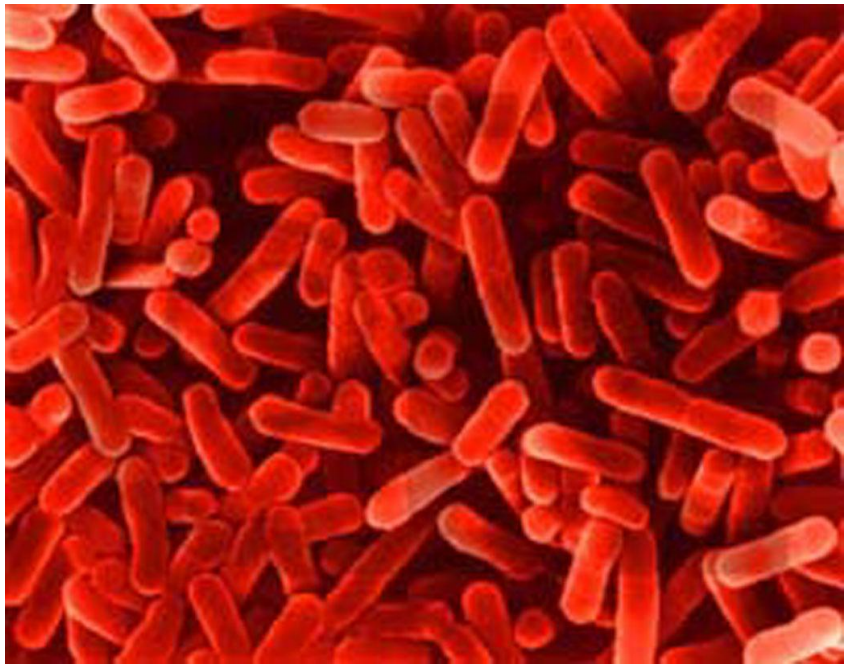
Primer on Metagenomics



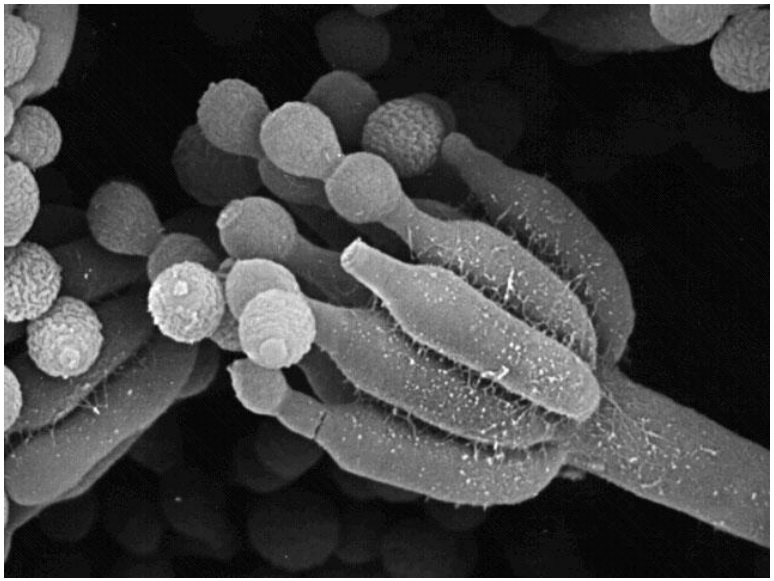
Special Topics in Bioinformatics, SS10

A. Regl, 7055213

Microbes are everywhere!

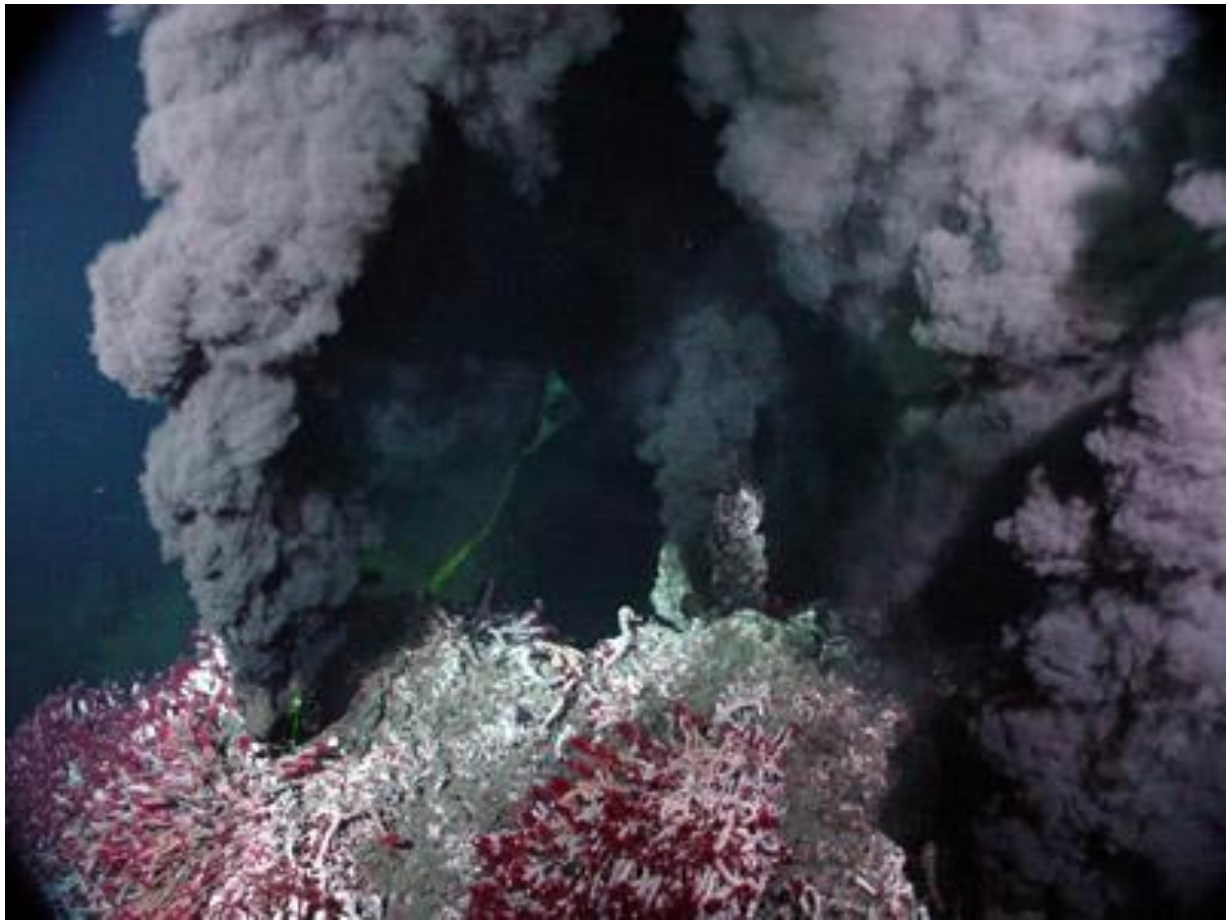
- Estimation: 5 x 10^{30} prokaryotes on our planet
  (= 1.000.000.000.000.000.000.000 times number of humans!)

- One human consists of
  10^{13} somatic cells, and
  10^{14} bacteria
- Astonishing diversity of habitat:
  deep sea vents (340 °C)
  6 km deep boreholes



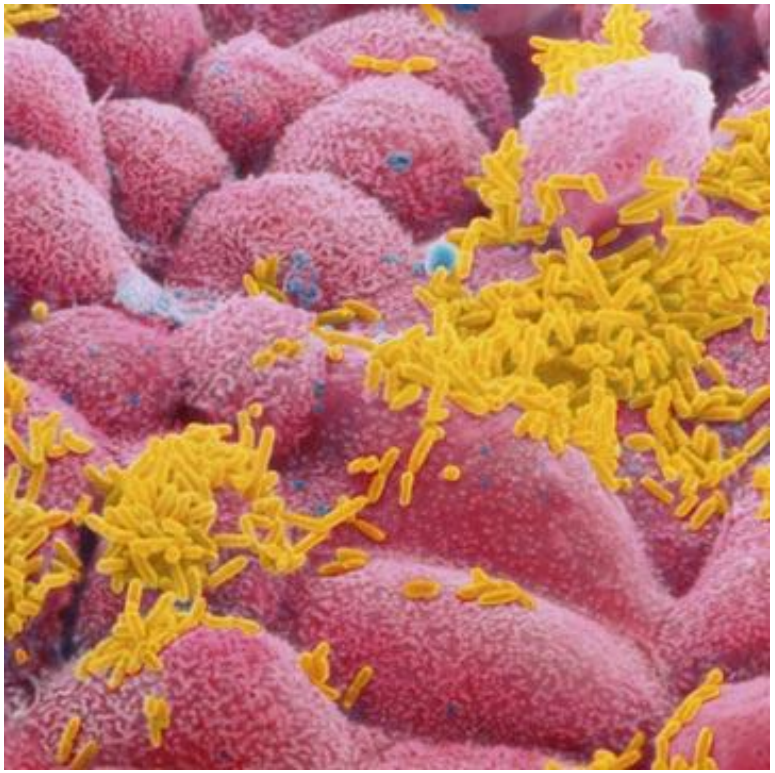
- Microbes affect human condition deeply:
  intestines, farm animals, crop pandemics, food industry, beer brewing, medicine, ...

Deep Sea Vent



Study of Microbes - up to now

- 1929: Penicillin (A. Fleming)
- Late 1970ies: sequencing of a bacteriophage
- 1995: sequencing of *haemophilus influenzae*.



- Today, GenBank contains:
  1000 bacteria,
  2000 viruses,
  100 archaea

- Limits for single-organism-studies:
  Cloning is necessary, (only few species can be cultured)
  True state in Nature includes full habitat and host organisms

What is „Metagenome sequencing“?

Classical sequencing

<--- Bacteria in habitat --->

1) Isolation

2) Cloning

3) Sequencing

AATGCTGAGATCCAAGG ...

Metagenome sequencing

1) Sequencing

AATGCTGAGATCCAAGG ...

The Difference

Classical sequencing

- mostly complete
- location of genes, operons etc can be identified
- well ordered picture of genome is possible

Metagenome sequencing

- fragmented and incomplete
- no link to individual or species
- mostly no reconstruction possible
- „big picture“ versus „accuracy“
- data sets are larger by several orders of magnitude
- computational challenges: new, huge, exciting

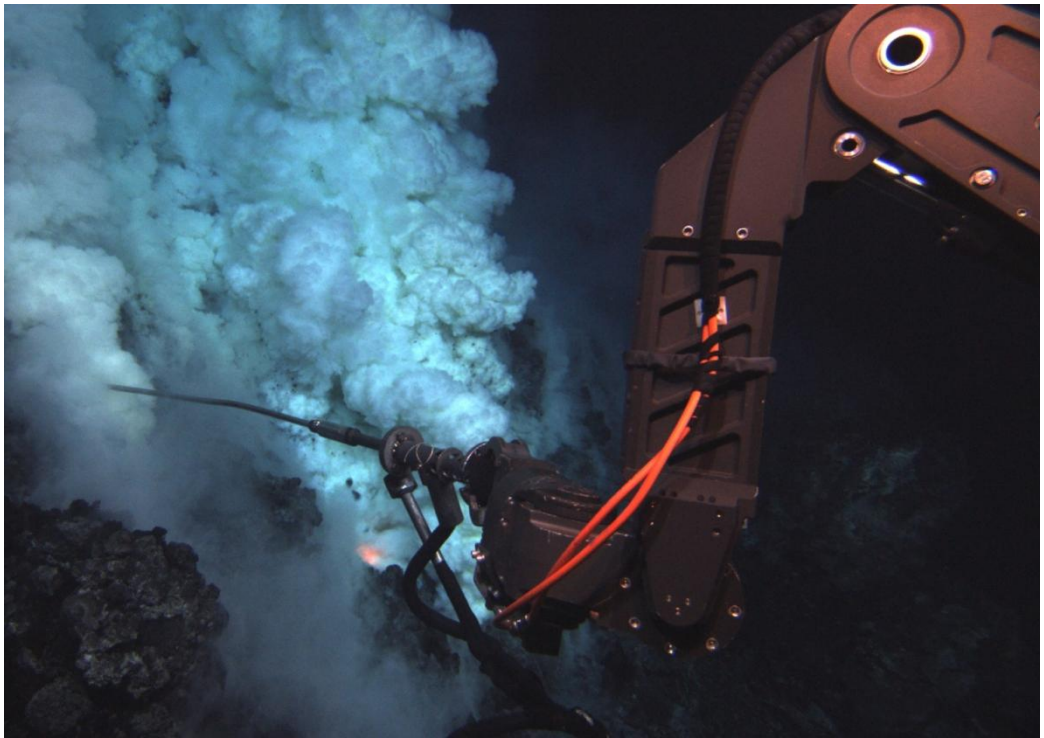
Sampling, Filtering

- Sampling is difficult - individuals are invisible
- Species are characterized by 16S rRNA subunit („ribotype“)

- remove junk from sample
- separation by size is not perfect, some junk remains

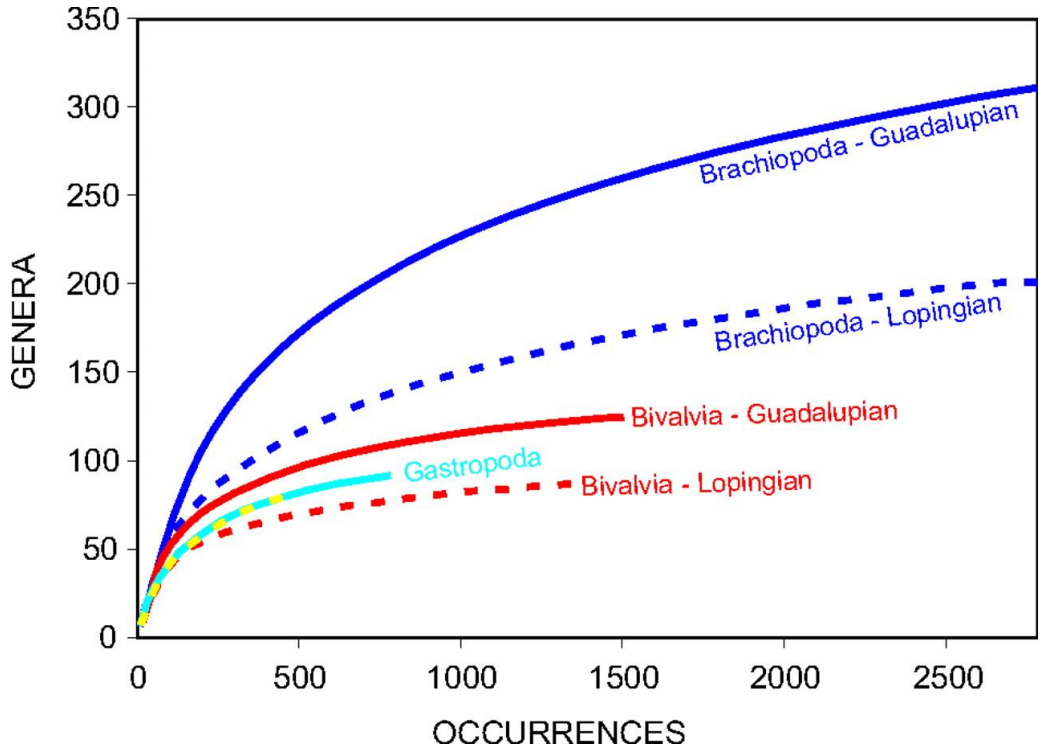
- „computational filtering“ has to be done after sequencing (using similarity searches in data bases)
- But: be aware of false negatives!
  (Metagenomics will yield many new sequences with no homologs in the data bases)

Sampling



Did we collect enough samples?

- The rarefaction curve gives the answer:
  steep: we need more samples
  flat: we are done, almost.



The Importance of Metadata

- Metadata:
  where is the sample from?
  when was it taken?
  what temperature did we have?
  pH of the water in the pond?
  ...
- Metadata is as important as the data!
  (no data mining without metadata)

- Ontological form desirable
- Consider MIGS/MIMS (like MIAME for Micro-Arrays)
- Under construction: Data Markup Language GCDML

GDDML example

```
/environmental_sample
/dev_stage = „text“
/organelle = „<organelle_value>“
/organism = „<organism_value>“
/specimen_voucher = „[<institution-code>;[<collection-
   code;]]<specimen_id>“
/PCR_primers = „[fwd_name: XXX.1,] fwd_seq: xxxx.1, ...“
...
```

First Generation Sequencing

Classical

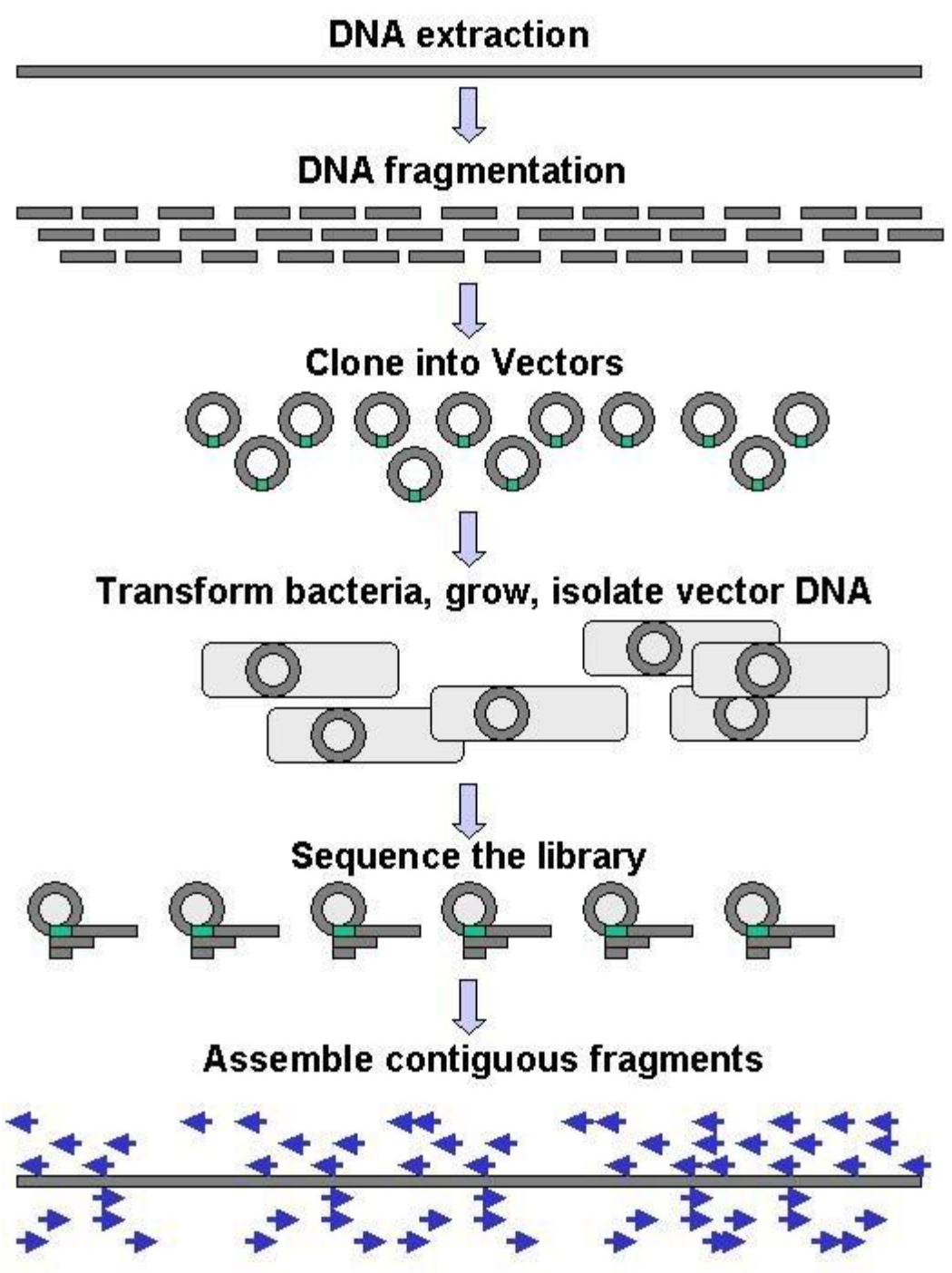
- Sanger shotgun (fragmentation, cloning, sequencing, assembly)
- Works up to 5 Mbps
- Large repeats are a problem
- Cloning bias

For Metagenomes

- Same as „classical“, but...
- Additional complexity (sequences from >1 species)
- Bias to larger populations
- Choosing primers is tricky

- ESS = „Environmental Shotgun Sequencing“

Sanger Shotgun Sequencing



Second Generation Sequencing

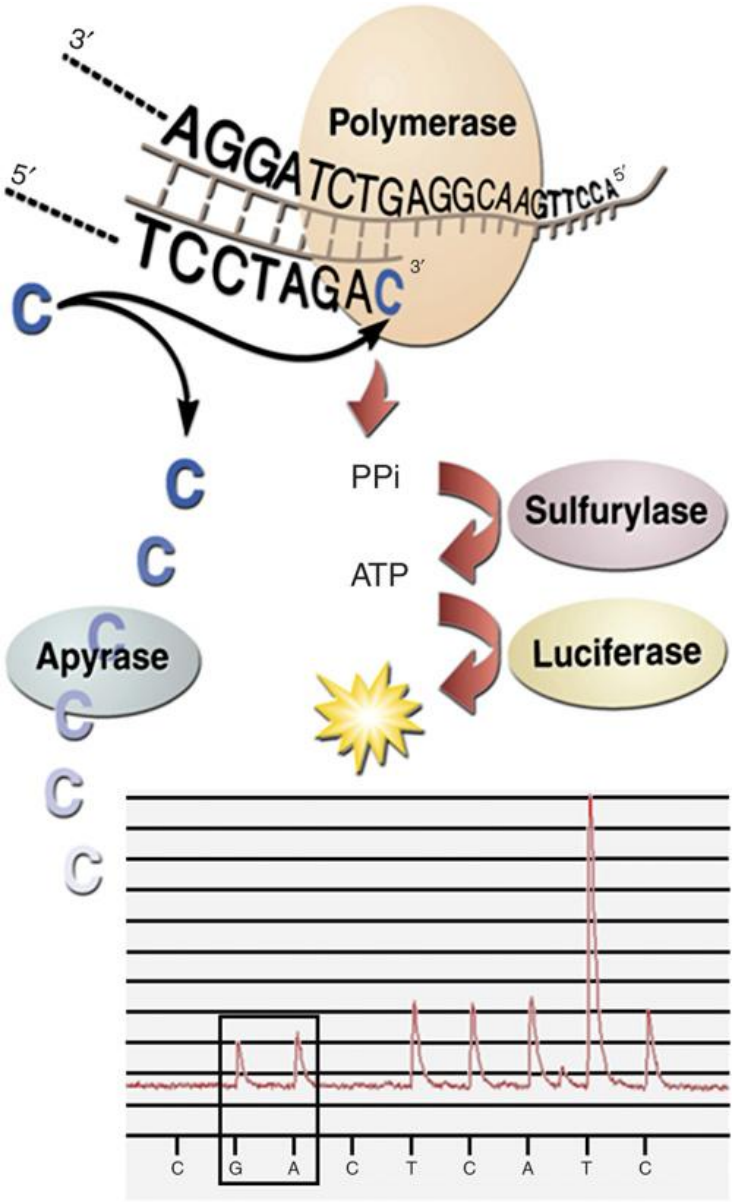
- Many different approaches to improve 1G-Sequencing (be careful when selecting one)
- Rapidly replacing Sanger for small genomes + metagenomics

- Massively parallel
- Very high troughput, but shorter sequences

- PCR creates „polonies“ (= PCR colonies) from single molecules
- Those „amplicons“ are the templates for sequencing

Pyrosequencing

- Watch DNA-Polymerase in action!
- $(DNA)_n$ + dNTP ----> $(DNA)_{n+1}$ + PPi
- The pyrophosphate released at the addition step is used as an indicator
- 300 - 500 bps maximum (half of Sanger)
- A 10 hour run can produce...
- ... 400 million (!) nucleotides

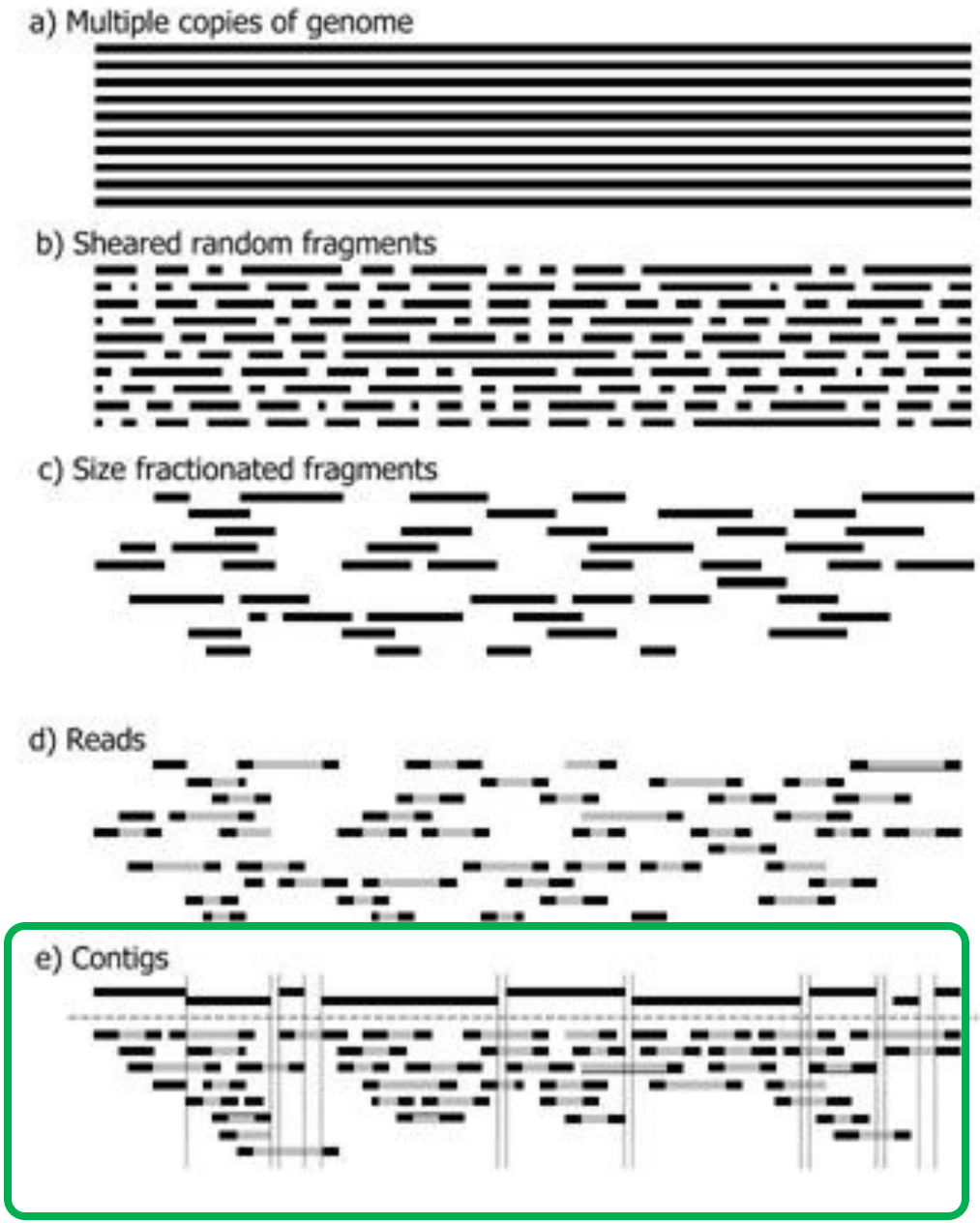


Sample Coverage

- How often is one single nucleotide sampled (in average)?
- Ideal case: one single read for whole genome,
  SC = 1 would be enough
- Shorter reads require more coverage to ensure a
  (more or less) complete overlap
- Assumption: shearing and sequencing is random
- Then C = (L, read length) x (N, nr of reads) / (G, genome length)
  C = 200 x 20.000 / 1.000.000 = 4
- Fraction of sequence covered:
  $P_0 = 1 - e^{-(LN/G)}$ ($P_0 = 1 - e^{-4} = 1 - 0{,}0183 = 98{,}17\ \%$)
- In metagenomics: account for many species, with different genome sizes and with different frequencies

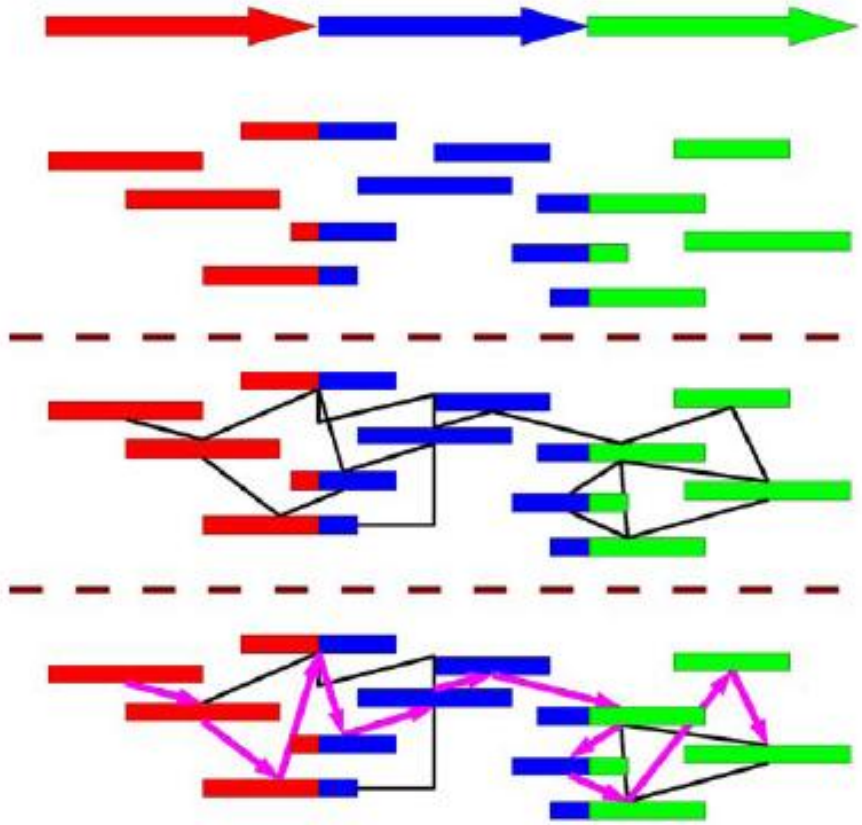
Genome Assembly (1)

- Note: single genome!
- Assembly is last step, after sequencing the fragments
- Big question: how do the fragments fit together? (usually, we have two or more possibilites for assembling two reads)



Genome Assembly (2)

- Solution:
  1) note all possible assemblies
  2) represent this as a graph
     (nodes are reads, edges are possible overlaps)
  3) Look for the Hamiltonian path in this graph
- Hamiltonian path: a path through the graph where every node is visited exactly once



- Small problem:
  Hamiltonian is NP-complete (expontial growth with n!)

Metagenome Assembly (1)

- Coverage is very likely incomplete
- Danger of interspecies chimeras
  (reads from different species will pair)

- Reads are much shorter
- Reads are much more numerous
- Reads will be redundant

- Therefore:
  Hamiltonian will be too difficult to compute

Metagenome Assembly (2)

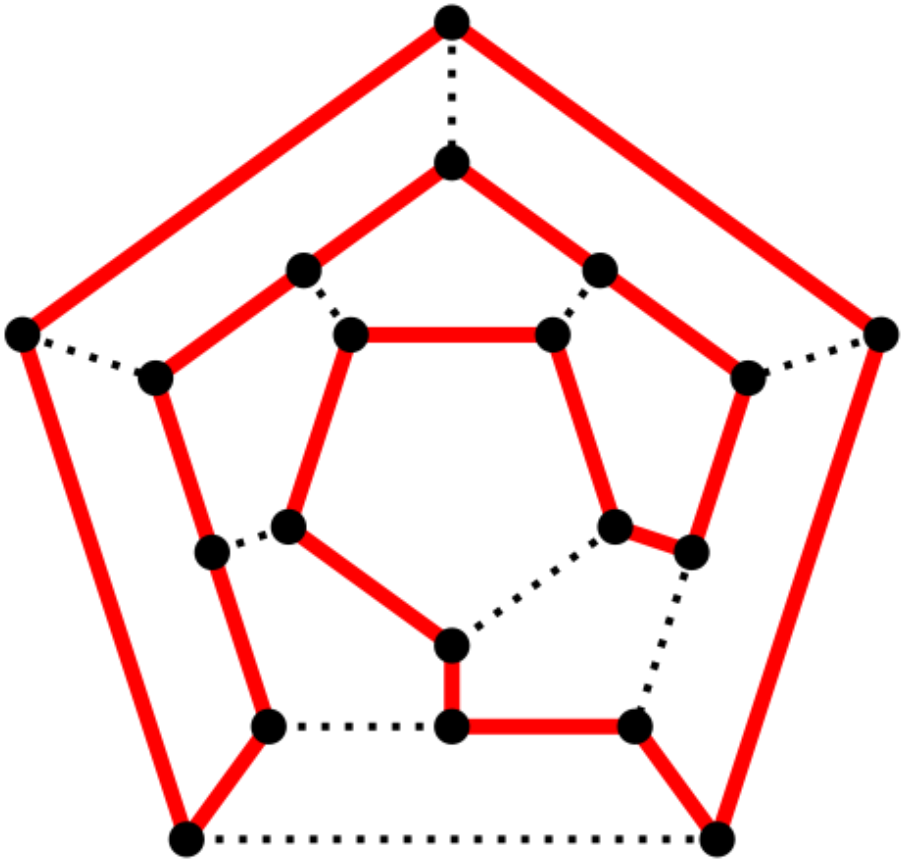
- Solution:

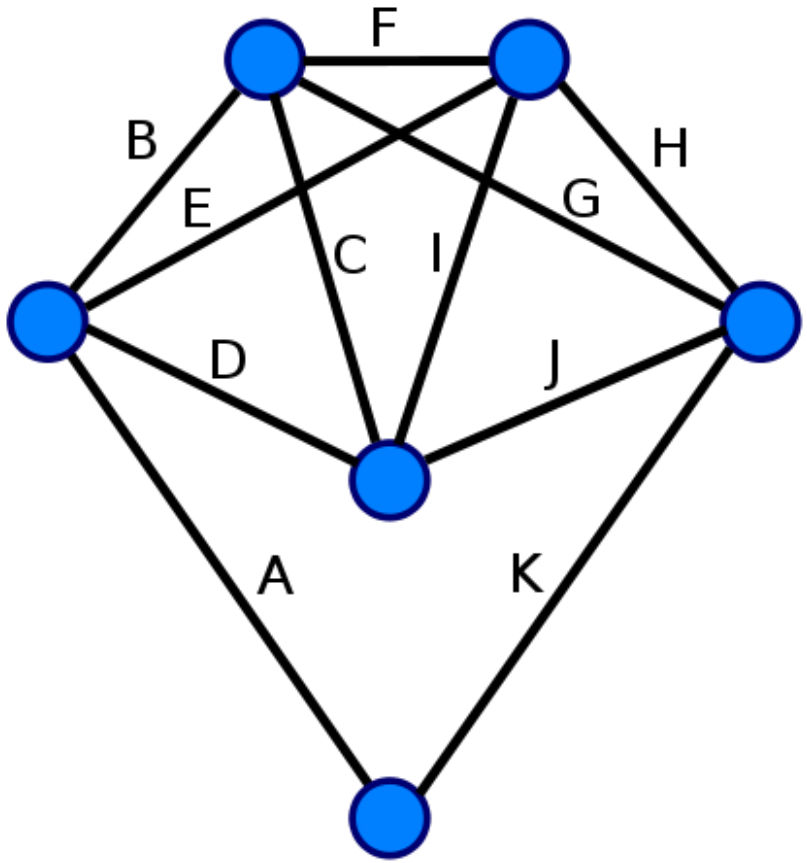
- A node represents a k-mer (not a read)
- An edge reprents a read (not a possible assembly)

- Now one has to find a path where each edge (not node) is visited exactly once („Eulerian path“)

- Euler path has to be found
  can be done in linear time
  (*remark: really???*)

Hamiltonian Path versus Euler Path





Gene Calling

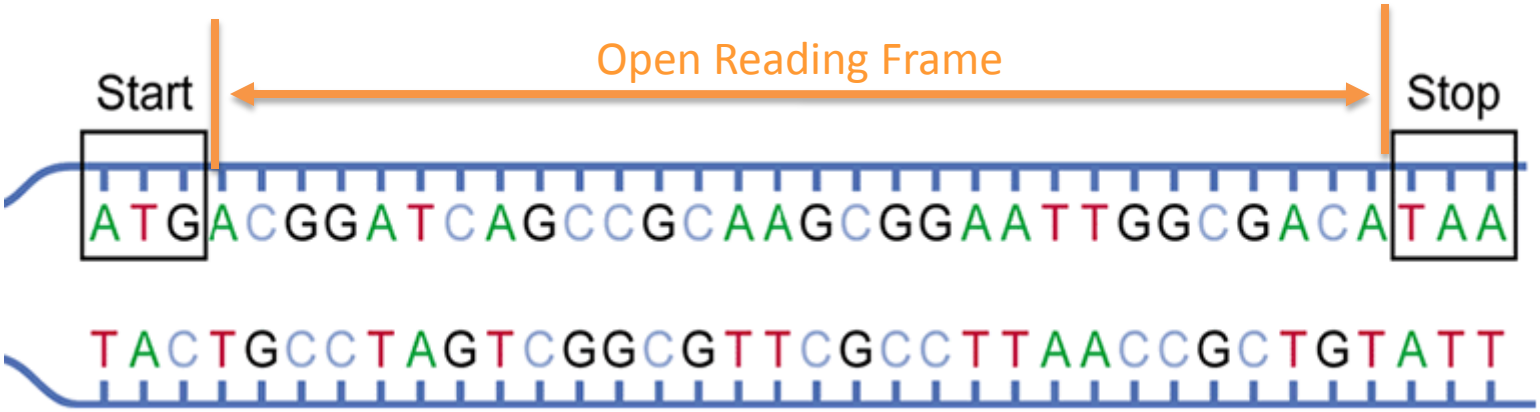
- „Gene Calling“ = Finding Genes within the DNA strand
  (actually, the name is patented by CuraGen for a certain method, but in this context: „looking for DNA stretches that are genes“)

- Gene = Gene or larger functional units (operons, funct.networks)
  (Operon = collection of genes that are controlled together)

- Again, troubles due to fragementary nature of metagenomic data
- Important thing: ORFs

Open Reading Frame (ORF)

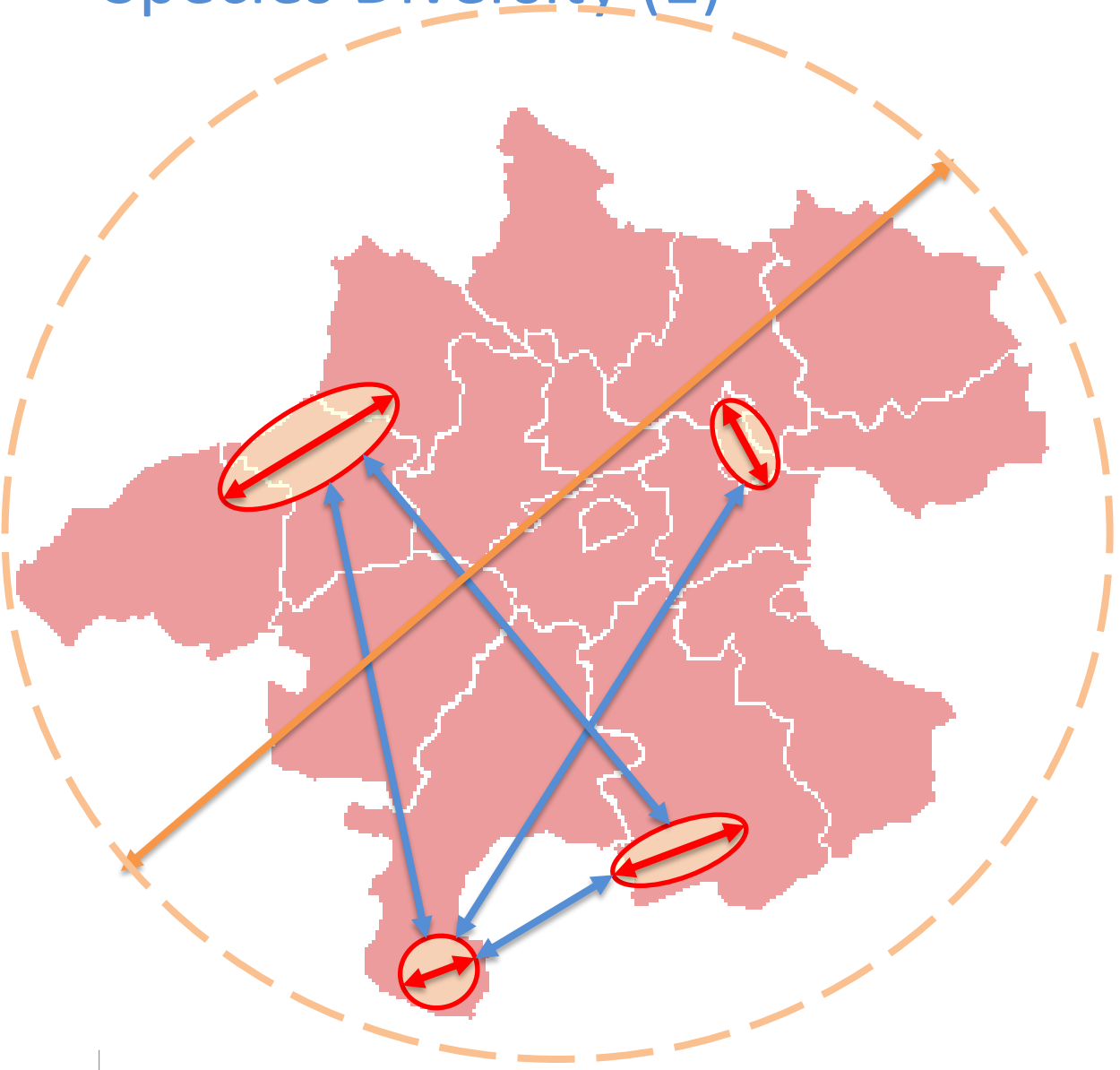
- ORF = everything between Start and Stop Codon (including introns, if any)
- We have 6 possible RFs in a double stranded DNA
- ORFs are flanked by untranslated (but important) regions
- Those will be (partly) translated to mRNA, but not transcribed to proteins
- In bacteria, ORFs sometimes overlap and are controlled together („Operons“)



Finding Genes in Metagenomes

- Finding ORFs is difficult in metagenomics
- Nevertheless, 85 - 90 % can be achieved
- BLASTing sequences: only if homologs already exist
  (not a good idea for metagenomes)
- Ab initio gene recognition: pattern recognition methods
  (HMMs, supervised learning, ...)
- Software example: `Genemark.hmm` (uses monocodon frequs)
- Method used by Yooseph et al:
  1) Identify longer ORFs
  2) Cluster nonredundant ORFs using BLAST (???)
  3) Eliminate false ORFs (overlapping ORFs due to reading frame)
  4) Eliminate ORFs that show no signs of evolutionary pressure
- False negatives are likely (excessive removal of ORFs), but no thorough study yet

Species Diversity (1)



Habitat, ecosystem

α-diversity

β-diversity

γ-diversity

We are talking about α-diversity here (mainly)

Several definitions of diversity exist, e.g.
Shannon: $+/-p_i \ln p_i$
or
Simpson: $+/(n_i/N)^2$
$(= +/p_i^2)$

a.regl

What is a species? (1)

- In higher organisms, this is easy:
  Individuals that can mate belong to the same species

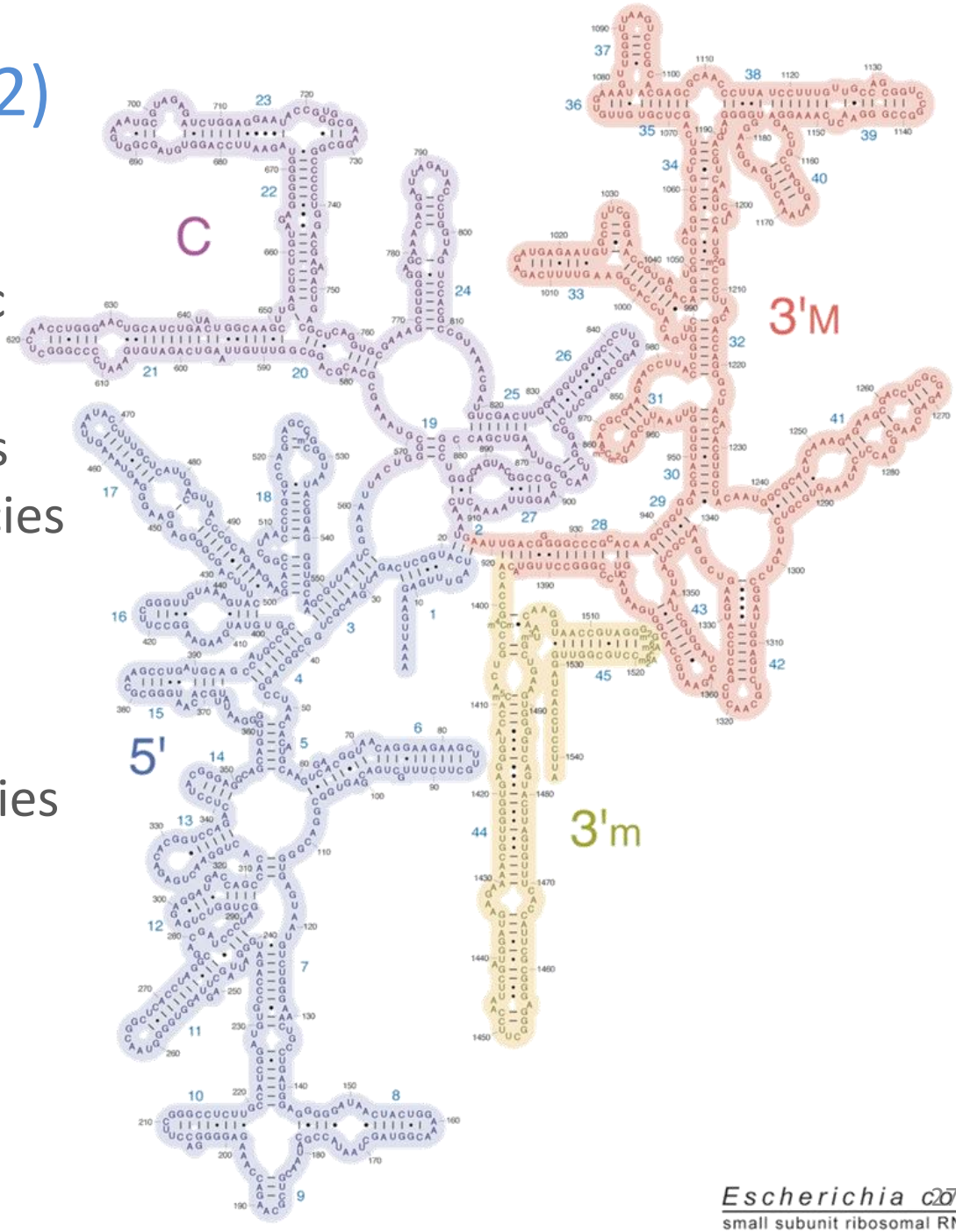

No


Yes

- But what about bacteria? Do they „mate“? If yes, how?

What is a species? (2)

- Species = „OTU“
  (= Operational Taxonomic Unit)
- Normally, the 16S rDNA is used as a marker for species identity
- Problems:
  1) HGT!
  2) multiple sequence copies
  3) large CNV
- Solution:
  use other genes, too



Escherichia coli
small subunit ribosomal RNA

„Binning“

- Which sequence belongs to which OTU?
- Two algorithms are used:

Composition based

- Rough classification: based on GC content
- Finer details (needed for metagenomics): tetramers or k-mers
- Problems: close species give rise to numerous misclassifications

Similarity based

- Similarities to reference sequences
- Good if the sample composition is roughly known
- Distance measure is used to build a similarity tree

Functional Annotation

- What are they doing?
  Or: what biological functions can be assigned to the ORFs?
- Great challenge for „normal“ genomics already, even more so for metagenomics.
- Genes alone are not enough, biological networks like biochemical pathways etc are interesting.

Two strategies:

1. Simplify the gene calling step (look for long 6-frame translations and scan DBs for motifs and other signatures)
2. Use unassembled reads for that purpose

Comparative Metagenomics

- Compare the results of two or more metagenomics data pools
- Traits like GC content, genome size etc are used
- Statistics tools like PCA are used to visualize and extract the important factors
- Again: metadata is very important.
- Common vocabulary is necessary to be able to compare.
- For the time being, there is no automatic method to do that.

Software

- Many packages for each step of metagenomics research (assembly, gene calling, binning, functional annotation, ...)
- Roughly 30 - 40 packages are mentioned in the article
- Broad spectrum of functionality
- Some are web-based, most are to download and install
- Most packages are in public domain or Open Source

Applications (1)

- An exemplary overview:

Correlations to metadata

- Leeuwenhoek: *„the animalcules in my mouth disappear when I drink hot coffee“*
- How does habitat influence microbial life there (and vice versa)?
- E.g.: In obese mice, the gut bacteria are more energy efficient and have more active carbohydrate enzymes

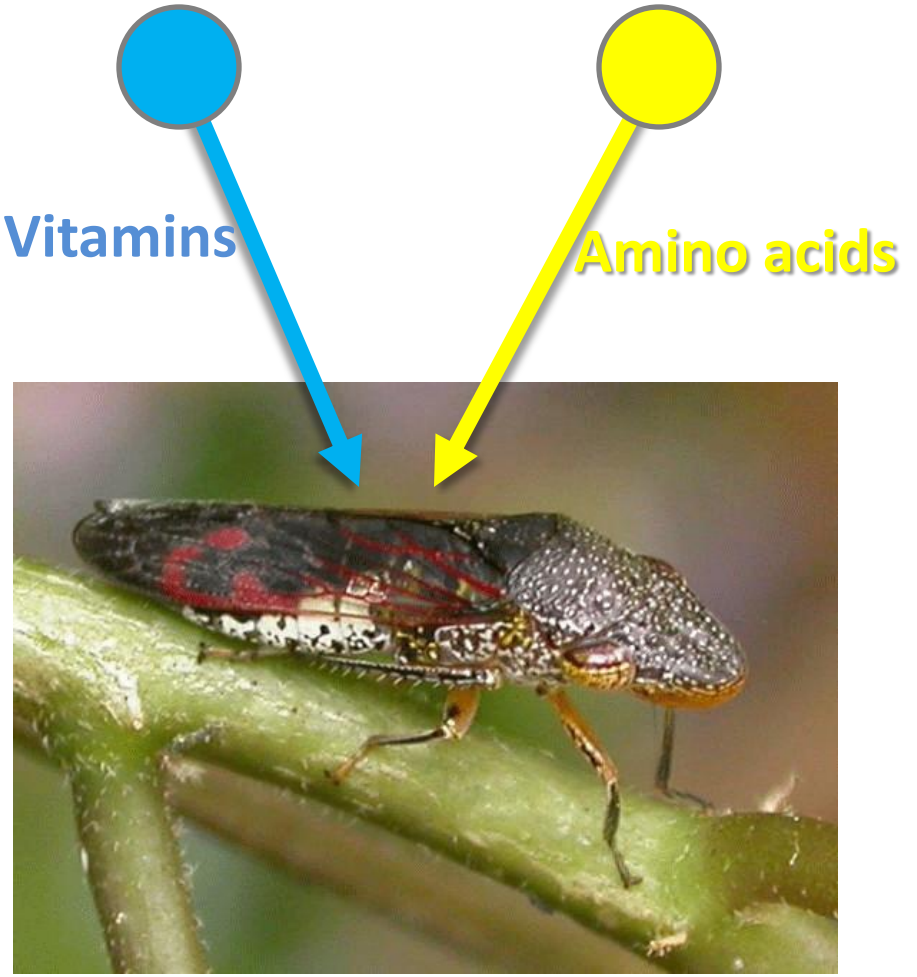
Applications (2)

Understanding symbiosis

- Interaction between host and symbionts
- (relatively) easy to handle

- Glassy-winged sharpshooter: two bacteria provide necessary ingredients
- Marine gutless worm: intricate network with four symbionts

- Metagenomics only!



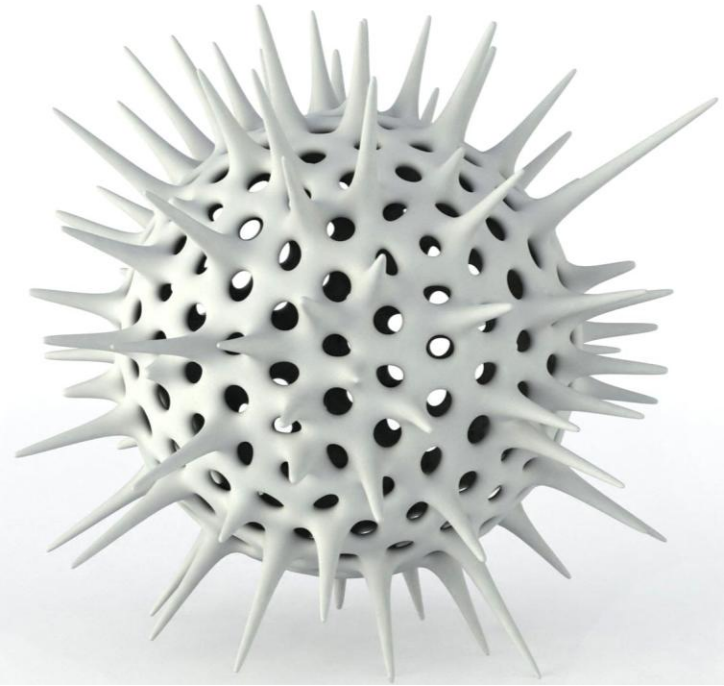
Applications (3)

Enriching gene families

- Search for new members of known gene families in the flood of new sequence data

Environmental Virology

- Viruses outnumber bacteria by far (~10^{30} in biosphere)
- 90 % of their sequences have no counterpart in GenBank etc.
- Transduction: important contribution to evolution, but poorly understood
- Virus genomes provide new challenges (prophages? species distinction without 16S?)



The Future

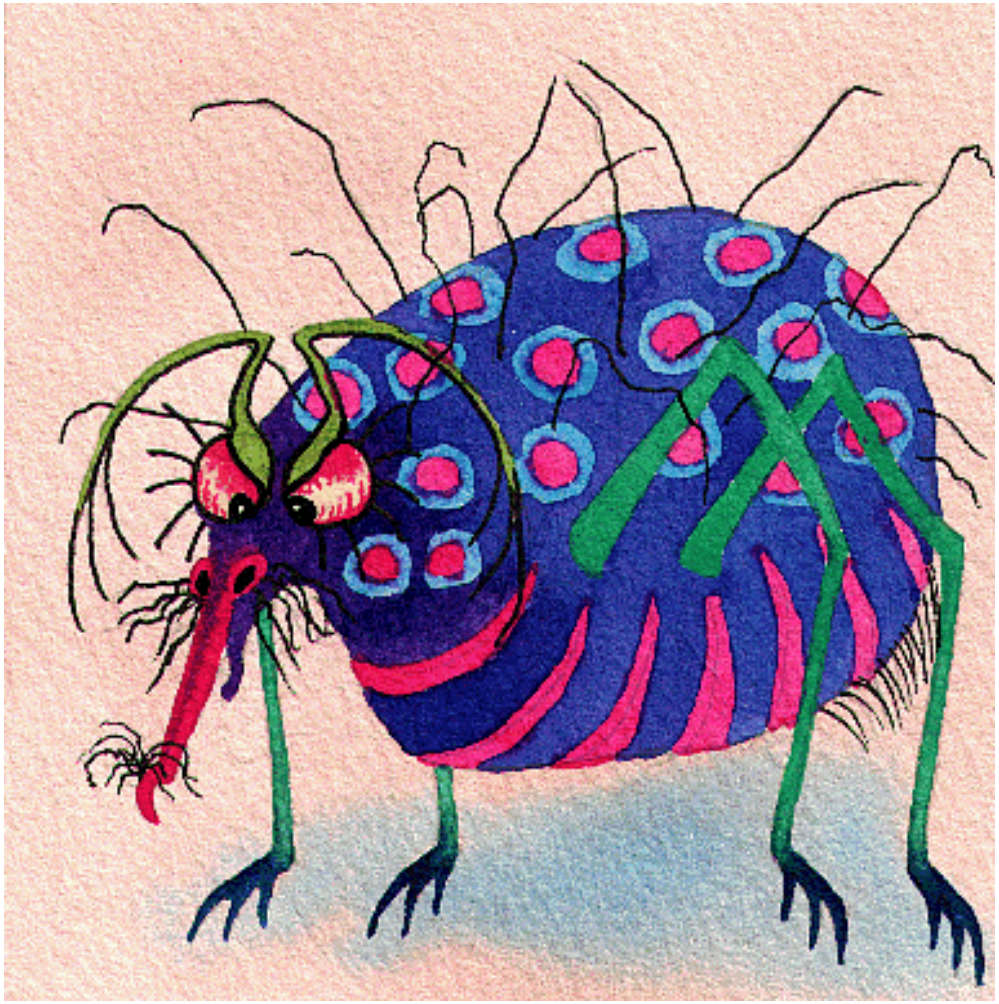
- Machinery of the last 3 years has provided more sequence data than Sanger for 30 years - and more to come!
- Main problem: overwhelming flood of data
- Needs for IT power grows faster than IT technology

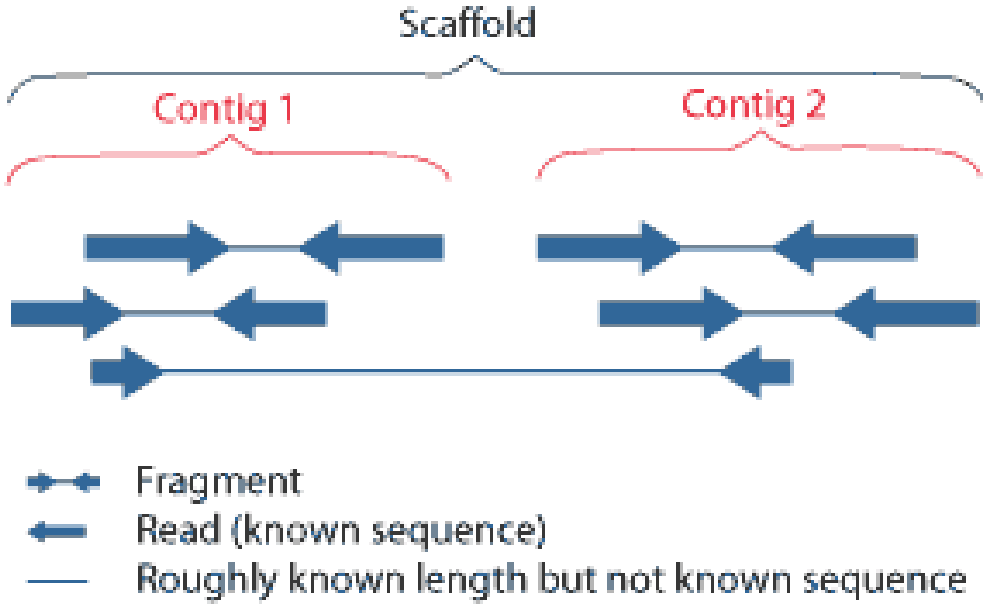
- 3G sequencing: no more short reads - a single read for a chromosome without fragments seems to be feasible

- Distinguishing methylated nucletides (gene regulation!) (BioInf is not prepared for that)

- „Meta-transcription“, „Meta-translation“ as next step

Any Questions?




Scaffold
Contig 1
Contig 2
Fragment
Read (known sequence)
Roughly known length but not known sequence

a.regl

Gene Calling by CuraGen