips if the length of the tip is < 2 x k



Bubble collapsing

- detect redundant path through graph
- compare the paths using sequence alignment
- if similar, merge the paths



Remove low coverage nodes

- remove errorneous nodes and connections using the „coverage cutoff“
- genuine short nodes will have high coverage

Zerbino & Birney (2008)

Evaluating an assembly

Assembly	final.contigs.fa
# contigs (>= 0 bp)	2670
# contigs (>= 1000 bp)	1021
# contigs (>= 5000 bp)	455
# contigs (>= 10000 bp)	265
# contigs (>= 25000 bp)	117
# contigs (>= 50000 bp)	40
Total length (>= 0 bp)	12958914
Total length (>= 1000 bp)	12177316
Total length (>= 5000 bp)	10827408
Total length (>= 10000 bp)	9482694
Total length (>= 25000 bp)	7172558
Total length (>= 50000 bp)	4430125
# contigs	1537
Largest contig	282362
Total length	12541412
GC (%)	39.44
N50	32048
N75	10401
L50	86
L75	258
# N's per 100 kbp	0.00

N50

N50 statistic defines assembly quality in terms of contiguity. Given a set of contigs, the N50 is defined as the sequence **length of the shortest contig at 50% of the total genome length**.

Example: 1 Mbp genome

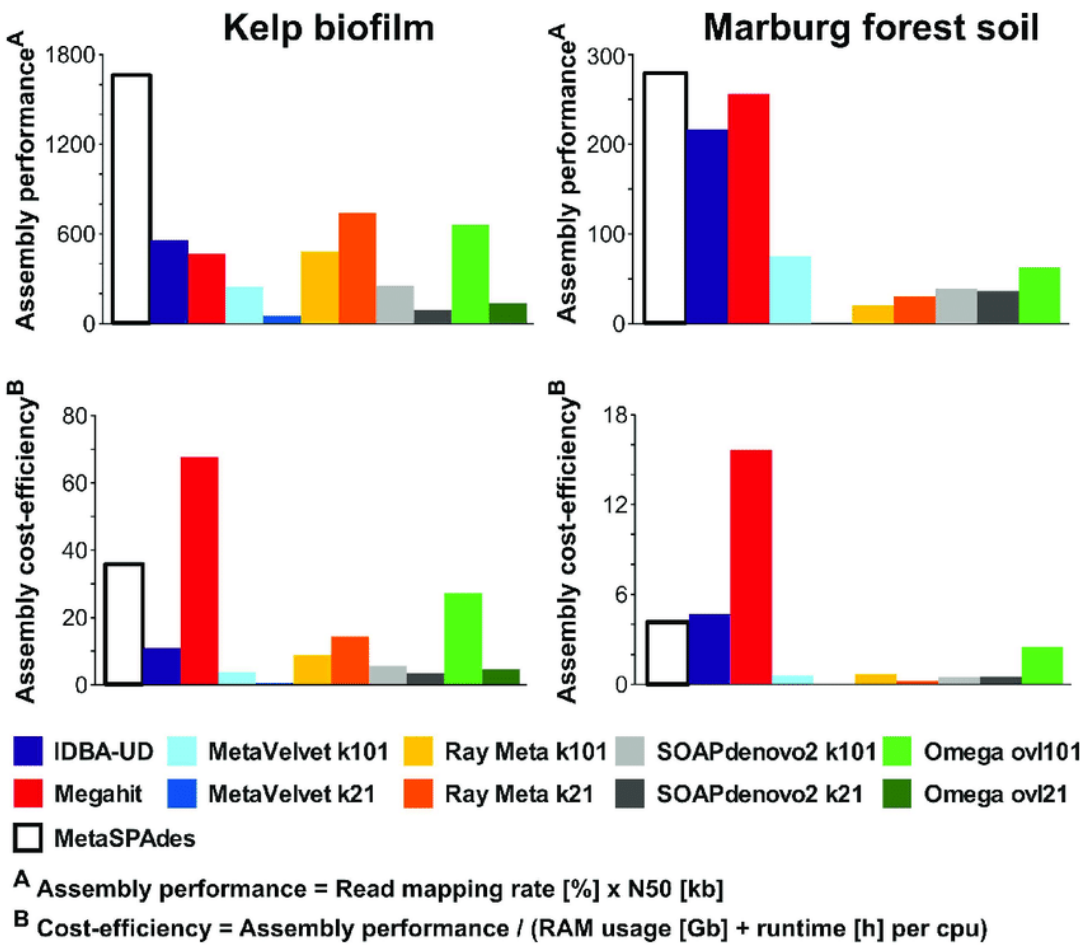
N50 size = 30 kbp
(300k+100k+45k+45k+30k = 520k >= 500kbp)

Note:
N50 values are only meaningful to compare when base genome size is the same in all cases

L50

Given a set of contigs, each with its own length, the L50 count is defined as the **smallest number of contigs whose length sum makes up half of genome size**. From the example above the L50=5.

PLOS ONE



RESEARCH ARTICLE

Comparing and Evaluating Metagenome Assembly Tools from a Microbiologist's Perspective - Not Only Size Matters!

John Vollmers, Sandra Wiegand, Anne-Kristin Kaster*

Leibniz Institute DSMZ - German Collection of Microorganisms and Cell Cultures, Braunschweig, Germany

* a.kaster@dsmz.de

Generally, metaSPAdes, IDBA-UD and MegaHit show the highest sensitivity and best genome recovery rates with more than 50% of each reference genome being reconstructed already 3x read coverage. Almost complete genome reconstruction was achieved at 6x coverage…since Megahit showed more favorable assembly cost-efficiencies than IDBA-UD and has all options clearly documented, it can be highly recommended.

Comparison of assembly performance and cost-efficiency. Kelp biofilm (KBF) assemblies are shown on the left and Marburg forest soil (MFS) assemblies on the right. Assemblers utilizing single k-mer lengths were tested with two different values for k, 21 and 101. Assembly performance was defined as the product of the respective read mapping rate (representing information content) and the respective N50. Cost efficiency was defined as the quotient of assembly performance and the sum of RAM usage and runtime per CPU, required for the respective assemblies.

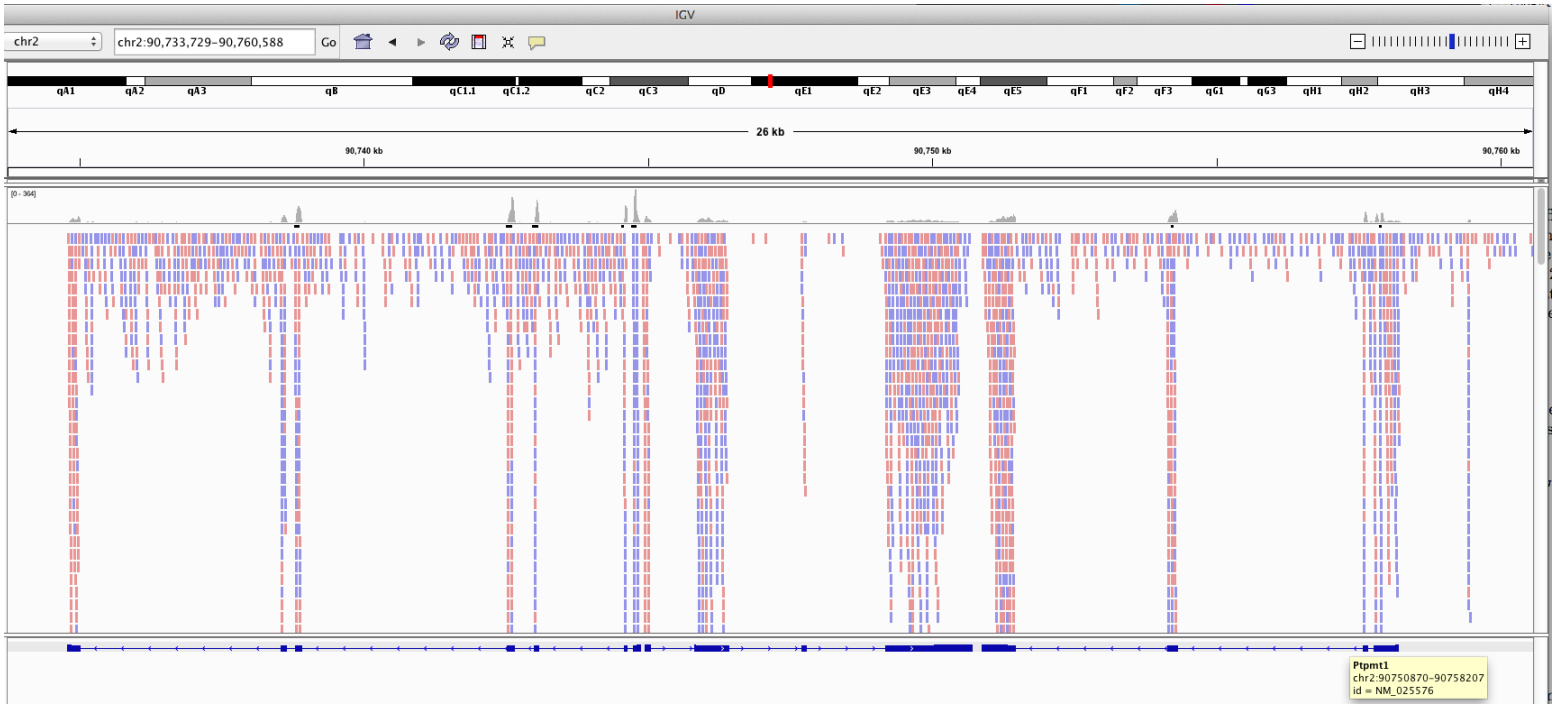
4. Map reads to assembly

Bowtie 2
Fast and sensitive read alignment

JOHNS HOPKINS UNIVERSITY

Mapping reads and quantifying genes

… Finally we have got the number of genes and annotations in the (assembled) contigs. Because these annotations are predicted from assembled reads we have lost the quantitatve information for the annotations. So to actually **quantify** the genes, we will map the input reads back to the assembly.



Mapping reads to assembly (Bowtie2)

Bowtie 2 is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences. It is particularly good at aligning reads of about 50 up to 100s or 1,000s of characters to relatively long (e.g. mammalian) genomes. Bowtie 2 indexes the genome with an FM Index (based on the Burrows-Wheeler Transform or BWT) to keep its memory footprint small.

Sequence Alignment Map (**SAM**) is a text-based format for storing biological sequences aligned to a reference sequence. The header section must be prior to the alignment section if it is present. Headings begin with the '@' symbol, which distinguishes them from the alignment section. Alignment sections have 11 mandatory fields, as well as a variable number of optional fields.

The binary representation of a SAM file is a **Binary Alignment Map** (**BAM**) file, which is a compressed SAM file.

Col	Field	Type	Brief Description
1	QNAME	String	Query template NAME
2	FLAG	Int	bitwise FLAG
3	RNAME	String	References sequence NAME
4	POS	Int	1- based leftmost mapping POSition
5	MAPQ	Int	MAPping Quality
6	CIGAR	String	CIGAR String
7	RNEXT	String	Ref. name of the mate/next read
8	PNEXT	Int	Position of the mate/next read
9	TLEN	Int	observed Template LENgth
10	SEQ	String	segment SEQuence
11	QUAL	String	ASCII of Phred-scaled base QUALity+33

http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml

Mapping reads to assembly

After mapping the reads back to contigs obtained "mapping table" needs to be normalized:

1. by contig and read length
2. by sample sizes

	length [bp]	Sample01_R1	Sample01_R2	Sample02_R1	Sample02_R2	Sample03_R1	Sample03_R2	Sample04_R1	Sample04_R2
contig1	**3205**	326	365	20	15	800	784	225	234
contig2	**564**	2	3	566	554	100	124	0	0
contig3	**789**	454	563	0	0	0	0	874	865

Normalization
by contig length

x = contig
y = sample
readLength = 250 bp

Value[x,y] = (raw_Sequence_Counts[x,y] * readLength) / contigLength[x]

	Sample01_R1	Sample01_R2	Sample02_R1	Sample02_R2	Sample03_R1	Sample03_R2	Sample04_R1	Sample04_R2
contig1	25.43	28.47	1.56	1.17	62.40	61.15	17.55	18.25
contig2	0.89	1.33	250.89	245.57	44.33	54.96	0.00	0.00
contig3	143.85	178.39	0.00	0.00	0.00	0.00	276.93	274.08

Then later you should normalize your table by column (sample) to make samples comparable...

METAGENOMIC APPROACHES

ENVIRONMENTAL SAMPLE

EXTRACTED DNA

ISOLATES

Targeted Gene Sequencing

Shotgun Metagenome Sequencing

Single Cell Genome Sequencing

known genes
identity, diversity, distribution, dynamics

"all" genes
functional potential metabolic reconstruction

genomes
linking phylogeny & function, genome diversification

binning to *metagenome* assembled *genomes* (*MAGs*)

Gene-centric

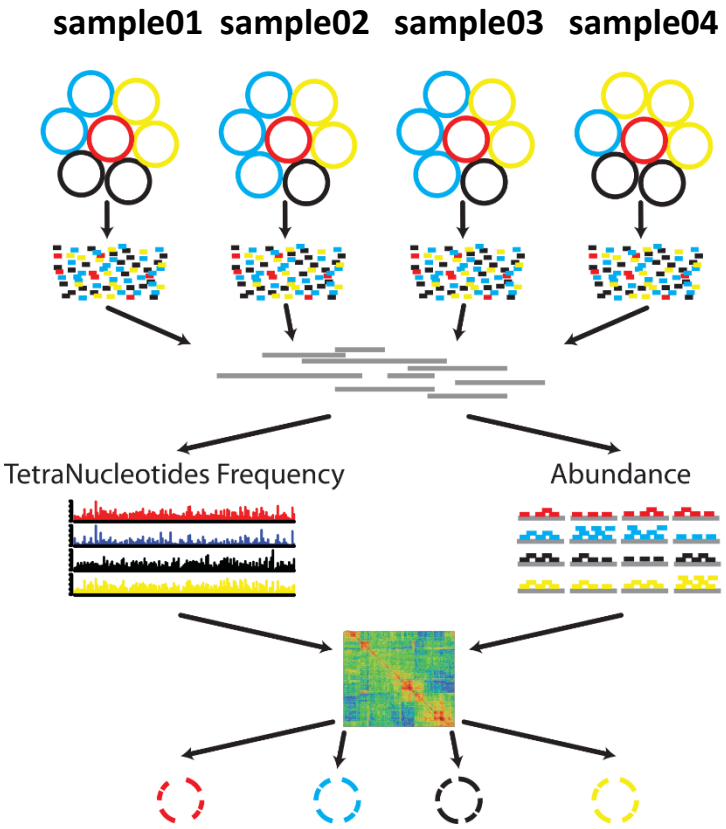
Genome-centric

Binning
Gouping nucleotide sequences belonging to individual/similar organism/s

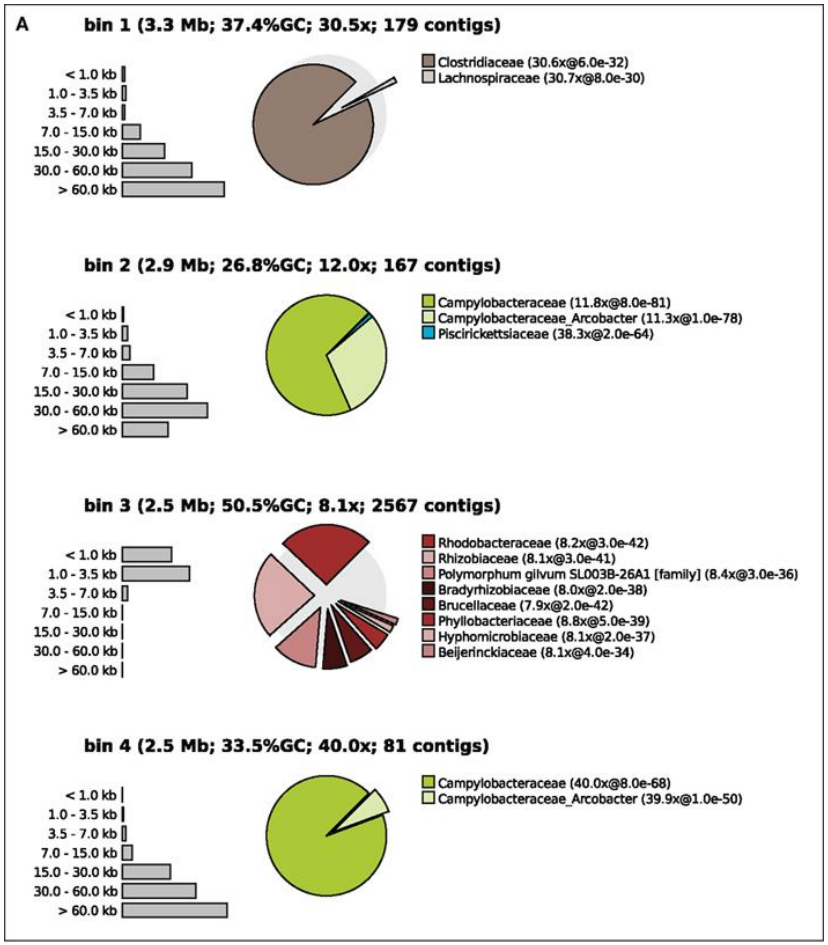
MetaBAT
MaxBin2

MetaBAT

4. Calculate TNF for each contig
5. Calculate Abundance per library for each contig
6. Calculate the pairwise distance matrix using pre-trained probabilistic models
7. Forming genome bins iteratively



Contig size distribution, sequencing coverage and taxonomic distribution of the four largest bins of sample 7 binned at medium confidence

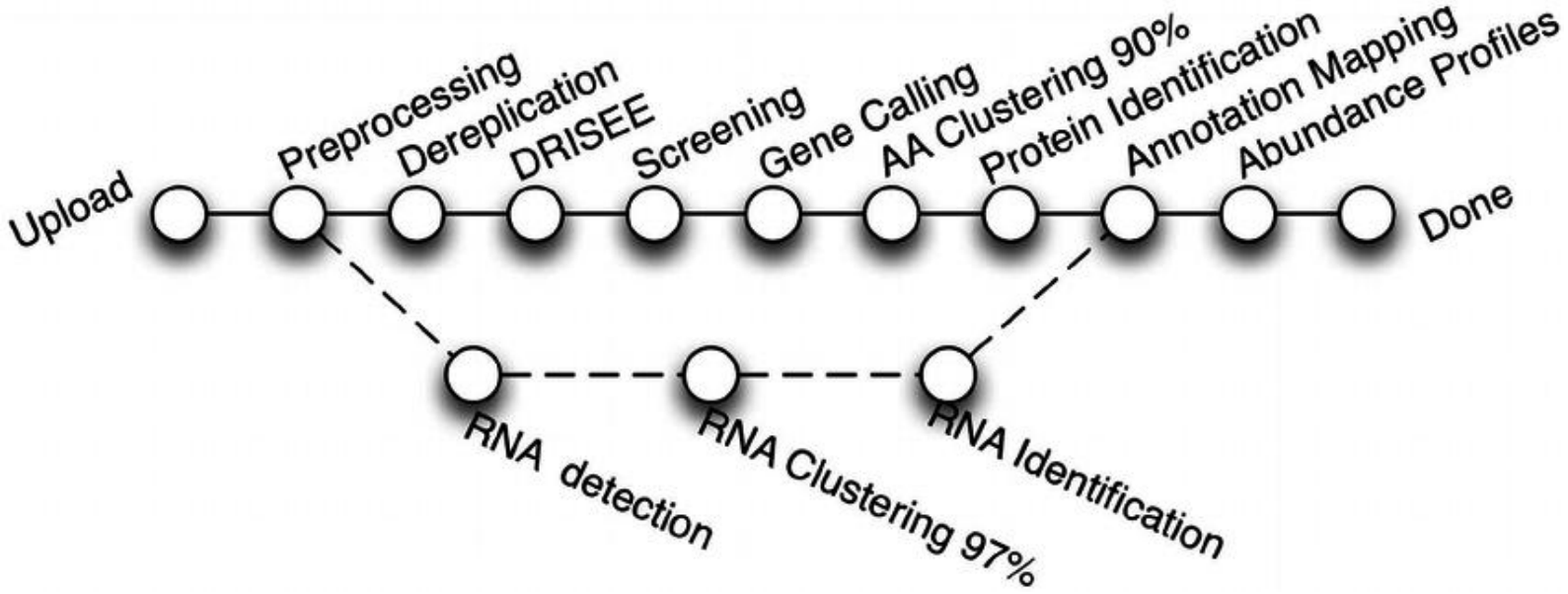


The exploded pies show the taxonomic distribution of the bins. The distance of each part from the center of the pie is a measure for the median e-value of the associated hits (the larger the e-value the larger the distance from the center). Coverage is shown for the bin as a whole and separately for each pie part.

Contigs annotation (MG-RAST)

The metagenomics RAST server is a SEED-based environment that allows users to upload metagenomes for automated analyses. The server is built as a modified version of the (Rapid Annotation using Subsystem Technology (RAST) server. The RAST technology was originally implemented to allow automated high-quality annotation of complete or draft microbial genomes using SEED data and has been adapted for metagenome analysis. The server provides the annotation of sequence fragments, their phylogenetic classification, functional classification of samples, and comparison between multiple metagenomes. The server also computes an initial metabolic reconstruction for the metagenome and allows comparison of metabolic reconstructions of metagenomes and genomes.

Different stages of the MG-RAST automated pipeline. In the annotation mapping stage, functions and taxonomic units from the M5nr are mapped to the MD5 identifiers found in the similarity search.



Keegan et al. (2016) MG-RAST, a Metagenomics Service for Analysis of Microbial Community Structure and Function. Methods Mol Biol. 2016;1399:207-33.

MG-RAST
metagenomics analysis server

upload › submit › progress

upload data

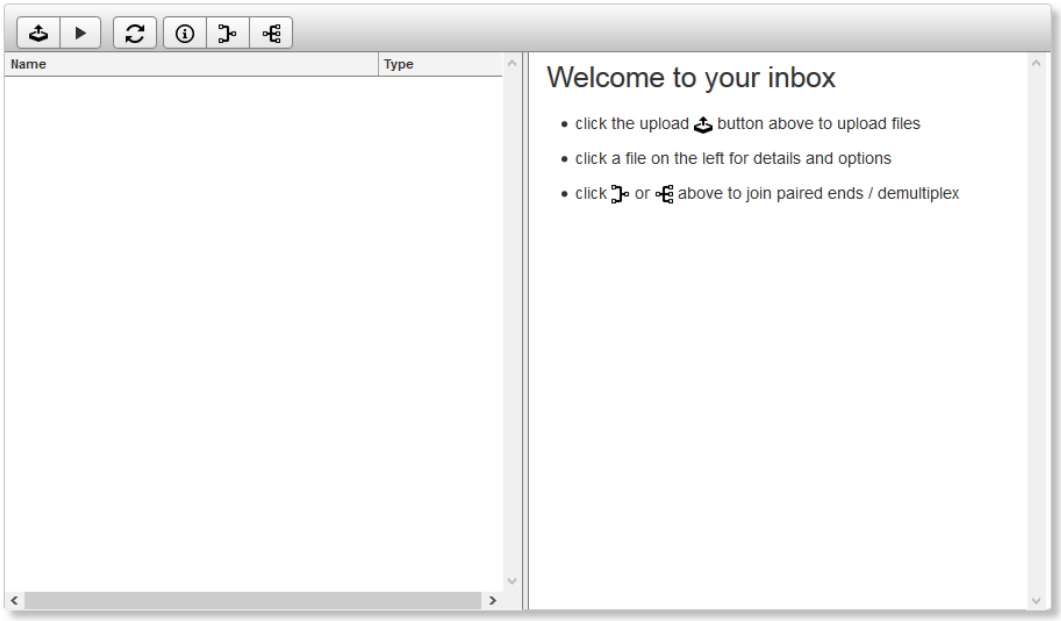
webkey | upload

Data submission is a two-step process. As the **first step**, data is uploaded into your private inbox on the MG-RAST server. This area is write only and accessible only to you. From the inbox data can then be submitted. Use the upload function, or use our API to upload your data. To view your webkey required for using the API, click here.

metadata

MetaZen tool | Excel template

Submission of multiple files, sharing of data, or data publication requires metadata which you can provide via an Excel template or our Metazen helper tool.

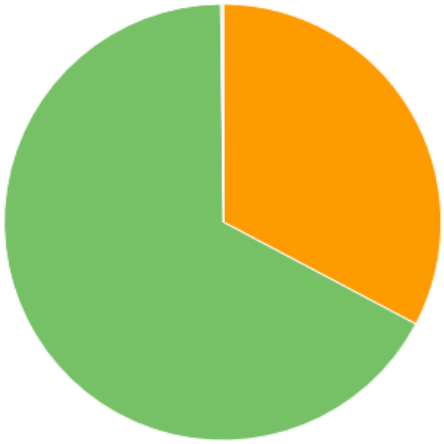


next

Contigs annotation (MG-RAST)

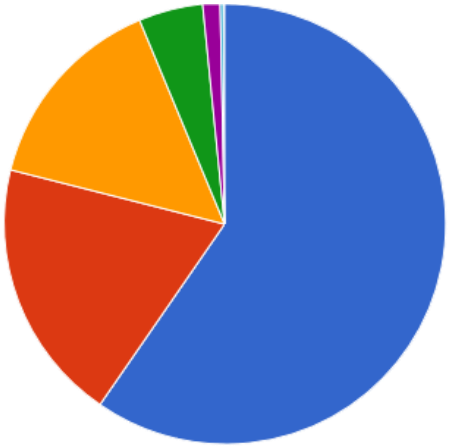
metagenomes

MG-RAST ID	name	bp count	seq. count
e685242d976d676d34373635303536 2e33	final.contigs	5,672,583,505	9,168,354



Predicted Features

- unknown protein - 2,749,165 (32.77%)
- **annotated protein - 5,625,518 (67.05%)**
- ribosomal RNA - 15,297 (0.18%)

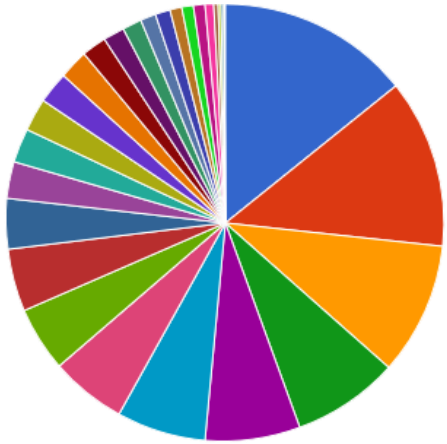


KO

- Metabolism - 1,770,361 (59.56%)
- Genetic Information Processing - 574,612 (19.33%)
- Environmental Information Processing - 440,029 (14.80%
- Cellular Processes - 142,392 (4.79%)
- Human Diseases - 36,044 (1.21%)
- Organismal Systems - 9,167 (0.31%)

Functional Category Hits Distribution

The pie charts below illustrate the distribution of functional categories for COGs, KOs, NOGs, and Subsystems at the highest level supported by these functional hierarchies. Each slice indicates the percentage of reads with predicted protein functions annotated to the category for the given source.



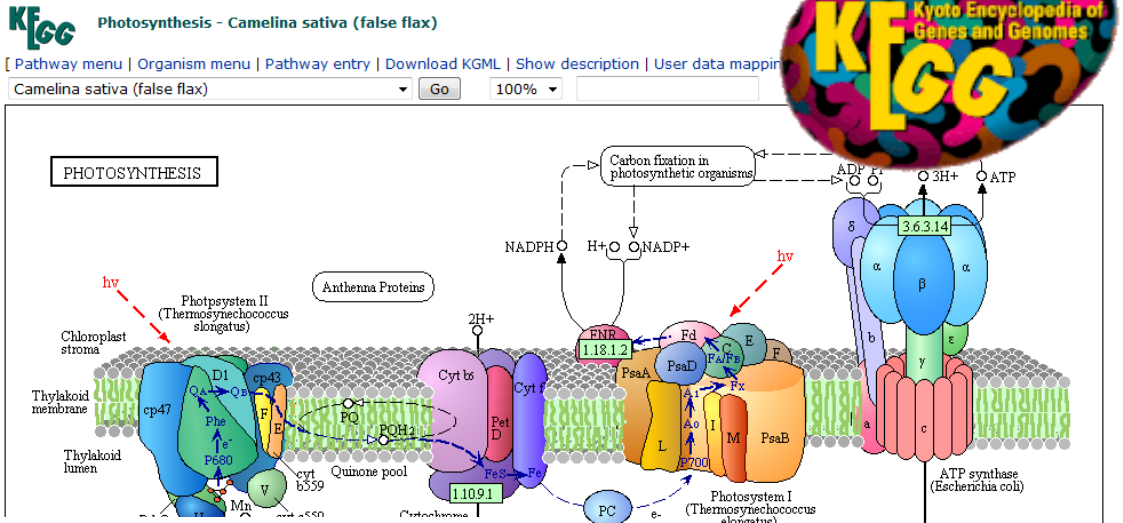
Subsystems

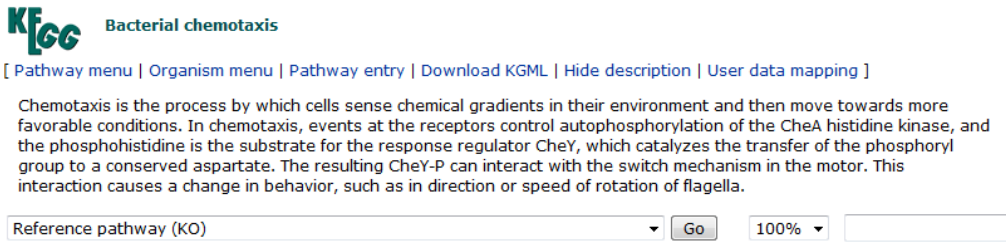
- Clustering-based subsystems - 1,218,346 (14.34%)
- Carbohydrates - 1,058,166 (12.45%)
- Amino Acids and Derivatives - 830,930 (9.78%)
- Miscellaneous - 665,022 (7.83%)
- Protein Metabolism - 598,791 (7.05%)
- Cofactors, Vitamins, Prosthetic Groups, Pigments - 555,4
- RNA Metabolism - 477,521 (5.62%)
- DNA Metabolism - 405,532 (4.77%)
- Cell Wall and Capsule - 399,674 (4.70%)
- Fatty Acids, Lipids, and Isoprenoids - 311,782 (3.67%)
- Virulence, Disease and Defense - 233,845 (2.75%)
- Nucleosides and Nucleotides - 208,501 (2.45%)
- Respiration - 207,286 (2.44%)
- Membrane Transport - 201,214 (2.37%)

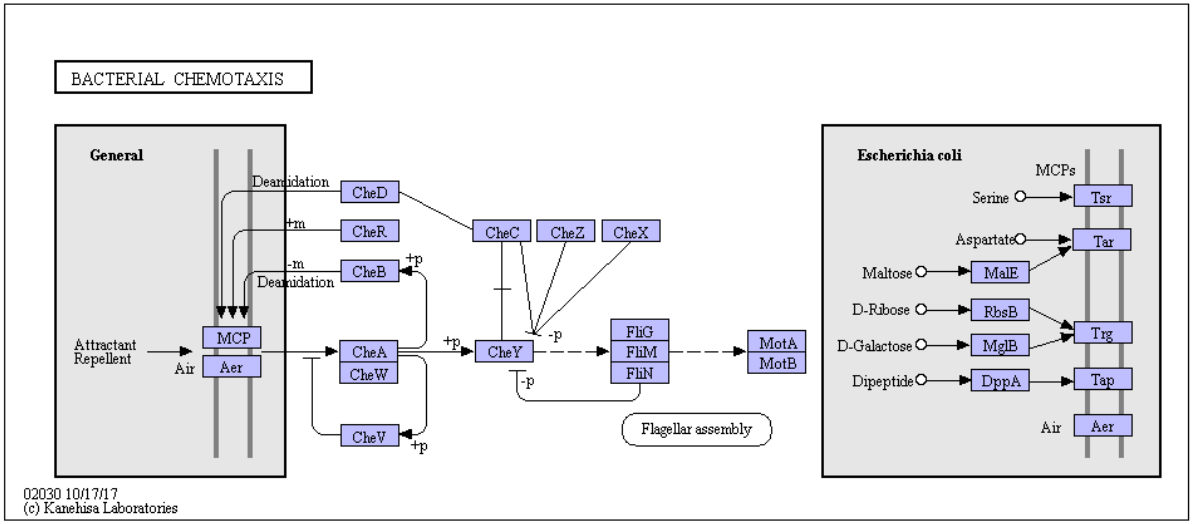
https://www.mg-rast.org/

MG-RAST - KEGG mapper

KEGG (**Kyoto Encyclopedia of Genes and Genomes**) is a collection of databases dealing with genomes, biological pathways, diseases, drugs, and chemical substances.





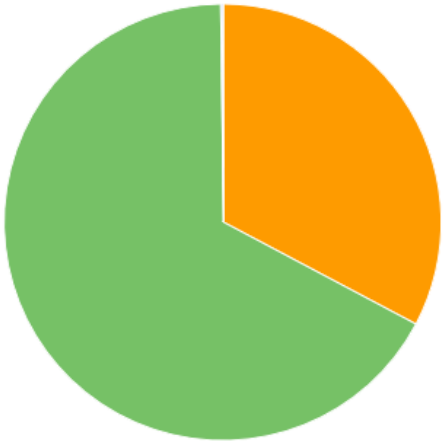


www.genome.jp/kegg

Contigs annotation (MG-RAST)

metagenomes

MG-RAST ID	name	bp count	seq. count
e685242d976d676d343736353033536 2e33	final.contigs	5,672,583,505	9,168,354

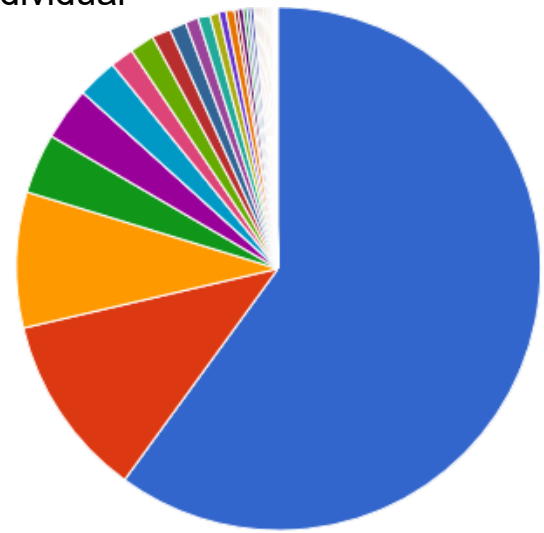


Predicted Features

- unknown protein - 2,749,165 (32.77%)
- **annotated protein - 5,625,518 (67.05%)**
- ribosomal RNA - 15,297 (0.18%)

Taxonomic Hits Distribution

The chart below represent the distribution of taxa using a contigLCA algorithm finding a single consensus taxonomic entity for all features on each individual sequence.



Phylum

- Proteobacteria - 2,064,628 (59.97%)
- Bacteroidetes - 388,798 (11.29%)
- Actinobacteria - 291,128 (8.46%)
- Verrucomicrobia - 125,085 (3.63%)
- Firmicutes - 113,822 (3.31%)
- Planctomycetes - 84,076 (2.44%)
- Chloroflexi - 53,529 (1.55%)
- Acidobacteria - 51,305 (1.49%)
- Gemmatimonadetes - 38,077 (1.11%)
- Cyanobacteria - 37,661 (1.09%)
- unclassified (derived from Bacteria) - 27,351 (0.79%)
- Euryarchaeota - 23,559 (0.68%)
- Chlorobi - 19,090 (0.55%)
- Nitrospirae - 17,301 (0.50%)

https://www.mg-rast.org/

Contigs annotation alternative – Integrated Microbial Genomes

JGI IMG/M INTEGRATED MICROBIAL GENOMES & MICROBIOMES

Quick Genome Search: Go

My Analysis Carts**: 0 Genomes | 0 Scaffolds | 0 Functions | 0 Genes

Home | Find Genomes | Find Genes | Find Functions | Compare Genomes | OMICS | My IMG | Data Marts | Help

IMG Content

Datasets	JGI	All
Bacteria	12511	61273
Archaea	483	1441
Eukarya	370	698
Plasmids	1	1190
Viruses	0	8382
Genome Fragments	0	91
Metagenome	5962	10805
Cell Enrichments	915	915
Single Particle Sorts	2970	2984
Metatranscriptome	2337	2361
Total Datasets		90157

Last Datasets Added On:

Genome 2018-05-02
Metagenome 2018-05-23

Project Map
Metagenome Projects Map
System Requirements
Microbial Genomics & Metagenomics Workshop

Tweets by JGI

The **Integrated Microbial Genomes (IMG)** system serves as a community resource for analysis and annotation of genome and metagenome datasets in a comprehensive comparative context. The **IMG data warehouse** integrates genome and metagenome datasets provided by IMG users with a comprehensive set of publicly available genome and metagenome datasets.

IMG provides users with tools (IMG UI Map) for analyzing publicly available genome datasets and metagenome datasets (Nucleic Acids Research, October 13, 2016).

IMG Statistics **Data Usage Policy**

	Isolates		SAGs		MAGs	
Sequenced at:	JGI	All	JGI	All	JGI	All
Bacteria	6684	53331	1952	3058	3847	4852
Archaea	173	740	230	339	80	362
Eukarya	369	697	0	0	1	1
Viruses	0	8303	0	53	0	0

(Only data sets with GOLD metadata were counted.)

Combined assembly data sets were excluded in the following metagenome and metatranscriptome table statistics.

https://img.jgi.doe.gov/cgi-bin/m/main.cgi