

Úprava NGS dat a Metagenomika

Petr Dvořák

Katedra botaniky PřF UP

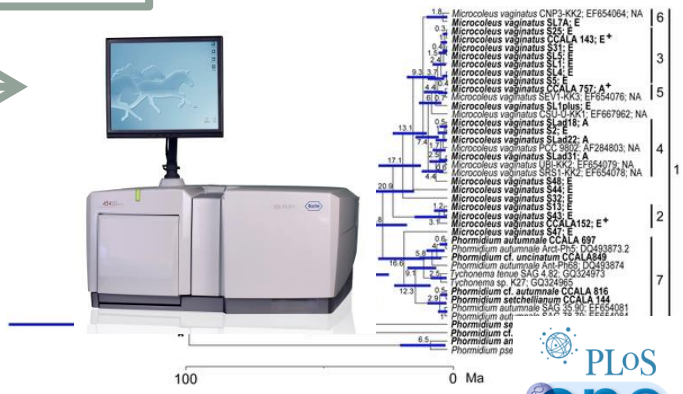
„Propojení výuky oborů Molekulární a buněčné biologie a Ochrany a tvorby životního prostředí “

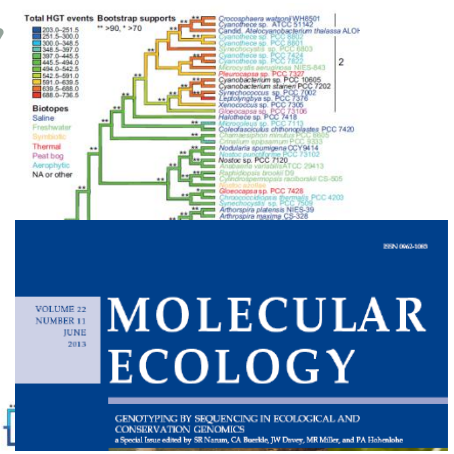
Reg. č.: CZ.1.07/2.2.00/28.0032

Krátce o mně

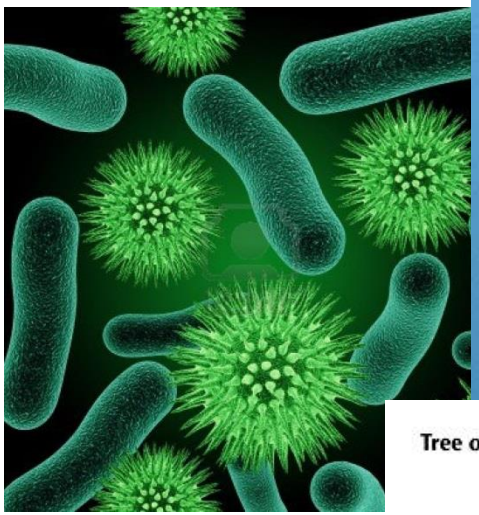
Sinice (Cyanobacteria)

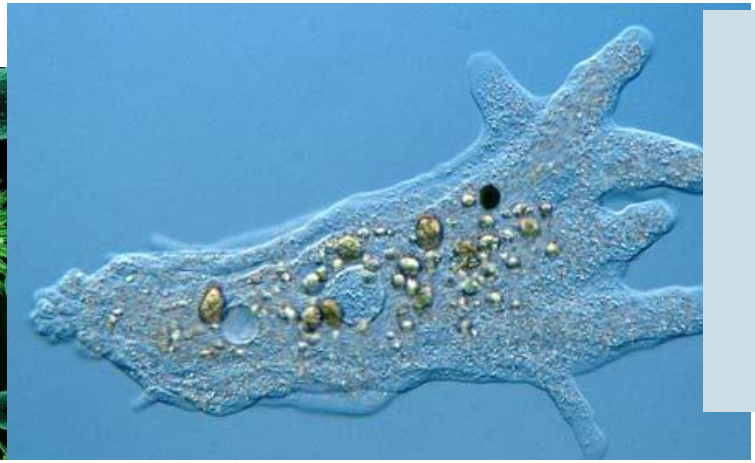
Metanogeny

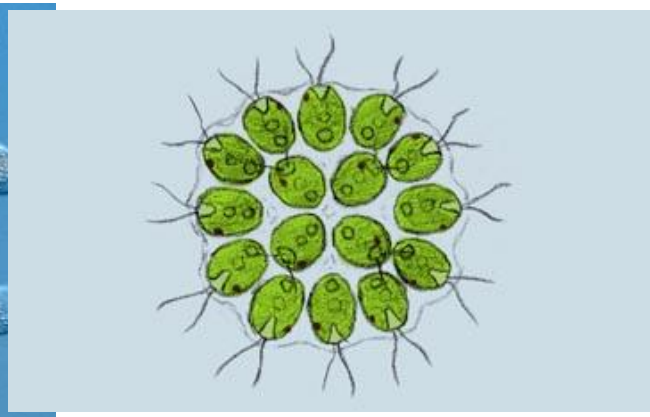


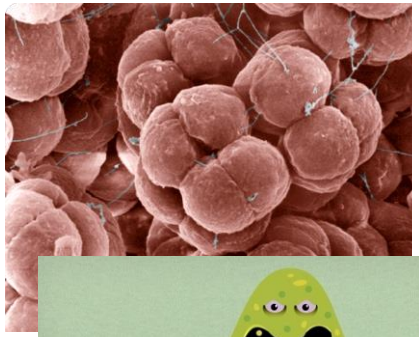


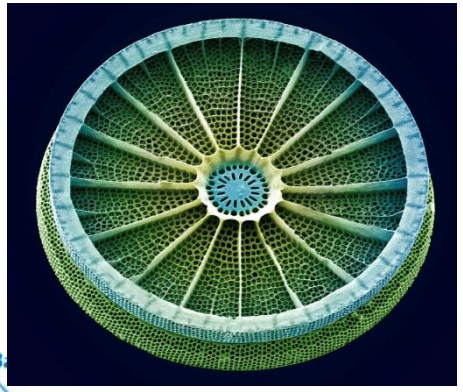
Co je to mikroorganismus?

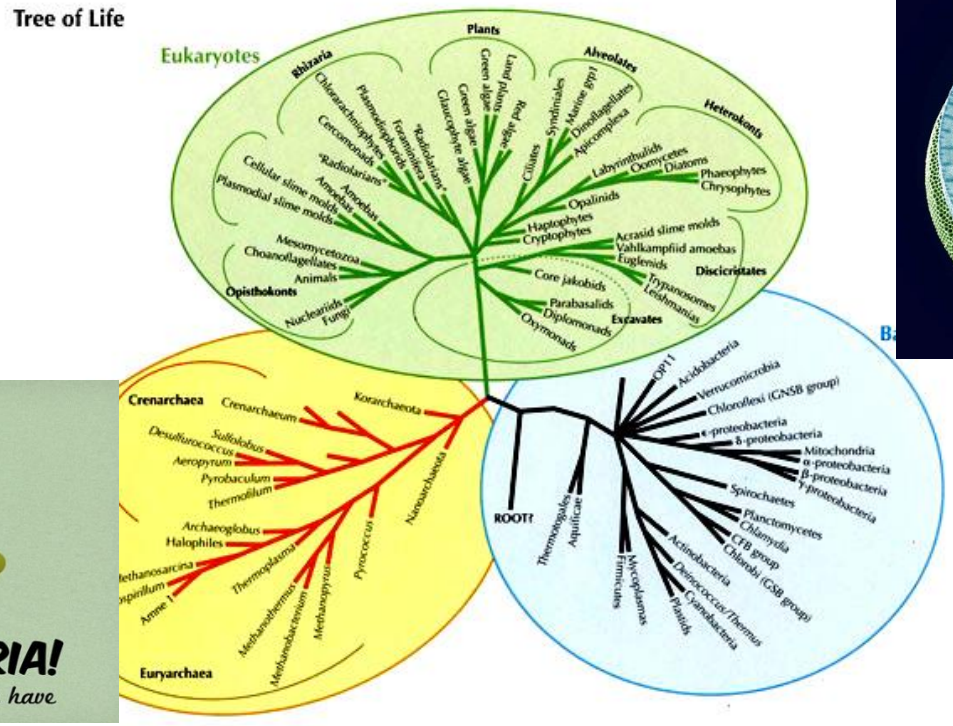








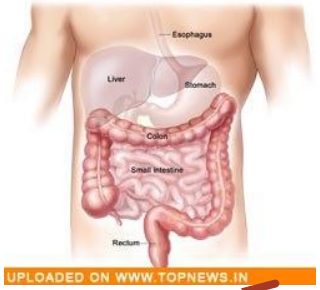




Kde všude najdeme mikroorganismy?

Horké prameny

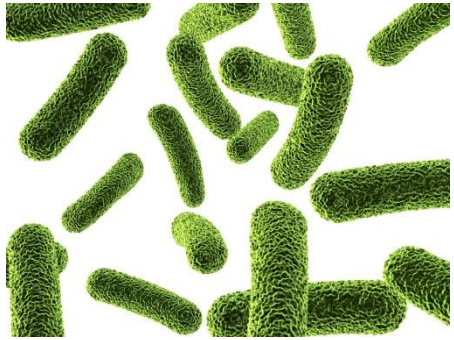
Trávicí trakt



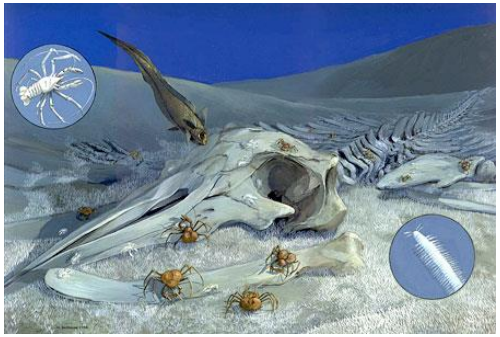
Sladká voda

Atmosféra

Slaná voda



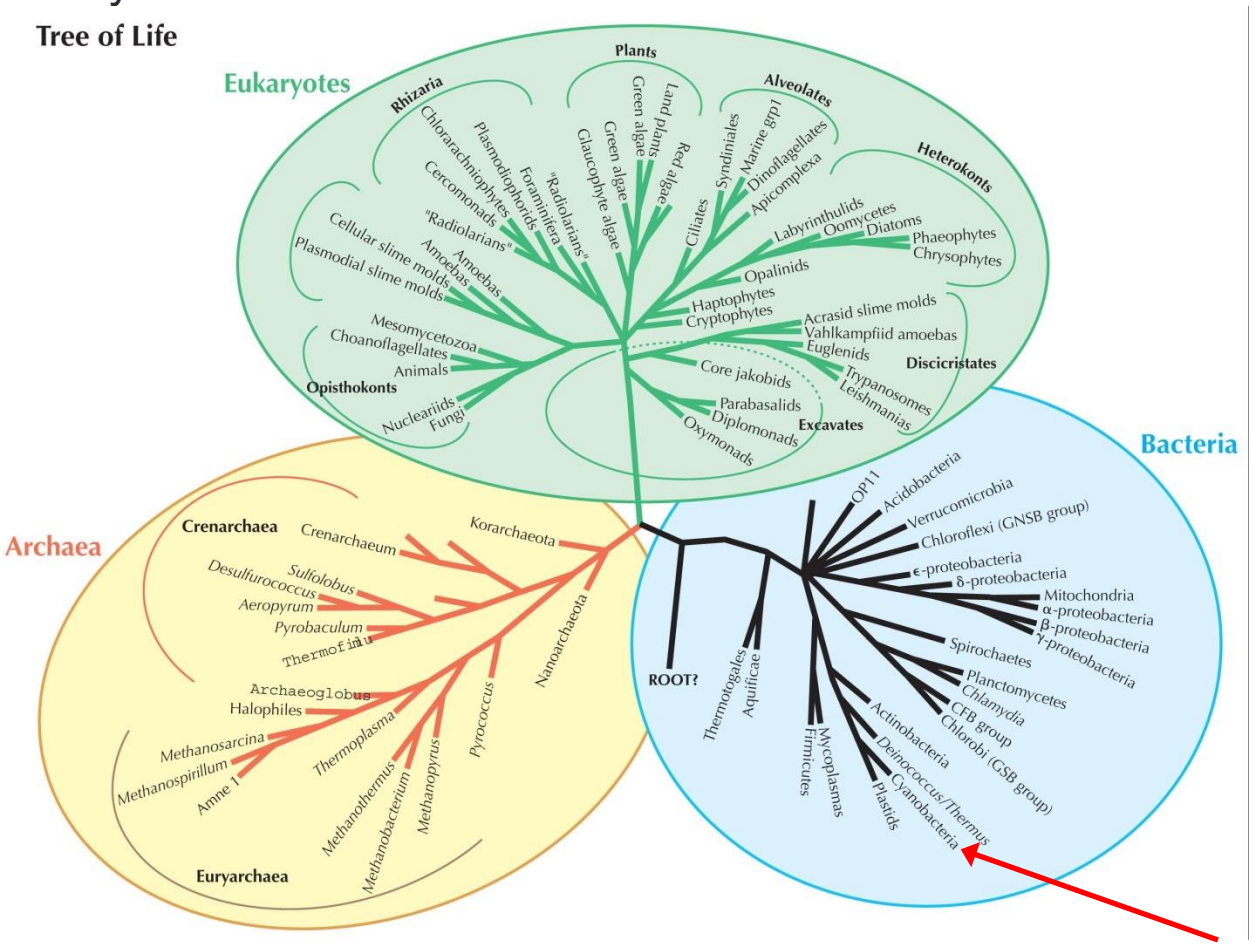
Mrtvá velryba



Půda

Úvod: co jsou sinice?

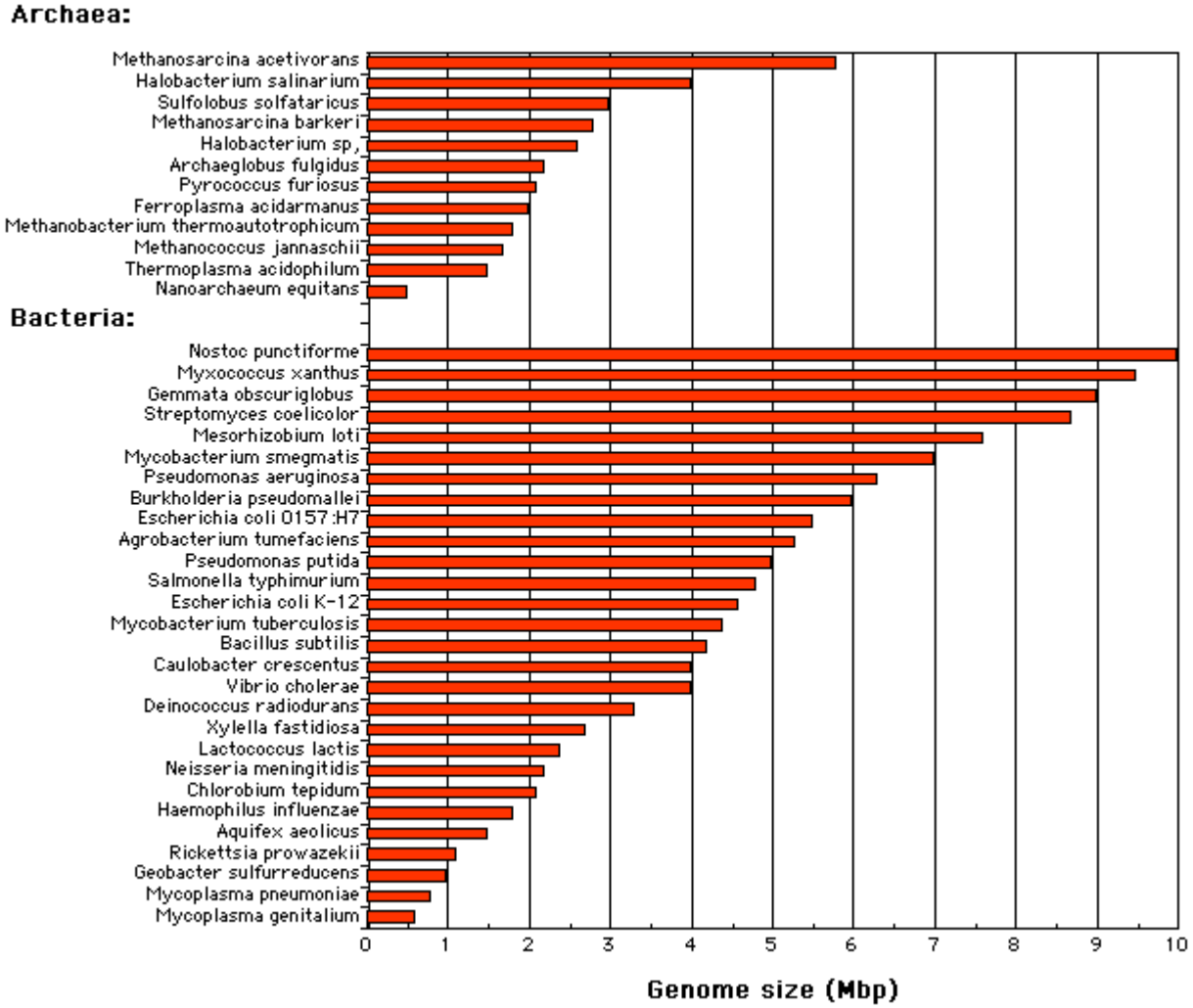
Prokaryotní organismy
Oxygenní fotosyntéza



http://phylogenomics.blogspot.cz/2008/10/twisted-tree-of-life-award-2-science.html

Velikost genomu sinic

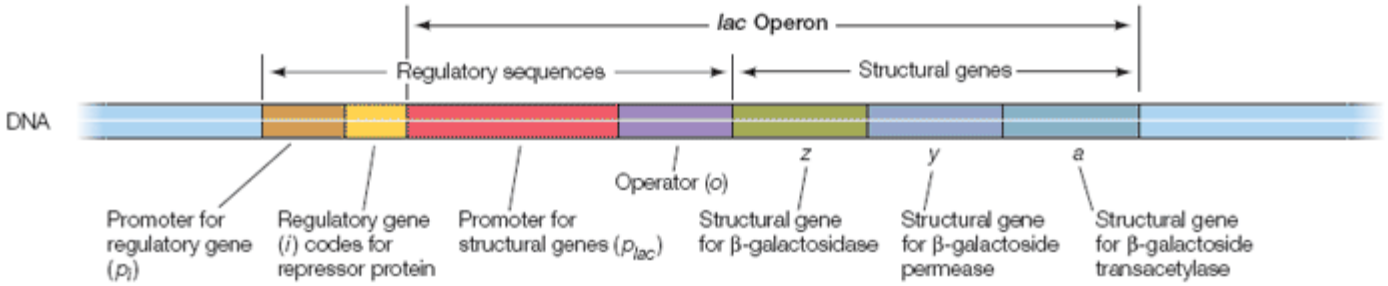
Člověk – 2.9 Gb



www.sci.sdsu.edu/~smaloy/MicrobialGenetics/topics/chroms-genes-prots/genomes.html

Genom sinic (i ostatních prokaryot)

- Většinou jeden chromozóm (možná polyploidie)
- Velká část kódující, málo „Junk DNA“
  - *Synechococcus* sp. WH8102 86%, 2526 protein kódujících genů
- Velikost genomu často koreluje s počtem genů – 1kb = 1 gen
- (Skoro) žádné introny!
- Chromozóm v supercoiled formě
- Transkripce a translace probíhá zároveň



http://www.nature.com/scitable/topicpage/genome-packaging-in-prokaryotes-the-circular-chromosome-9113

Porovnání NGS a Sanger

Dnes 300 – 600 (pair end) bp

Sequencer	454 GS FLX	HiSeq 2000	SOLiDv4	Sanger 3730xl
Sequencing mechanism	Pyrosequencing	Sequencing by synthesis	Ligation and two-base coding	Dideoxy chain termination
Read length	700 bp	50SE, 50PE, 101PE	50 + 35 bp or 50 + 50 bp	400~900 bp
Accuracy	99.9%*	98%, (100PE)	99.94% *raw data	99.999%
Reads	1 M	3 G	1200~1400 M	—
Output data/run	0.7 Gb	600 Gb	120 Gb	1.9~84 Kb
Time/run	24 Hours	3~10 Days	7 Days for SE 14 Days for PE	20 Mins~3 Hours
Advantage	Read length, fast	High throughput	Accuracy	High quality, long read length
Disadvantage	Error rate with polybase more than 6, high cost, low throughput	Short read assembly	Short read assembly	High cost low throughput

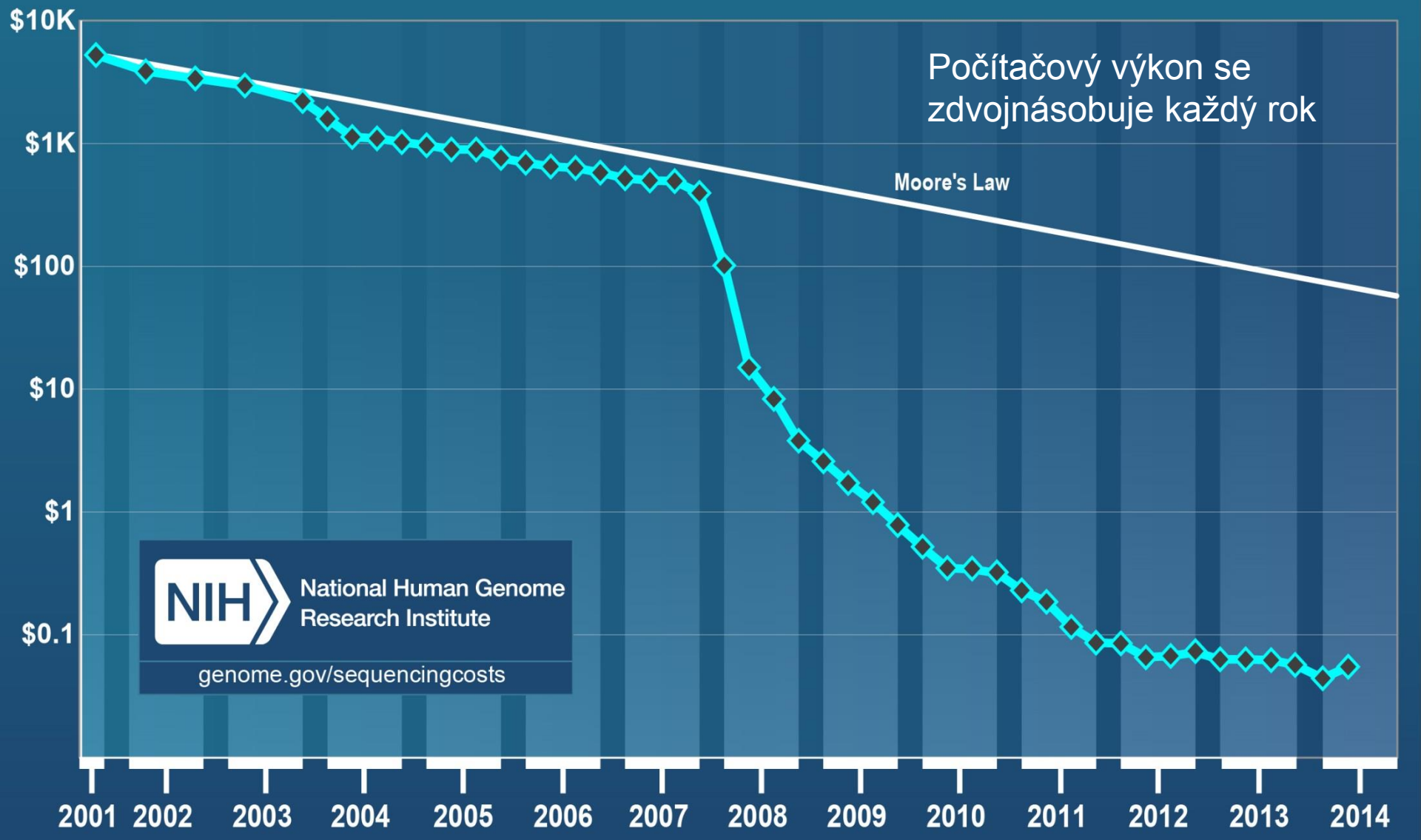
Liu et al. 2012

Porovnání NGS a Sanger

Sequencers	454 GS FLX	HiSeq 2000	SOLiDv4	3730xl
Instrument price	Instrument $500,000, $7000 per run	Instrument $690,000, $6000/(30x) human genome	Instrument $495,000, $15,000/100 Gb	Instrument $95,000, about $4 per 800 bp reaction
CPU	2* Intel Xeon X5675	2* Intel Xeon X5560	8* processor 2.0 GHz	Pentium IV 3.0 GHz
Memory	48 GB	48 GB	16 GB	1 GB
Hard disk	1.1 TB	3 TB	10 TB	280 GB
Automation in library preparation	Yes	Yes	Yes	No
Other required device	REM e system	cBot system	EZ beads system	No
Cost/million bases	$10	$0.07	$0.13	$2400

Liu et al. 2012

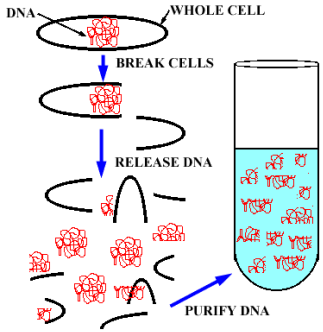
Cost per Raw Megabase of DNA Sequence



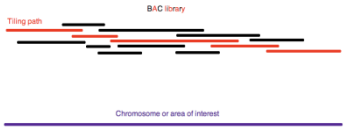
Projekt: sekvenování genomu sinice

- Cíl: získat kompletní genom sinice
- Postup:

Získání monoklonální kultury



Izolace DNA

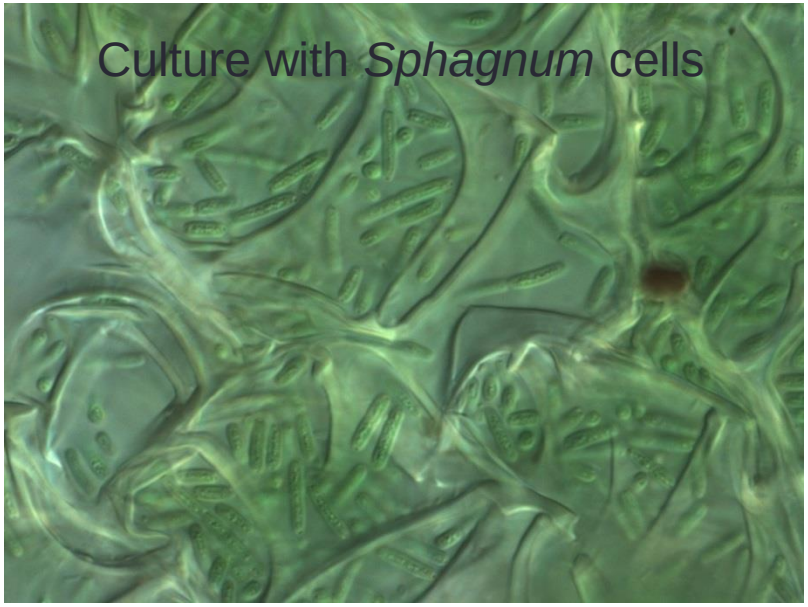


Sekvenátor – 454, Illumina, ….

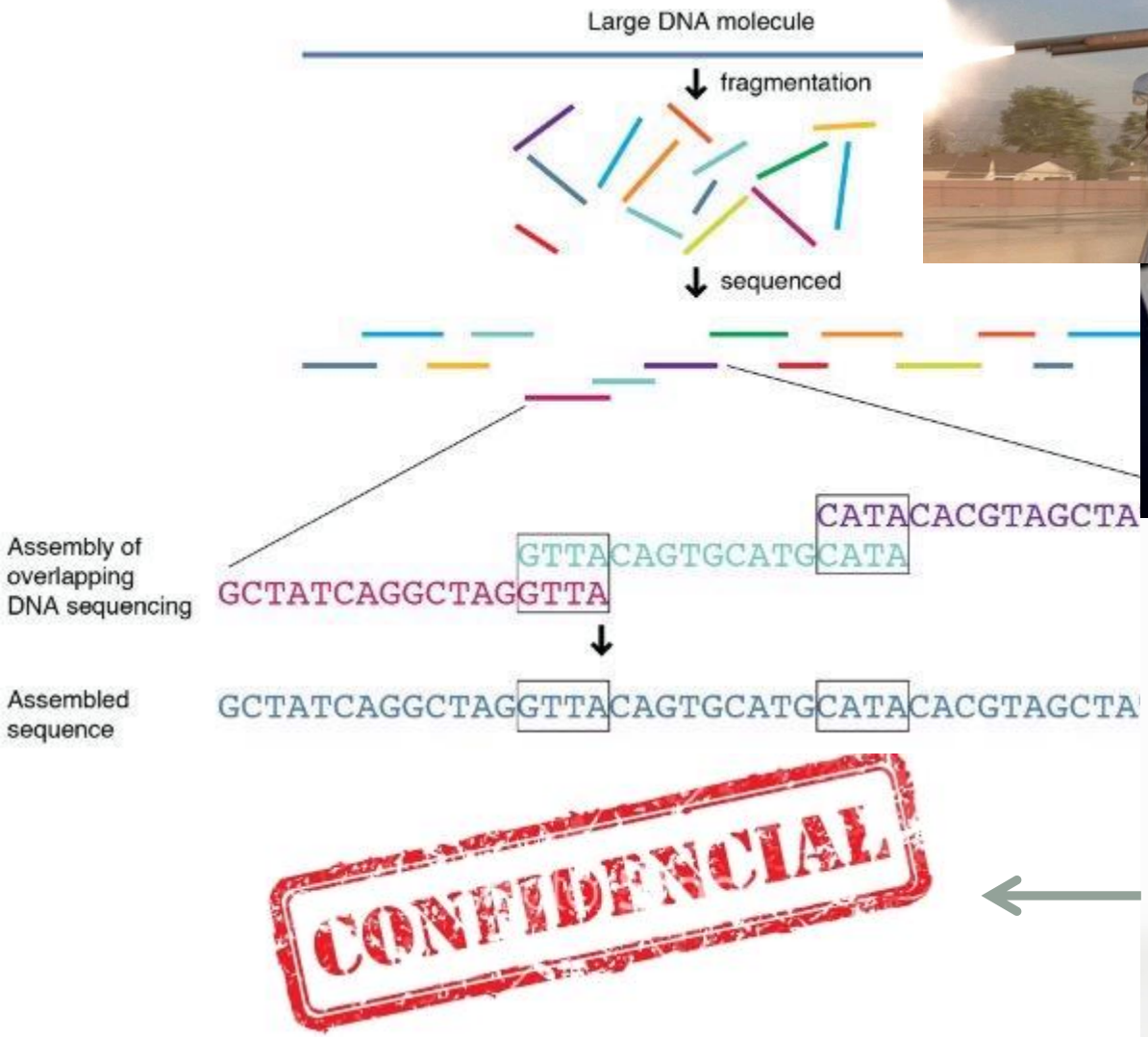
Bioinformatická analýza

Neosynechococcus sphagnicola

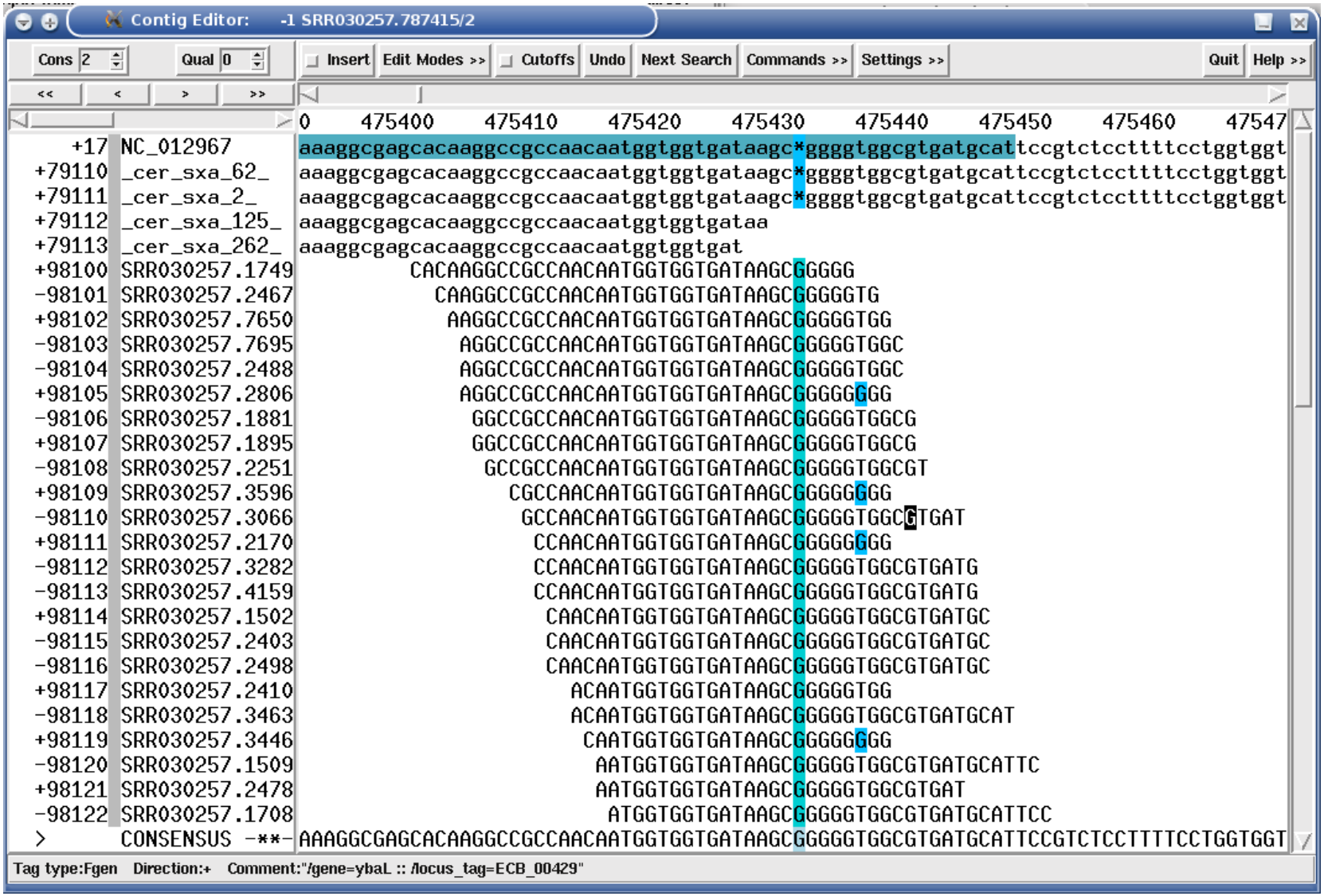
Life style – *Sphagnum* style



Shotgun sequencing

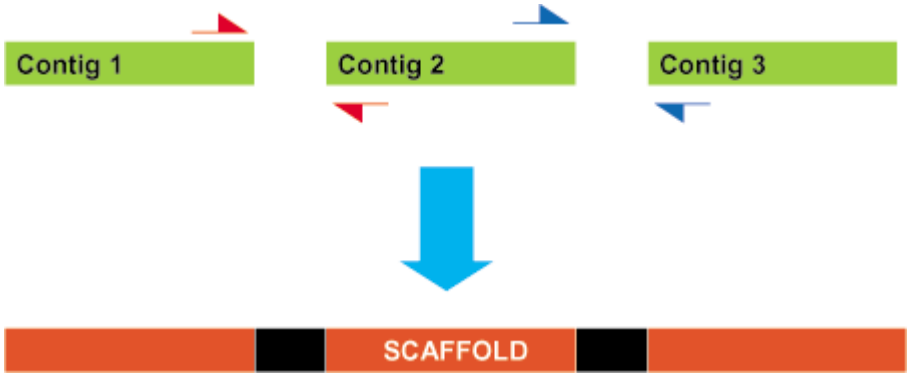


Assembly – poskládání



Assembly

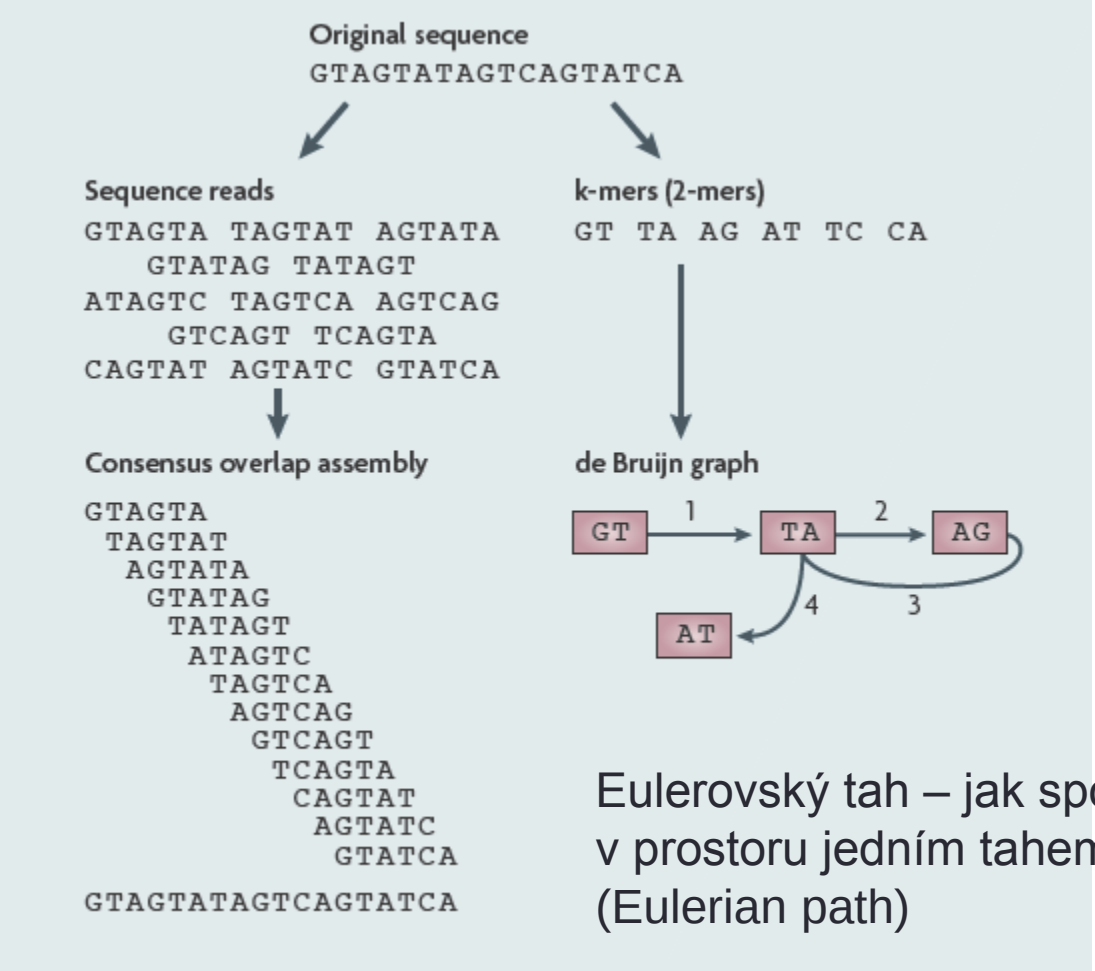
- Jednotlivé sekvence (reads) – contig – scaffold



Software: SOAP, MIRA…

Assembly – poskládání

- Výpočetně náročnější
- Nevýhodné pro krátká čtení

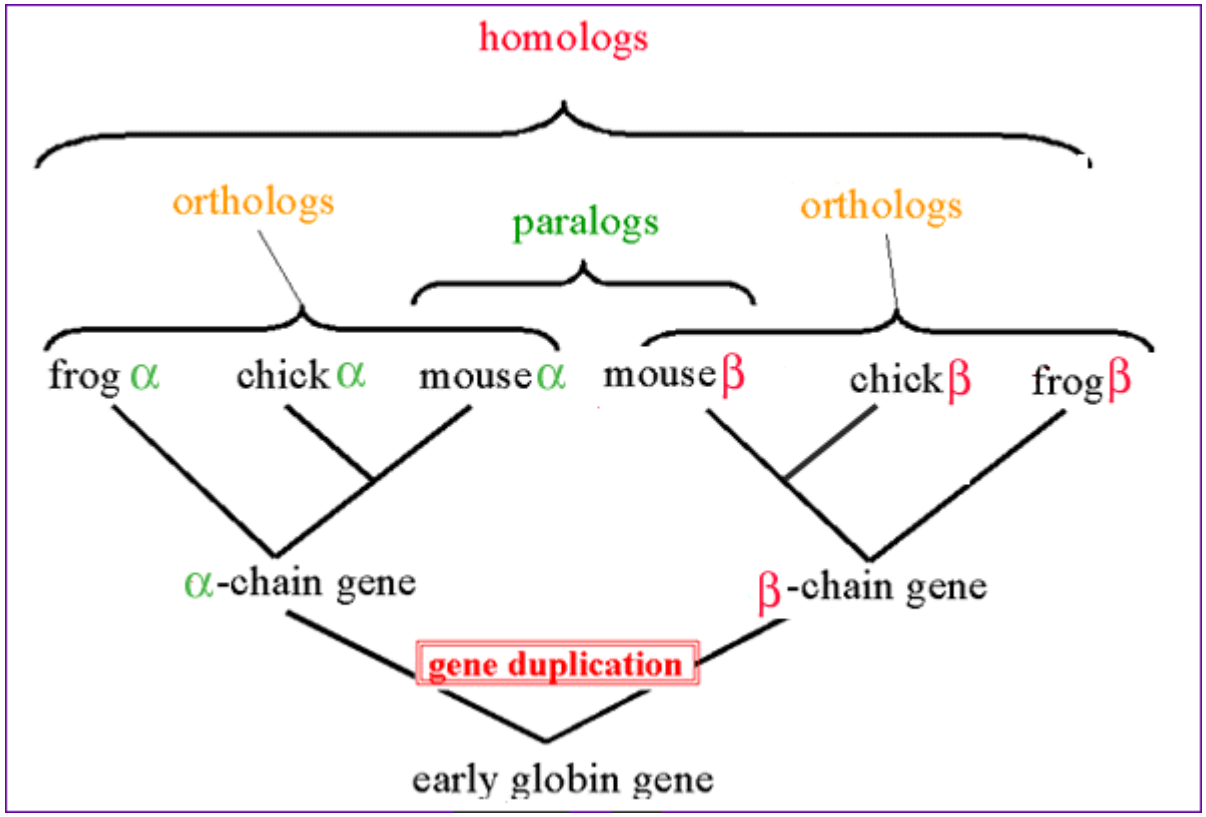


- Výpočetně Méně náročné
- Výhodné pro krátká čtení

Eulerovský tah – jak spojit všechny body v prostoru jedním tahem? (Eulerian path)

http://gcat.davidson.edu/phast/debruijn.html

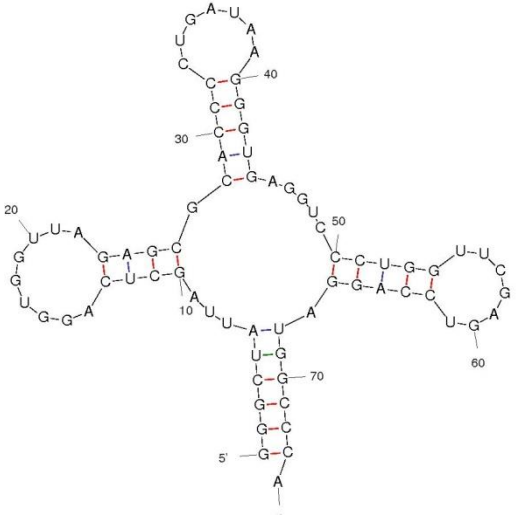
Ortolog versus paralog



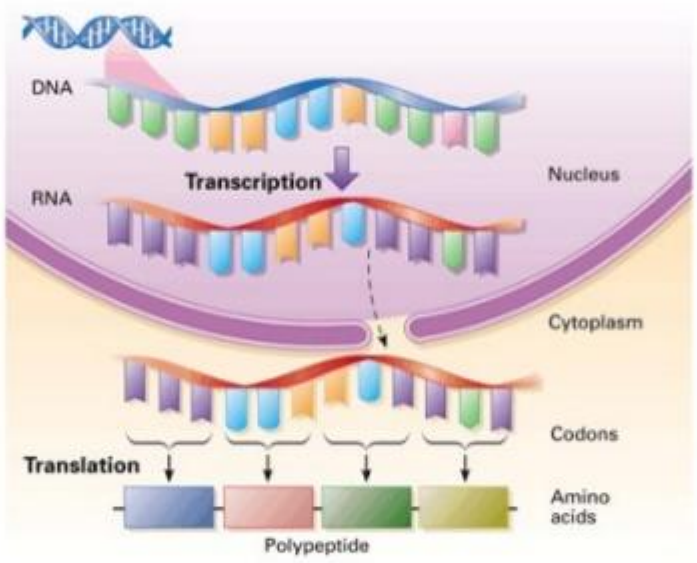
Anotace sekvencí

- Predikce genů – kde jsou v sekvenci geny a kde nekódující sekvence
- Predikce tRNA – v rámci nekódujících úseků – typická sekundární struktura
- Identifikace ortologních sekvencí pomocí algoritmu BLAST v databázi GenBank (http://www.ncbi.nlm.nih.gov/)

NCBI BLAST



The Central Dogma of



Molecular Biology

Predikce genů

aatgcatgcggctatgctaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggc
tatgctaatgcatgcggctatgcaagctgggatccgatgactatgctaagctgggatccgatgaca
atgcatgcggctatgctaatgaatggtcttgggatttaccttggaatgctaagctgggatccgatgac
aatgcatgcggctatgctaatgaatggtcttgggatttaccttggaatatgctaatgcatgcggctatg
ctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatc
cgatgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctgggatccgatgaca
atgcatgcggctatgctaatgcatgcggctatgcaagctgggatcctgcggctatgctaatgaatg
gtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgcggctatgctaatgaat
ggtcttgggatttaccttggaatatgctaatgcatgcggctatgctaagctgggaatgcatgcggcta
tgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctggg
atccgatgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctcatgcggctatg
ctaagctgggaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgcta
atgcatgcggctatgcaagctgggatccgatgactatgctaagctgcggctatgctaatgcatgcg
gctatgctaagctcggctatgctaatgaatggtcttgggatttaccttggaatgctaagctgggatcc
gatgacaatgcatgcggctatgctaatgaatggtcttgggatttaccttggaatatgctaatgcatgc
ggctatgctaagctgggaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggc
tatgctaatgcatgcggctatgcaagctgggatccgatgactatgctaagctgcggctatgctaatg
catgcggctatgctaagctcatgcgg

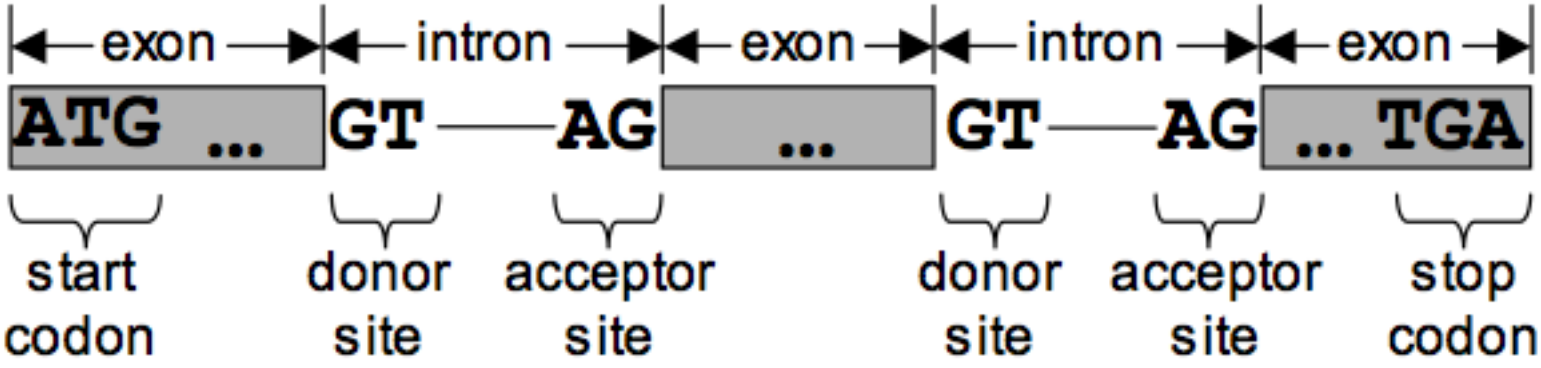
Predikce genů

aatgcatgcggctatgctaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggc
tatgctaatgcatgcggctatgcaagctgggatccgatgactatgctaagctgggatccgatgaca
atgcatgcggctatgctaatgaatggtcttgggatttaccttggaatgctaagctgggatccgatgac
aatgcatgcggctatgctaatgaatggtcttgggatttaccttggaatatgctaatgcatgcggctatg
ctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatc
cgatgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctgggatccgatgaca
atgcatgcggctatgctaatgcatgcggctatgcaagctgggatcctgcggctatgctaatgaatg
gtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgcggctatgctaatgaat
ggtcttgggatttaccttggaatatgctaatgcatgcggctatgctaagctgggaatgcatgcggcta
tgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctggg
atccgatgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctcatgcggctatg
ctaagctgggaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgcta
atgcatgcggctatgcaagctgggatccgatgactatgctaagctgcggctatgctaatgcatgcg
gctatgctaagctcggctatgctaatgaatggtcttgggatttaccttggaatgctaagctgggatcc
gatgacaatgcatgcggctatgctaatgaatggtcttgggatttaccttggaatatgctaatgcatgc
ggctatgctaagctgggaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggc
tatgctaatgcatgcggctatgcaagctgggatccgatgactatgctaagctgcggctatgctaatg
catgcggctatgctaagctcatgcgg

GEN

Predikce genů

- CDS – coding DNA sequence
- Identifikace start kodonu, stop kodonu
- V případě eukaryot – identifiace exonů a intronů



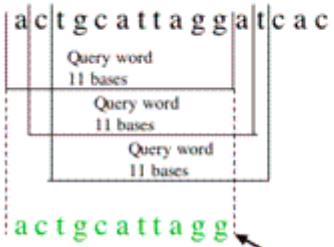
Anotace genů – BLAST

Label Title Line Comment

```
>fig|282458.1.peg.1 Chromosomal replication initiator protein dnaA
MSEKEIWEKVLEIAQEKLSAVSYSTFLKDTELYTIKDGEAIVLSSIPFNANWLNQQYAEI
IQAILFDVVGYEVKPHFITTEELANYSNNETATPKEATKPSTETTEDNHVLGREQFNAHN
TFDTFVIGPGNRFPHAASLAVAEAPAKAYNPLFIYGGVGLGKTHLMHAIGHHVLDNNPDA
KVIYTSSEKFTNEFIKSIRDNEGEAFRERYRNIDVLLIDDIQFIQNKVQTQEEFFYTFNE
LHQNNKQIVISSDRPPKEIAQLEDRLRSRFEWGLIVDITPPDYETRMAILQKKIEEEKLD
IPPEALNYIANQIQSNIRELEGALTRLLAYSQLLGKPITTELTAEALKDIIQAPKSKKIT
IQDIQKIVGQYYNVRIEDFSAKKRTKSIAYPRQIAMYLSRELTDFSLPKIGEEFGGRDHT
TVIHAHEKISKDLKEDPIFKQEVENLEKEIRNV
```

Data Lines

Query



Index (Neighborhood words that are present in or have been extracted from database sequences)

a c t t c a t t a g c

c c t g c a t t a g g

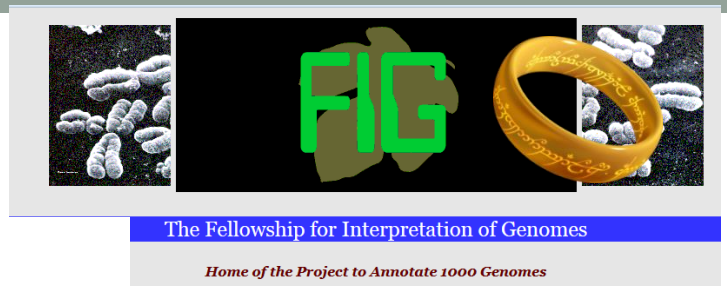
a c t g c a t t a g g

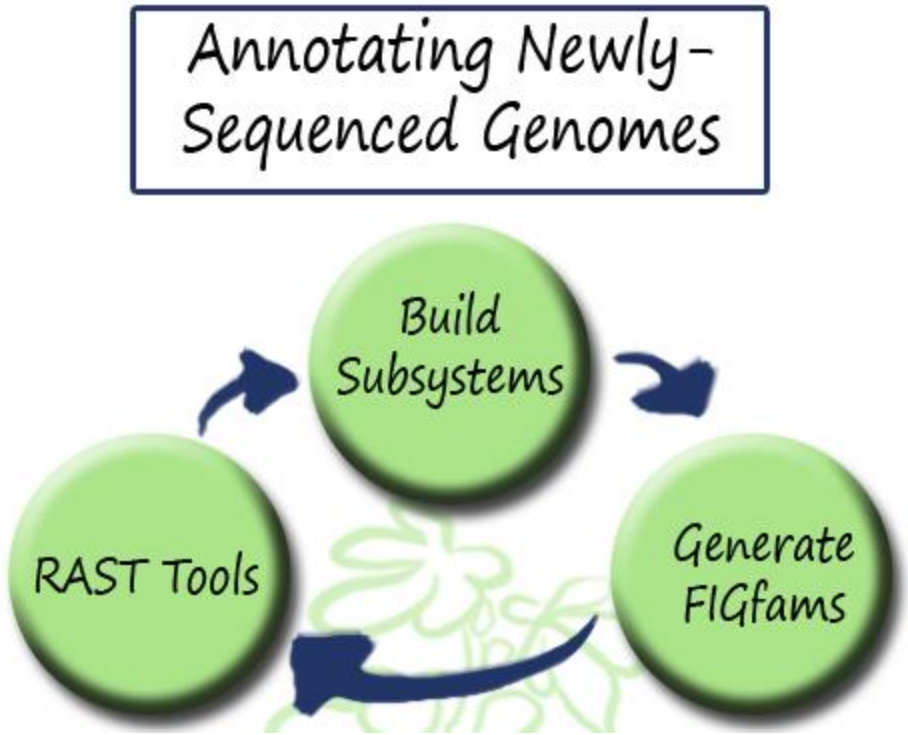
g a c g c a t c t a a

Exact match to an 11-base "neighborhood word" that is present in a database sequence.

Program	Query	Database
blastp	protein	protein
blastn	nucleotide	nucleotide
blastx	nucleotide	protein
tblastn	protein	nucleotide
tblastx	nucleotide	nucleotide

Anotace genů – RAST





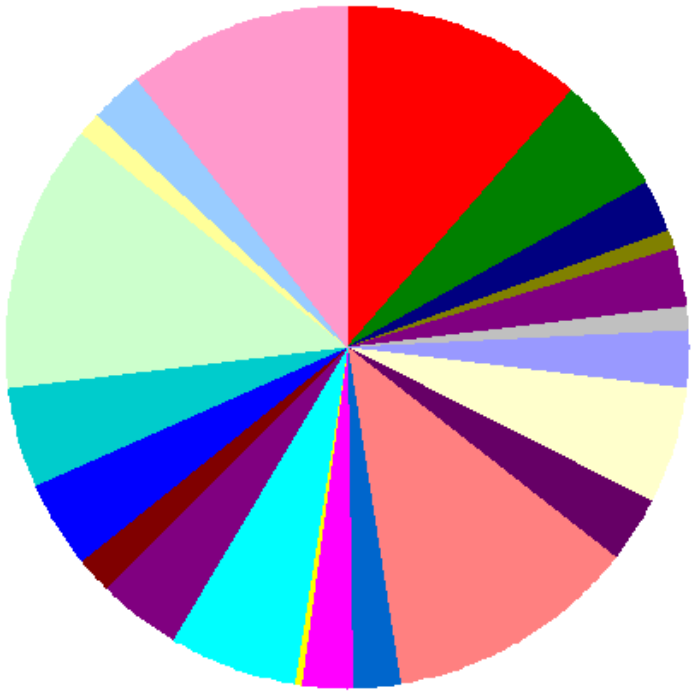
Rapid Annotation using Subsystem Technology

Proteiny, které mají u všech organismů stejnou funkci

- 70% shodné – BLAST

RAST – subsystems

Features in Subsystems



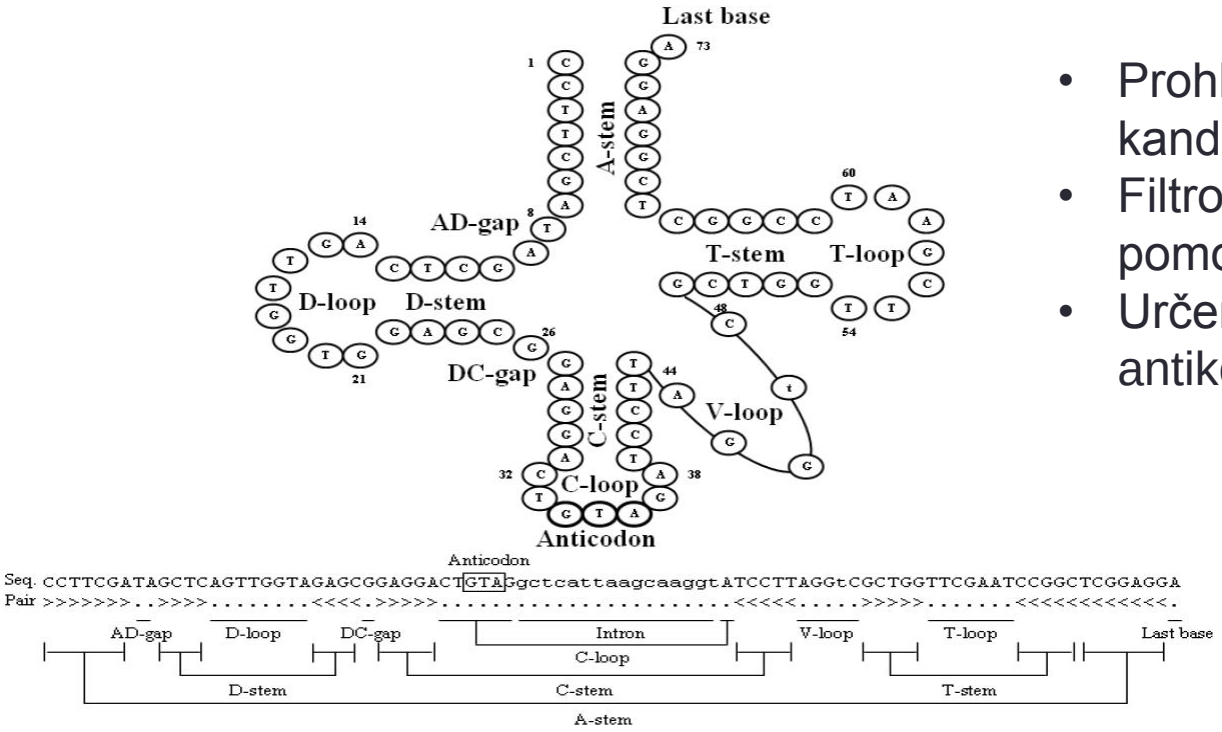
Subsystem Feature Counts

- Cofactors, Vitamins, Prosthetic Groups, Pigments (235)
- Cell Wall and Capsule (113)
- Virulence, Disease and Defense (50)
- Potassium metabolism (16)
- Photosynthesis (58)
  - Light-harvesting complexes (19)
  - Photosynthesis - no subcategory (0)
  - Electron transport and photophosphorylation (39)
    - Photosystem II (24)
    - Photosystem I (15)
- Miscellaneous (22)
- Phages, Prophages, Transposable elements, Plasmids (1)
- Membrane Transport (56)
- Iron acquisition and metabolism (5)
- RNA Metabolism (112)
- Nucleosides and Nucleotides (65)
- Protein Metabolism (243)
- Cell Division and Cell Cycle (42)
- Motility and Chemotaxis (0)
- Regulation and Cell signaling (51)
- Secondary Metabolism (8)
- DNA Metabolism (122)
- Fatty Acids, Lipids, and Isoprenoids (82)
- Nitrogen Metabolism (33)
- Dormancy and Sporulation (3)
- Respiration (82)
- Stress Response (98)
- Metabolism of Aromatic Compounds (2)
- Amino Acids and Derivatives (257)
- Sulfur Metabolism (23)
- Phosphorus Metabolism (50)
- Carbohydrates (203)

Subsystém – soubor genů se společnou funkcí

Genome	Synechococcus elongatus SP08
Domain	Bacteria
Taxonomy	Bacteria; Synechococcus elongatus SP08
Neighbors	View closest neighbors
Size	4,331,368 bp
Number of Contigs (with PEGs)	118
Number of Subsystems	343
Number of Coding Sequences	4598
Number of RNAs	50

Predikce tRNA

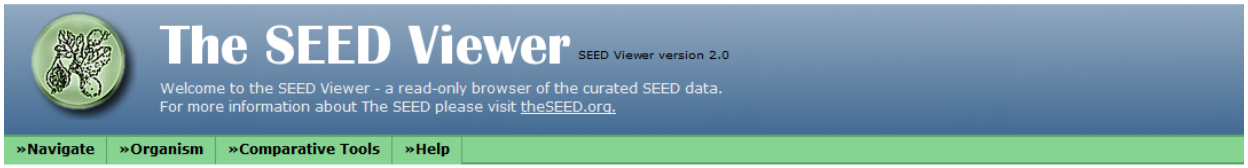


- Prohledání sekvence pro kandidátní tRNA
- Filtrování kandidátních tRNA pomocí sekundární struktur
- Určení typu tRNA podle antikodonu

tRNA -scan
Lowe & Eddy (1997)
http://lowelab.ucsc.edu/tRNAscan-SE/

Výsledek – vizualizace anotace

Vizualizace – genome browser



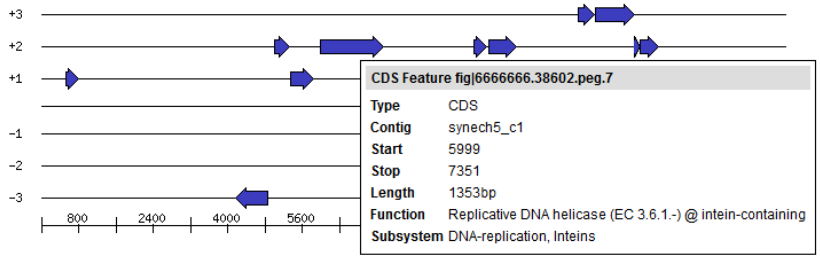
http://www.theseed.org/wiki/Main_Page

Browse Genome: Synechococcus elongatus SP08 (6666666.38602)

Location | Focus | Upload List

contig	synech5_c1 (136,391 bp)
start base	0
window	16,000 bp
Color features	by focus

<== draw ==>



CDS Feature fig|6666666.38602.peg.7

Type	CDS
Contig	synech5_c1
Start	5999
Stop	7351
Length	1353bp
Function	Replicative DNA helicase (EC 3.6.1.-) @ intein-containing
Subsystem	DNA-replication, Inteins

export table | clear all filters

display 15 items per page

displaying 1 - 15 of 4166 next» last»

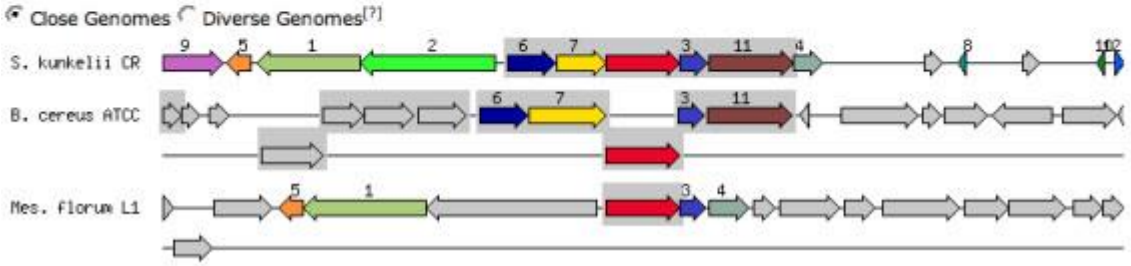
Feature ID	Type	Contig	Start	Stop	Length (bp)	Function	Subsystems	Region
	all	all	<	<	<			
fig|6666666.38602.peg.1	CDS	synech5_c1	523	810	288	transposase	- none -	show
fig|6666666.38602.peg.4	CDS	synech5_c1	4862	4182	681	NADH dehydrogenase (EC 1.6.99.3)	Respiratory dehydrogenases 1, Riboflavin synthesis cluster	show
fig|6666666.38602.peg.5	CDS	synech5_c1	4997	5320	324	Transcriptional regulator, ArsR family	- none -	show
fig|6666666.38602.peg.6	CDS	synech5_c1	5350	5853	504	LSU ribosomal protein L9p	Ribosome LSU bacterial	show
fig|6666666.38602.peg.7	CDS	synech5_c1	5999	7351	1353	Replicative DNA helicase (EC 3.6.1.-) @ intein-containing	DNA-replication, Inteins	show
fig|6666666.38602.peg.8	CDS	synech5_c1	9197	7416	1782	Fibronectin/fibrinogen-binding protein	- none -	show
fig|6666666.38602.peg.9	CDS	synech5_c1	9293	9565	273	FIG003307: hypothetical protein	- none -	show
fig|6666666.38602.peg.10	CDS	synech5_c1	9608	10204	597	Guanylate kinase (EC 2.7.4.8)	Purine conversions	show
fig|6666666.38602.peg.11	CDS	synech5_c1	10489	10806	318	hypothetical protein	- none -	show
fig|6666666.38602.peg.12	CDS	synech5_c1	11535	11873	339	hypothetical protein	- none -	show
fig|6666666.38602.peg.13	CDS	synech5_c1	11892	12740	849	hypothetical protein	- none -	show
fig|6666666.38602.peg.14	CDS	synech5_c1	12737	12853	117	hypothetical protein	- none -	show
fig|6666666.38602.peg.15	CDS	synech5_c1	12857	13264	408	VapC toxin protein	Toxin-antitoxin replicon stabilization systems	show
fig|6666666.38602.peg.16	CDS	synech5_c1	14123	13521	603	hypothetical protein	- none -	show

Porovnání regionů genomů

Protein	fig|273035.4.peg.1008		
Function	Cysteine desulfurase (EC 2.8.1.7)	**EC Number**	2.8.1.7
Organism	Spiroplasma kunkelii CR2-3x	**Taxonomy ID**	273035

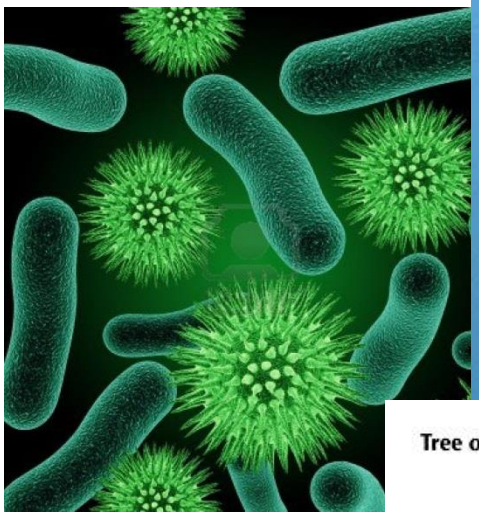
This page offers access to the data relating to the protein encoded by a gene. Each protein implements one or more functional roles, which are themselves components of cellular subsystems (e.g., pathways or complexes).

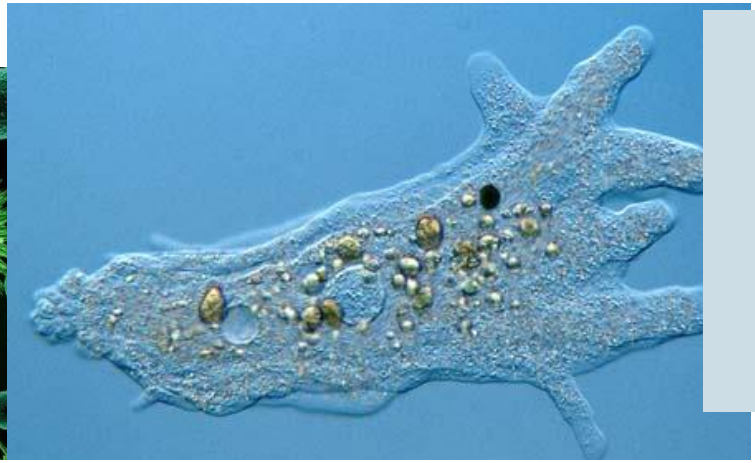
Genomic Context and Compared Regions[?]

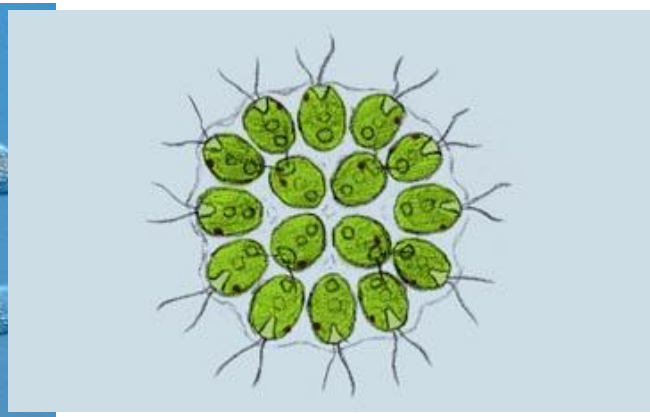


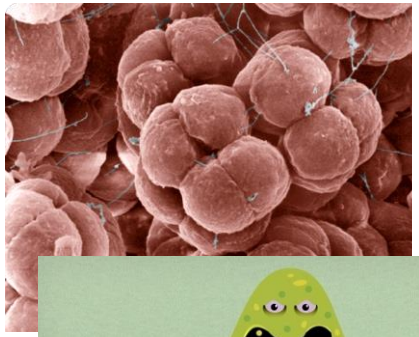
http://www.theseed.org/wiki/Main_Page

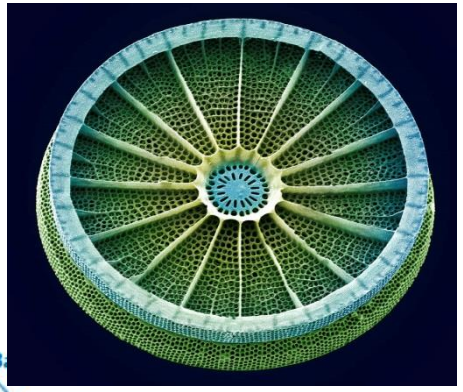
Co je to mikroorganismus?

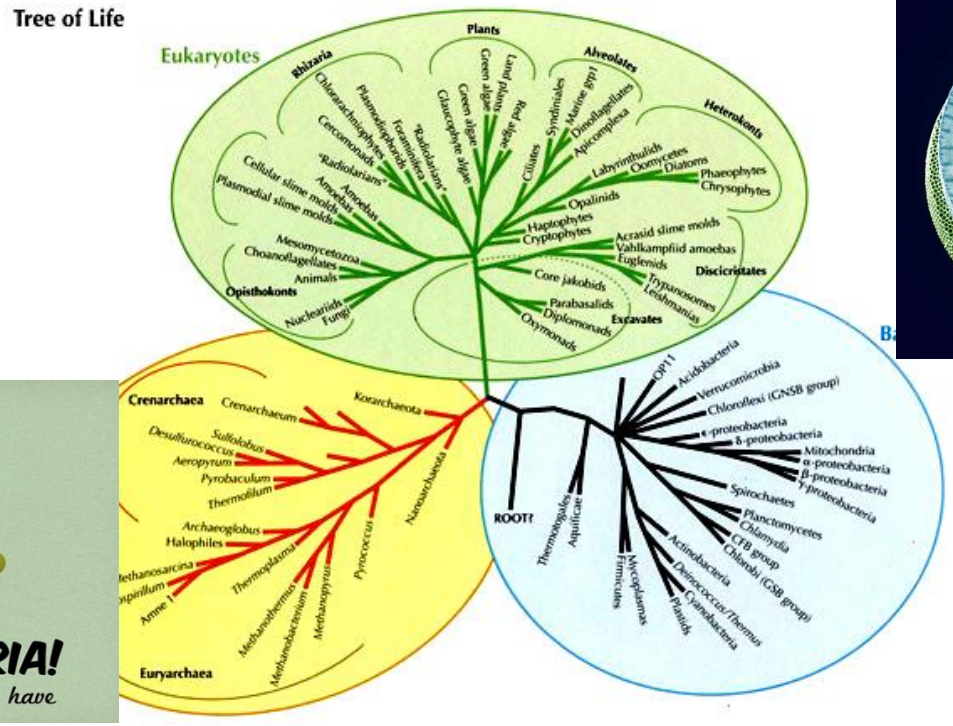








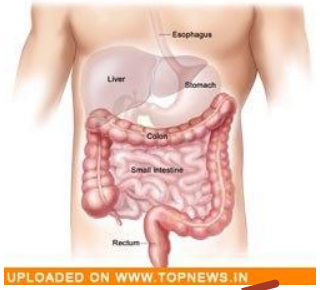




Kde všude najdeme mikroorganismy?

Horké prameny

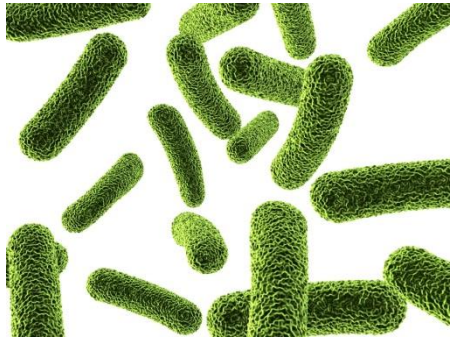
Trávicí trakt



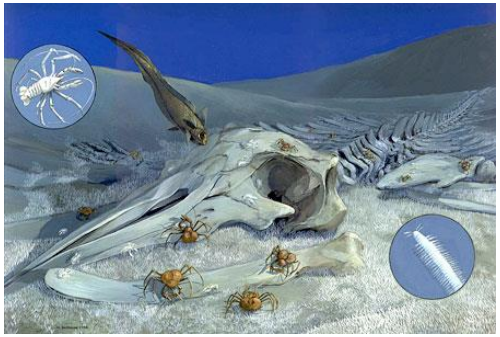
Sladká voda

Atmosféra

Slaná voda



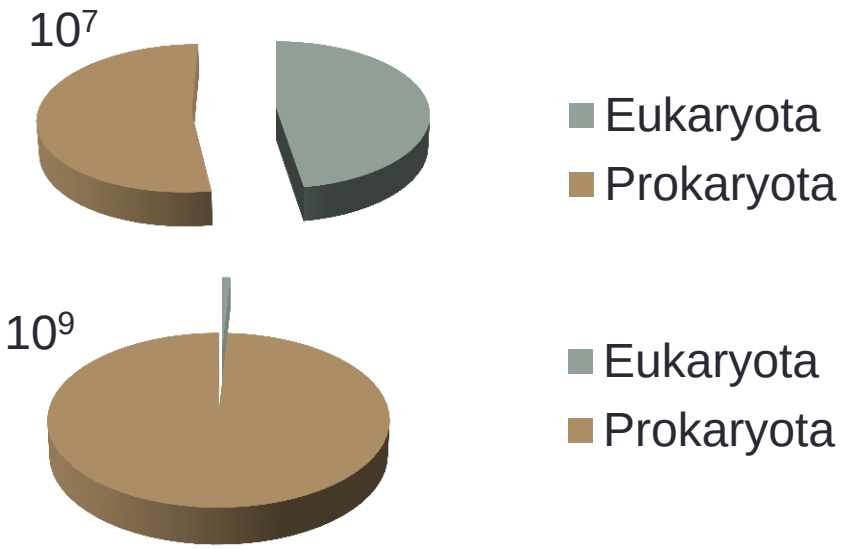
Mrtvá velryba

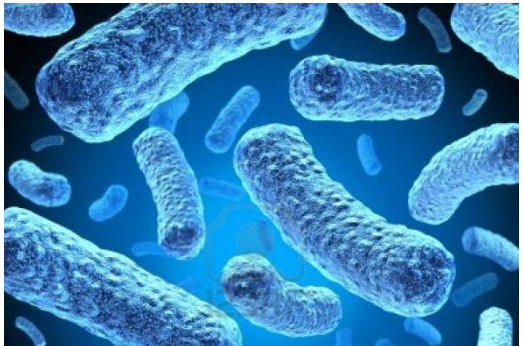


Půda

Kolik druhů mikroorganismů existuje?

- 8,7 miliónů druhů eukaryot (http://www.nature.com/news/2011/110823/full/news.2011.498.html)
  - 20% popsáno
- 10^7 až 10^9 druhů prokaryot (Curtis et al. 2002,Dykhuizen 1998) a 10^{30} buněk (Whittman et al. 1998)





Ale…

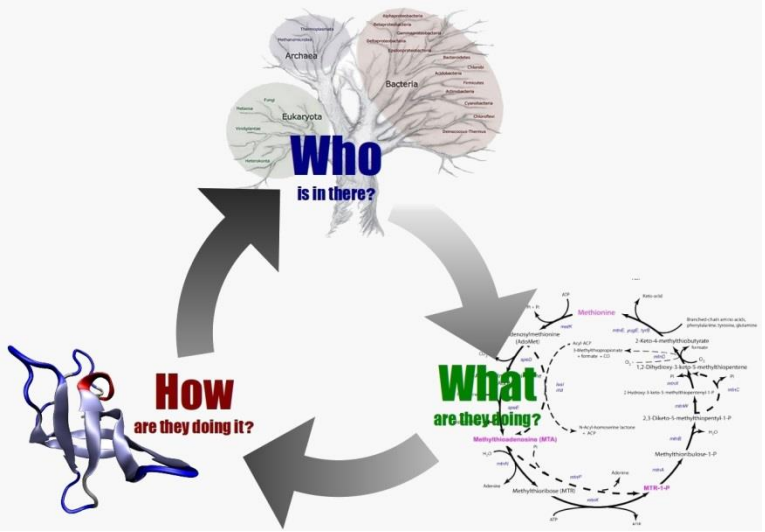
- 90% mikroorganismů se nedá konvenčními způsoby kultivovat

- Jak tedy popsat druhovou či metabolickou diverzitu mikroorganismů???

?

Odpověd: metagenomika

- Kdo?
- Co?
- Jak?

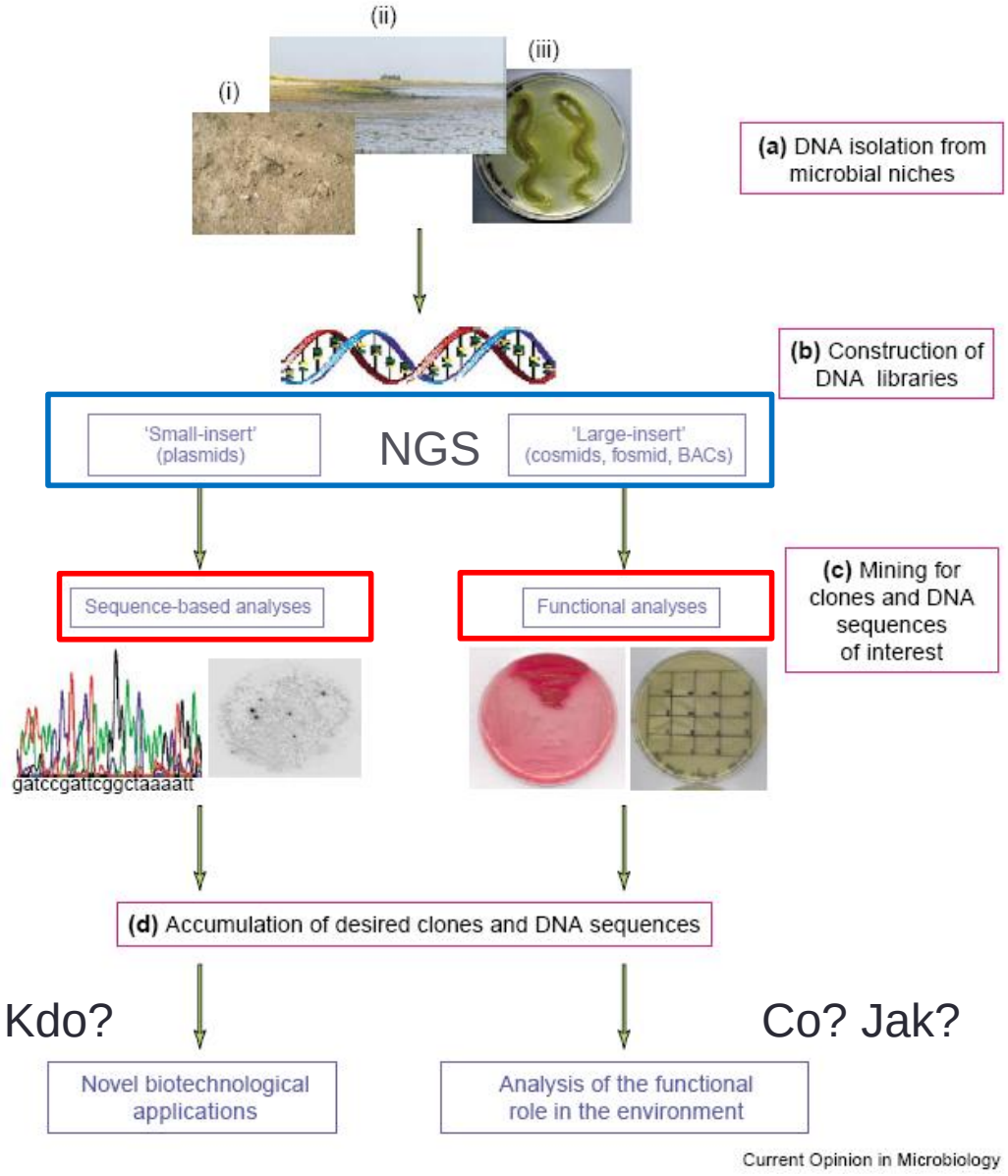


http://www.cbs.dtu.dk/researchgroups/metagenomics/

METAGENOMICS Gilbert & Dupont (2010) – definice

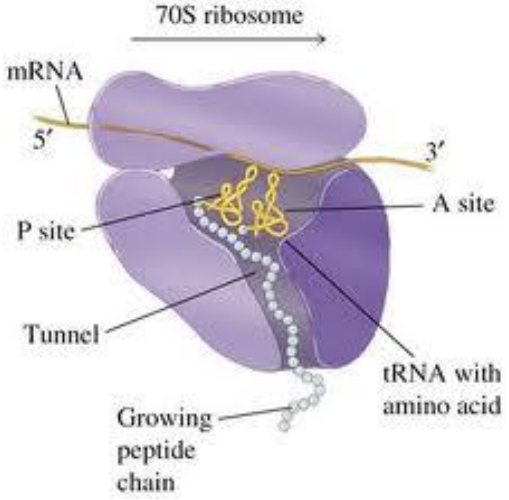
The basic definition of metagenomics is the analysis of genomic DNA from a whole community; this separates it from genomics, which is the analysis of genomic DNA from an individual organism or cell. In fact, the most appropriate translation of meta in Greek is "beyond," and hence the term literally means "beyond the single genome study." The term was first published in 1998 in a study of soil microbes using random cloning of environmental DNA (Handelsman et al. 1998). Subsequently, definitions have varied to include any study whereby a whole community is analyzed, e.g., directed studies of 16S rDNA diversity from an environment to isolation and analysis of total DNA from environmental samples without prior cultivation (Chen & Pachter 2005). It could be argued that prior cultivation of communities, in the case of enrichment studies or community cell-encapsulation cultures, can also be analyzed using metagenomics, and hence such definitions must be kept broad.

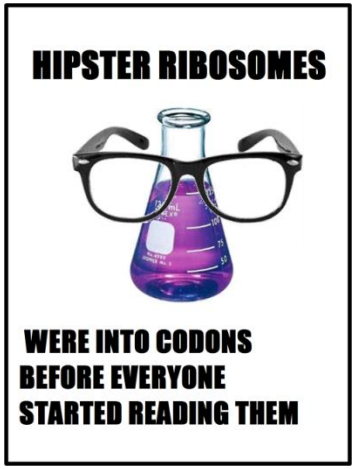
Dva přístupy v metagenomice



Molekulární markery používané v metagenomice

- Prokaryota – 16S rRNA, mcrA
- Eukaryota – 18S rRNA, 28S rRNA, COI, rbcL
- 16S rRNA, 18S rRNA – ribozomální RNA
  - Podléhá málo mutacím – základ metabolismu
  - V každém organismu
  - Nekóduje protein – často se vyskytují mezery





RIBOSOMAL DATABASE PROJECT

RDP Release 10, Update 32 :: May 14, 2013 :: 2,765,278 16S rRNAs The Ribosomal Database Project (RDP) provides ribosome related data and services to the scientific community, including online data analysis and aligned and annotated Bacterial and Archaeal small-subunit 16S rRNA sequences.

Cite RDP's NAR article

RDP Release 10 brings two major changes to the RDP:

- RDP10 provides new **Bacterial** and **Archaeal** alignments with several significant enhancements over the previous RDP 9 alignments.
- Use of the ***Infernal*** secondary-structure based aligner that provides better support for short partial sequences and handles certain sequencing artifacts in a more intuitive manner.

Explore our online analysis tools:

- myRDP
- BROWSERS
- CLASSIFIER
- LIB COMPARE
- SEQ MATCH
- PROBE MATCH
- TREE BUILDER
- PYROSEQUENCING PIPELINE
- ASSIGNMENT GENERATOR
- TAXOMATIC
- RDP_MIMARKS

HOVER over any tool item in the menu to see a brief popup description of its features;

CLICK on the tool menu item to begin working with it.

Try our NEW PROCEDURAL TUTORIALS to use our site to your fullest advantage.

Be sure to view the video tutorials and visit each tool's help file if needed.

Sponsors:

DOE Office of Biological and Environmental Research

NIEHS Superfund Research Program

myRDP

RDP News

10/09/2013 FunGene article published
The article describing our FunGene data and tools is published in Frontier in Microbiolgy.

10/09/2013 RDP FrameBot article published
The article describing RDP FrameBot (a frameshift correction tool) is published in the journal mBio

10/01/2013 RDP Staff and Poster
5th Argonne Soil Metagenomics Meeting

09/25/2013 Campus internet interruptions
RDP is back online.

08/16/2013 Power outage alert
RDP sites are now back online

06/06/2013 Amplicon chimera checking with uchime
The functional gene pipeline now offers a tool to check amplicon sequencing datasets for chimeras powered by

05/14/2013 RDP 10, update 32 released

Hierarchy View:

rootrank Root (0/1258845/0) (selected/total/search matches) [options]

- domain Bacteria (0/1235311/0)
  - phylum "Actinobacteria" (0/180827/0)
  - phylum "Aquificae" (0/945/0)
  - phylum "Bacteroidetes" (0/141057/0)
  - phylum "Caldiserica" (0/220/0)
  - phylum "Chlamydiae" (0/504/0)
  - phylum "Chlorobi" (0/1058/0)
  - phylum "Chloroflexi" (0/20447/0)
  - phylum "Chrysiogenetes" (0/12/0)
  - phylum "Deferribacteres" (0/378/0)
  - phylum "Deinococcus-Thermus" (0/1844/0)
  - phylum "Dictyoglomi" (0/22/0)
  - phylum "Elusimicrobia" (0/172/0)
  - phylum "Fibrobacteres" (0/308/0)
  - phylum "Fusobacteria" (0/9378/0)
  - phylum "Gemmatimonadetes" (0/1256/0)
  - phylum "Lentisphaerae" (0/1707/0)
  - phylum "Nitrospira" (0/1433/0)
  - phylum "Planctomycetes" (0/11233/0)
  - phylum "Proteobacteria" (0/341501/0)
  - phylum "Spirochaetes" (0/9606/0)
  - phylum "Synergistetes" (0/1120/0)
  - phylum "Tenericutes" (0/3214/0)
  - phylum "Thermodesulfobacteria" (0/107/0)
  - phylum "Thermotogae" (0/582/0)
  - phylum BRC1 (0/399/0)
  - phylum OD1 (0/153/0)
  - phylum OP11 (0/97/0)
  - phylum SR1 (0/234/0)
  - phylum TM7 (0/2167/0)
  - phylum WS3 (0/529/0)
  - phylum "Armatimonadetes" (0/1016/0)
  - phylum "Verrucomicrobia" (0/9586/0)
  - phylum "Acidobacteria" (0/14151/0)
  - phylum Firmicutes (0/420503/0)
  - phylum Cyanobacteria/Chloroplast (0/21330/0)
  - ▶ Archaea Outgroup (0/1/0)
  - ▶ unclassified_Bacteria (0/36214/0)
- domain Archaea (0/23532/0)
  - phylum "Crenarchaeota" (0/7123/0)
  - phylum "Euryarchaeota" (0/12628/0)
  - phylum "Korarchaeota" (0/92/0)
  - phylum "Nanoarchaeota" (0/138/0)
  - phylum "Thaumarchaeota" (0/0/0)
  - ▶ Bacteria Outgroup (0/1/0)
  - ▶ unclassified_Archaea (0/3550/0)
- ▶ unclassified_Root (0/2/0)

silva
comprehensive ribosomal RNA databases

SILVA

Welcome to the SILVA rRNA database project

A comprehensive on-line resource for quality checked and aligned ribosomal RNA sequence data.

SILVA provides comprehensive, quality checked and regularly updated datasets of aligned small (16S/18S, SSU) and large subunit (23S/28S, LSU) ribosomal RNA (rRNA) sequences for all three domains of life (*Bacteria*, *Archaea* and *Eukarya*).

SILVA are the official databases of the software package **ARB**.

For more background information → Click here

ARB

The **software package ARB** represents a graphically-oriented, fully-integrated package of cooperating software tools for handling and analysis of sequence information.

The **ARB project** has been started more than 15 years ago by Wolfgang Ludwig at the Technical University in Munich, Germany, see www.arb-home.de.

The MEGX.net data portal

Visit our partner site www.megx.net, the data portal for **Marine Ecological Genomix**, to get a feeling how your research can be improved using integrated databases.

SILVA Terms of Use/License Information

SILVA uses a **dual licensing model.** In short, browsing and deploying the SILVA database content displayed on the SILVA webportal is free for all (academic and non-academic users) whereas all downloads are only free for academic users. Both, academic and non-academic users should have a look at the → SILVA Terms of Use/License Information

News

23.08.2013

SILVA 115 released

A nearly endless trip to get release 115 done. Highlights: Improved taxonomy, especially for the Eukaryotes, and ALL sequences in SILVA are now classified based on the SILVA taxonomy!

16.06.2013

Preview SILVA Release 115 Statistics

More than 4 Mio SSU and LSU sequences...

06.06.2013

Working towards SILVA release 115

Preparation of SILVA release 115 has started. SILVA 115 will be a full release with updated taxonomy and trees, as well as ARB files. The release is planned for July 2013.

03.03.2013

Meet ARB & SILVA at VAAM 2013

Talk to the ARB and SILVA developers at VAAM 2013 (10.03-13.03) in Bremen, Germany. Follow the link to see the sessions where you will find us.

go to Archive ->

SILVA 115 - full release

	SSU Parc	SSU Ref	SSU Ref NR	LSU Parc	LSU Ref
Minimal length	300	1200/900	1200/900	300	1900
Quality filtering	basic	strong	strong	basic	strong
Guide Tree	no	no	yes	no	yes
Release date	23.08.13	23.08.13	23.08.13	23.08.13	23.08.13
Aligned rRNA sequences	**3,808,884**	1,426,414	479,726	**361,874**	39,412

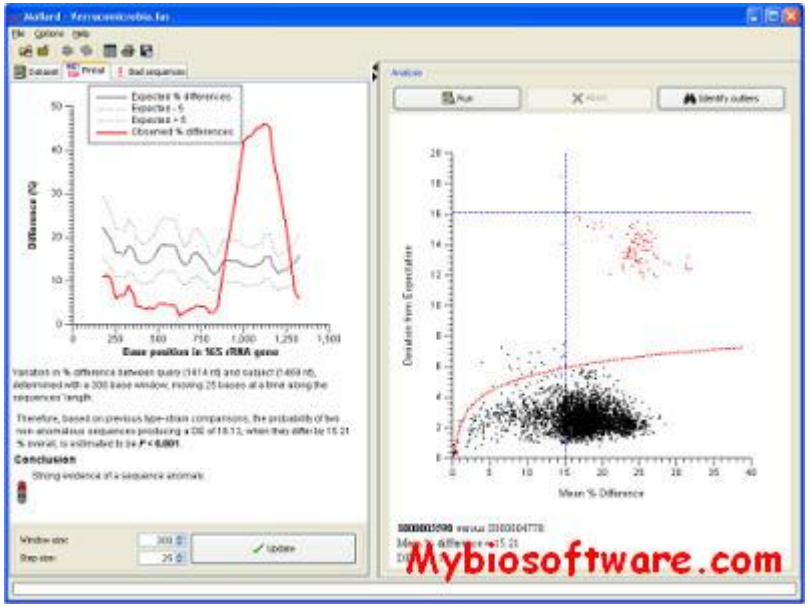
SILVA on Twitter!

Follow us on Twitter to get the latest news ...

Chiméry

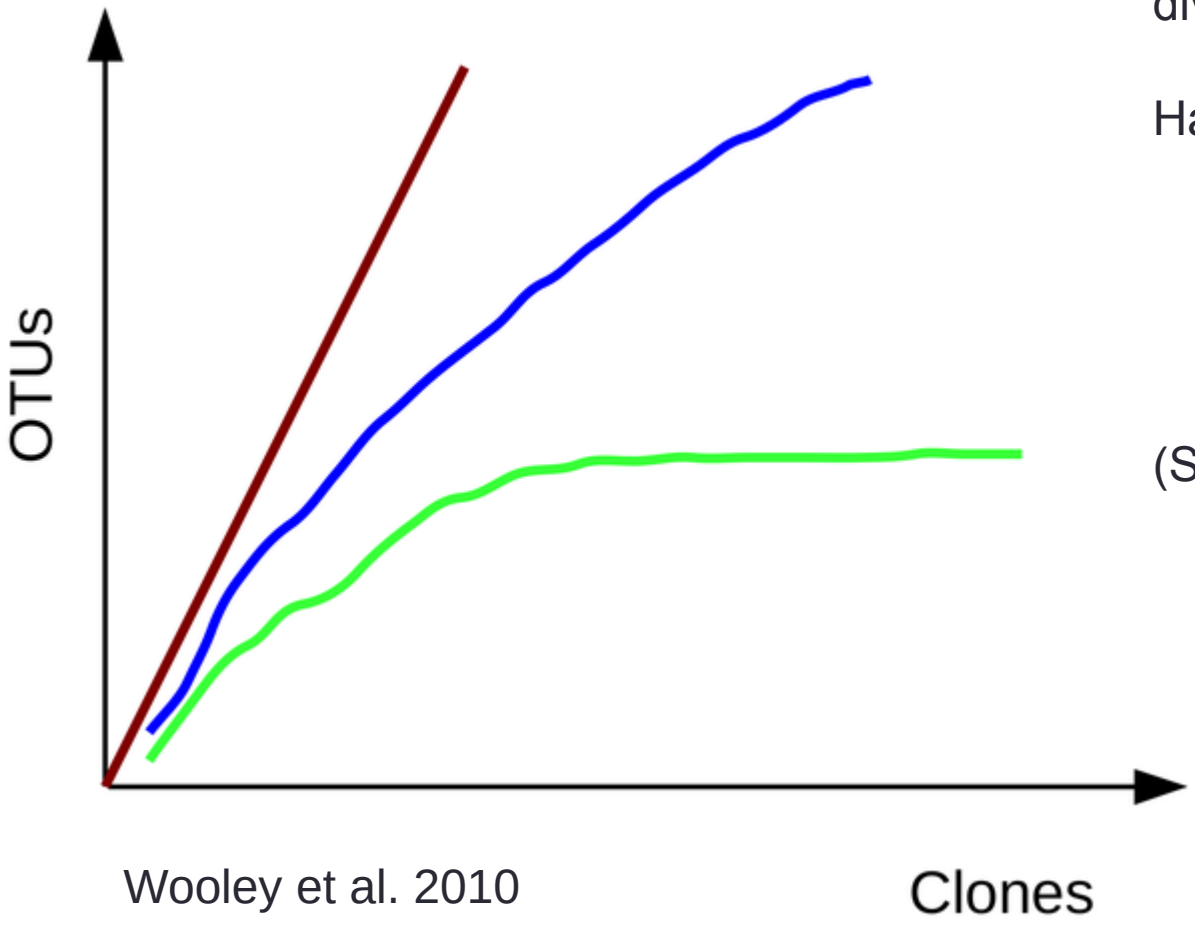
- Artefakty vzniklé při PCR
- Většinou vznikají nekompletní extenzí primerů
- Takto vzniklé prodloužené primery se navazují na různé templáty a vznikají chimerické sekvence
- Další minoritní důvody vzniku:
  - Nízká procesivita polymerázy (málo navázaných bází)
  - Špatně vložené nukleotidy
  - A další….
- Jak odhalit chiméru? Bioinformatika…

Chiméry



Detekce chimér pomocí progra Mallard (Ashelford et al. 2006)

Kolik je dost? Kolik sekvencí je potřeba k popisu společenstva



Wooley et al. 2010

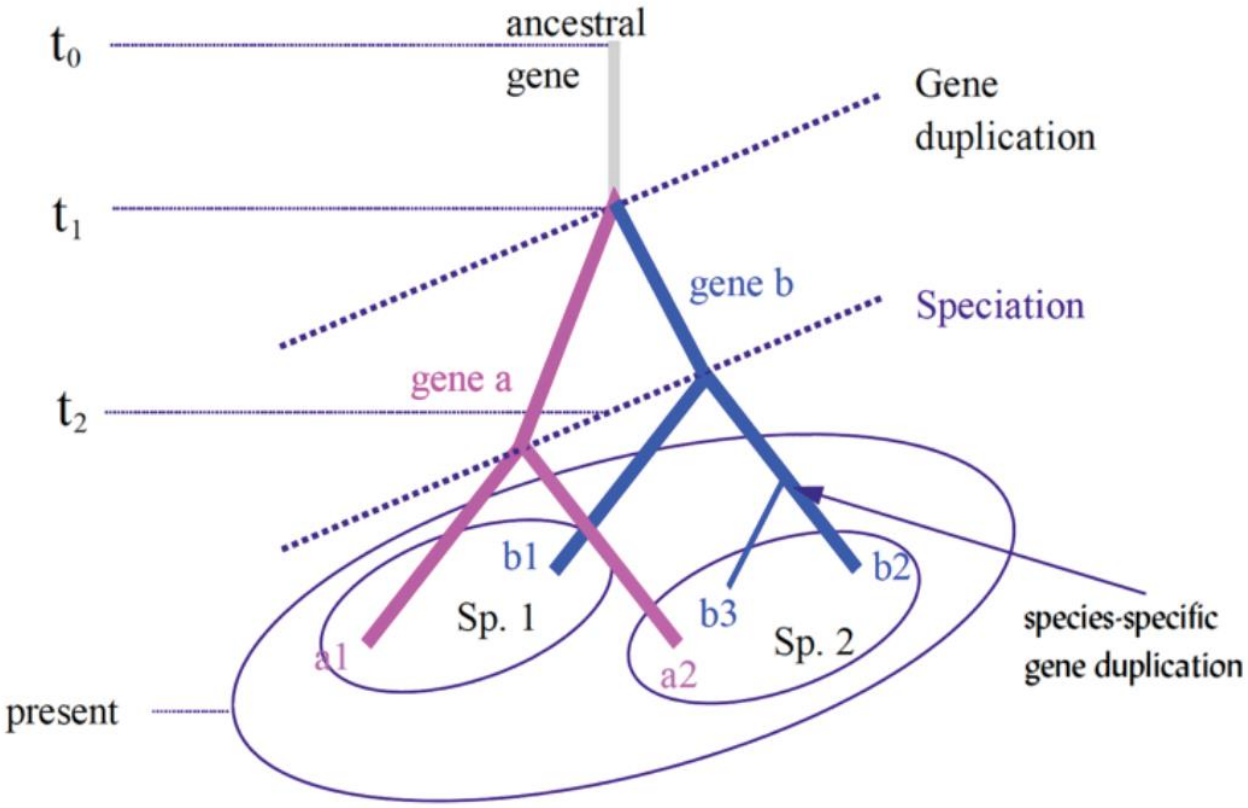
Nedostatečné vzorkování, většina diversity nepopsaná

Habitat nedostatečně vzorkován

(Skoro) všechny druhy popsány

Fundamentální podstata fylogenetiky

- Teorie neutrální evoluce – Kimura (1968)

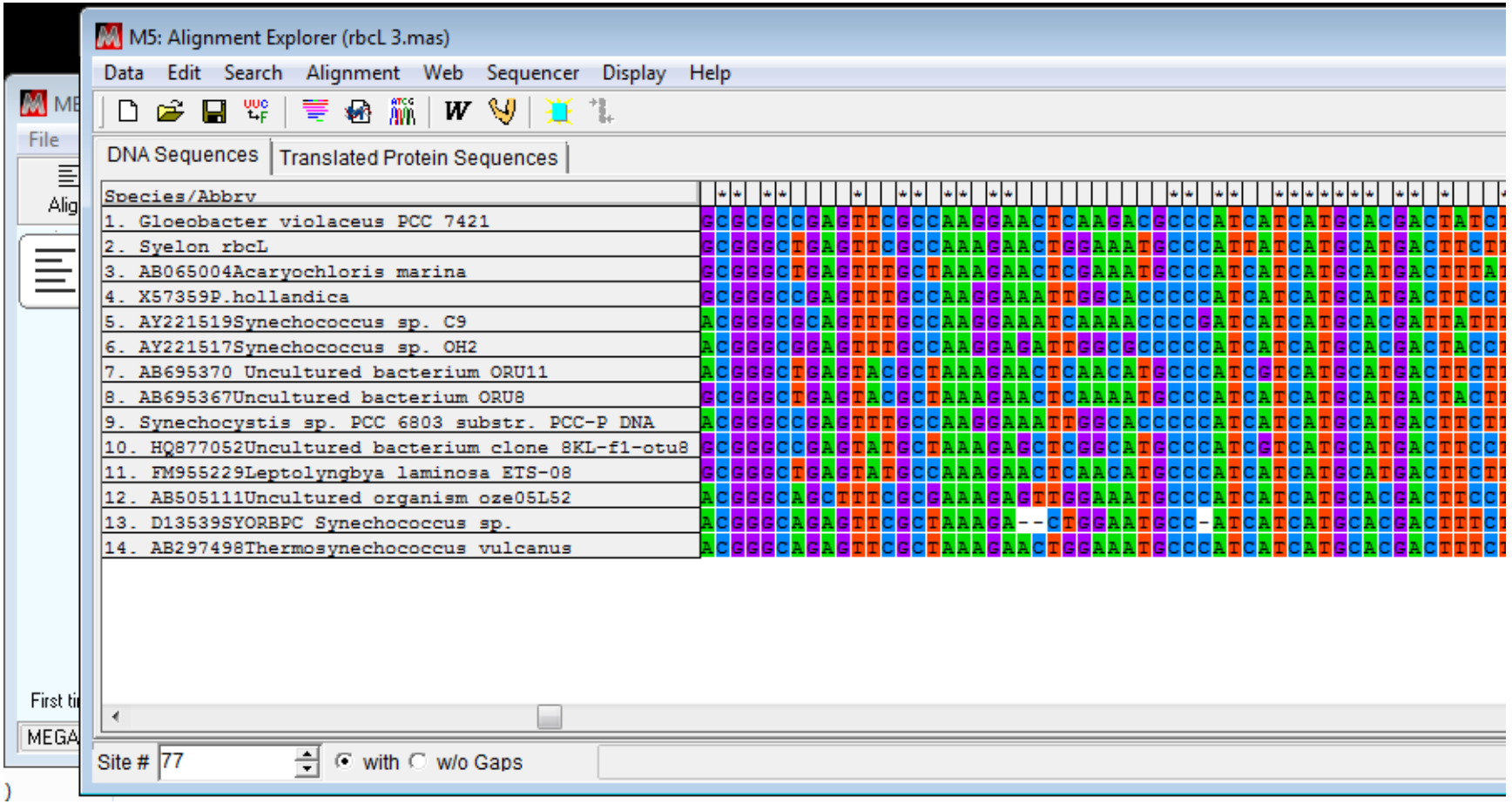


Fundamentální podstata fylogenetiky

- Vznik variability popsatelné molekulárními markery
  - Binární – AFLP, RFLP, DGGE,…
  - **Vícestavové – sekvence DNA,…**
- Předpoklad homologie genů – ortologie x paralogie, homoplasie – konvergentní evoluce

Multiple Sequence alignment

- Uspořádání sekvencí do sloupců



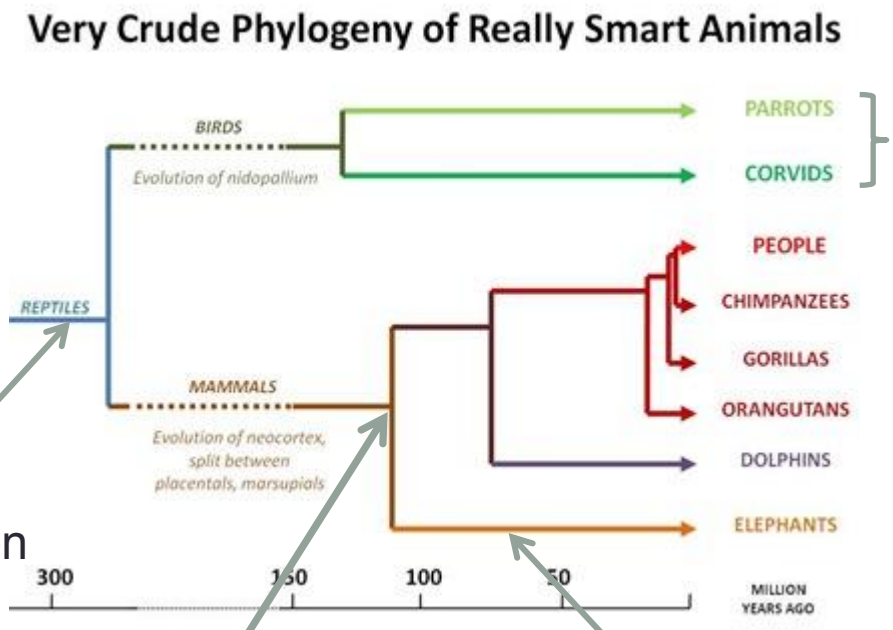
Multiple Sequence alignment algoritmy

- **Progressive alignment construction – ClustalW**
  - Fylogenetický strom pro přesnější vyhledání podobných sekvencí
- Iterative method – Muscle

MUSCLE

  - Podobné, ale je možné se vrátit zpět a alignment score vylepšit
- HMM – Hidden Markov Chain
  - Pravděpodobnostní metoda, přiřazování mezerám a kombinacím bází pravděpodobnosti
- …

Fylogenetický strom



Clade – klad

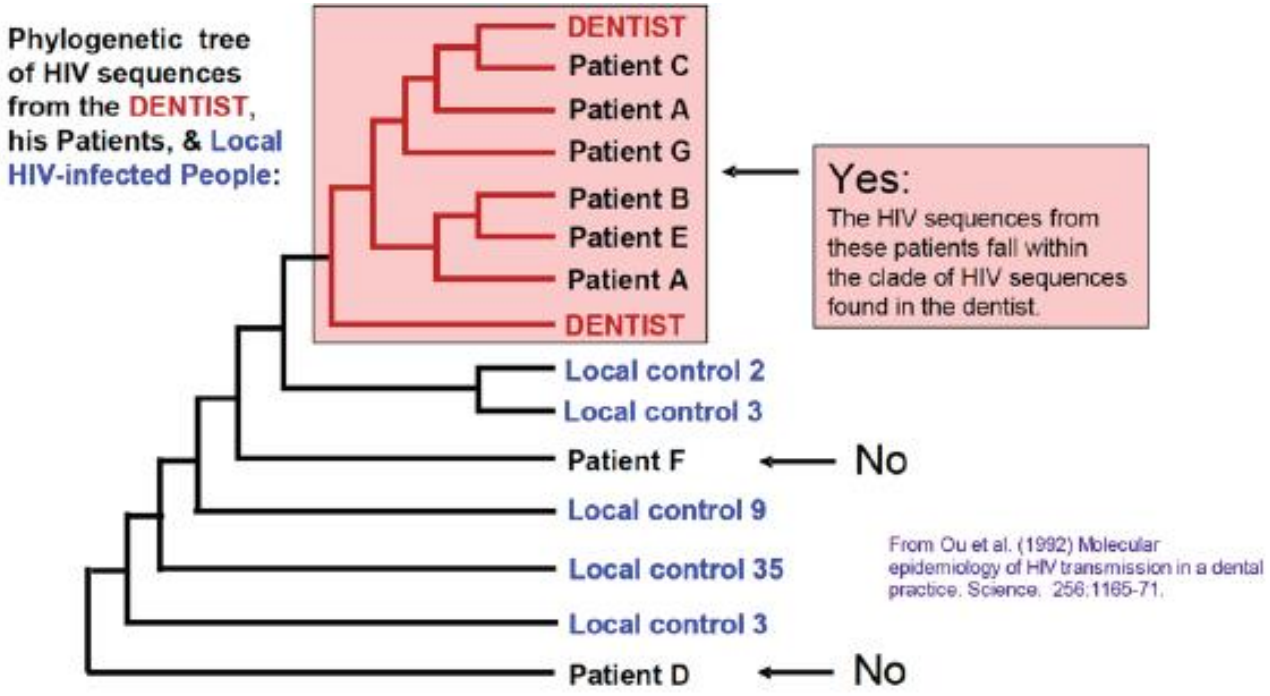
Root – kořen

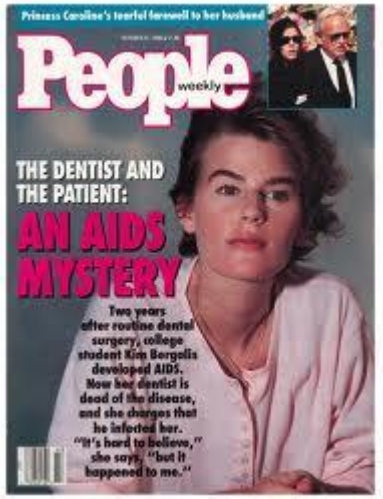
Node – uzel

Branch – větev

http://watchingtheworldwakeup.blogspot.cz/2009_11_01_archive.html

Využití fylogenetického stromu v kriminalistice







METAGENOMIKA II

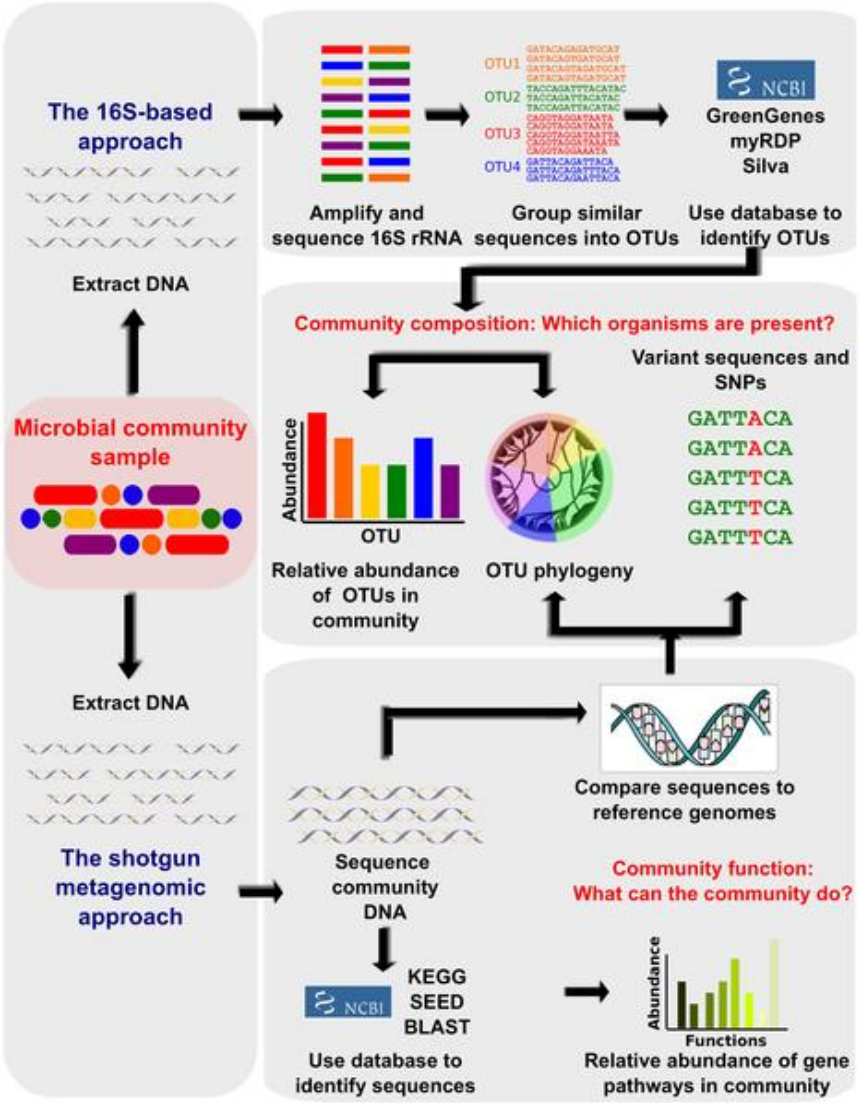
Petr Dvořák

Katedra botaniky PřF UP

„Propojení výuky oborů Molekulární a buněčné biologie a Ochrany a tvorby životního prostředí “

Reg. č.: CZ.1.07/2.2.00/28.0032

Dva přístupy v metagenomice

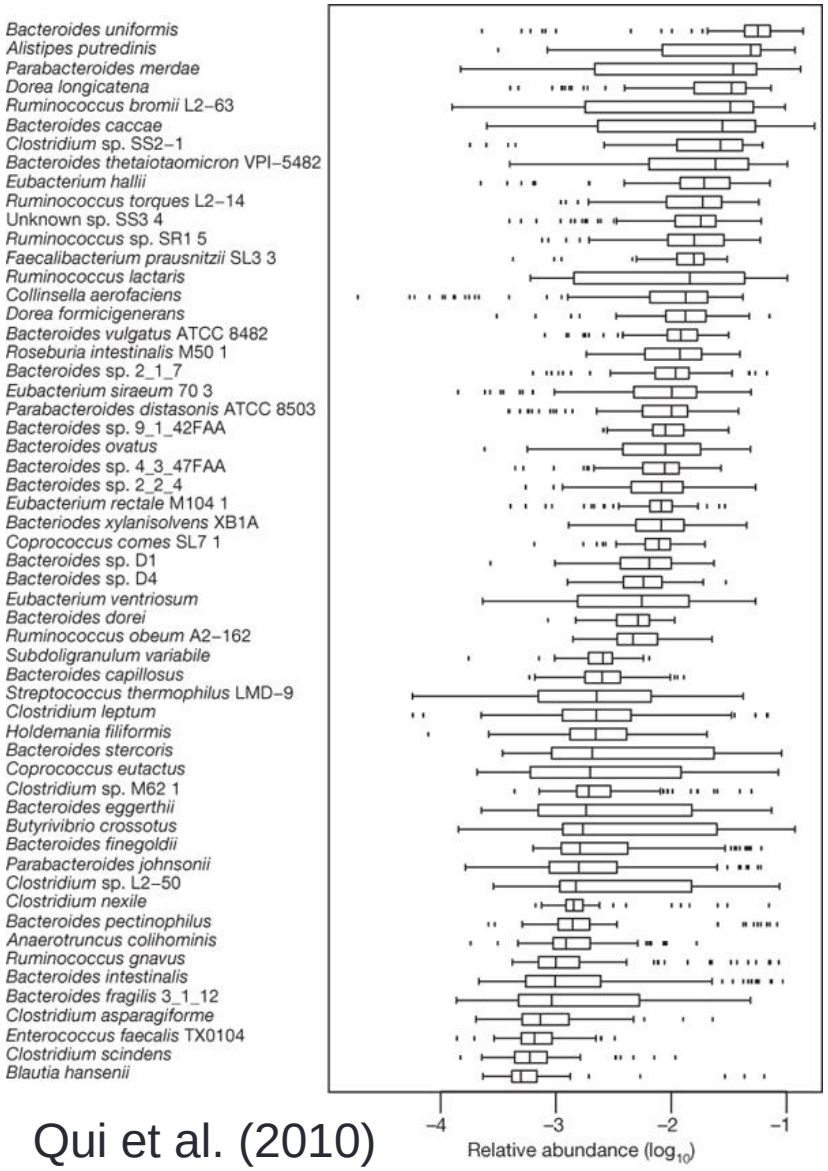


Kdo?

Co? Jak?

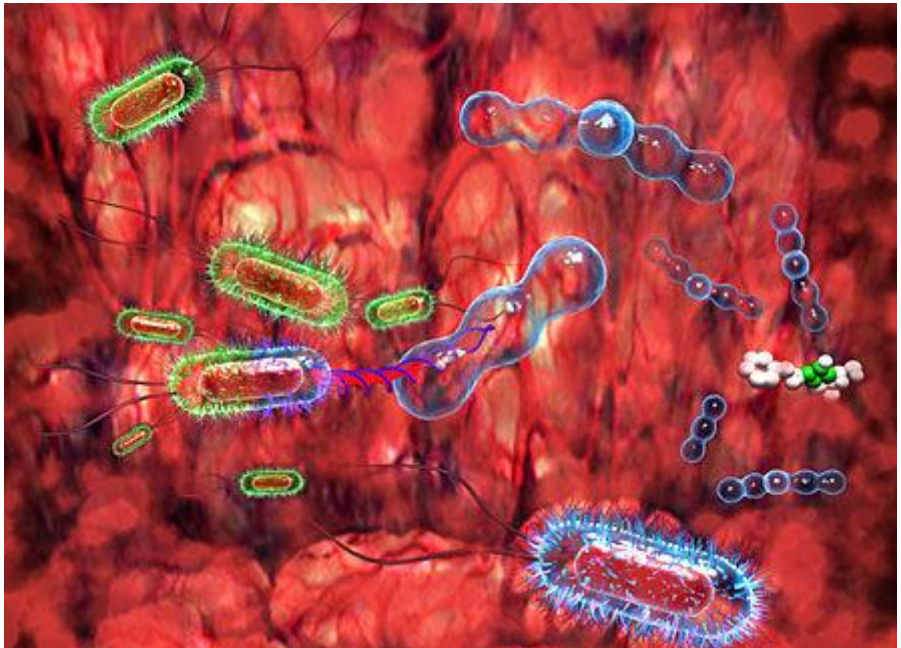
Morgan & Huttenhower (2012)

Analýza střevní mikroflóry



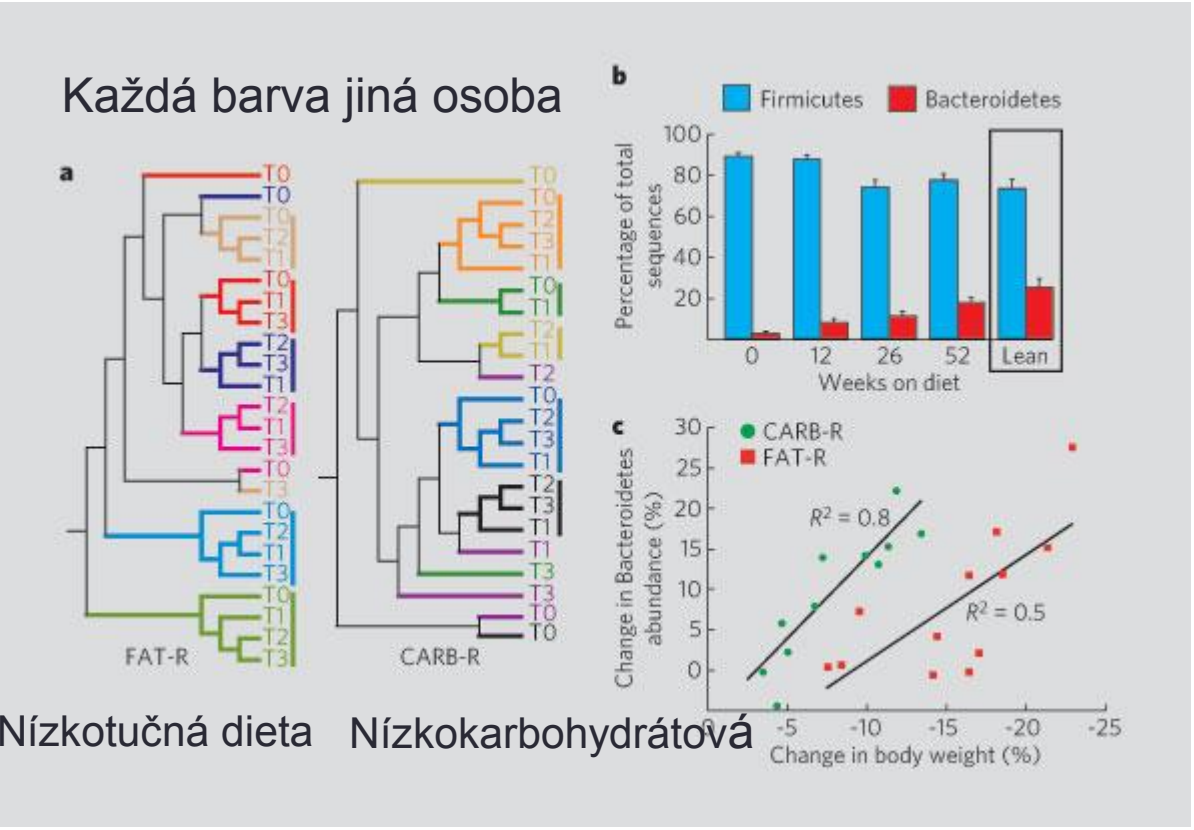
Qui et al. (2010)

- 10^{13} až 10^{14} mikroorganismů
- Obrovská druhová diversita
- 100x více genů než genom člověka
- Pomáhají zpracovat nestravitelné polysacharidy



Změna střevní mikrobioty při hubnutí

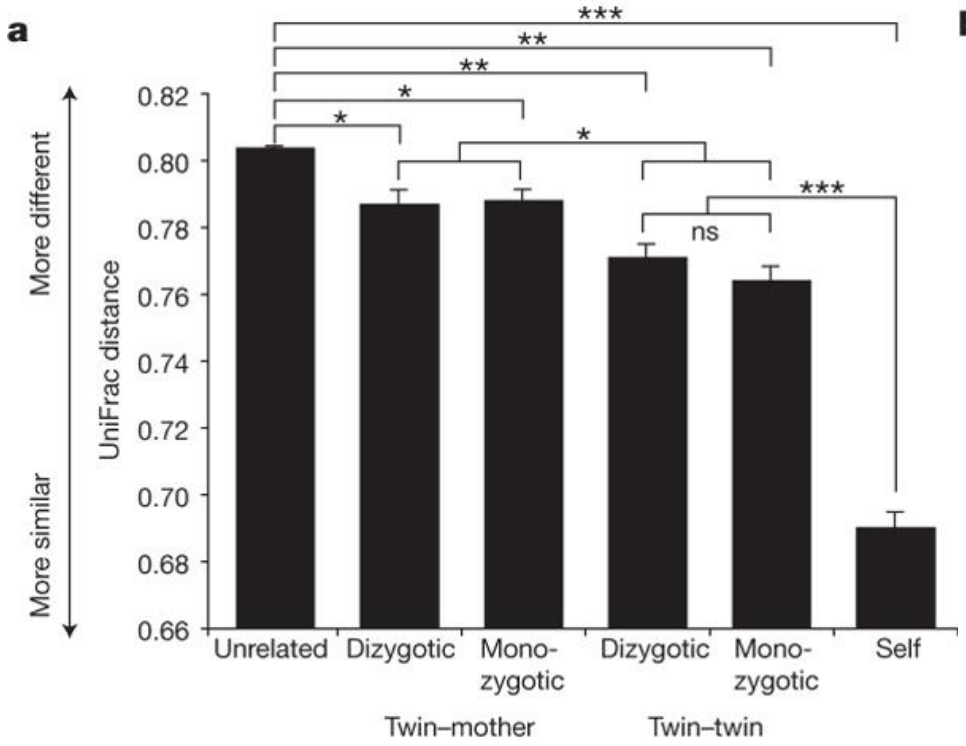
Ley et al. (2006)

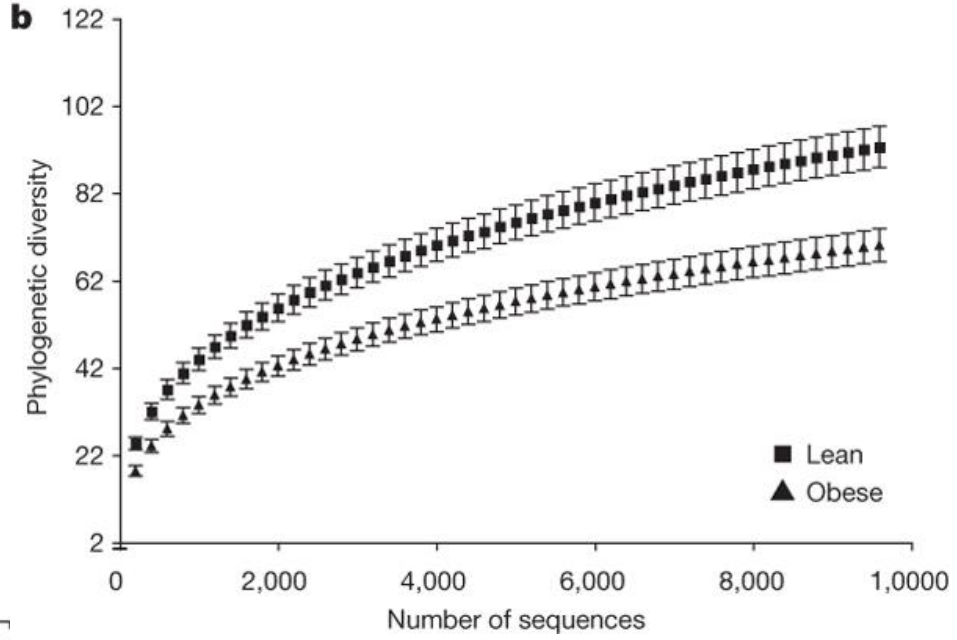


- Mikrobiota specifická pro určitou osobu (a)
- Bacteroidetes převažují při hubnutí

- T0 – začátek experimentu
- T1 – 12 týdnů
- T2 – 26 týdnů
- T3 – 52 týdnů

Mikrobiální střevní diverzita dvojčat





Turnbaugh et al. (2009)

- Největší rozdíl v diverzitě nepříbuzných
- Jednovaječná menší diverzita než dvojvaječná

megx
marine ecological genomics

geographic-blast

result

faq | contact | news

home | tools | browse | edutainment | portal

Sampling Sites

- Prokaryote Genomes
- Metagenomes
- Phage Genomes
- rRNA Sequences

Physicochemical & Biological Layers

Select Layer

Annual

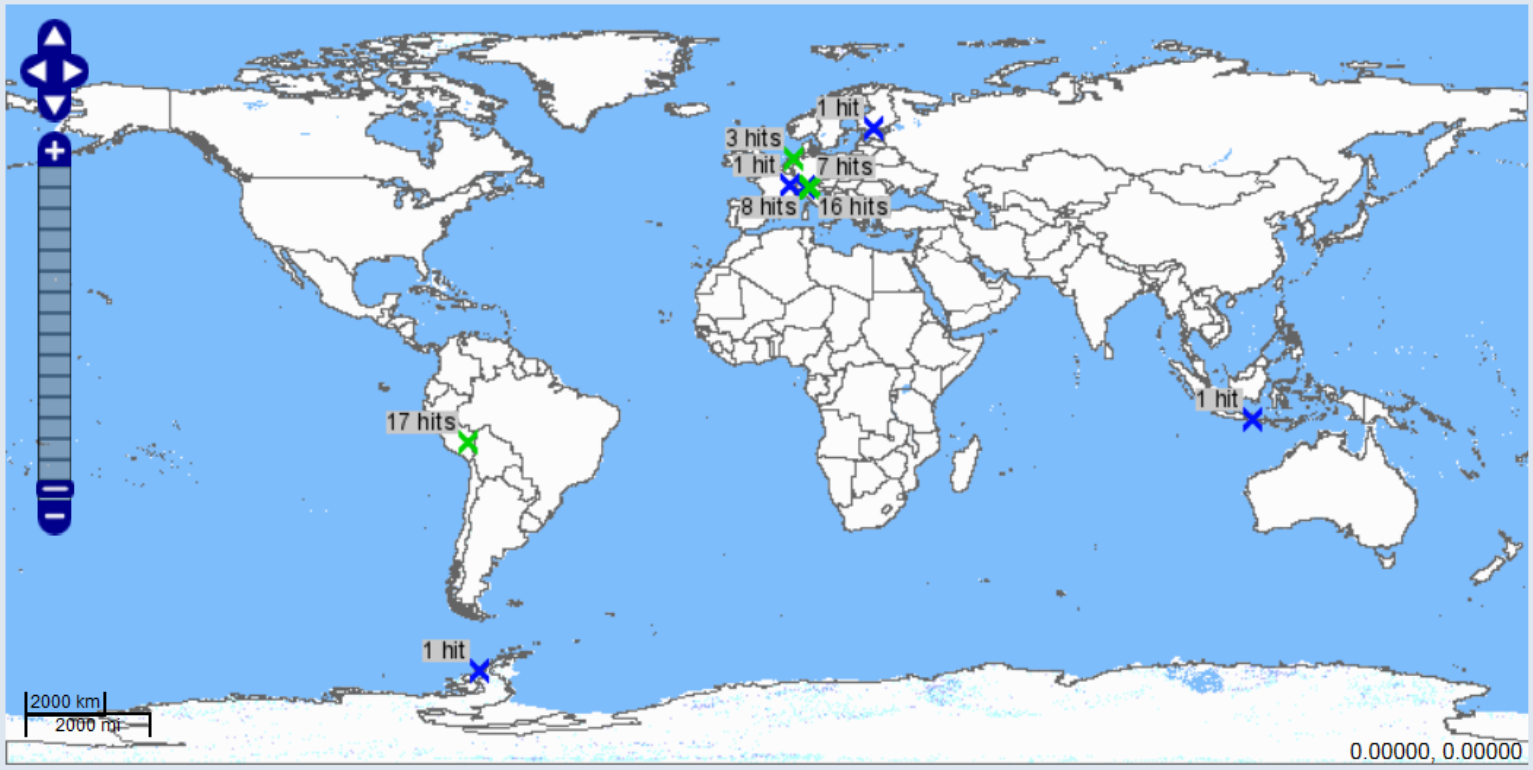
0m

Map it

Base Layers

- Default
- Bathymetry
- Boundaries
- Limits of Oceans & Seas 1953
- Coordinates
- Lakes

Click on a sampling site (dot) to retrieve environmental data!

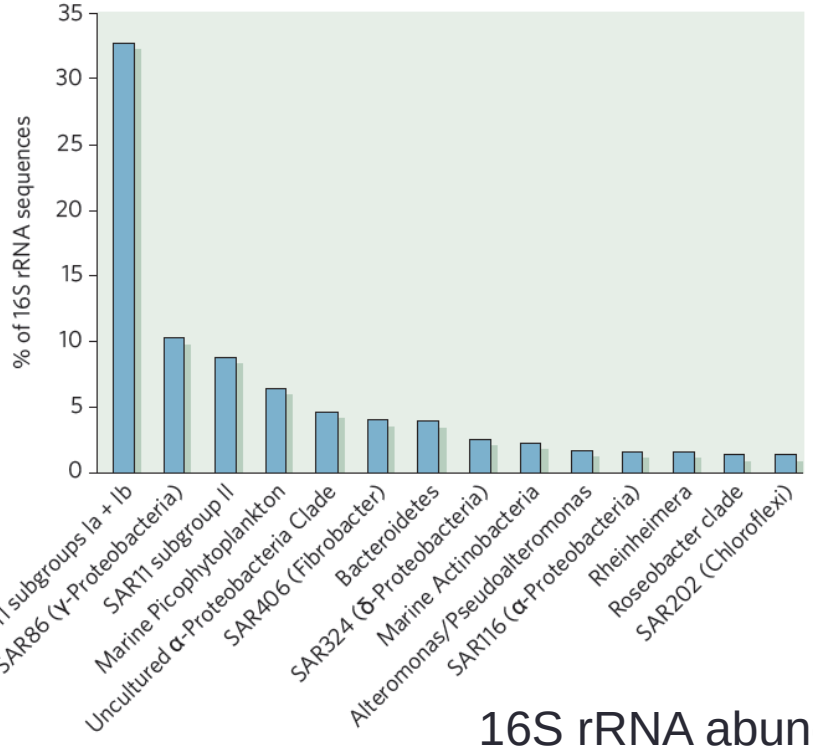


Legend

SAR11 – nejpočetnější organismus

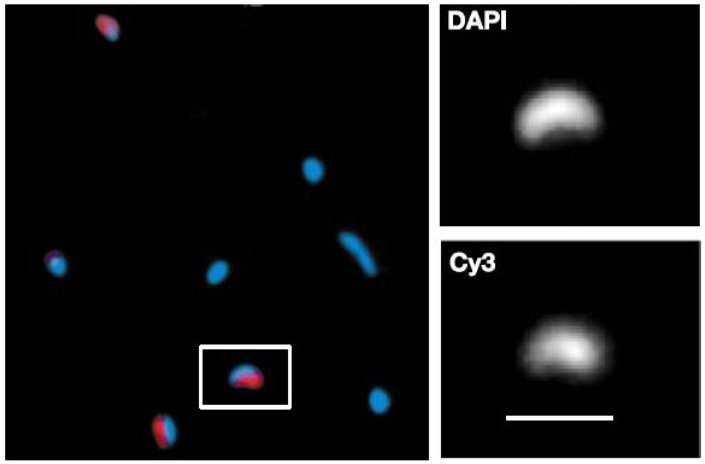
- Nejpočetnější organismus v oceánech
- Nekultivovatelný
- Alfa-proteobakterie
- Často 50% komunity
- Po celém světě

Giovanonni & Stingl (2005)



16S rRNA abundance

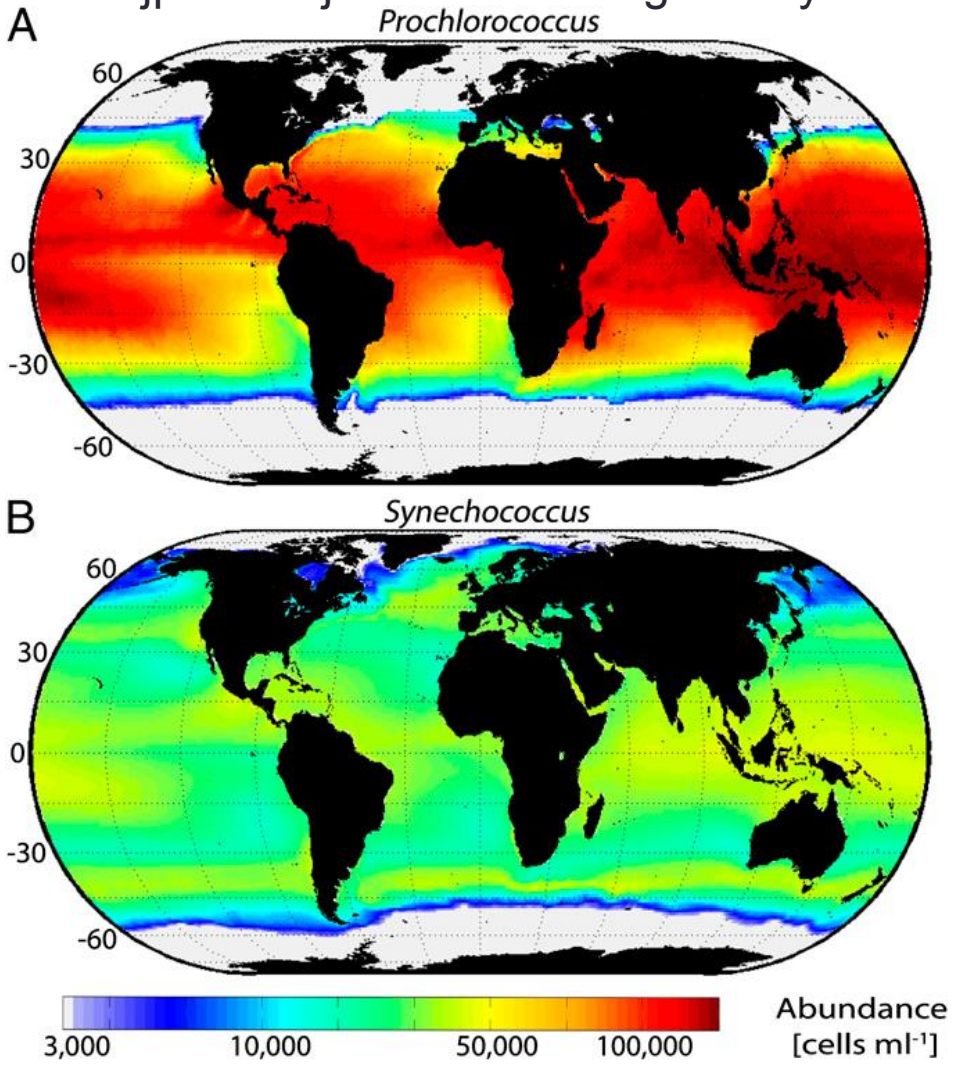
Morris et al. (2002)

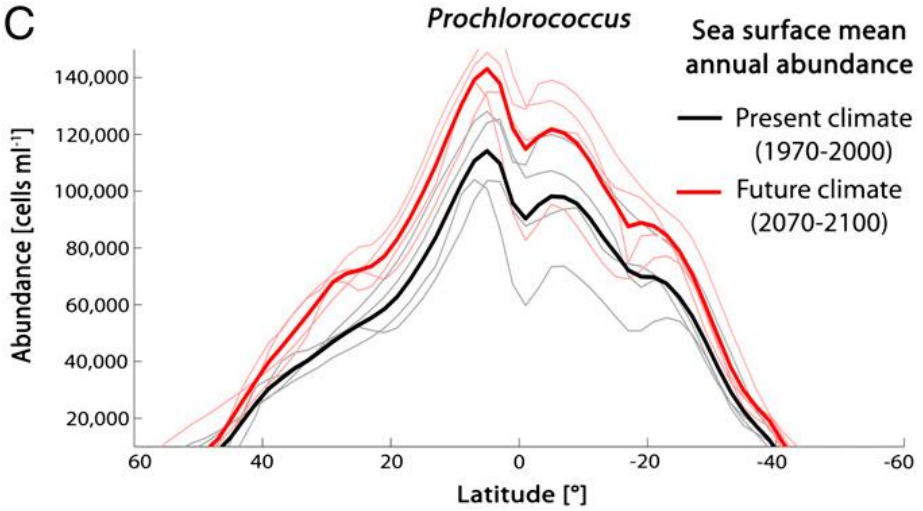


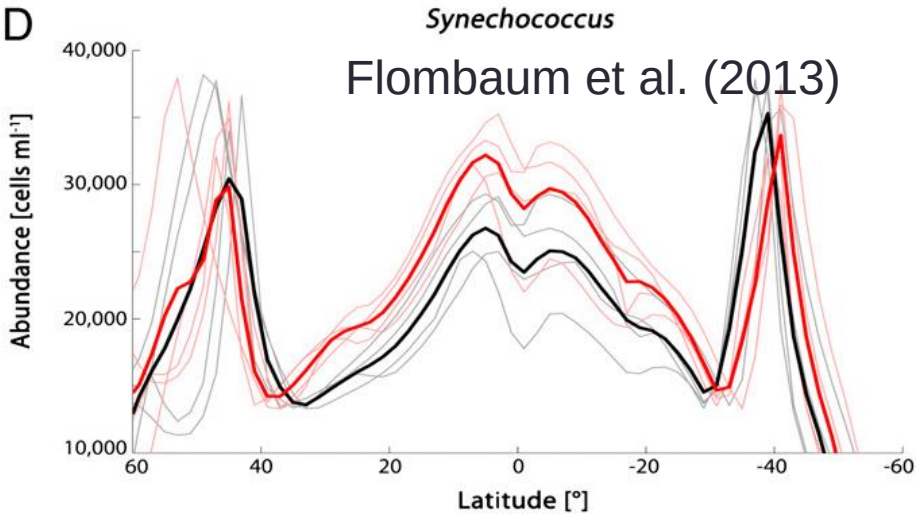
FISH - Cy3 barví 16S rRNA

Synechococcus/Prochlorococcus

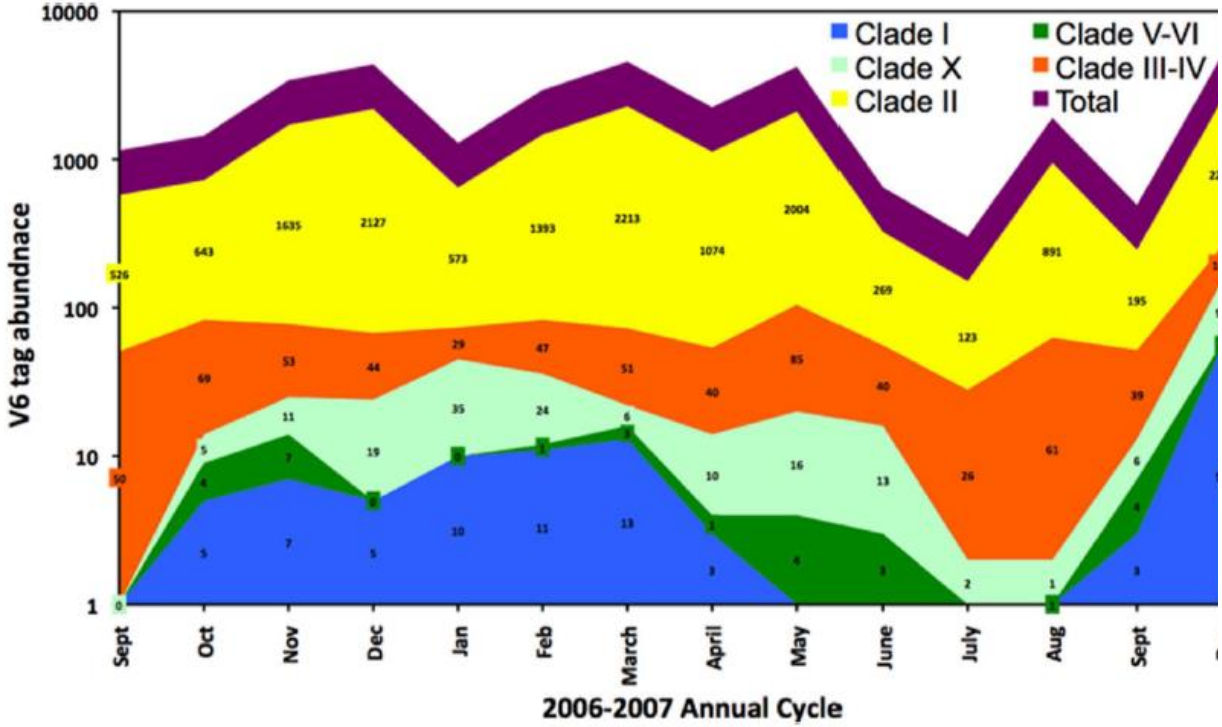
Nejpočetnější autotrofní organismy

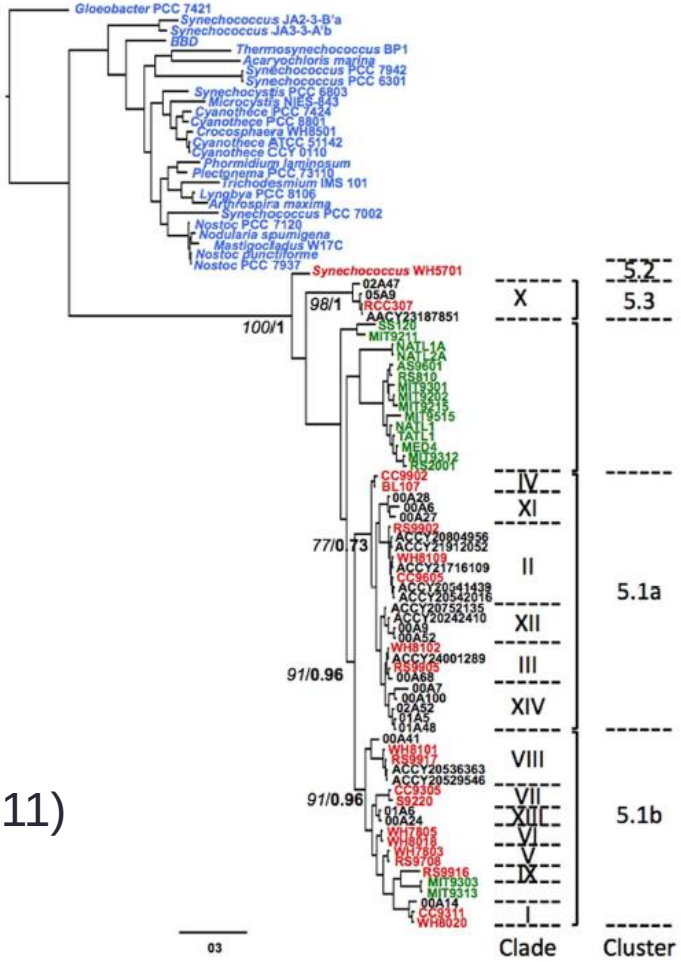






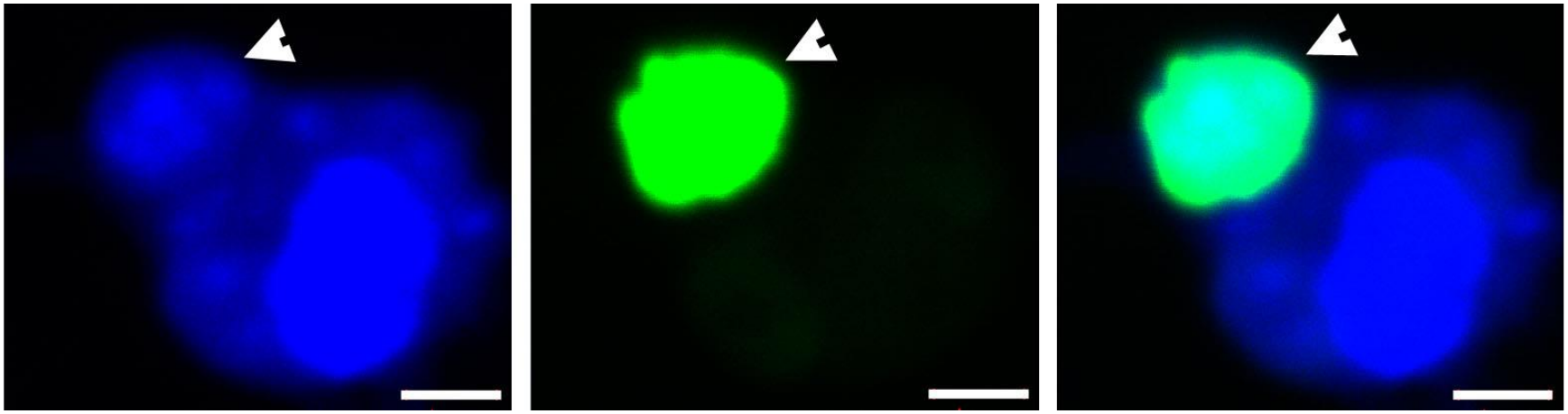
Diverzita *Synechococcus* v povrchové vrstvě oceánu v průběhu roku





Post et al. (2011)

UCYN-A: nový druh symbiózy sinice a Haptophyta



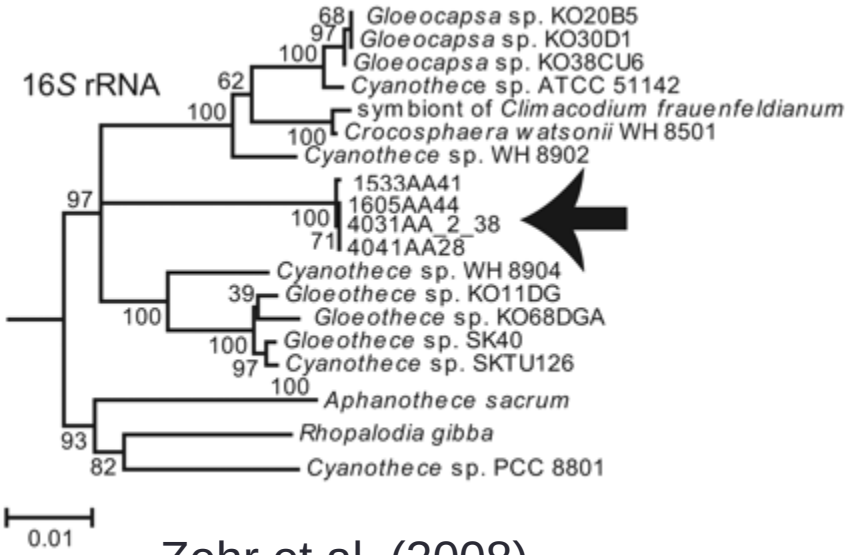
www.mpi-bremen.de/en/Unusual_symbiosis_discovered_among_marine_microorganisms.html

Braarudosphaera bigelowii – Haptophyta
Chrysochromulina parkeae – Haptophyta

Schopnost fixovat dusík, ale chybí fotosystém II (Zehr et al. (2008)

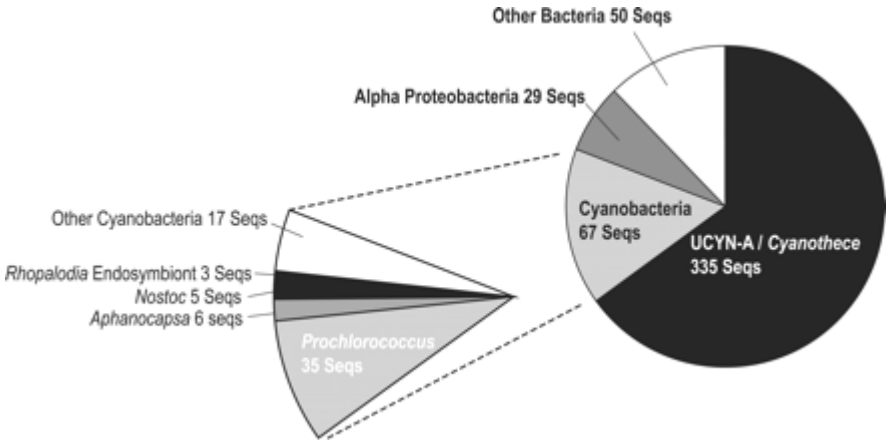
nannotax.org/content/braarudosphaera-bigelowii

UCYN-A: nový druh symbiózy



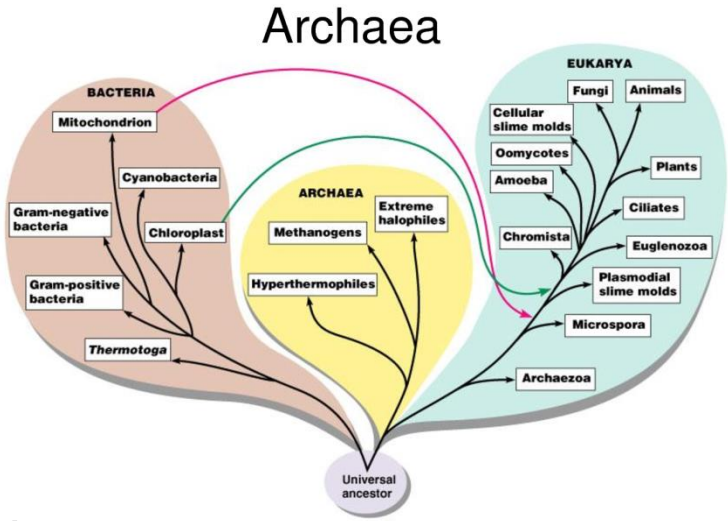
Zehr et al. (2008)

- Globálně rozšířená sinice
- Nekultivovatelná
- Izolace průtokovou cytometrií

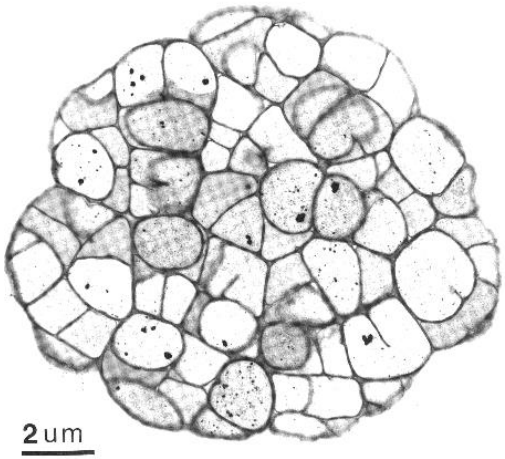


- Významná součást fytoplanktonu oceánů
- Složení 16S rRNA sekvencí na stanici ALOHA Hawaii

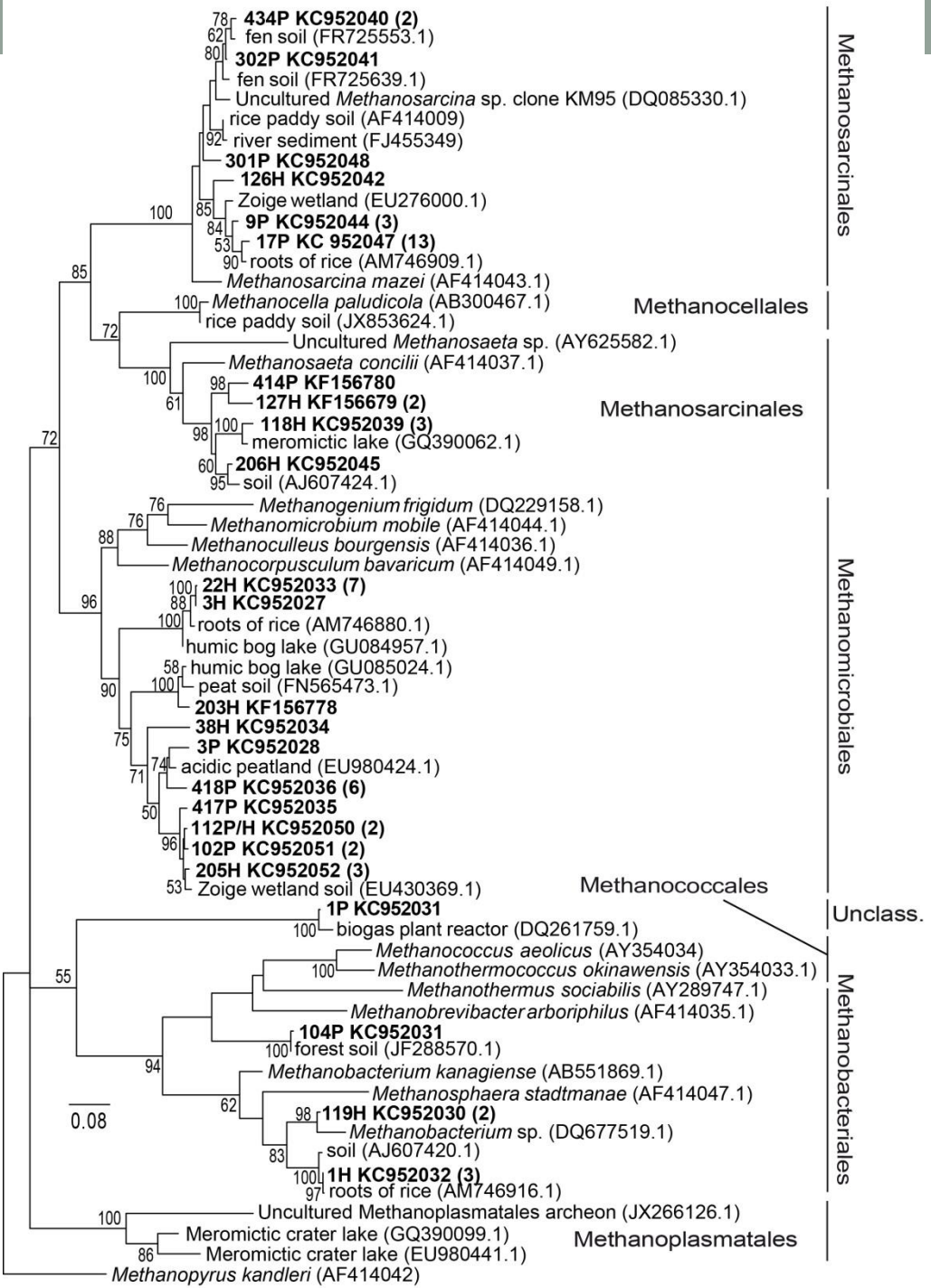
Metanogeny



- Archea – euryarcheota
  - Zvláštní buněčná stěna – pseudomurein
  - Unikátní struktura bičíků a ribozómů
- Produkce metanu – metyl-koenzym M reduktáza – *mcr*
- 40 – 50% produkce metanu v sedimentech (sladkovodních), mokřady, rýžoviště
- Metan – 25x větší efekt na globální oteplování než CO_2

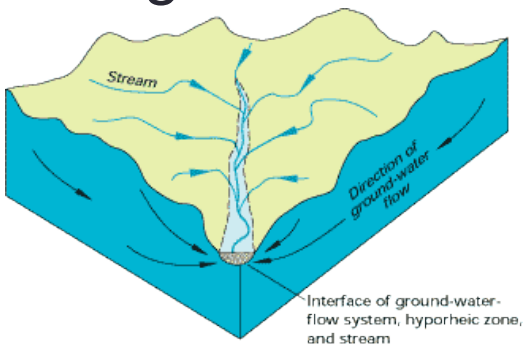


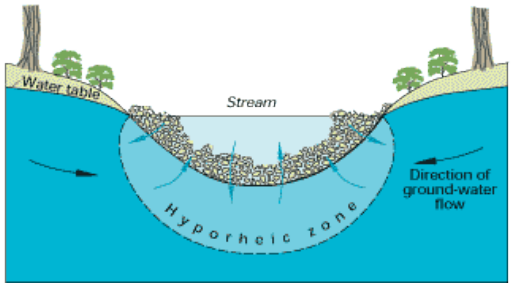
Methanosarcina



Metanogeny v říčním sedimentu

- Hyporeický sediment
- Významná komunita metanogenů





Buriánková et al. (2013)

DNA barcoding – základní principy

- Každá buňka má DNA
- Sekvence DNA se postupem času mění – evoluce, mutace
- Takže pro každý druh můžeme najít specifickou sekvenci DNA – barcode
- Definice: použití krátkého fragmentu DNA místo morfologie (Herbert et al. (2003)
- Kód produktu: 10 číslic, 11 pozic = 10^{11} kombinací
- Ideální DNA barcode 100 pozic = $\mathbf{4^{100}}$ kombinací
- Odhadem $\mathbf{10^6}$ (Hammond 1992) druhů, nepočítají se mikroorganismy

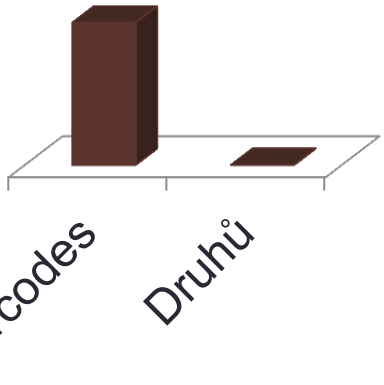
Morfologie

Sekvence DNA

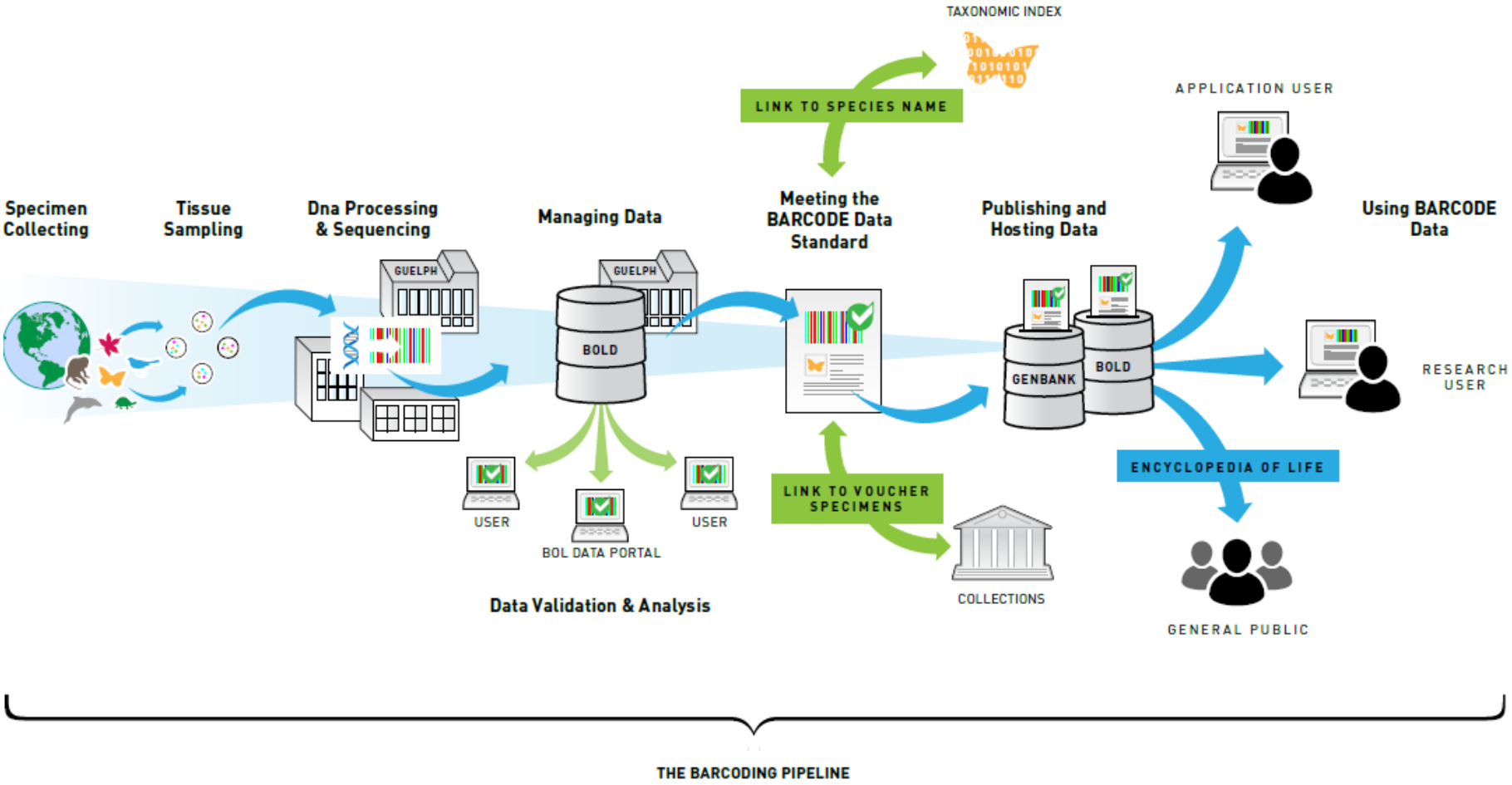
X AAACCTGGGTTT

DNA barcode

AAACCTGGGTTTT

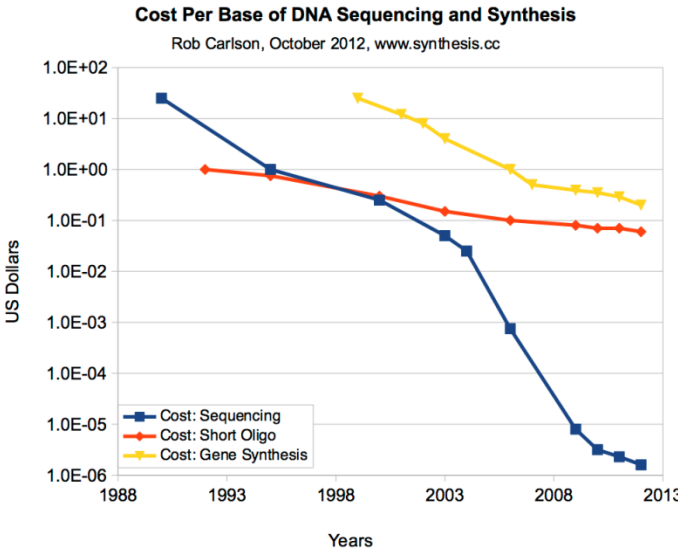


DNA barcoding

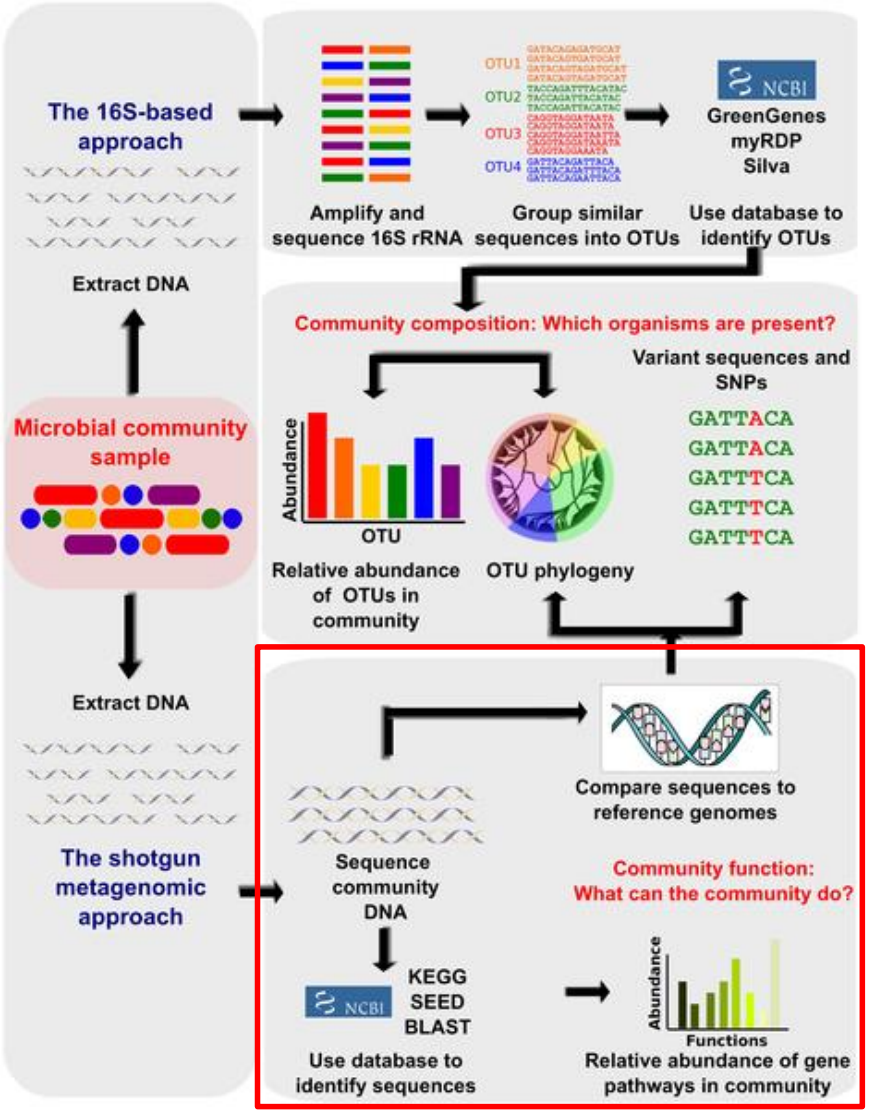


Výhody DNA barcodingu

1. Práce s krátkými **DNA sequencemi** – jednoduché zpracování – nejčastěji cytochrom oxydáza (zvířata)
2. Rychlejší identifikace
3. Redukce chyby při identifikaci druhů
4. Rutinně levnější
5. Lepší přístup k datům – veřejné databáze
6. Analýza možná pro všechny stádia vývoje
7. Možná analýza kryptických druhů
8. V budoucnu – příruční analyzátor?



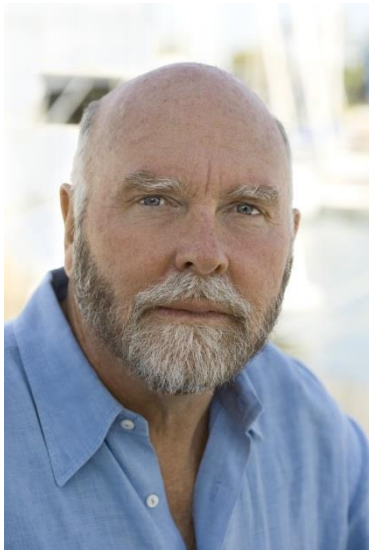
Function based metagenomika



Morgan & Huttenhower (2012)

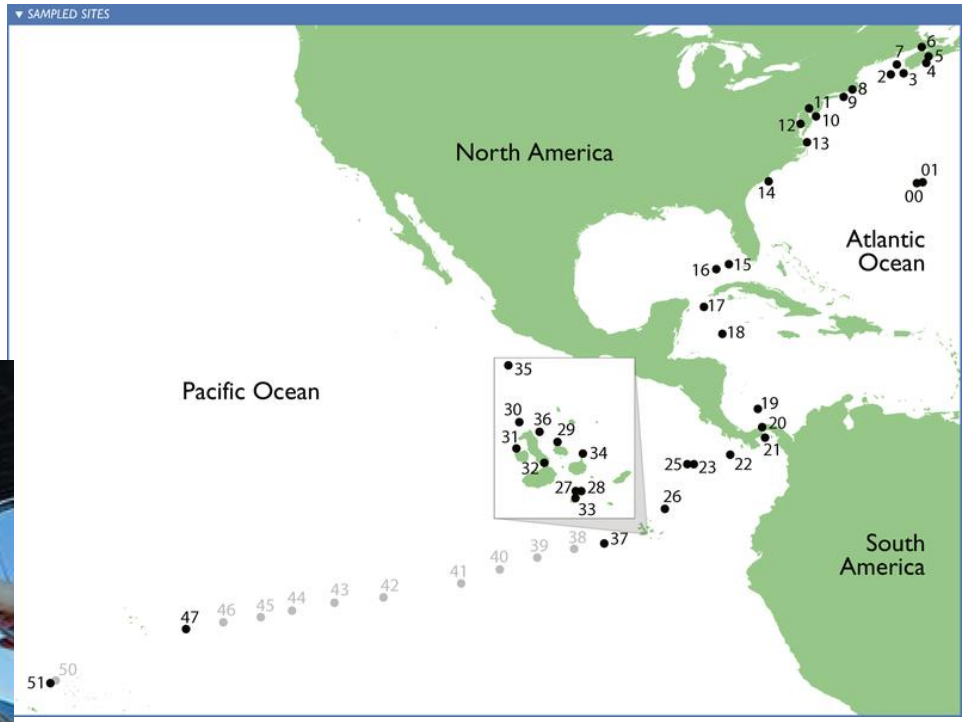
Craig Venter a jeho jachta

Bylo nebylo…



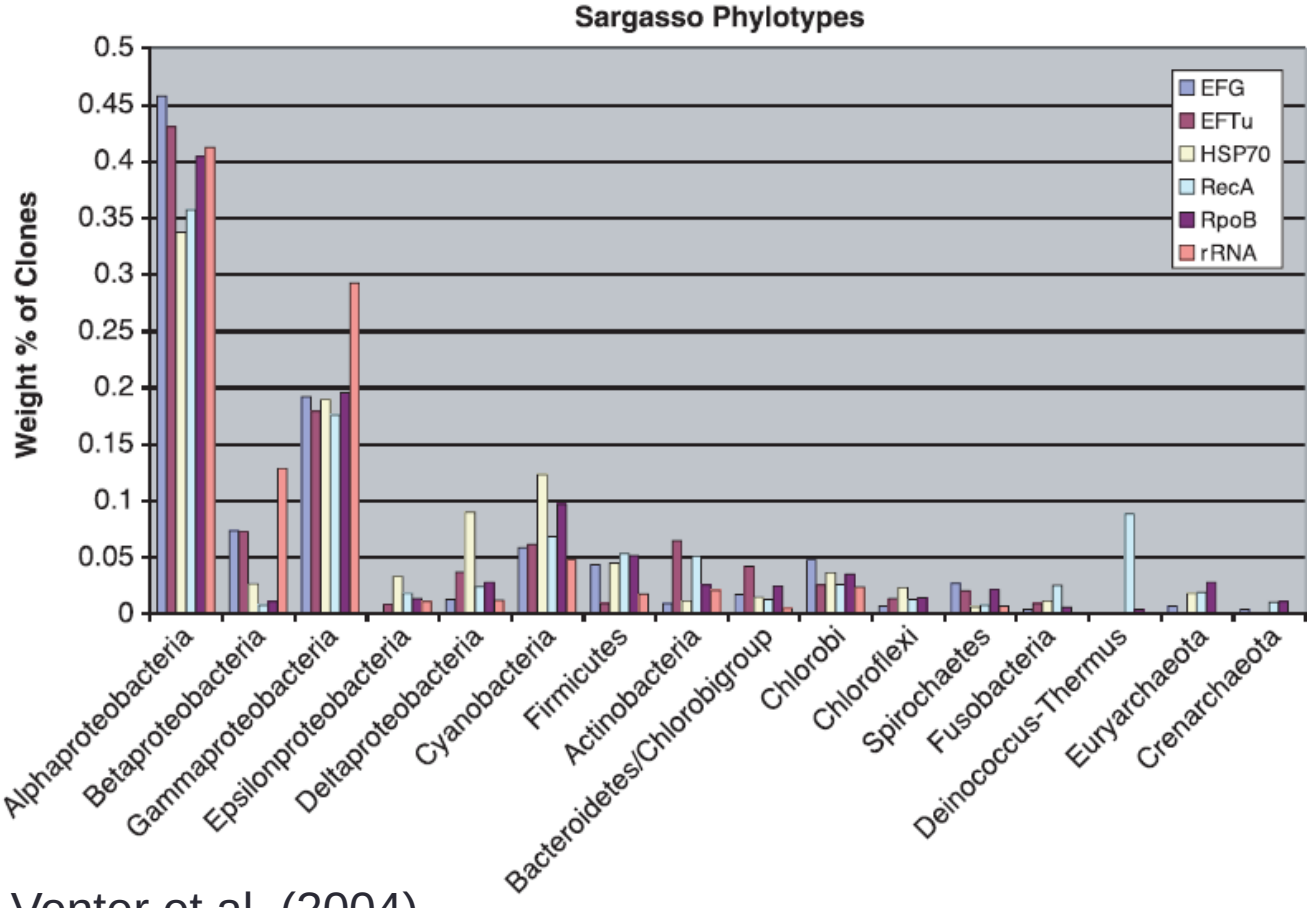
Global ocean sampling expedition

- Sběr vzorků vody, izolace DNA
- Shotgun sequencing
- Assembly
- Anotace



Sorcerer II

V sargasovém moři

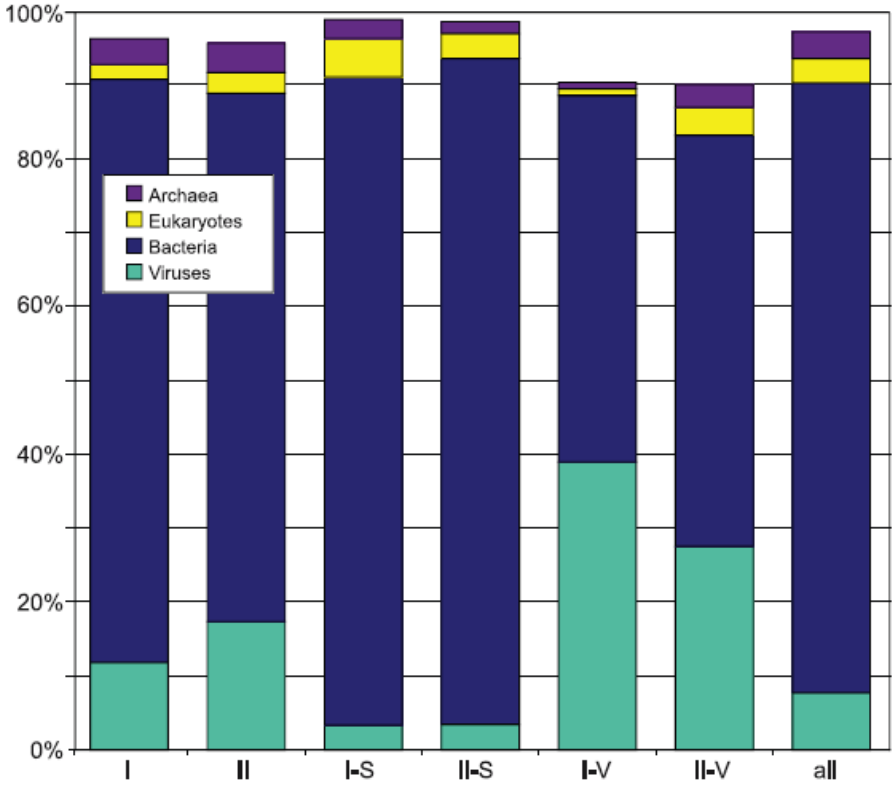


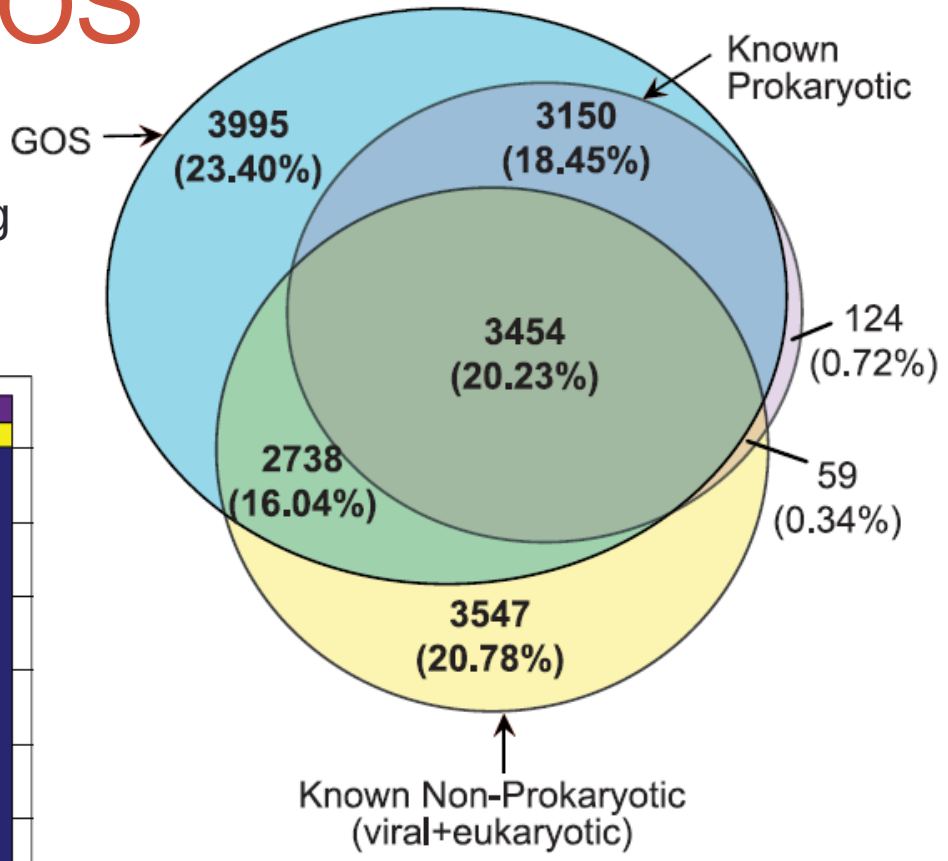
Venter et al. (2004)

Složení genů při GOS

Global ocean sampling

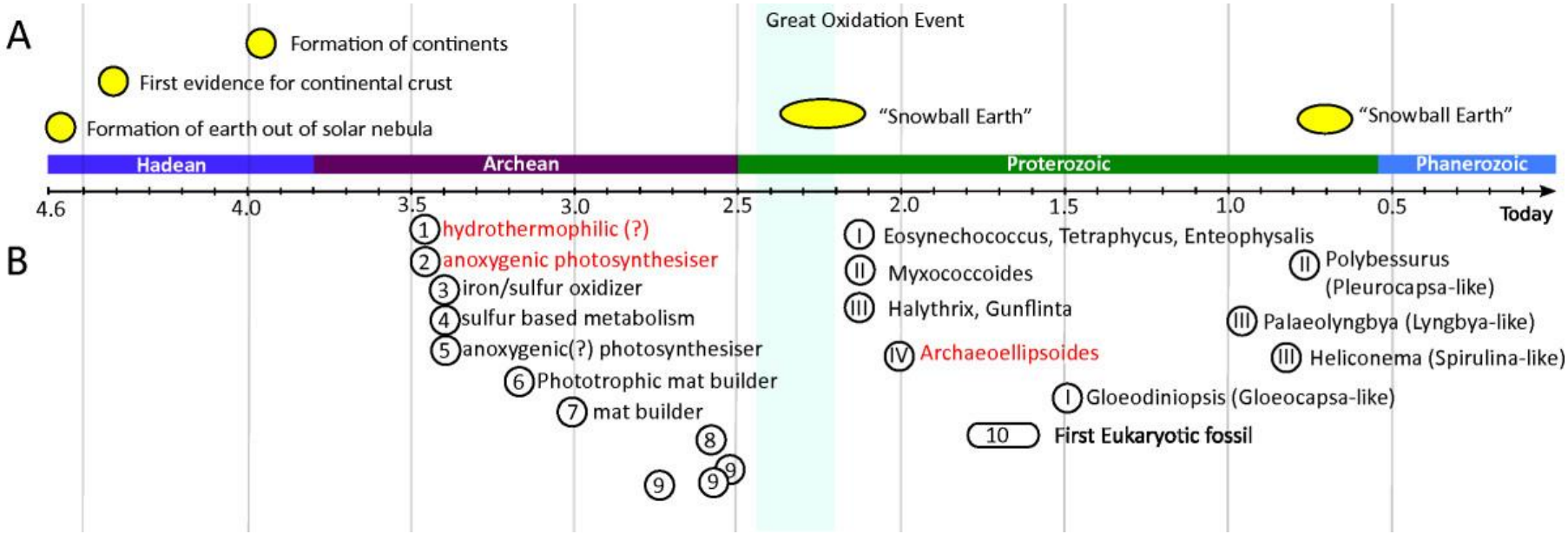
6.2 miliónů predikovaných genů





Yooseph et al. (2007)

Život v raném období evoluce

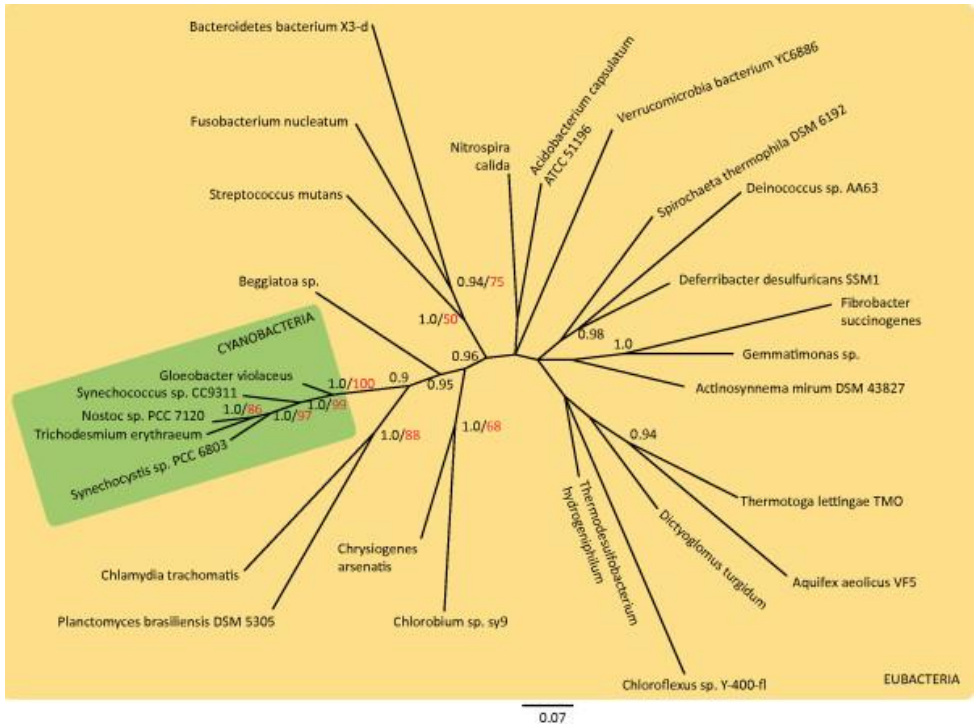


Schirrmeister et al. (2011)

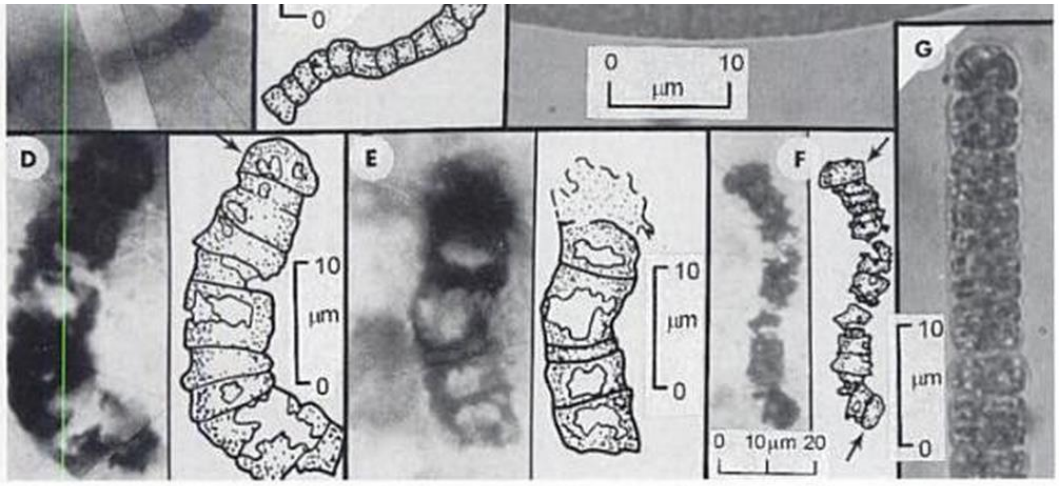
Sinice

Prokaryotní organismy
Oxygenní fotosyntéza

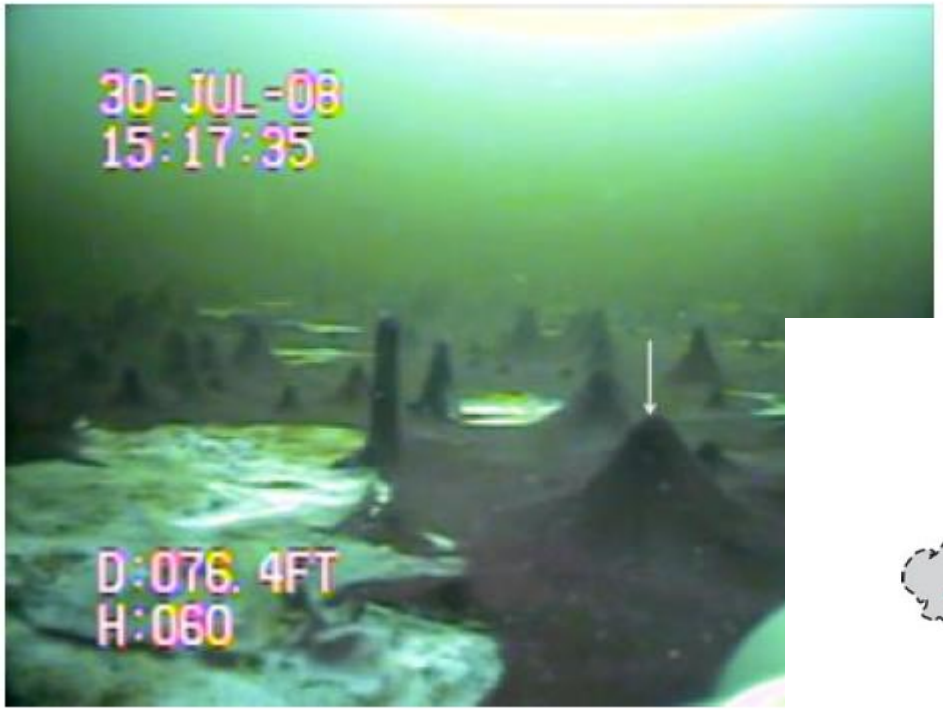
Schirrmeister et al. (2011)
16S rRNA



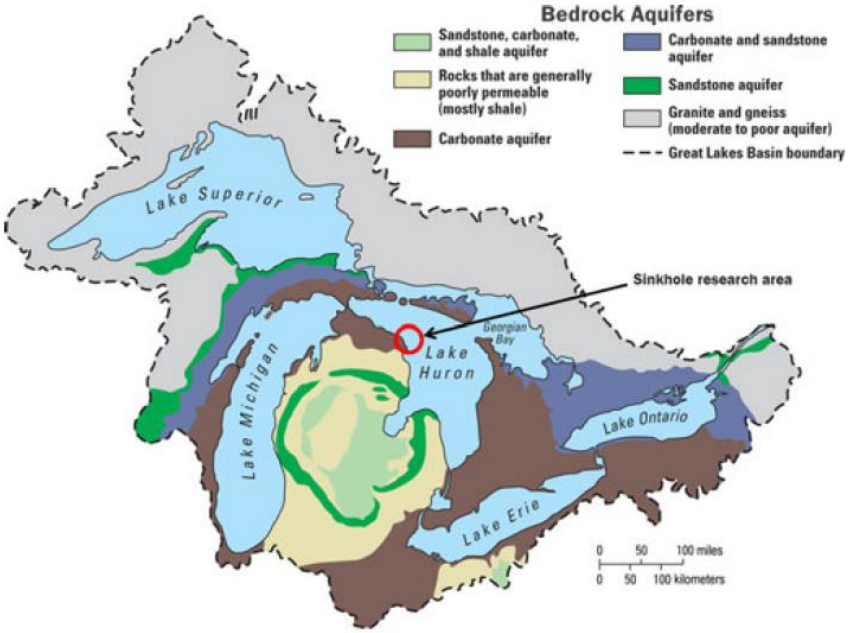
3.5 mld let staré fosílie (Schopf 1996)
Apex Chert v západní Austrálii



Život sinic při nízké koncentraci kyslíku

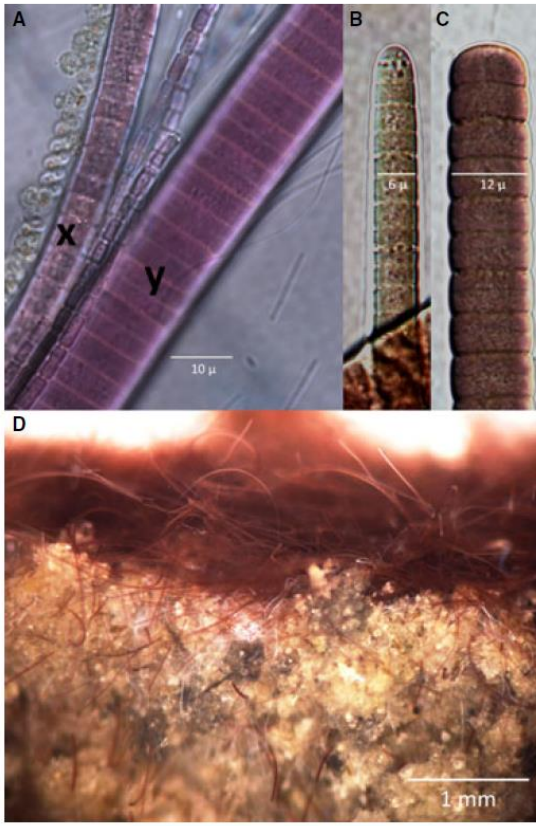


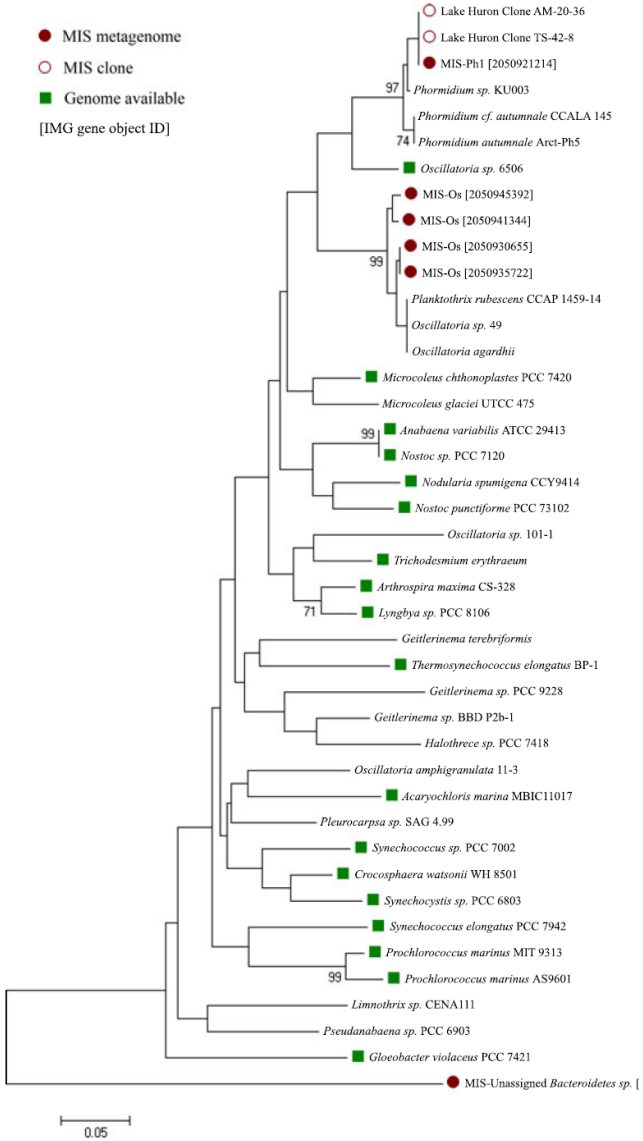
- Sink hole mat – lake huron
- Konstatní nízká koncentrace kyslíku
- Simulace počátku evoluce



Voorhies et al. (2012)

Život sinic při nízké koncentraci kyslíku

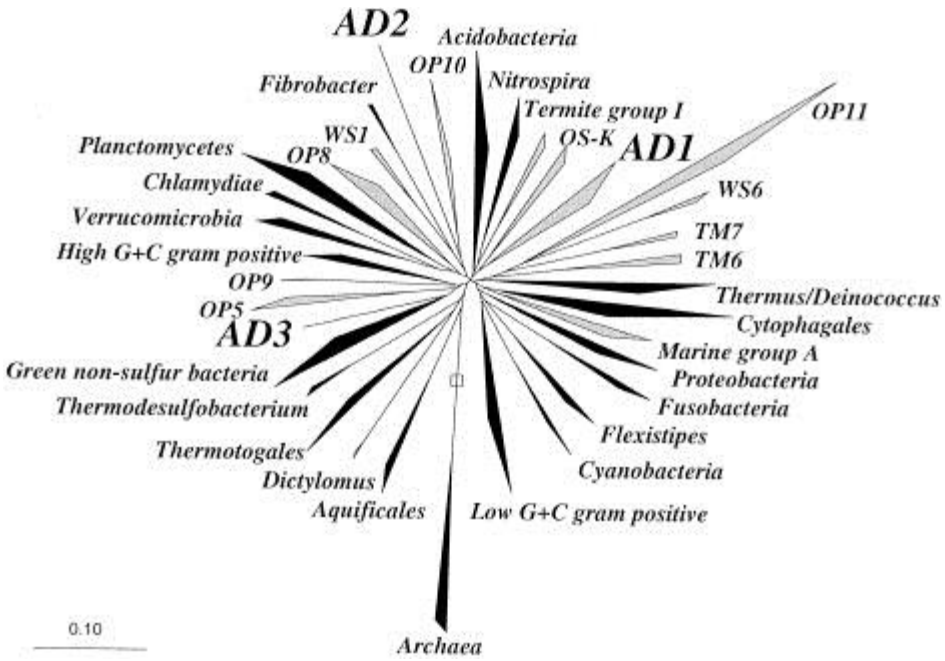




- Malá diversita komunity
- Metagenom obsahoval geny oxygenní i anoxygenní fotosyntézy
- Jen tři dominantní genotypy

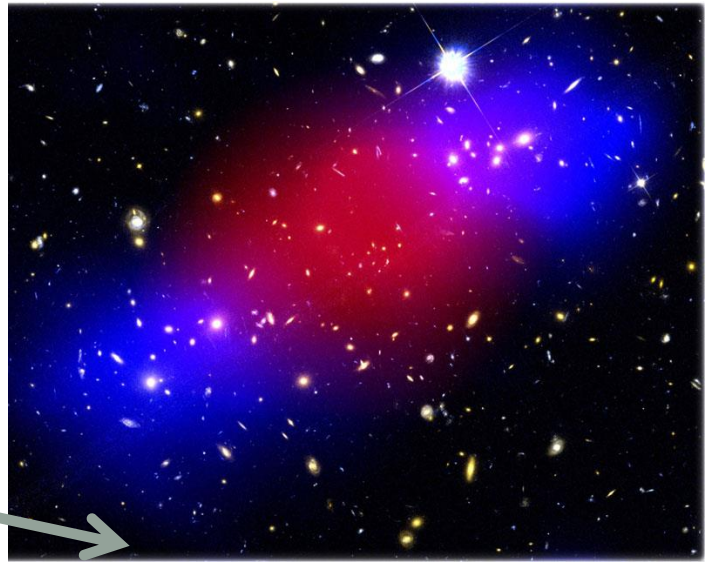
Voorhies et al. (2012)

Systematika bakterií

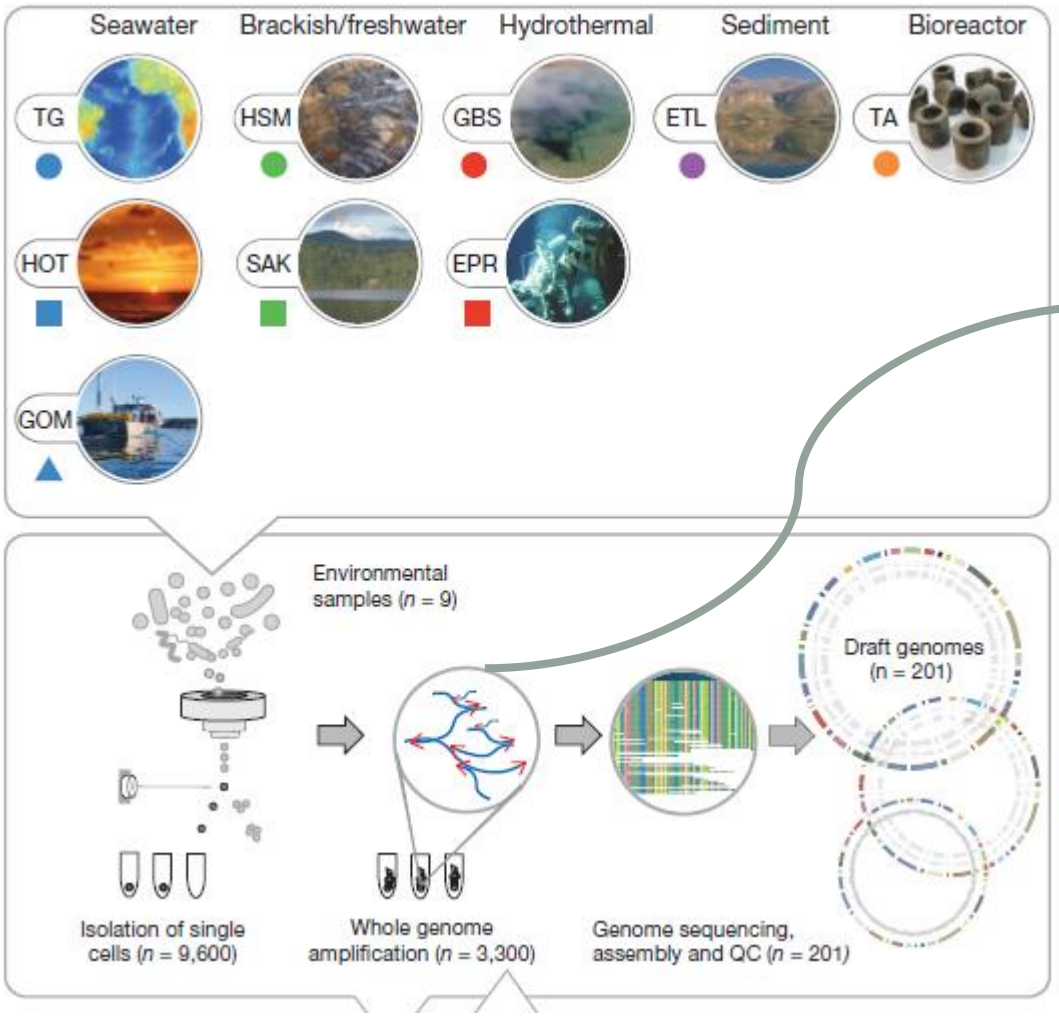


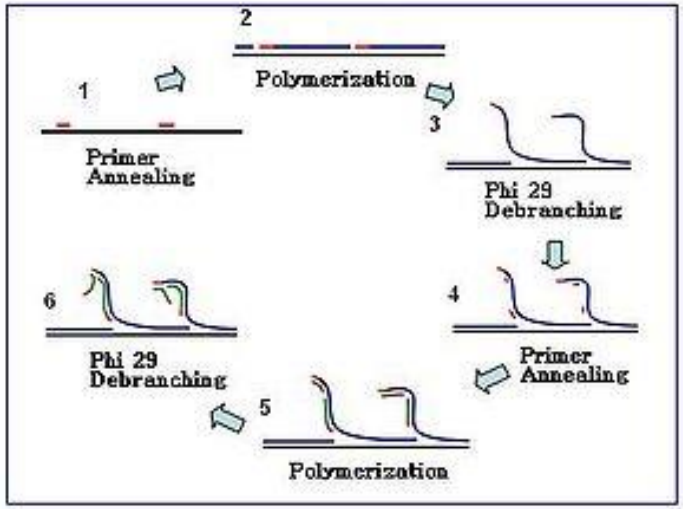
- Kandidátní divize v rámci bakterií
- Nemohou být popsány
- Není dostupná kultura
- Nikdo je nikdy neviděl…pouze jejich sekvenci DNA

Dark matter



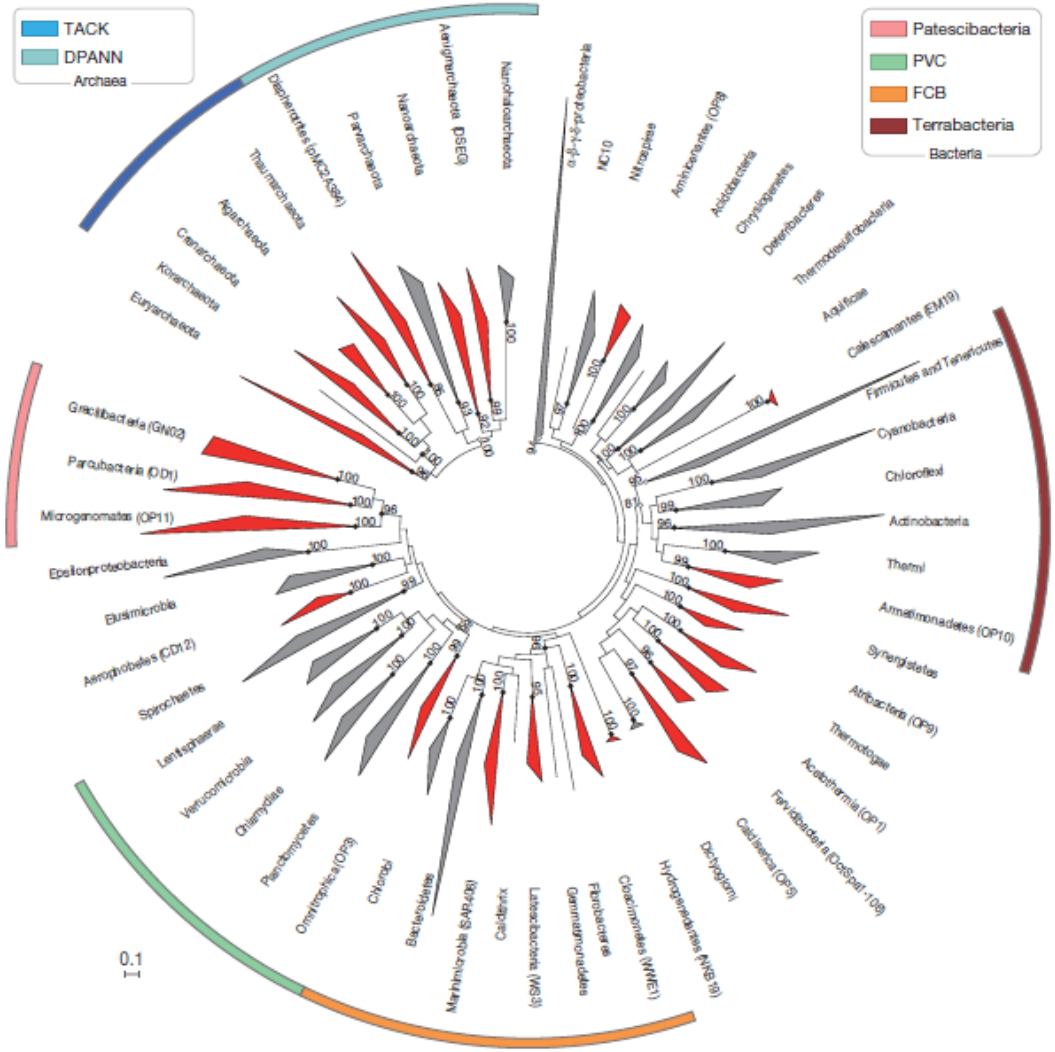
Mikrobiální „Dark matter“ bakterií





Rinke et al. (2013)

Mikrobiální „Dark matter“



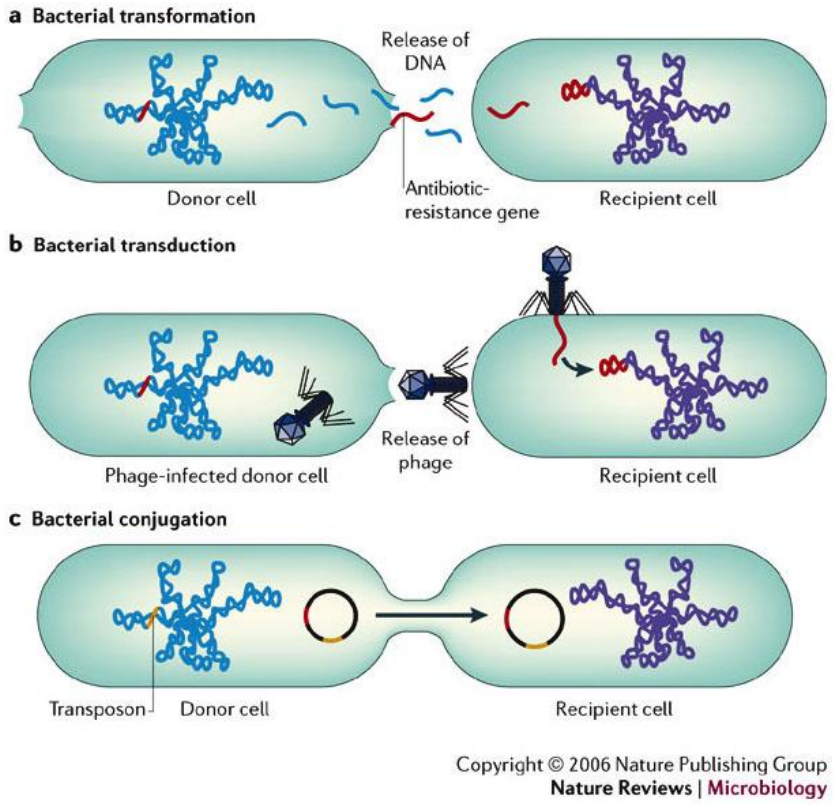
- Identifikace nových skupin, potvrzení existence kandidátních
- stop kodon UGA kóduje glycin u gracilobakterií
- Archeální syntéza purinů u bakterií
- A další..

Rinke et al. (2013)

Horizontální přenos genů

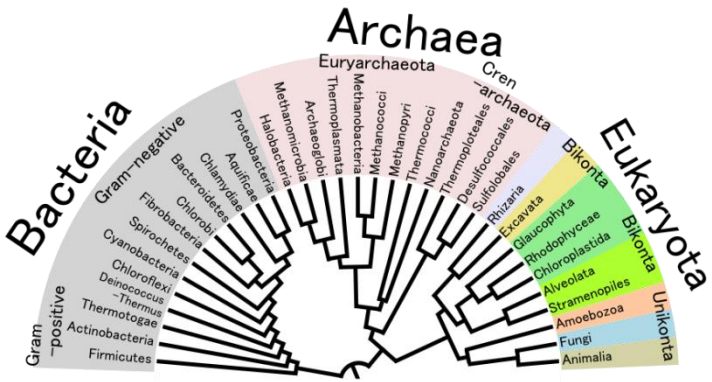
- Přenos genů mezi organismy jinak než pohlaním rozmnožováním
- Velice častý u prokaryot
- Všechny geny (resp. skupiny) v genomu mohou projít horizontálním transferem (Zhaxybayeva et al. 2006)
- Ale ke změnám v genomu nedochází náhodně, nýbrž v tvz. „genomic islands“ (Rodriguez-Valera et al. 2009)
- Častěji dochází k přenosu lokálně, ne na velké vzdálenosti

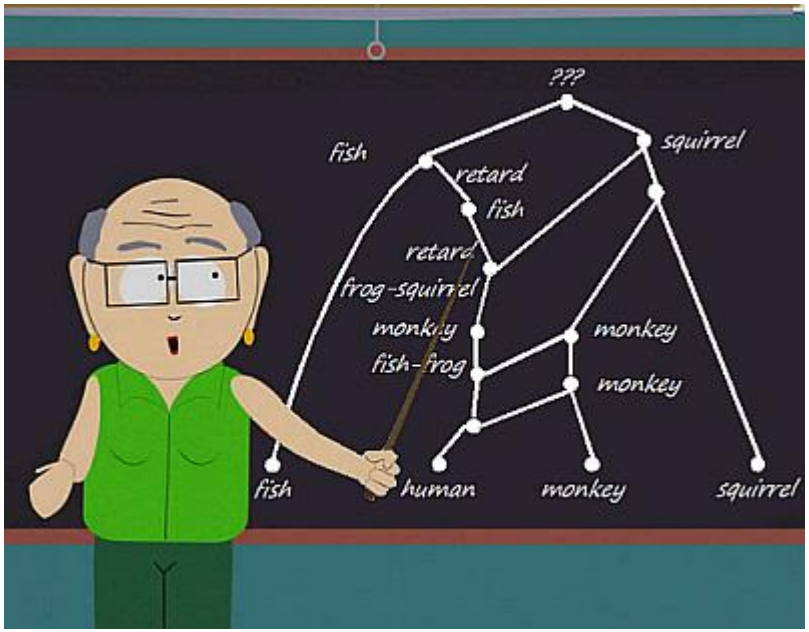
Waters (2001)

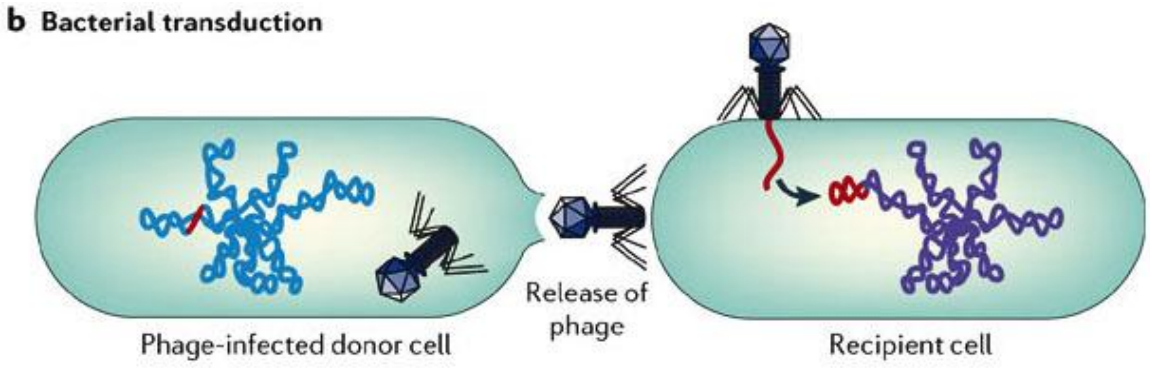


Furuya & Lowy (2006)

Horizontální transfer genů



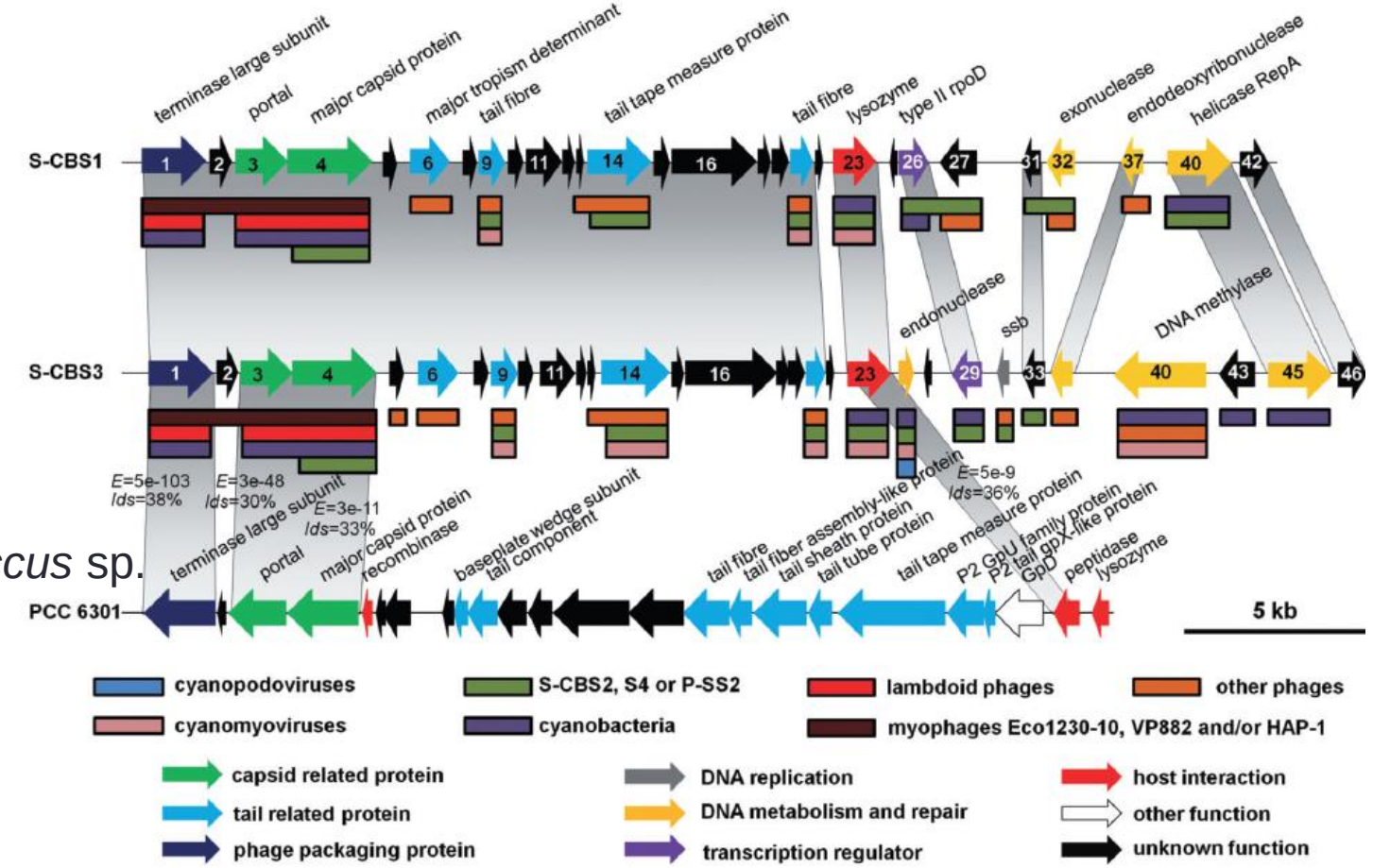




Horizontální transfer genů

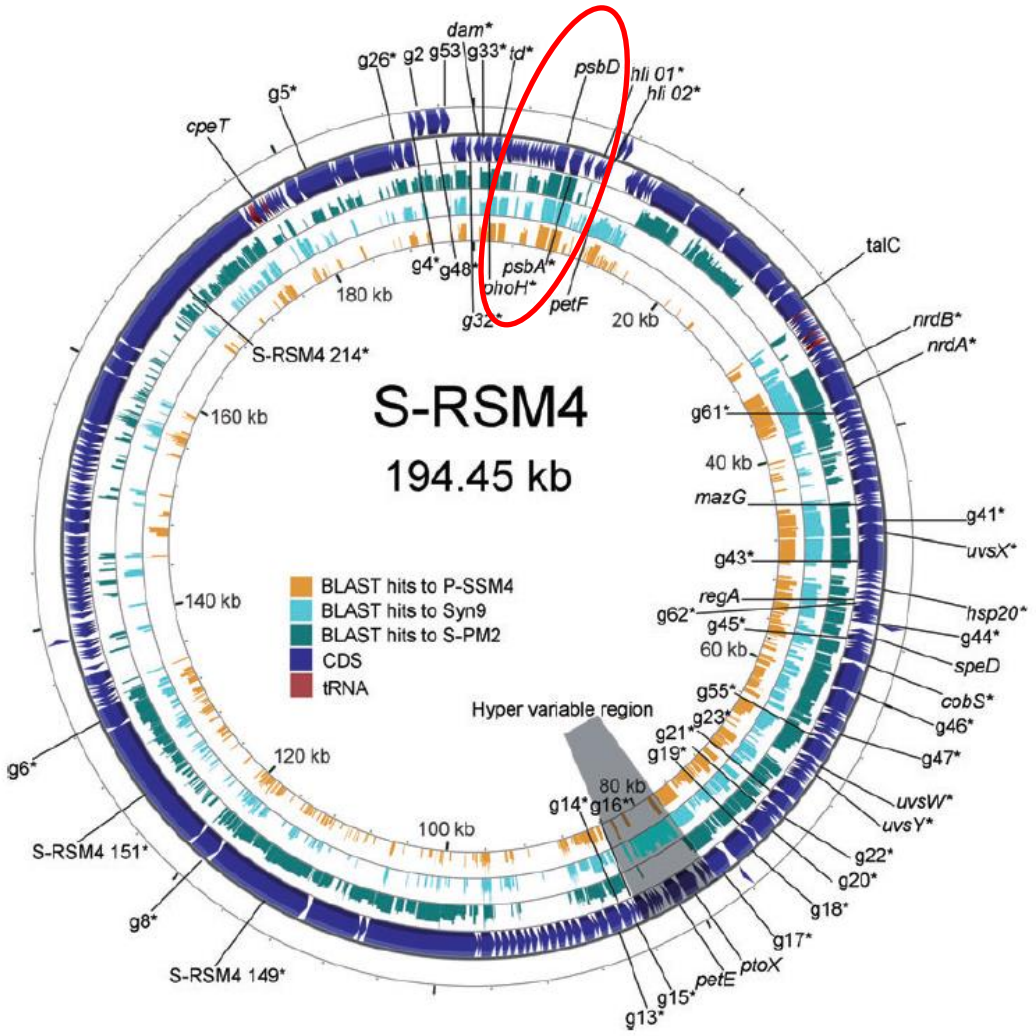
Siphnoviry

Synechococcus sp.



Huang et al. (2012)

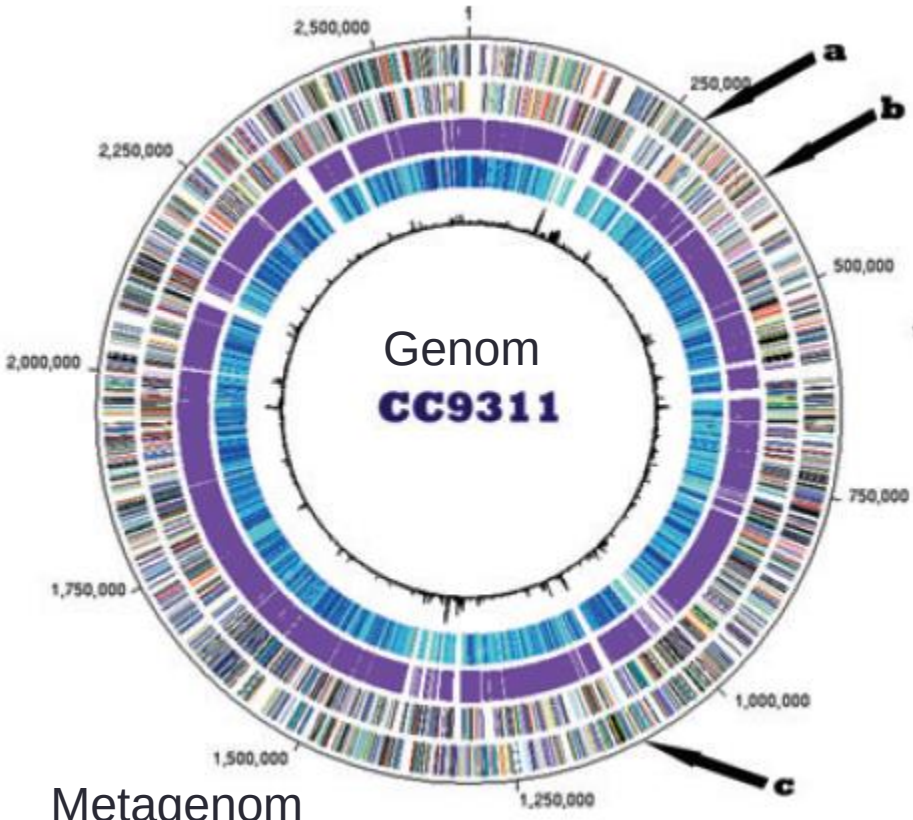
Horizontální transfer genů



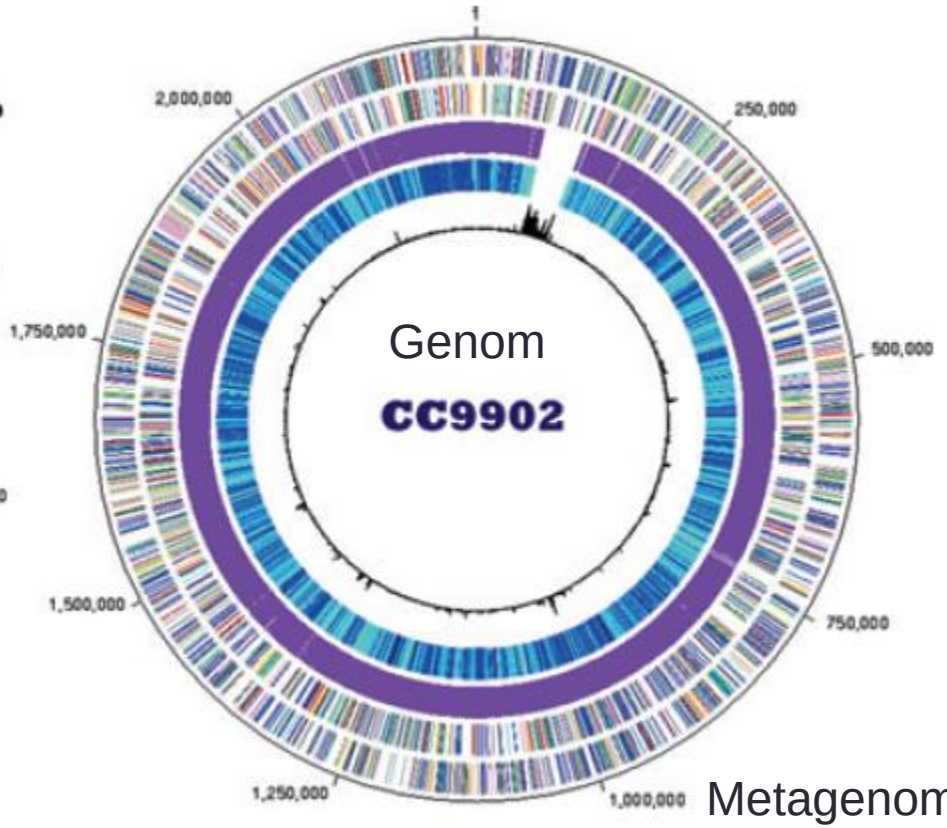
Millard et al. (2009)

Cyanomyovirus
- *psbA* a *psbD* geny fotosystému
- homology k hostiteli *Synechococcus* sp.

HGT u mořských sinic *Synechococcus*



Metagenom



Metagenom

- Celkem tři mobilní elementy (plazmidy) v metagenomu, ale ne v genomu sinice
- Horizontální transfer genů

Děkuji za pozornost!!!

- S metagenomikou nejste nikdy sami ☺