The Impact of Horizontal Gene Transfers on Prokaryotic Genome Evolution

Doctoral Dissertation Defense

Pascal Lapierre

Graduate Program in Genetics, Genomics and Bioinformatics
Molecular and Cell Biology Department

Tuesday May 29th, 2007

What is Horizontal Gene Transfer (HGT)?

Any process in which an organism transfers genetic material to another cell that is not its offspring. By contrast, vertical transfer occurs when an organism receives genetic material from its ancestor, e.g. its parent or a species from which it evolved.

(Wikipedia)

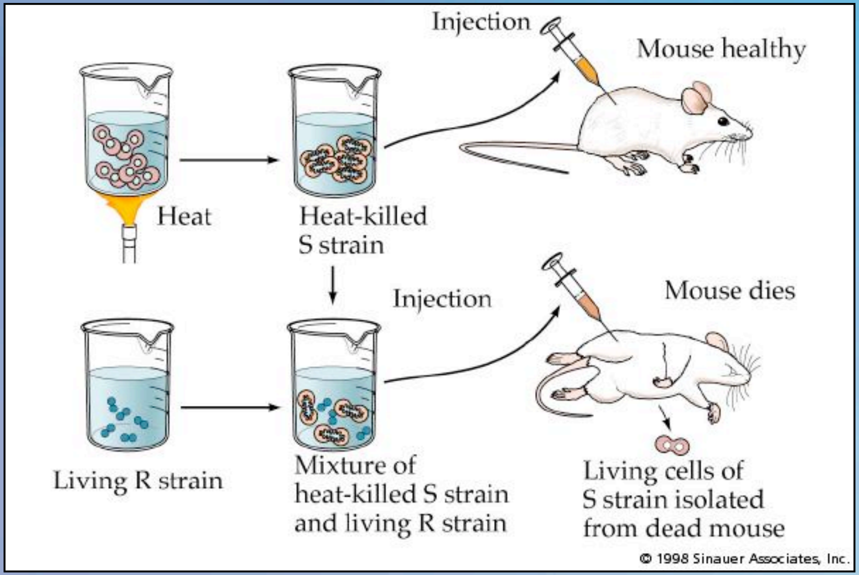
- Transformation (Uptake of DNA)

- Transduction (Phages)

- Conjugation (Bacteria-Bacteria)

First Evidence for HGT

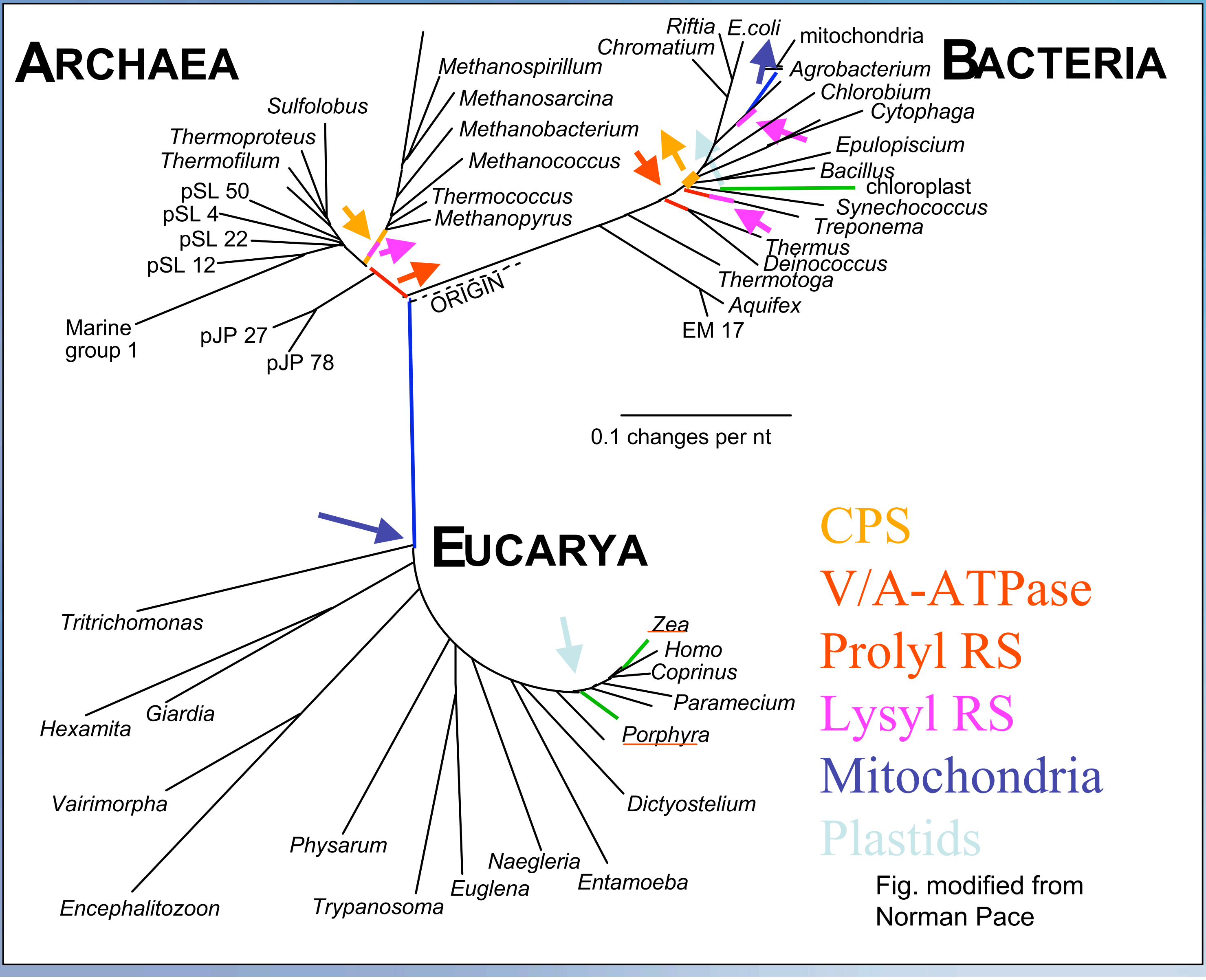
The Griffith's experiment (1928) :



(Taken from http://www.mie.utoronto.ca/labs/lcdlab/biopic/)

Avery, MacLeod, McCarty (1944) : DNA is most likely responsible for the transformation of the R strain cell

A Few Examples :



CPS

V/A-ATPase

Prolyl RS

Lysyl RS

Mitochondria

Plastids

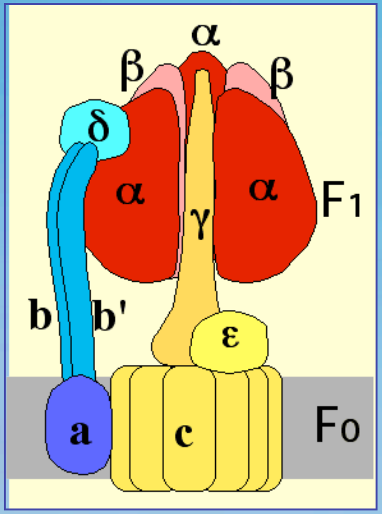
Fig. modified from
Norman Pace

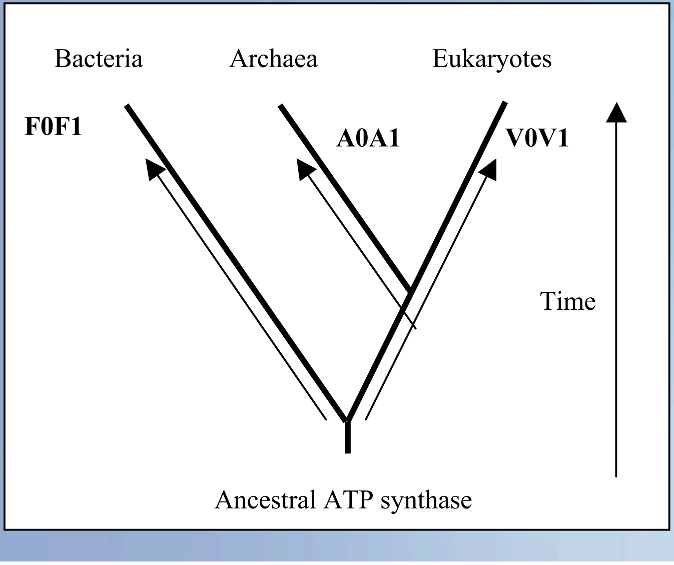
Part I :

Evolutionary history of the archaeal-type ATP synthase in the bacterial domain

ATP synthase - general characteristics

- Multisubunit proteins
- Found in all living cells
- Soluble part (F1) and transmembrane part (F0)
- Uses an ion gradient (H+ or Na+) to generate ATP molecules





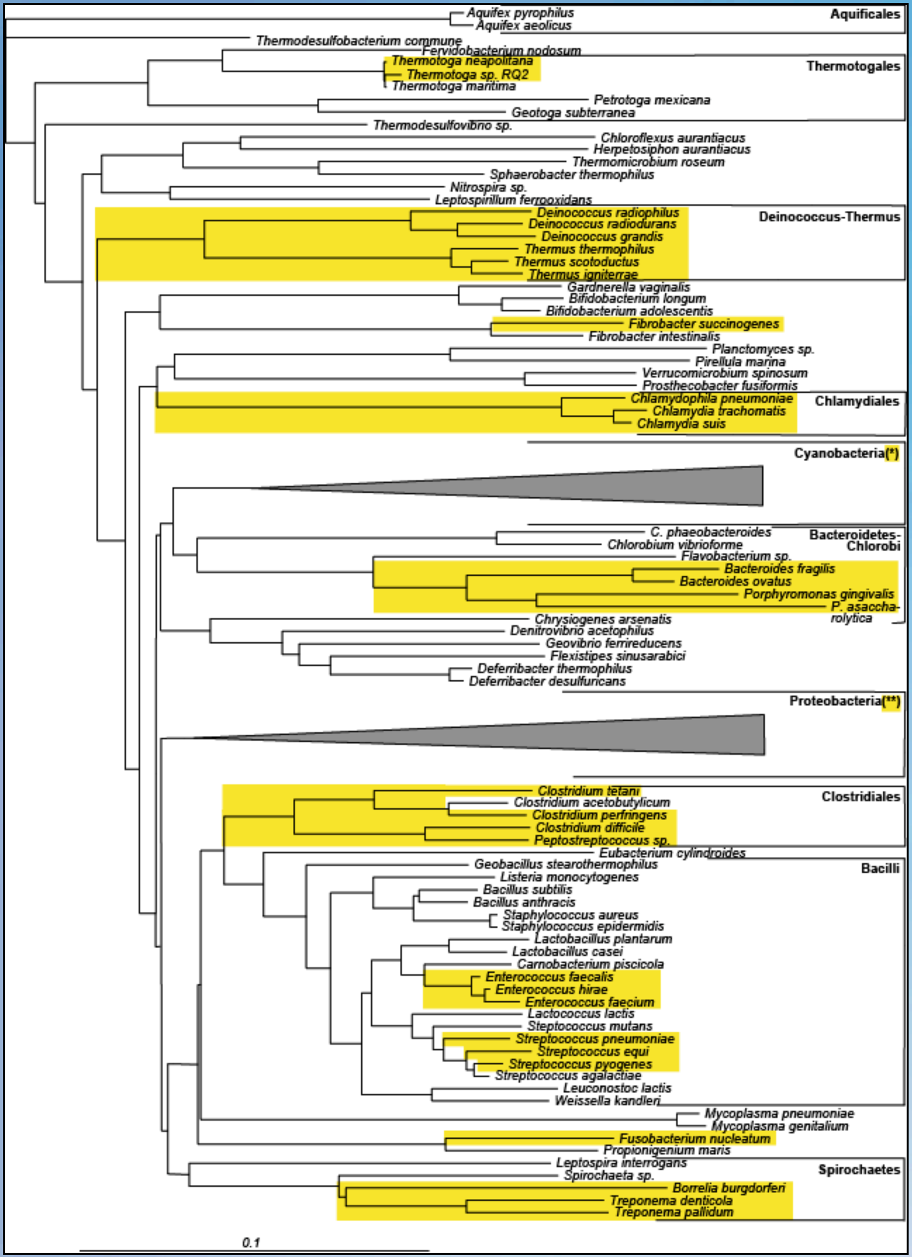
16s rRNA tree of the bacterial domain

Competing theories :

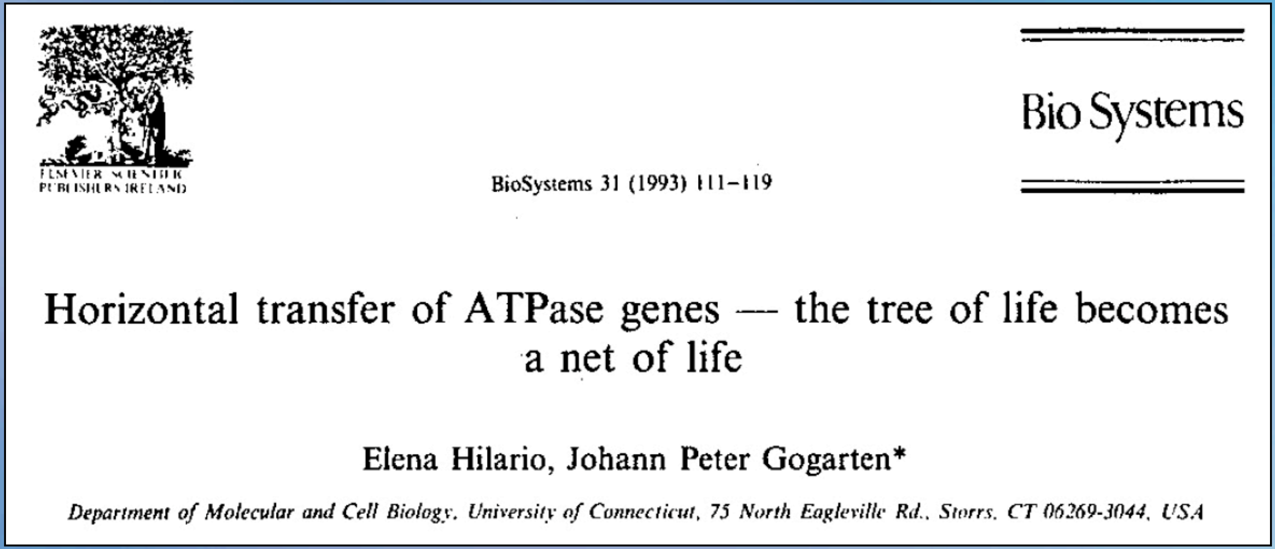
Both F- and A/V-type ATPase already present in LUCA

Or

Horizontal transfers from Archaea to Bacteria



Go to the expert!



ELSEVIER SCIENTIFIC PUBLISHERS IRELAND

BioSystems 31 (1993) 111–119

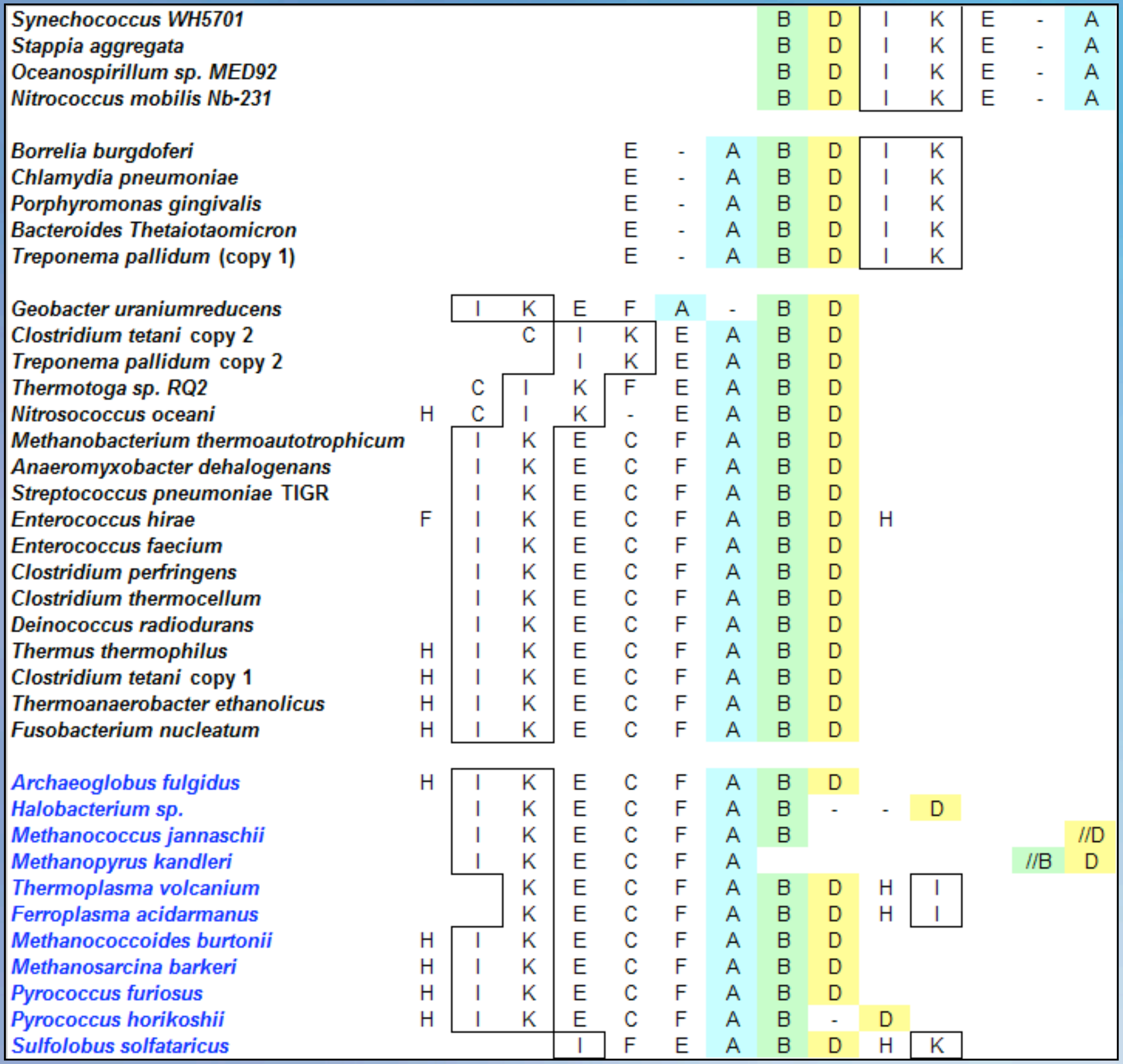
Bio Systems

Horizontal transfer of ATPase genes — the tree of life becomes a net of life

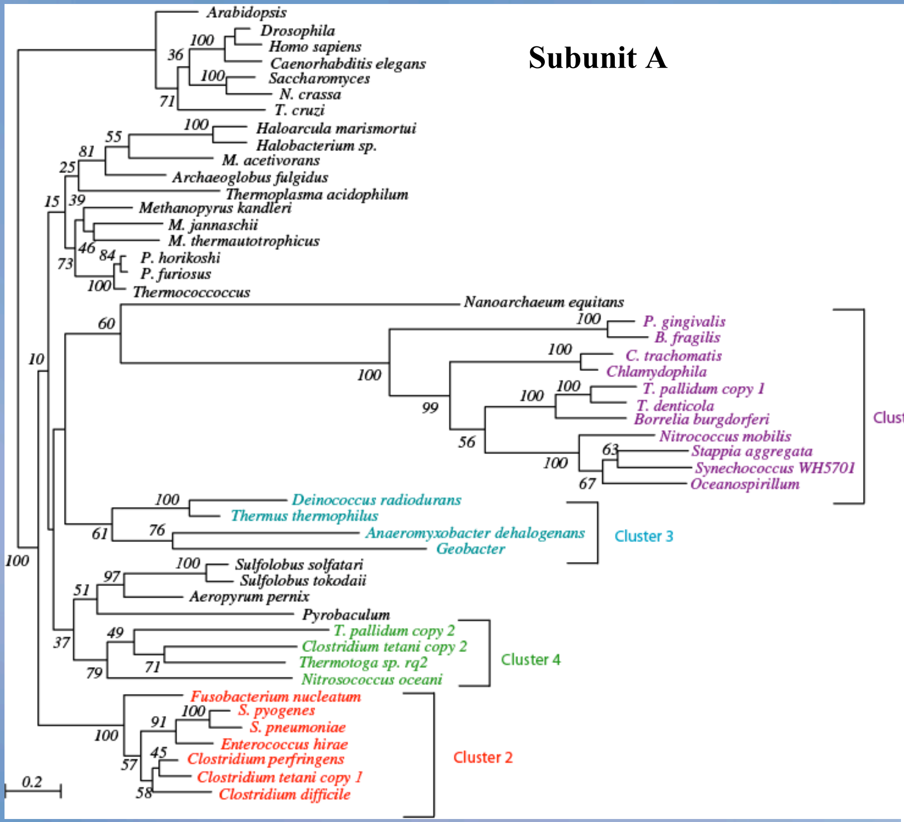
Elena Hilario, Johann Peter Gogarten*

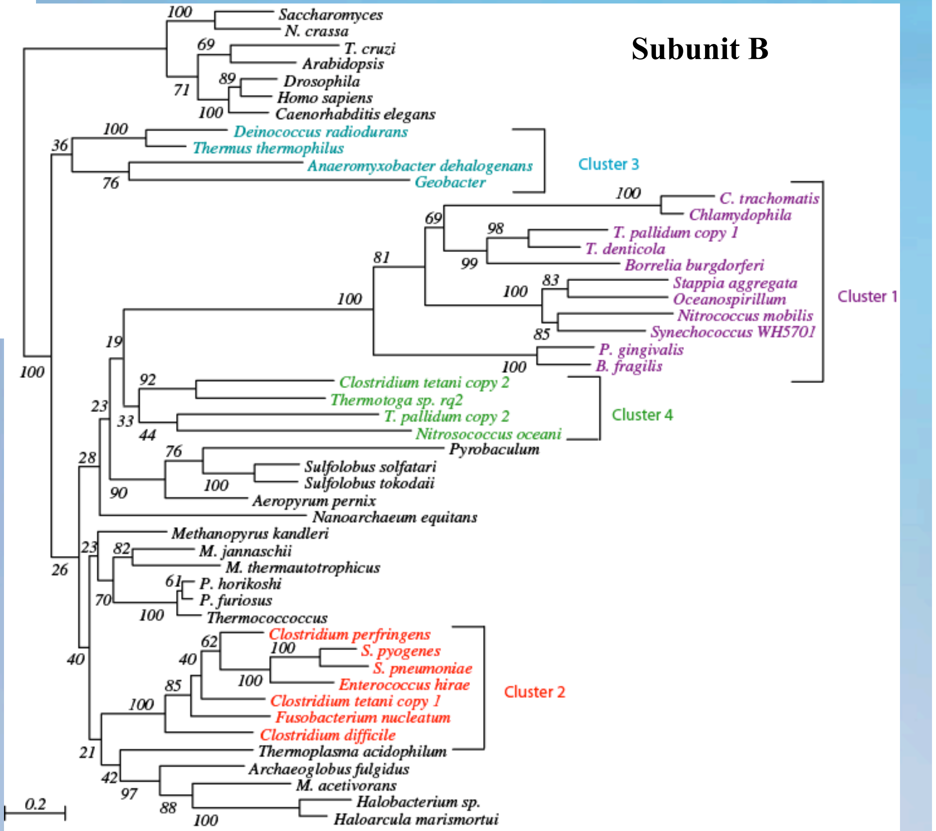
Department of Molecular and Cell Biology, University of Connecticut, 75 North Eagleville Rd., Storrs, CT 06269-3044, USA

Operon organization

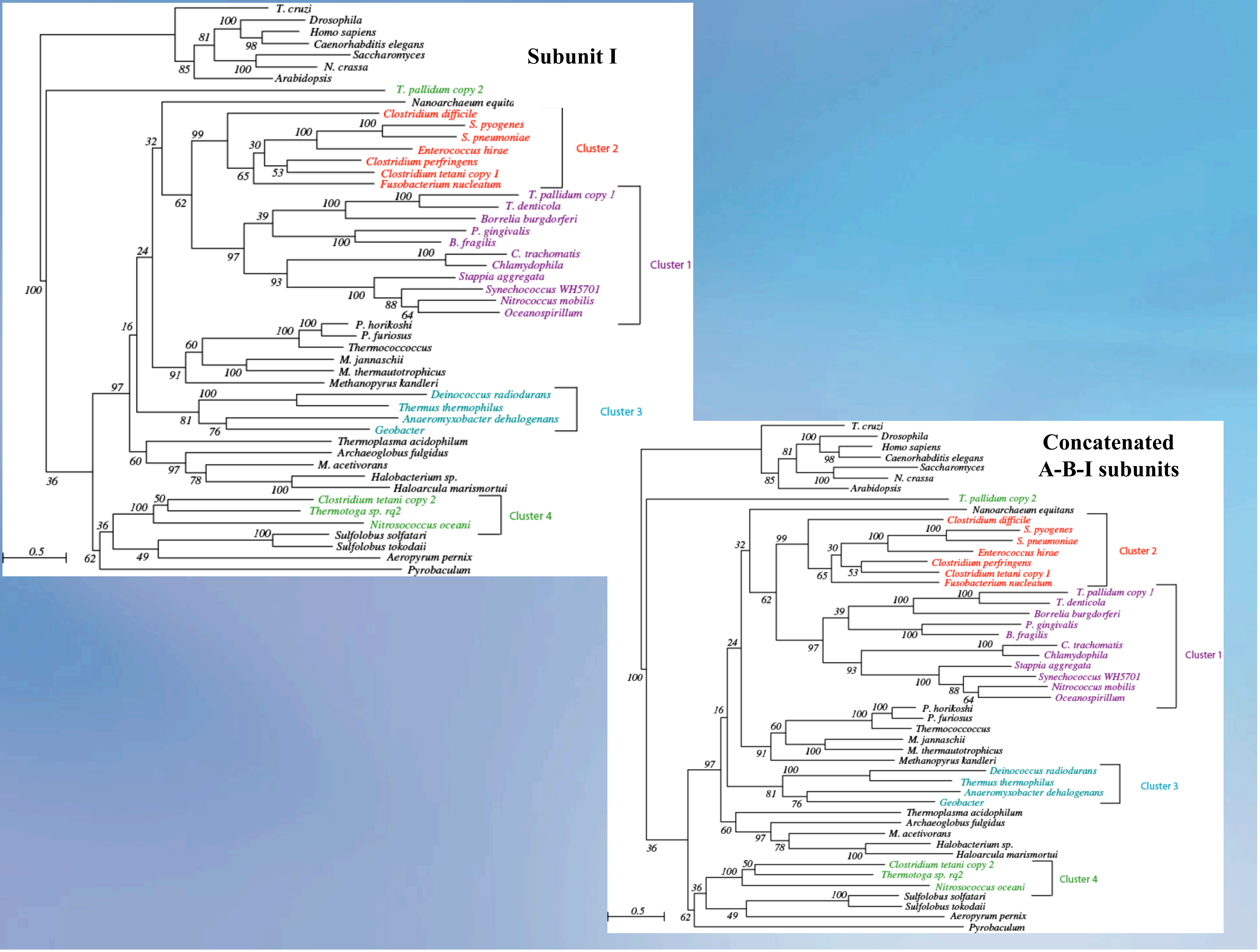


Synechococcus WH5701									B	D	I	K	E	-	A
Stappia aggregata									B	D	I	K	E	-	A
Oceanospirillum sp. MED92									B	D	I	K	E	-	A
Nitrococcus mobilis Nb-231									B	D	I	K	E	-	A
Borrelia burgdoferi						E	-	A	B	D	I	K			
Chlamydia pneumoniae						E	-	A	B	D	I	K			
Porphyromonas gingivalis						E	-	A	B	D	I	K			
Bacteroides Thetaiotaomicron						E	-	A	B	D	I	K			
Treponema pallidum (copy 1)						E	-	A	B	D	I	K			
Geobacter uraniumreducens		I	K	E	F	A	-	B	D						
Clostridium tetani copy 2			C	I	K	E	A	B	D						
Treponema pallidum copy 2				I	K	E	A	B	D						
Thermotoga sp. RQ2		C	I	K	F	E	A	B	D						
Nitrosococcus oceani	H	C	I	K	-	E	A	B	D						
Methanobacterium thermoautotrophicum		I	K	E	C	F	A	B	D						
Anaeromyxobacter dehalogenans		I	K	E	C	F	A	B	D						
Streptococcus pneumoniae TIGR		I	K	E	C	F	A	B	D						
Enterococcus hirae	F	I	K	E	C	F	A	B	D	H					
Enterococcus faecium		I	K	E	C	F	A	B	D						
Clostridium perfringens		I	K	E	C	F	A	B	D						
Clostridium thermocellum		I	K	E	C	F	A	B	D						
Deinococcus radiodurans		I	K	E	C	F	A	B	D						
Thermus thermophilus	H	I	K	E	C	F	A	B	D						
Clostridium tetani copy 1	H	I	K	E	C	F	A	B	D						
Thermoanaerobacter ethanolicus	H	I	K	E	C	F	A	B	D						
Fusobacterium nucleatum	H	I	K	E	C	F	A	B	D						
Archaeoglobus fulgidus	H	I	K	E	C	F	A	B	D						
Halobacterium sp.		I	K	E	C	F	A	B	-	-	D				
Methanococcus jannaschii		I	K	E	C	F	A	B							//D
Methanopyrus kandleri		I	K	E	C	F	A							//B	D
Thermoplasma volcanium			K	E	C	F	A	B	D	H	I				
Ferroplasma acidarmanus			K	E	C	F	A	B	D	H	I				
Methanococcoides burtonii	H	I	K	E	C	F	A	B	D						
Methanosarcina barkeri	H	I	K	E	C	F	A	B	D						
Pyrococcus furiosus	H	I	K	E	C	F	A	B	D						
Pyrococcus horikoshii	H	I	K	E	C	F	A	B	-	D					
Sulfolobus solfataricus				I	F	E	A	B	D	H	K				



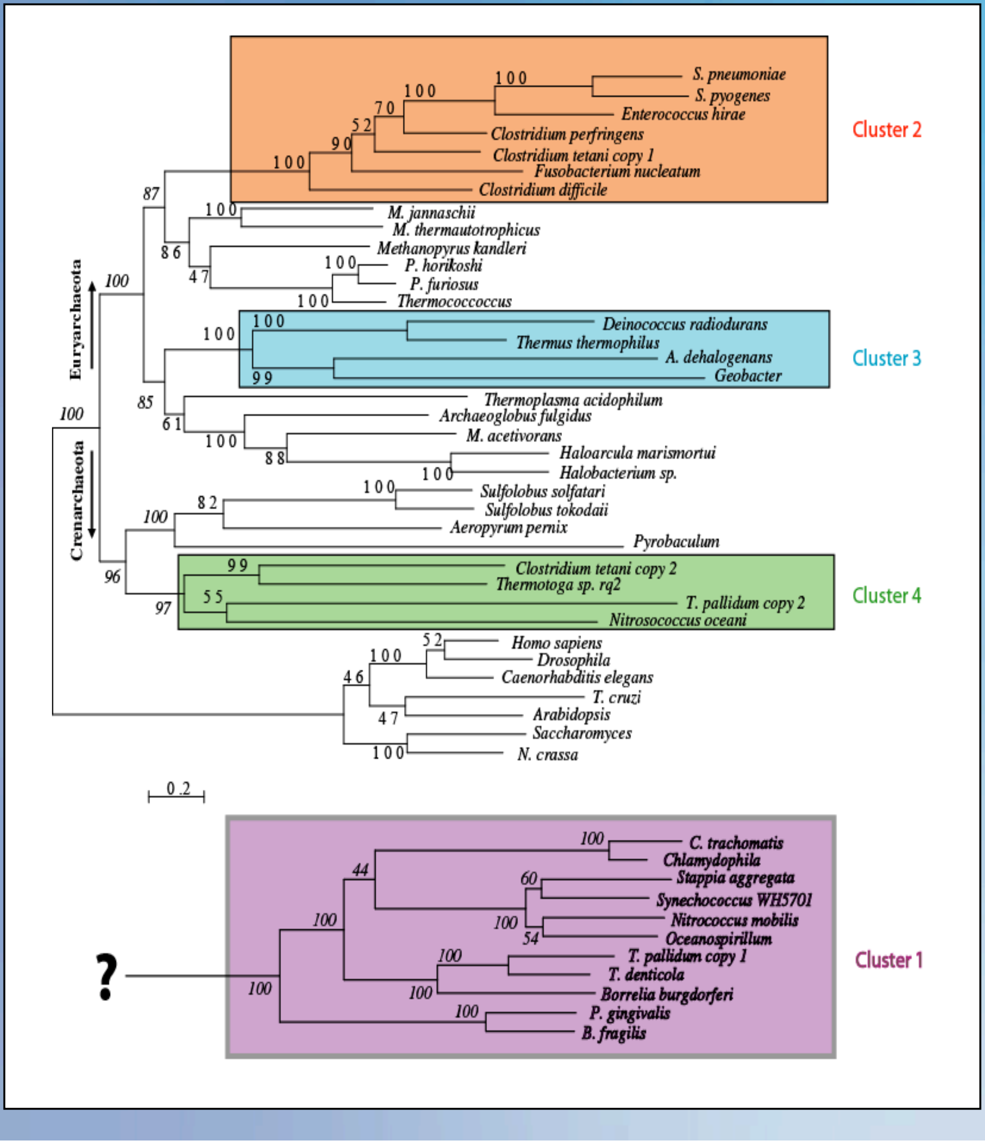


PhyML tree using WAG model, among site variations with 8 categories, estimated pinvar

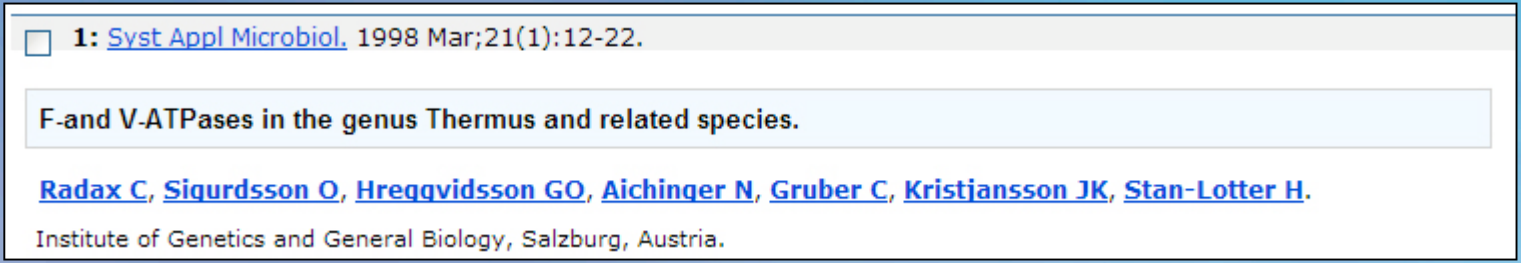

Subunit I

Concatenated A-B-I subunits

At least three ancient independent transfers



Why an Archaeal ATP synthases?



□ 1: Syst Appl Microbiol. 1998 Mar;21(1):12-22.

F-and V-ATPases in the genus Thermus and related species.

Radax C, Sigurdsson O, Hreggvidsson GO, Aichinger N, Gruber C, Kristjansson JK, Stan-Lotter H.

Institute of Genetics and General Biology, Salzburg, Austria.

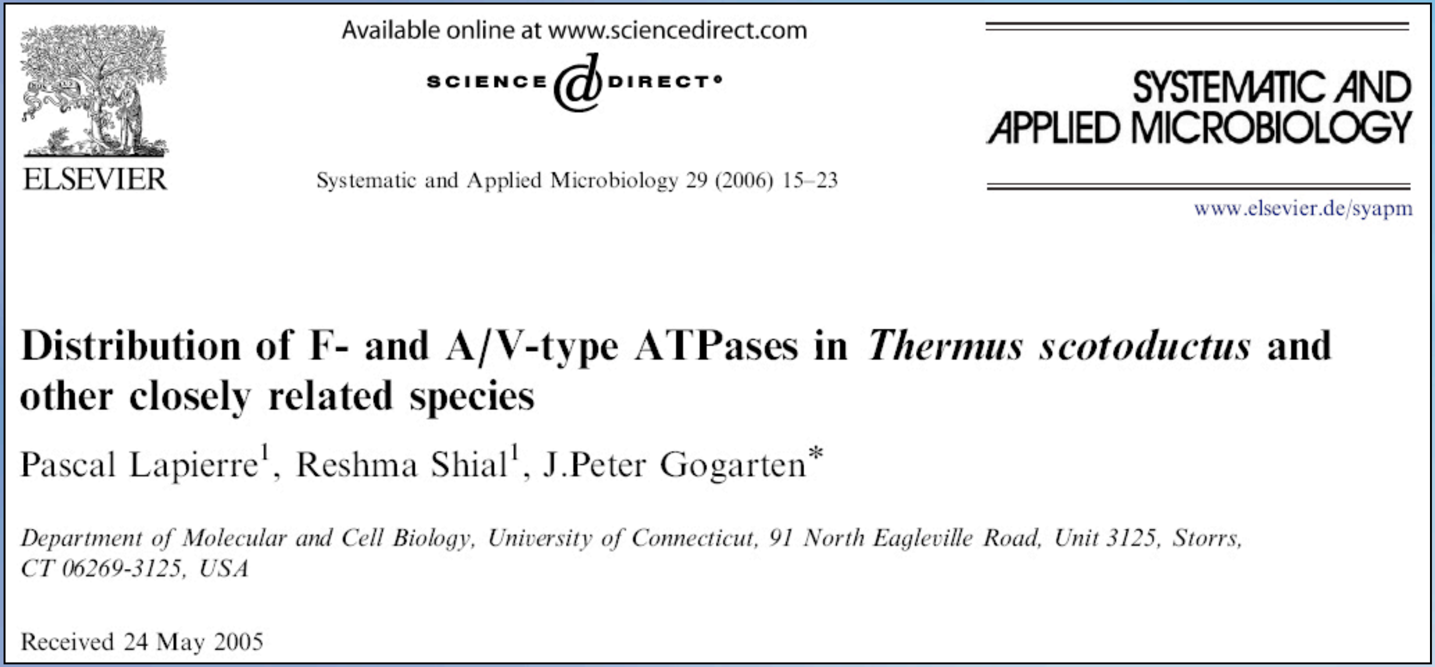
Few sequenced peptide residues were 100% identical to an F-ATPase from *Bacillus*

Compare *T. thermophilus* (V-type) and *T. scotoductus* (F- type) to find evolutionary reasons between having one or two different ATP synthase

Reshma Shial

Not so fast….

PCR amplification, sequencing and Northern blots have shown that *T. scotoductus* does not possess an F-type ATP synthase



Available online at www.sciencedirect.com

SCIENCE DIRECT®

ELSEVIER

Systematic and Applied Microbiology 29 (2006) 15–23

SYSTEMATIC AND APPLIED MICROBIOLOGY

www.elsevier.de/syapm

Distribution of F- and A/V-type ATPases in *Thermus scotoductus* and other closely related species

Pascal Lapierre[1], Reshma Shial[1], J.Peter Gogarten*

Department of Molecular and Cell Biology, University of Connecticut, 91 North Eagleville Road, Unit 3125, Storrs, CT 06269-3125, USA

Received 24 May 2005

General characteristics of Thermotogales

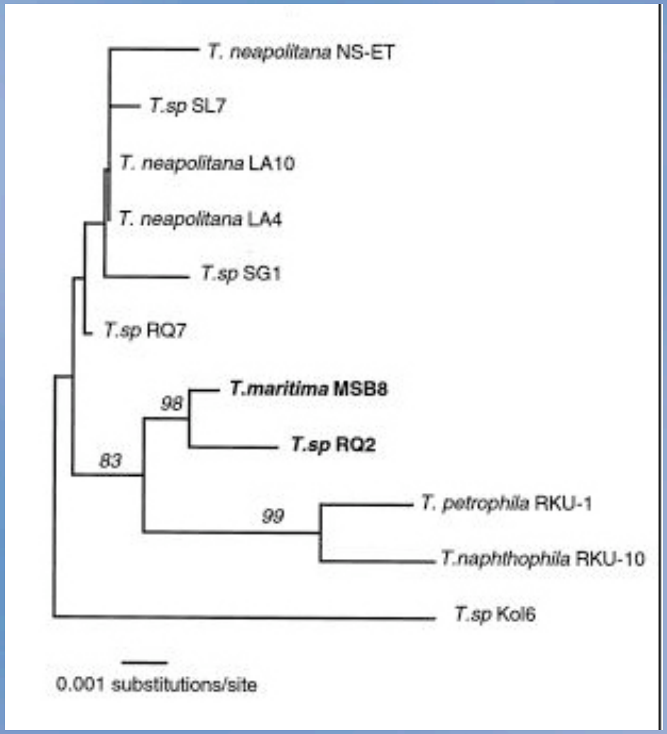
- Thermotogales are a group of deep branching bacteria that live at high temperatures (80 degrees C) near volcanic vents.

- They live around thermophilic Archaea. It has been estimated that **24%** of the genes were acquired from Archaea via HGT's (Based on data from *T. maritima* MSB8).

- New isolates show a mesophilic lifestyle
(C. Nesbo, J. Dipippo)

Strains used

Parsimony 16s rRNA tree



From Nesbø *et al.*, J Bacteriol. 2002 Aug;184(16):4475-88

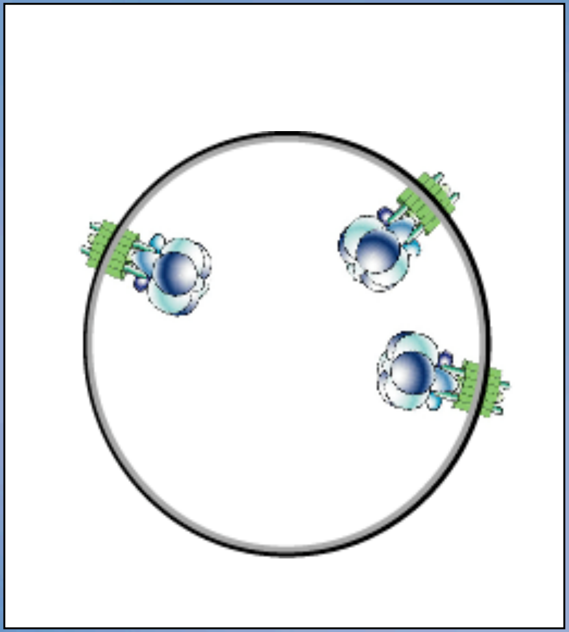
- **Strain MSB8 and RQ2 have 99.7% identity in the small-subunit rRNA sequence**

- **RQ2 possess an F- and A/V-type ATP synthase.**

- **MSB8 possess only an F-Type**

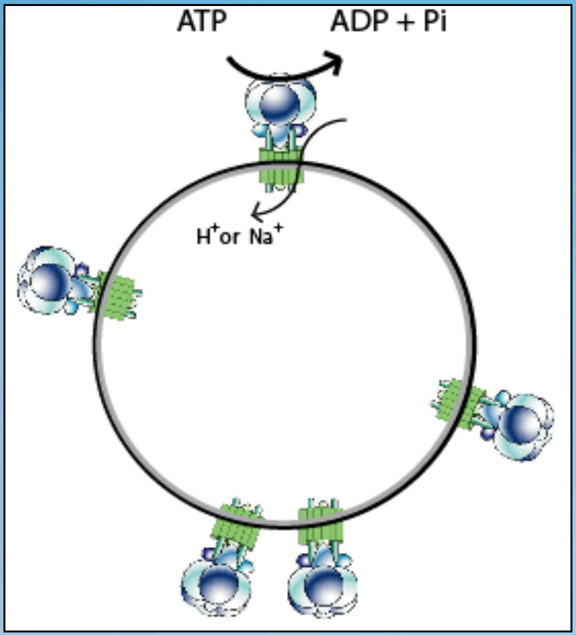
Inverted membranes

Normal Vesicles



AND

Inside-out Vesicles

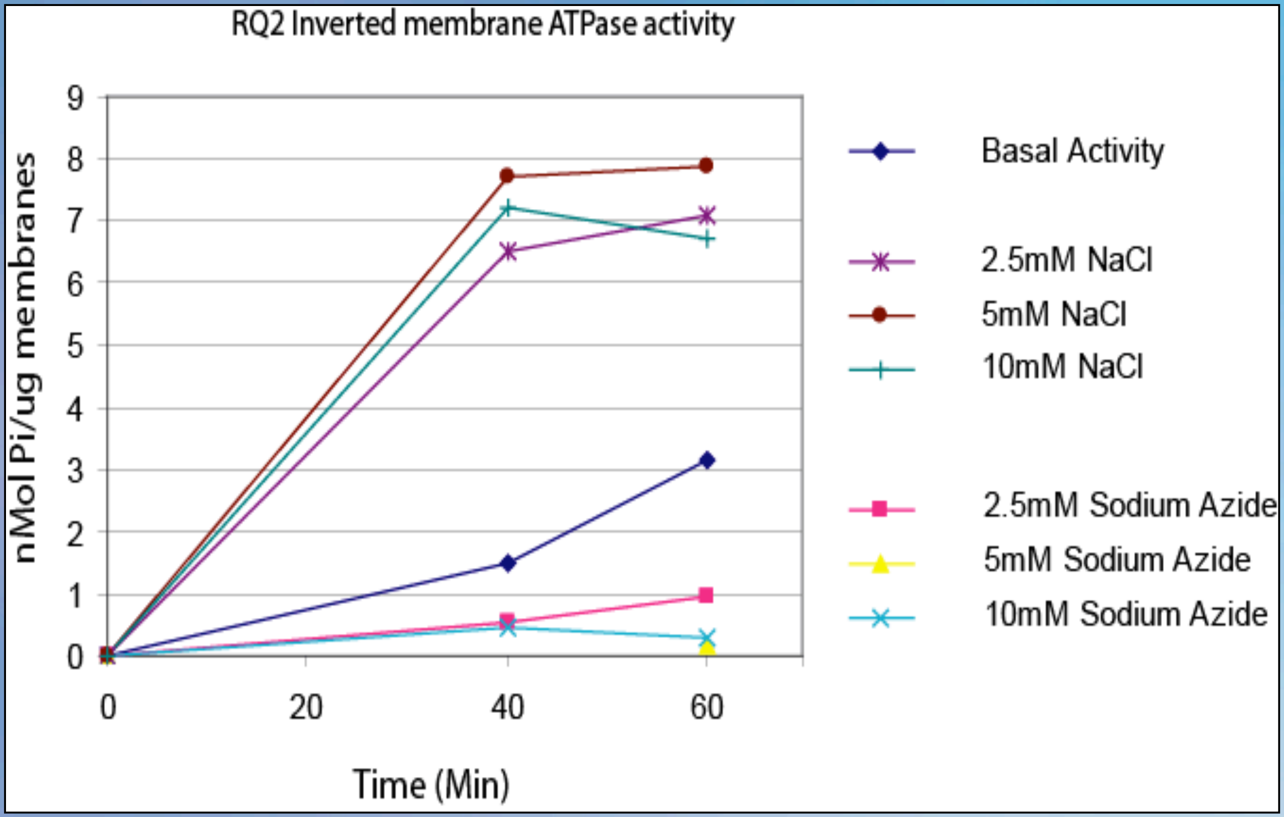


- **Malachite Green Assays:**

  - Release of free phosphate molecules (Pi) resulting from the ATP hydrolysis (ATPase activity) causes a change in absorbance of a colored phosphomolybdate malachite green complex measurable at 630nm.

Class of chemical	Effects on:	Mode of Action
Inhibitors:		
Sodium Azide (NaN_3)	F_0F_1	Stabilize an inactive complex between ADP and the F_0F_1 ATPase[13].
Diethylstilbestrol (DES)	F_0F_1, A_0A_1?	Mode of action unknown, uncoupling of ATP synthesis?[14,15].
N-ethylmaleimide (NEM)	V_0V_1, A_0A_1	React with the cysteine residues of the catalytic subunits[16].
Sodium Vanadate	V_0V_1, A_0A_1	Inhibit phosphorolated intermediate of the ATPase[17].
Bafilomycin	V_0V_1, A_0A_1	Bind to at least one protein of the V_0 sector[18,19].
Nitrate	V_0V_1, A_0A_1	Uncouples H^+ pumping from ATP hydrolysis[20].
DCCD	F_0F_1, V_0V_1, A_0A_1	Bind to the free carboxyl group of the proteolipid subunits in hydrophobic environments[21].
Oligomycin	F_0F_1	Bind to F_0, alter the ATP binding properties of F_1[22].
Ionophores :		
FCCP	H^+	Allow equilibration of H^+ across the membrane or vesicle[23].
Nigericin	Na^+	Allow exchange diffusion of Na^+/K^+ across the membrane or vesicle[24].

F-ATPase is activated in presence of Na^+

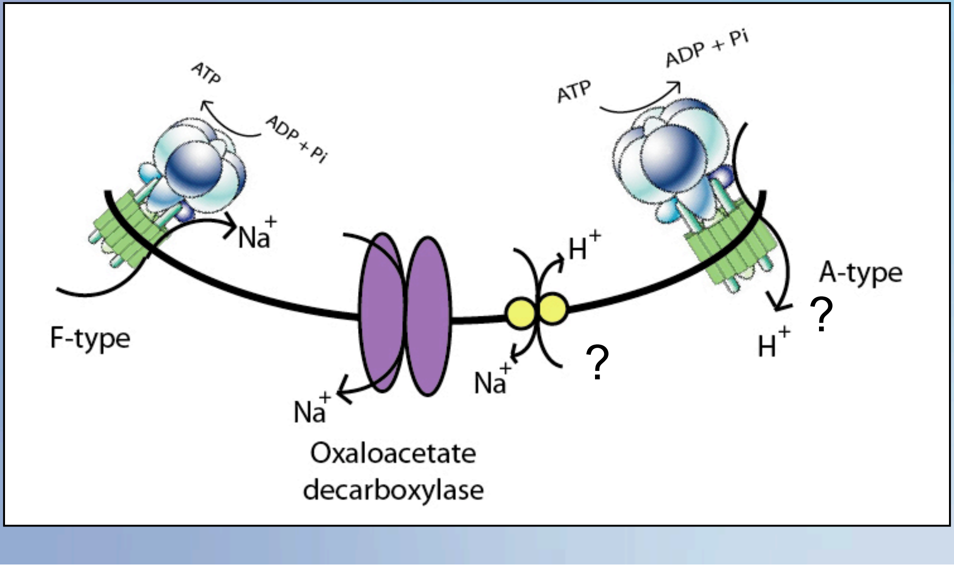


No activity from the A-type ATPase was detected!

Other work

New experiments are underway to directly measure by real-time PCR ATPase rRNA expression in growing culture under varying conditions (K. Swithers)

Nine strains of Thermotogales (including RQ2) are being sequenced (K. Noll). Sequence comparisons may provide further clues on the metabolisms of the different strains/species.



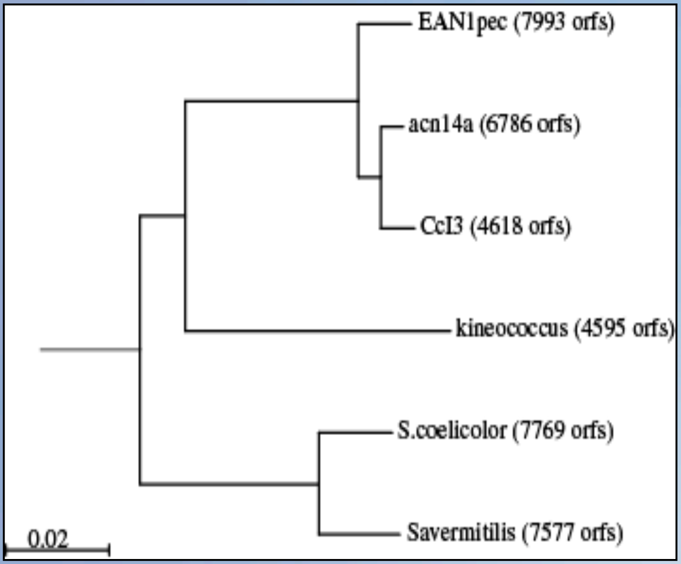
Part II :

Comparative analysis of three newly sequenced *Frankiacea* genomes

- *Frankia* sp. are nitrogen-fixing actinomycetes, high G+C gram-positive actinobacteria that form root nodules on ecologically important actinorhizal plants

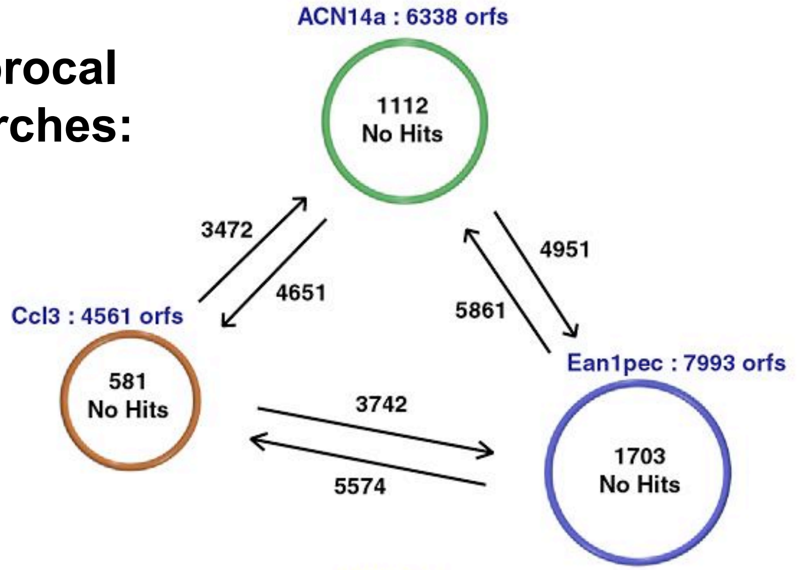
- 97.8% to 98.9% identity over the 16s rRNA

Strains	Length	Predicted ORFs	Seq. Center	Status
Frankia sp. strain HFPCcI3	4.53 Mbp	4618 orfs	JGI	Completed
Frankia alni strain ACN14a	7.50 Mbp	6786 orfs	Genoscope	Completed
Frankia sp. EAN1pec	9.04 Mbp	8026 orfs	JGI	Unfinished

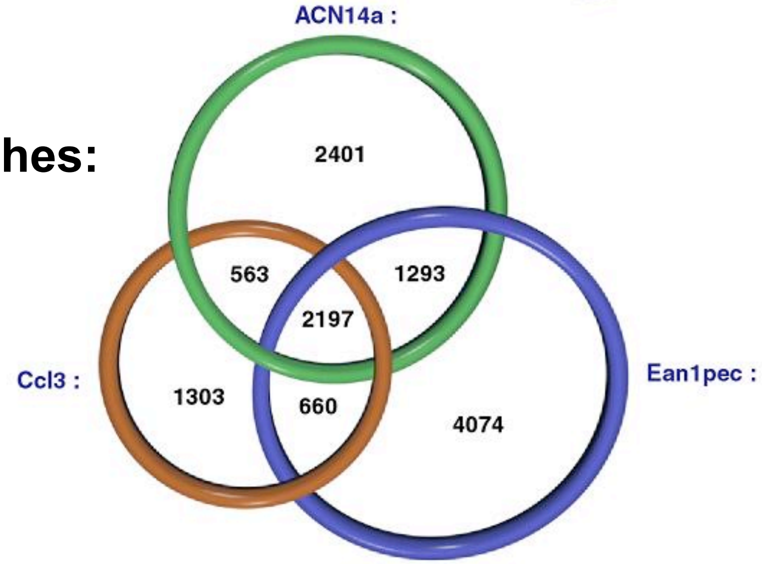


Blast comparisons using a bit score cutoff of 50 (~10e-04)

Non-reciprocal Blast searches:



Reciprocal Blast searches:



Comparison of Gene Families

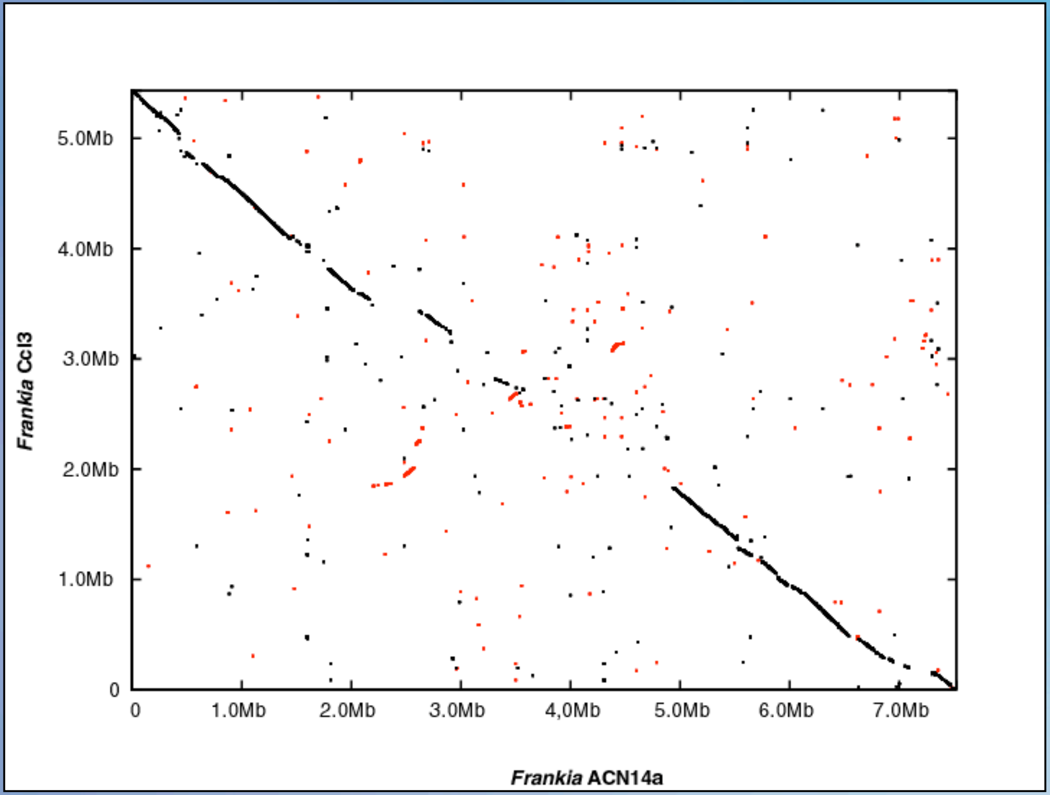
Result from BlastClust (25% identity over 40% of the length) : Equivalent results using TRIBE-MCL

Cci3	Acn	Ean	Total	Predicted function
20	101	131	252	Dehydrogenase
42	100	106	248	Putative ABC transporter ATP-binding protein
30	64	75	169	WD-40 repeat protein
20	47	41	108	FadD8
17	36	48	101	Putative membrane transport protein.
8	41	43	92	Putative acyl-CoA dehydrogenase
12	25	52	89	CYTOCHROME P450
12	21	45	78	Putative two-component system response-regulator
4	35	34	73	Putative enoyl-CoA hydratase
11	23	38	72	Multi-domain Polyketide synthases
13	25	24	62	Hypothetical protein
6	22	31	59	Putative Betaine Aldehyde Dehydrogenase (BADH)
2	23	33	58	Putative fatty acid-CoA racemase
11	15	29	55	Sensory box protein
…	…	…	…	…

155	**33**	**195**	**383**	**Transposases**
32	**13**	**74**	**119**	**Integrases**

Synteny between genomes

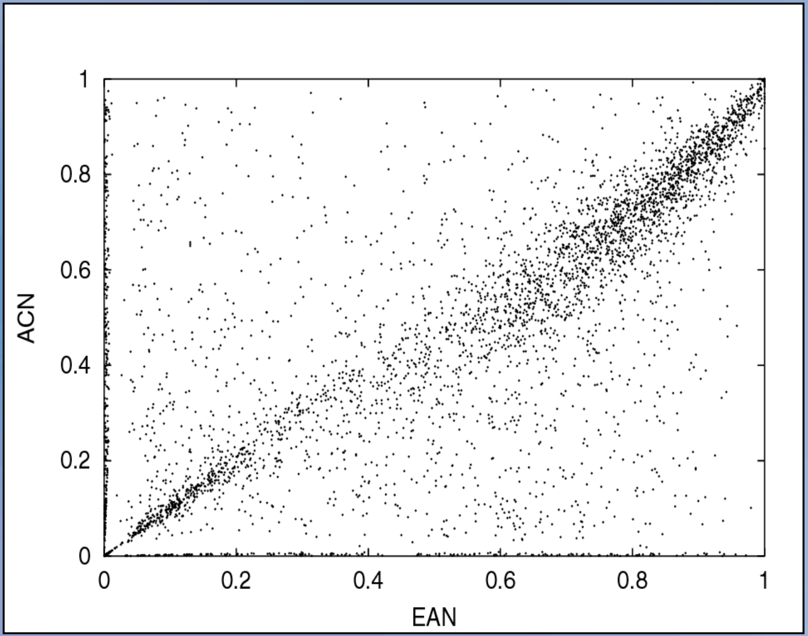
Nucleotide-nucleotide genome comparison using Mummer

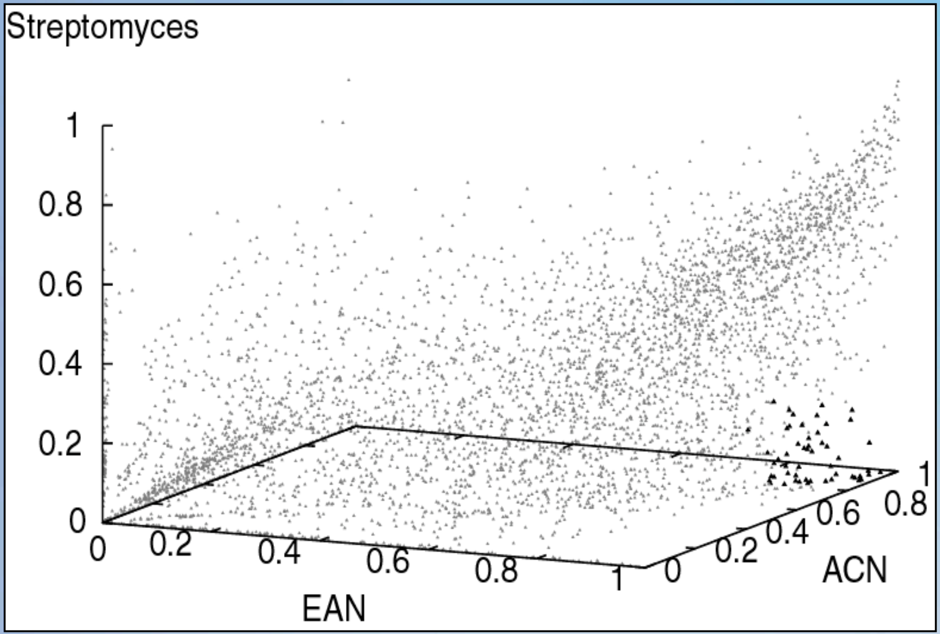


BLAST SCORE RATIO (BSR) PLOTS*

- **Blast each ORFs against itself from a reference genome (Ccl3) (Reference bit score)**

$$\text{BSR pair} = \left(\text{BSR}_1 = \frac{\text{Query}_1}{\text{Reference}},\ \text{BSR}_2 = \frac{\text{Query}_2}{\text{Reference}} \right)$$



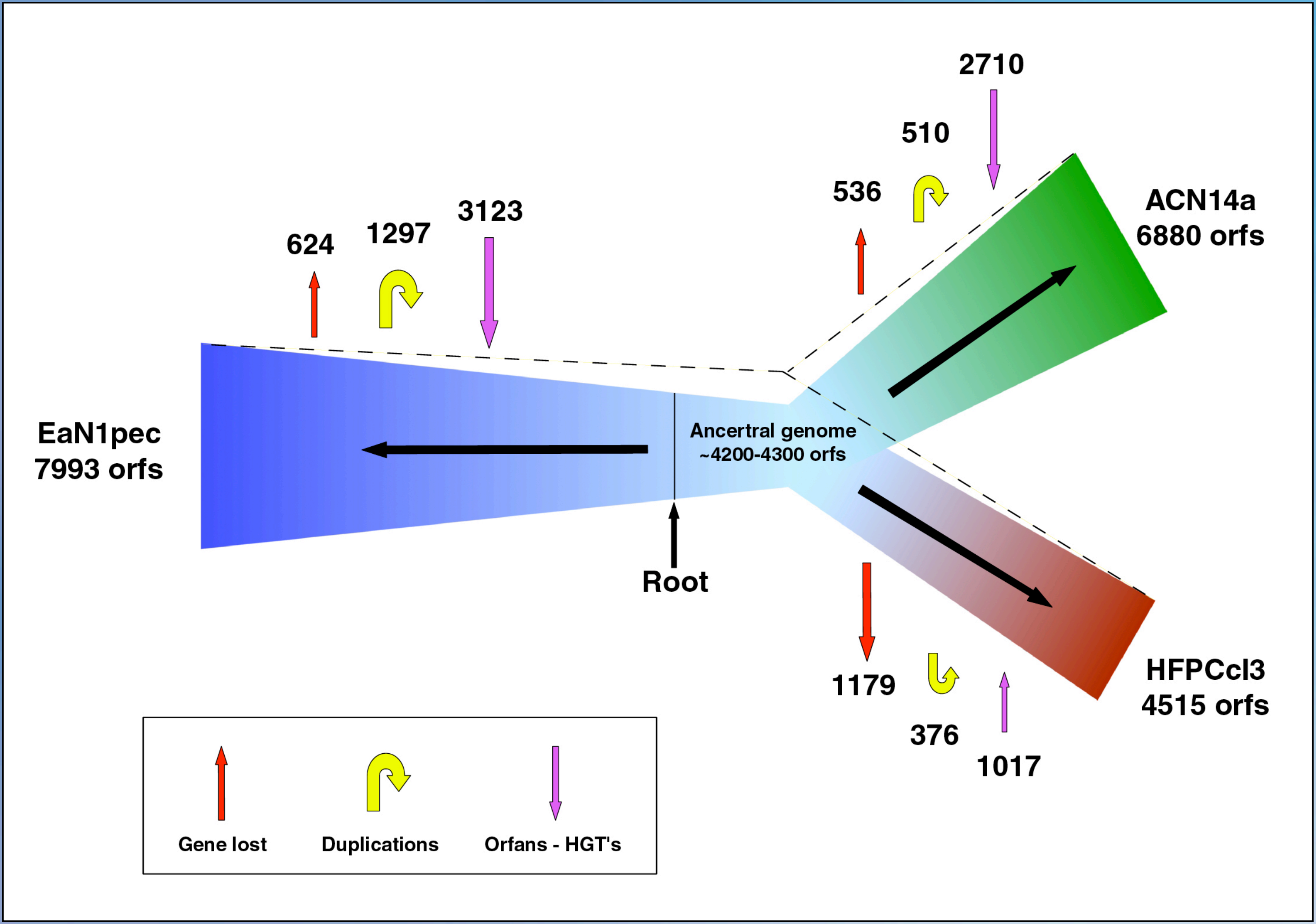


(Graphics generated in GNUplot)

***BMC Bioinformatics. 2005; 6: 2**

Estimation of the ancestral genome state

Using data obtain from self blasts, blasts against other *Frankia*'s and NR database

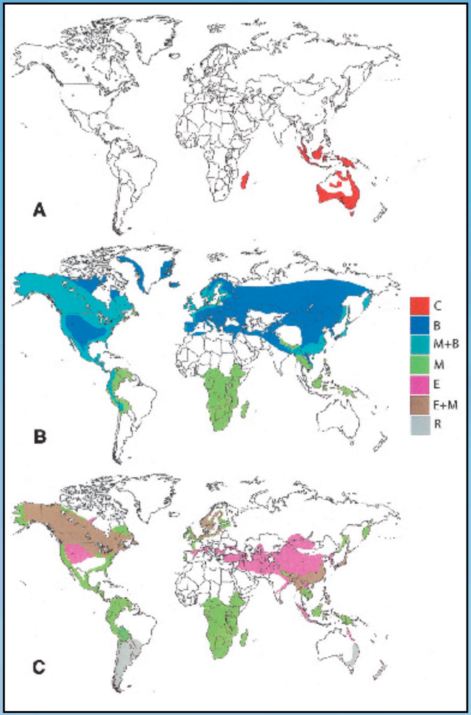


Conclusions

-The genome sizes correlate with the biogeographic distribution and host ranges of the *Frankia* sp. strains

-The reduce genome size of Ccl3 might be indicative that the strain is on his way to became an obligate symbionts

-The amounts transposable elements found in Ccl3 and EaN1pec may have play an important role in genome size differences



Genome characteristics of facultatively symbiotic *Frankia* sp. strains reflect host range and host plant biogeography

Genome Research, 2007 Jan;17(1):7-15

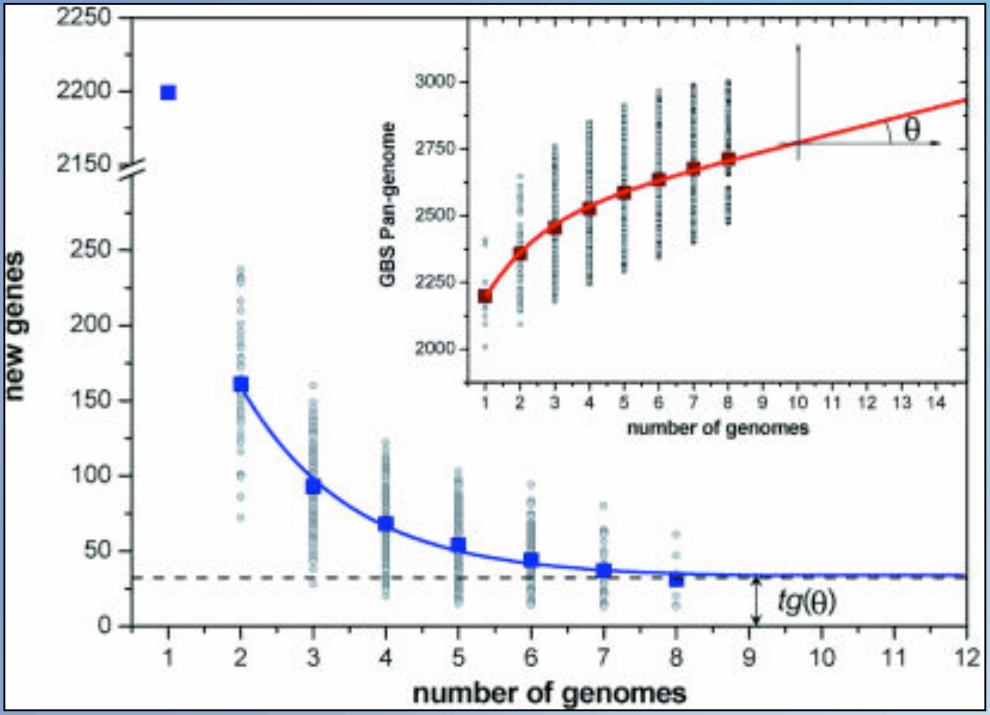
Philippe Normand, **Pascal Lapierre**, Louis S. Tisa, J. Peter Gogarten, Nicole Alloisio, Emilie Bagnarol, Carla A. Bassi, Alison M. Berry, Derek M. Bickhart, Nathalie Choisne, Arnaud Couloux, Benoit Cournoyer, Stephane Cruveiller, Vincent Daubin, Nadia Demange, M. Pilar Francino, Eugene Goltsman, Ying Huang, Olga R. Kopp, Laurent Labarre, Alla Lapidus, Celine Lavire, Joelle Marechal, Michele Martinez, Juliana E. Mastronunzio, Beth C. Mullin, James Niemann, Pierre Pujic, Tania Rawnsley, Zoe Rouy, Chantal Schenowitz, Anita Sellstedt, Fernando Tavares, Jeffrey P. Tomkins, David Vallenet, Claudio Valverde, Luis G. Wall, Ying Wang, Claudine Medigue, & David R. Benson

Part III :

The bacterial pan-genome

Description of the group B *Streptococcus* pan-genome

Genome comparisons of 8 closely related GBS strains



Tettelin, Fraser et al., PNAS 2005 Sep 27;102(39)

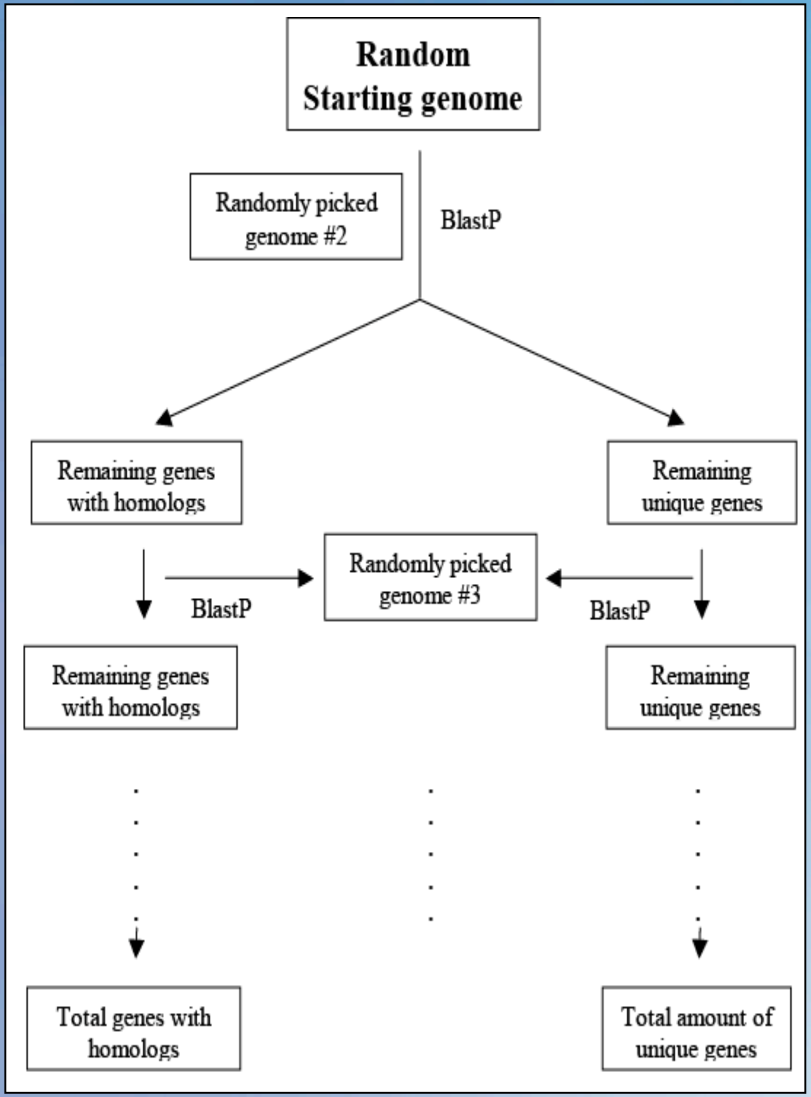
Goal

Using all the complete genome sequences, is it possible to describe the complete bacterial pan-genome using the same extrapolation methods?

Dataset :

- 293 completed bacterial genomes

Method

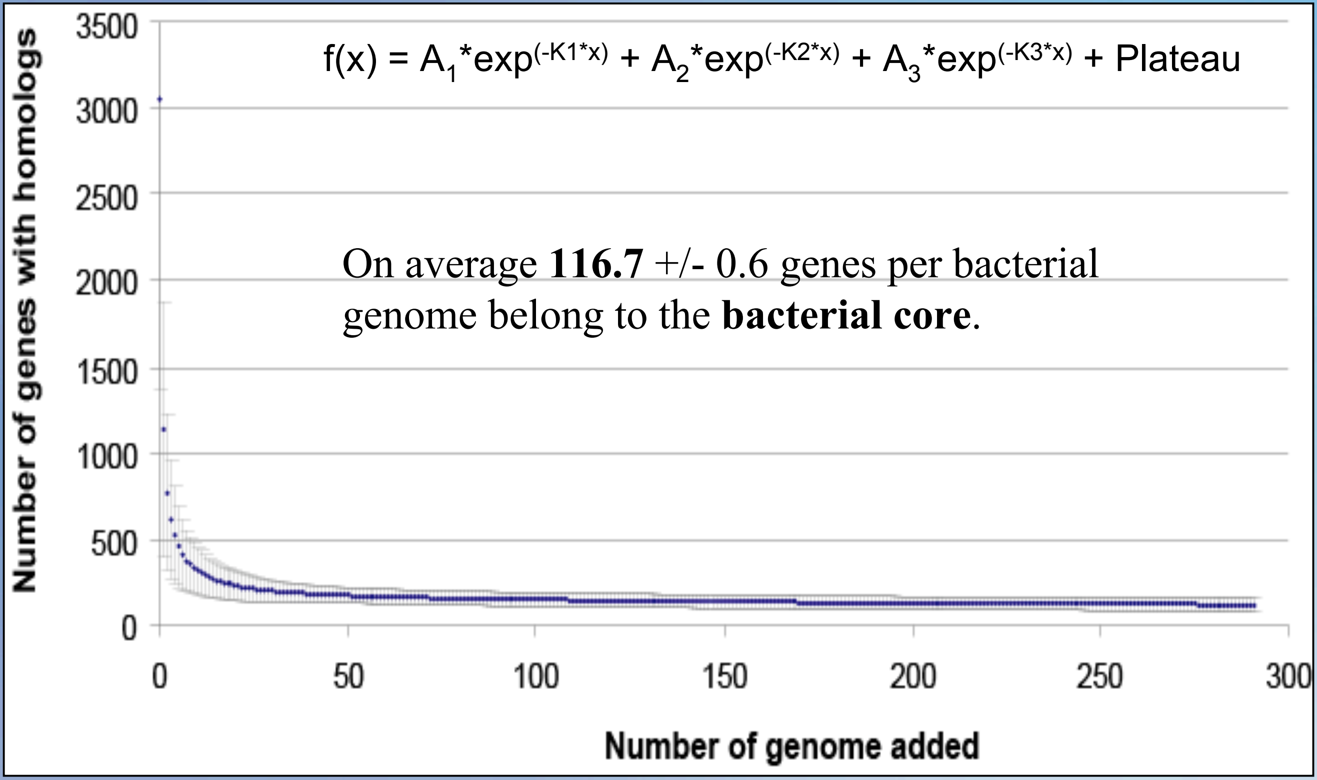


Total of 1011 sampling runs

The Bacterial Core

Genes that are shared among all bacteria

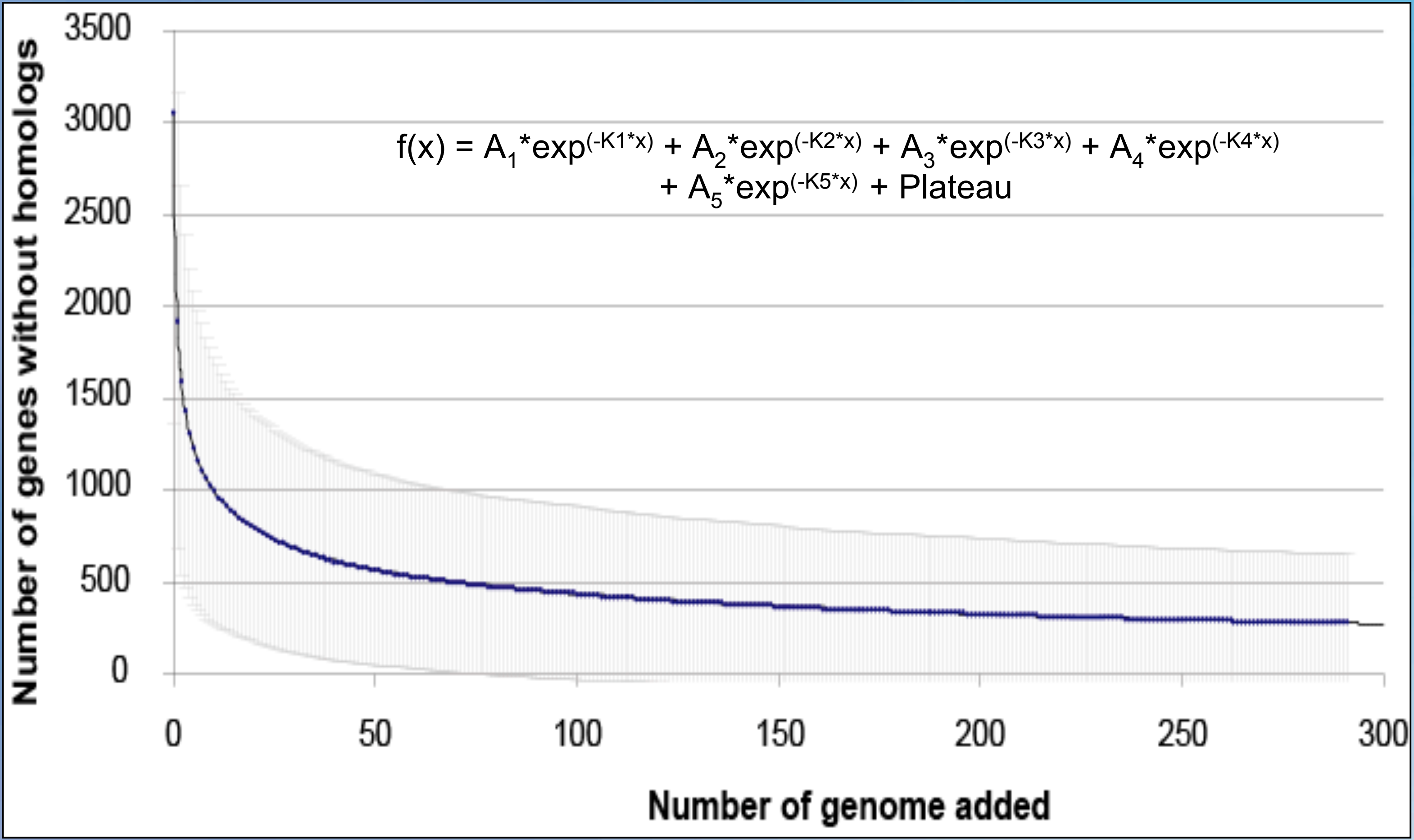
Bit score cutoff 50.0 (~10E^{-4})



$f(x) = A_1\text{*}\exp^{(-K1\text{*}x)} + A_2\text{*}\exp^{(-K2\text{*}x)} + A_3\text{*}\exp^{(-K3\text{*}x)} + \text{Plateau}$

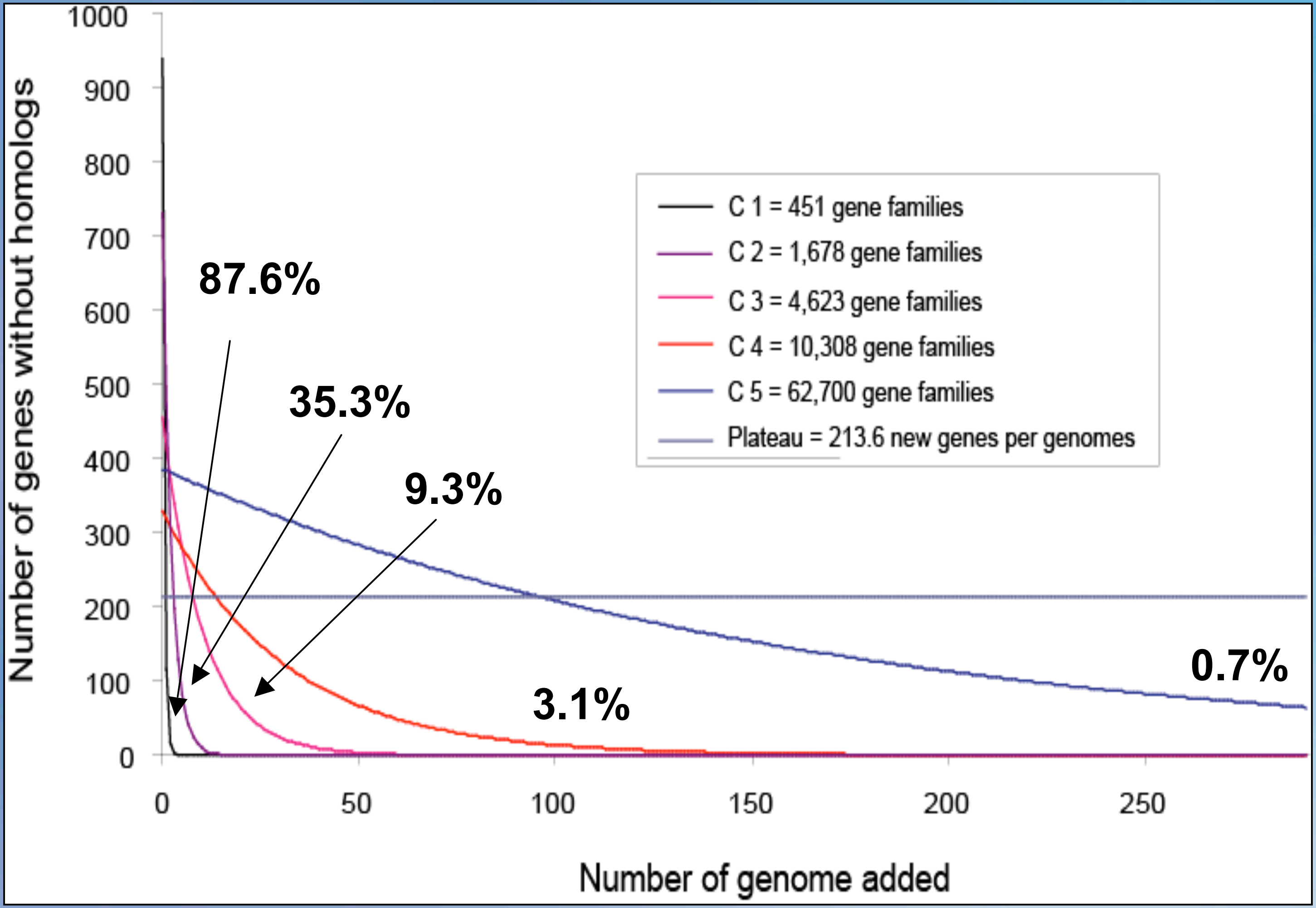
On average **116.7** +/- 0.6 genes per bacterial genome belong to the **bacterial core**.

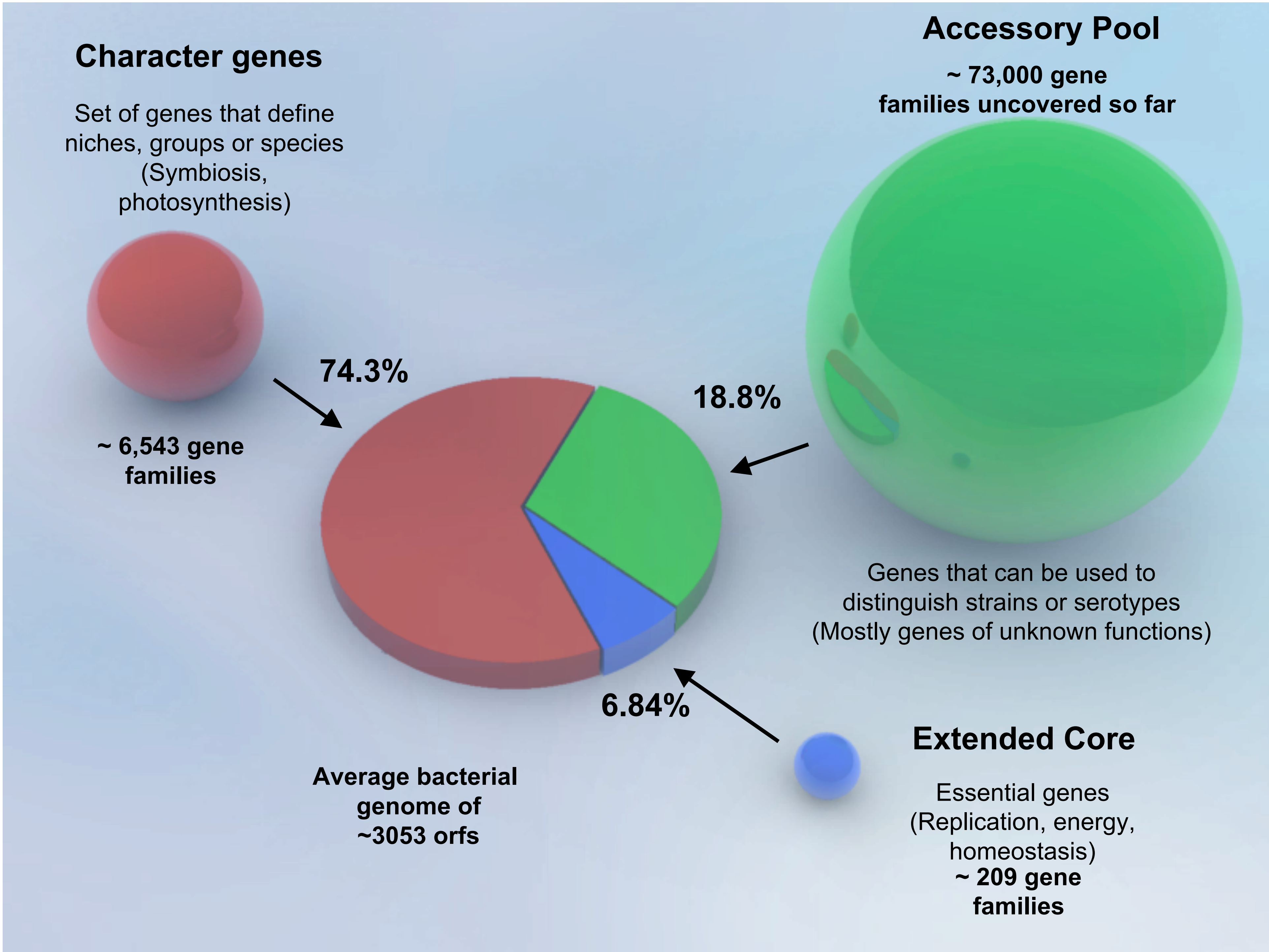
Genes without homologs



$$f(x) = A_1{*}\exp^{(-K1{*}x)} + A_2{*}\exp^{(-K2{*}x)} + A_3{*}\exp^{(-K3{*}x)} + A_4{*}\exp^{(-K4{*}x)} + A_5{*}\exp^{(-K5{*}x)} + \text{Plateau}$$

Decomposed function





Character genes

Set of genes that define niches, groups or species (Symbiosis, photosynthesis)

~ 6,543 gene families

Accessory Pool

~ 73,000 gene families uncovered so far

Genes that can be used to distinguish strains or serotypes (Mostly genes of unknown functions)

Extended Core

Essential genes (Replication, energy, homeostasis)
~ 209 gene families

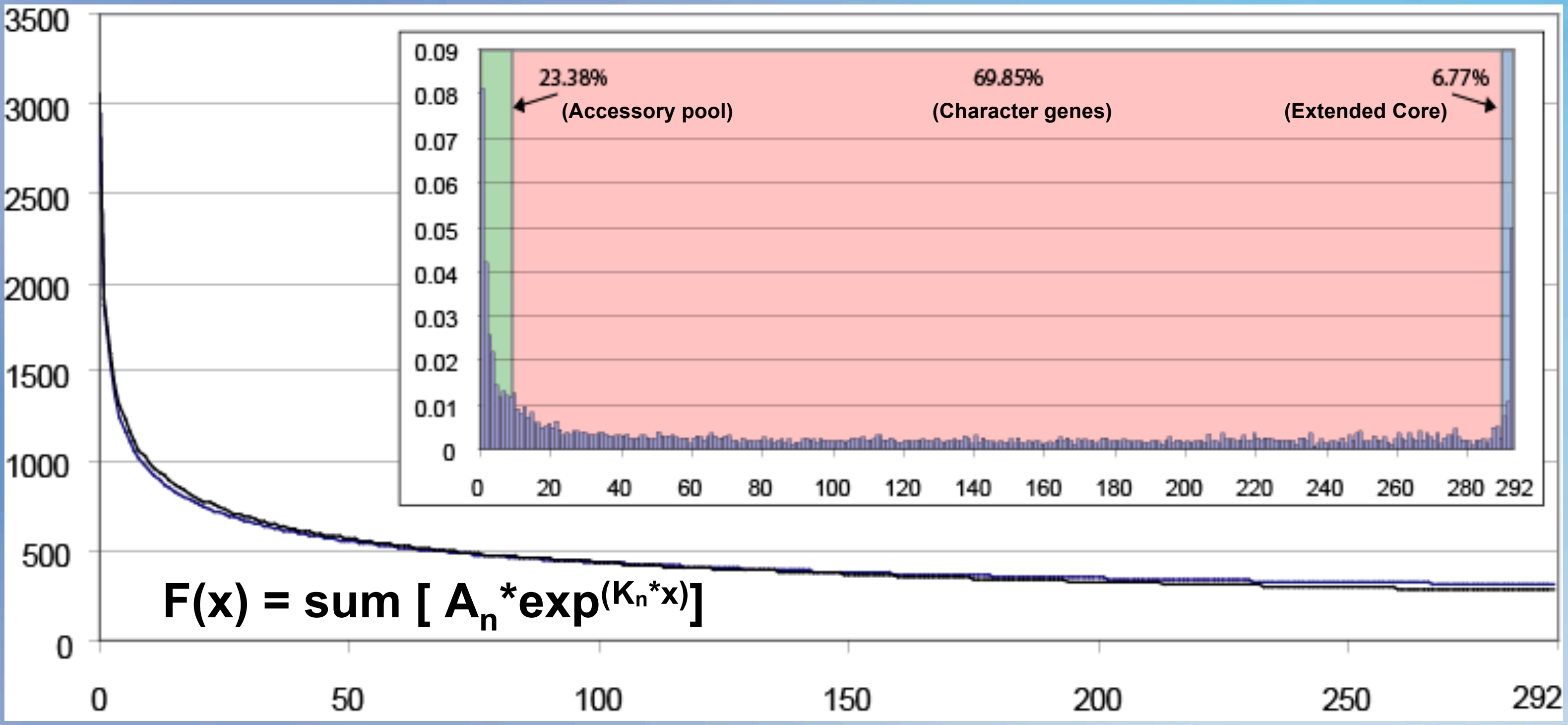
Average bacterial genome of ~3053 orfs

Gene frequency in a typical genome

-Pick a random gene from any of the 293 genomes

-Search in how many genomes this gene is present

-Sampling of 15,000 genes



$F(x) = \text{sum} [A_n * \exp^{(K_n * x)}]$

Evolutionary Mechanisms

Extended Core : - Very high selective pressure, drastic changes harmful
- Fine tuning of the active regions by point mutations

Character genes : - These proteins evolve through gene transfer, gene duplication and substitutions
- Acquisition of new functions using a "Lego" principle i.e., the reuse of already existing building blocks

Accessory Pool : - High turnover rates in genomes; they are not subject to strong selective pressures
- Frequently reside in phage and extrachromosomal genetic elements
- This pool may allow creation of new proteins from 'scratch'

Part IV :

Whole genome approach to estimate molecular clocks using a Bayesian framework

Collaborative works done with Dr. Lynn Kuo and Dr. Ming-Hui Chen from the UConn Department of Statistics

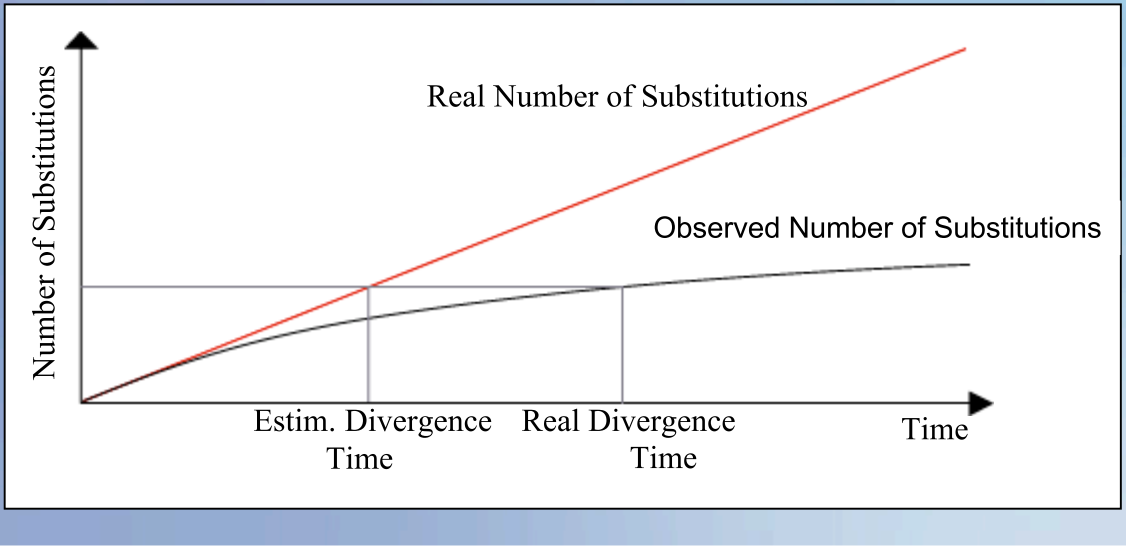
Molecular Clocks

-Using DNA substitution to estimate dates of past events

-Based on the assumption that substitutions occurred at a fairly constant rates (like the regular ticking of a clock)

Problems associated with molecular clocks

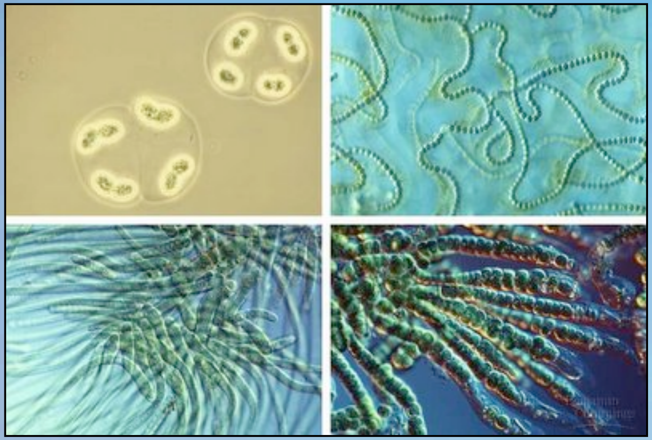
-Rates of mutations is not constant between species, saturation
-Accuracy and sparseness of the fossil records
-Difficulties of phylogenetic reconstruction, HGTs



Cyanobacteria

-The rise of oxygen on earth around 2.3 billion years ago

-Most likely, the cyanobacteria were already present before



Taken from http://scienceblogs.com/clock/2006/09 /circadian_clocks_in_microorgan.php

-Previous molecular clock estimates date cyanobacteria at 2.6 GyA

(BMC Evolutionary Biology 2001, 1:4)

-Biochemical evidences point toward 3.7 GyA

(Earth and Planetary Science Letters 217, 237-244)

Project overview

Traditional molecular methods use either a single molecular marker or concatenation of many genes for time estimates. Both methods can potentially include datasets with presence of HGTs.

Instead of using a **single gene** to date the divergence of the cyanobacteria, we calculate clock on **genomic set** of orthologous genes and **combine** the results under Bayesian probability framework. Only nodes compatible with a reference tree are used for the final time estimation.

-Build datasets of orthologous genes from cyanobacteria genomes

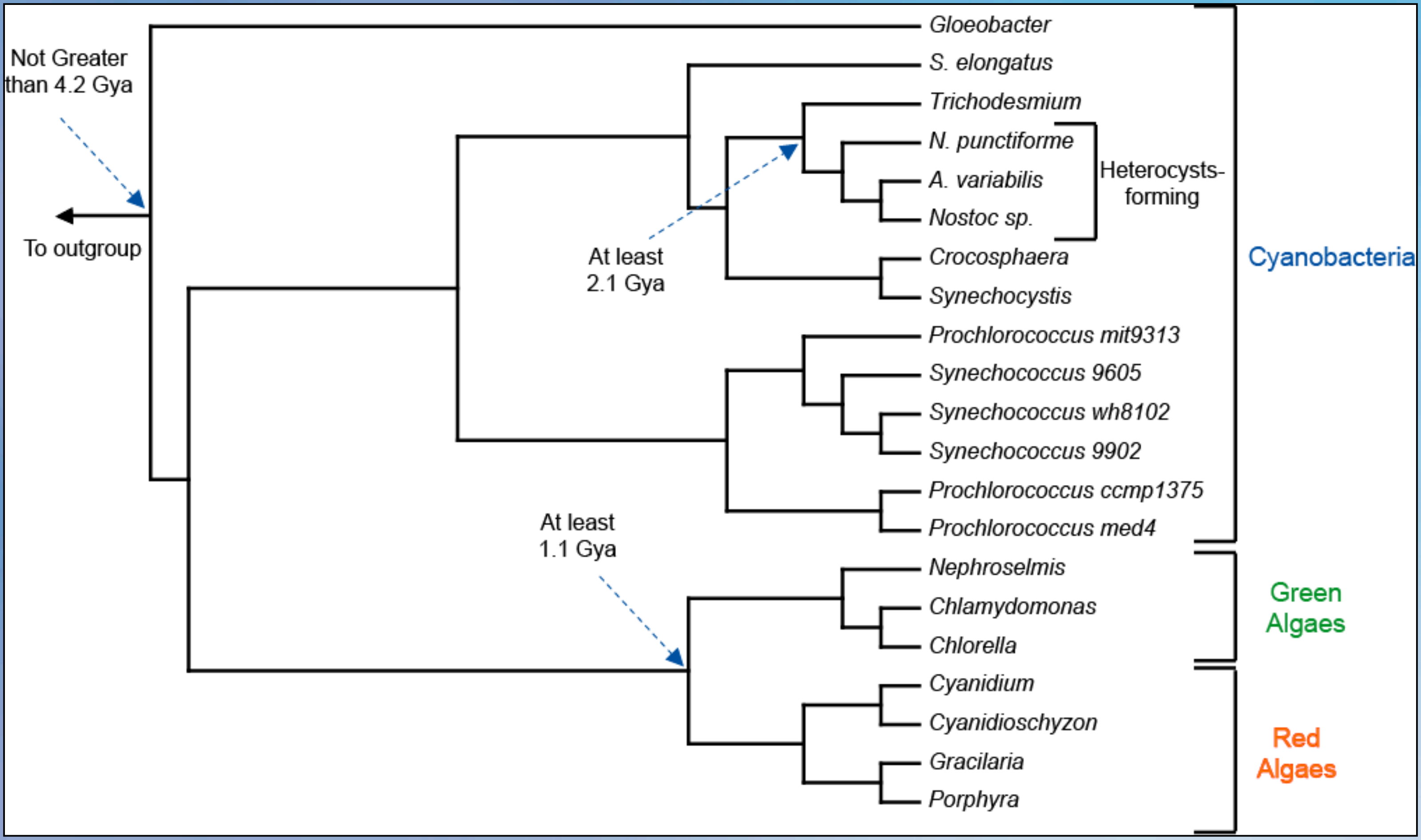
**-Calculate clock on individual datasets using the Thornian Time Traveler*
(Local clock model, Multiple calibration points, only allows hard priors)**

-Combine the posterior probability distributions of the time estimates into a final probability of time intervals for each nods of a consensus tree

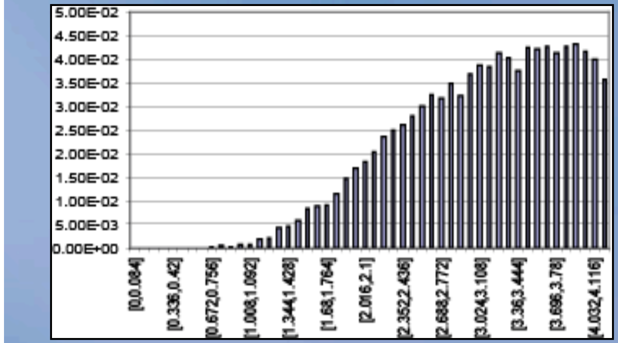
***Molecular Biology and Evolution, Vol 15, 1647-1657**

Times Estimation (Thornian Time Traveler)

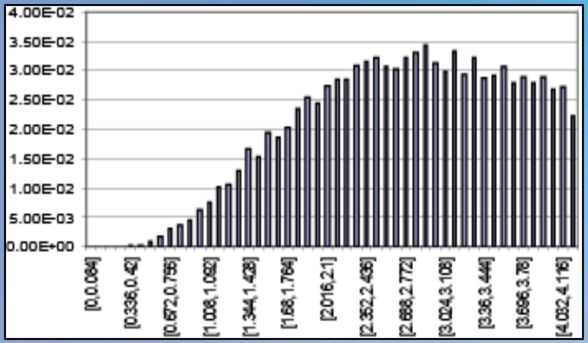
Multidivtime: Performed a Bayesian MCMC analysis to approximate the posterior distributions of subs. rates and divergence times.



Combining the posterior probabilities

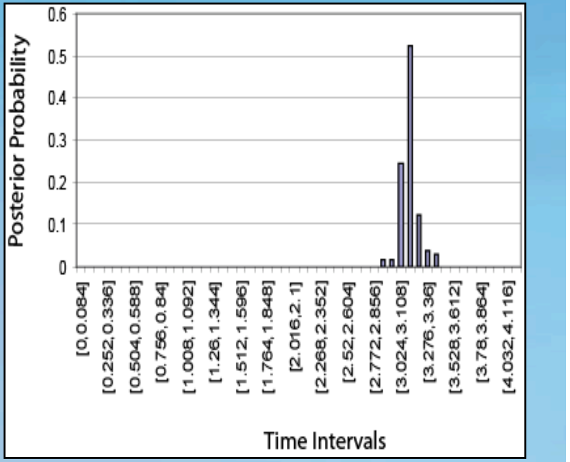


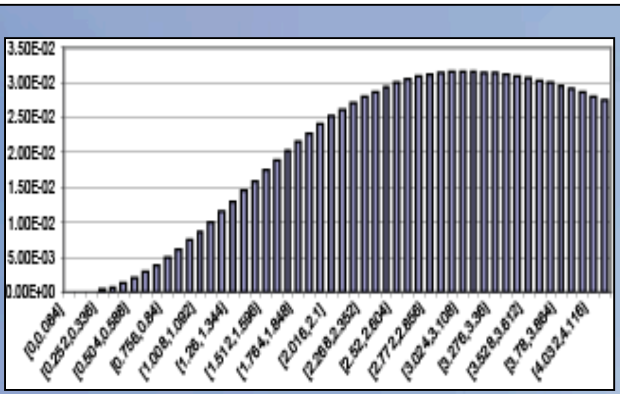
*



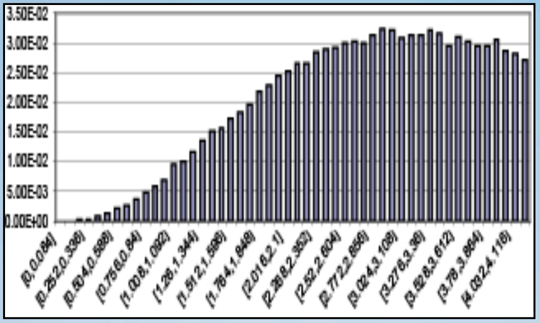
$$P(S|D_1, D_2 \ldots D_n) = \frac{P(S|D_1) * P(S|D_2) * \ldots * P(S|D_n)}{(P(S)^{(n - 1)} * k)}$$

Combined time estimates





Smoothed prior



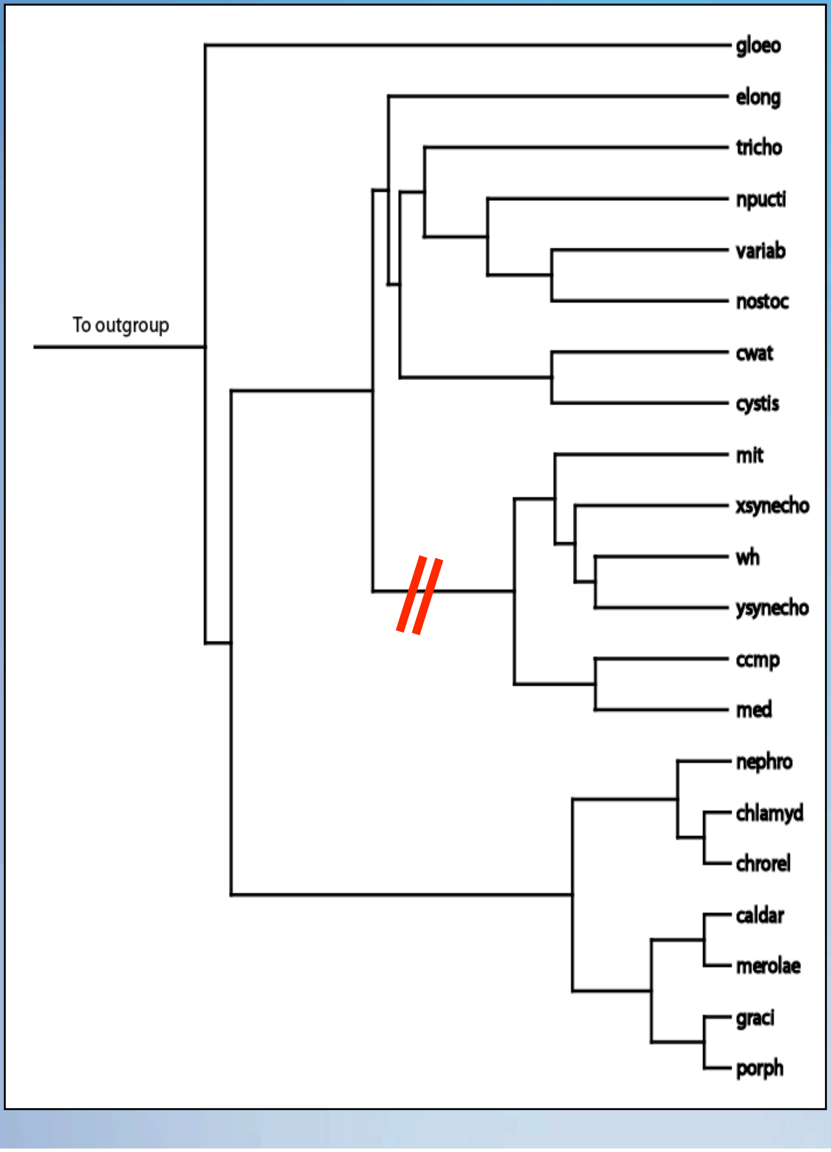
Original prior

Consensus Tree

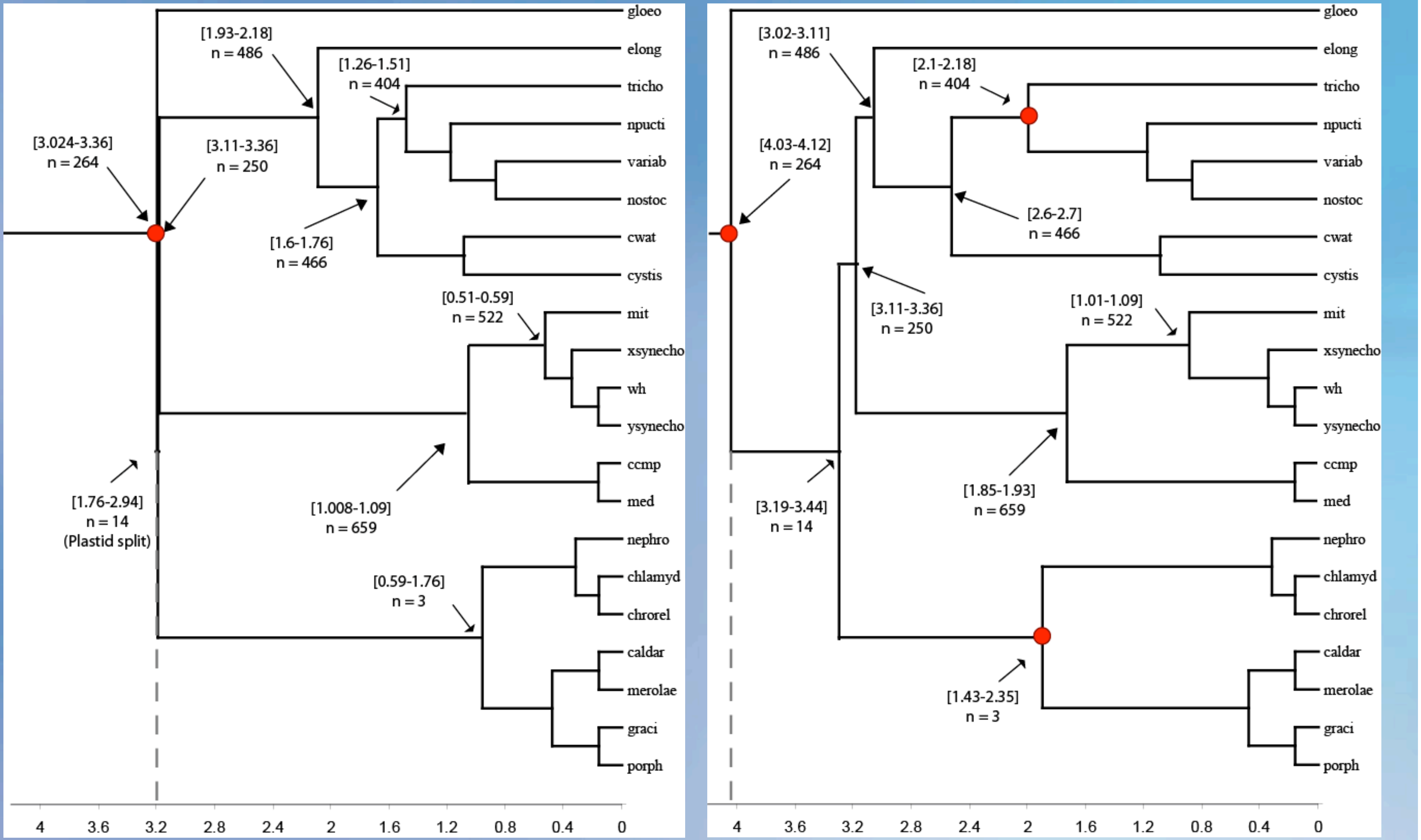
Inferred from MRP Supertree, Concatenated genes, Literature

For each node of interest, screen for corresponding bi-partition in each dataset

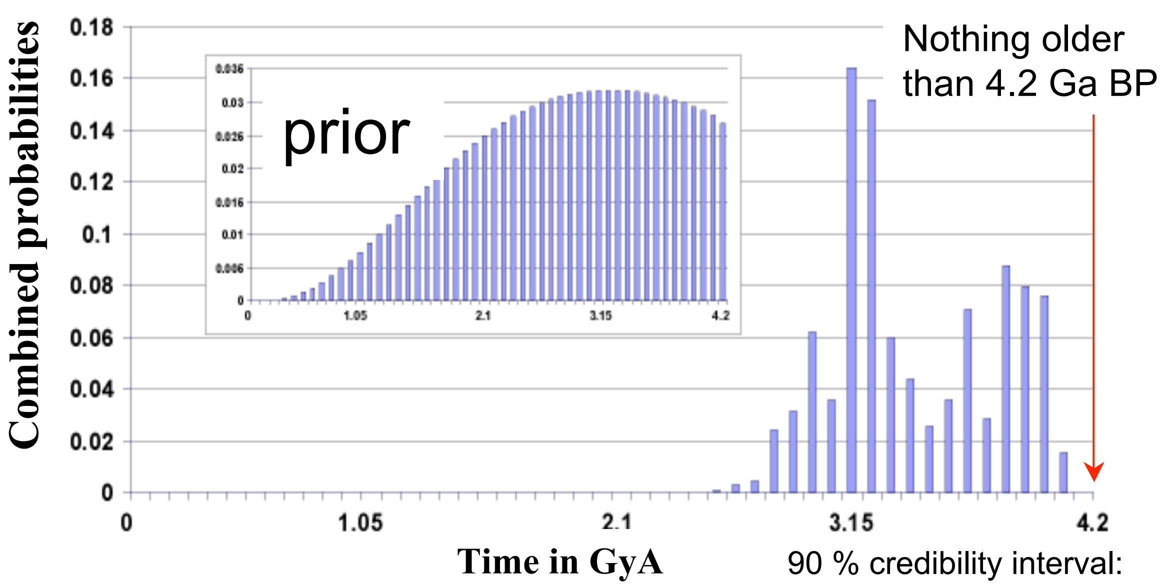
Minimize the effects of HGTs



Results combined time estimates

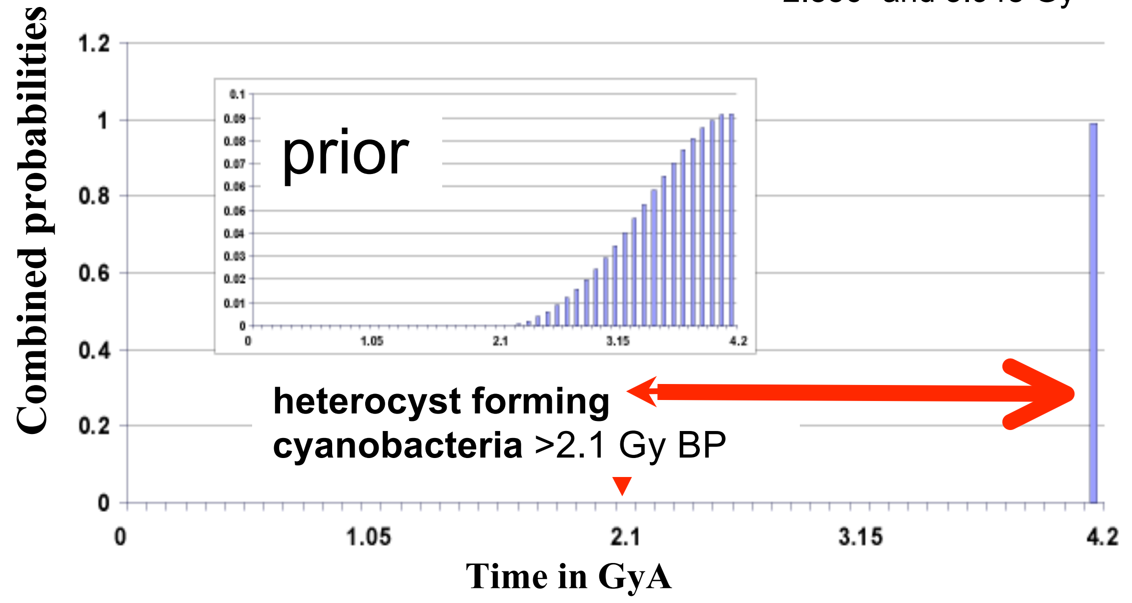


Deepest Split inside cyanobacteria



Nothing older than 4.2 Ga BP

90 % credibility interval:
2.856- and 3.948 Gy



heterocyst forming
cyanobacteria >2.1 Gy BP

Conclusions

Substitution rates in early evolution of life were higher than today.

These higher rates persisted until after the divergence of the bacterial phyla (cyanobacteria, Gram-positive, spirochaetes).

This described method can handle incongruence introduced by gene transfer events, only if the node itself does not reflect a gene transfer event.

General Conclusions

-Rather than being static over long period of time, prokaryotic genomes are composed of ever changing collection of genes.

-The pan-genome analyses has show that on the genome level of an organism, different evolutionary mechanisms exist and contribute to the incredible power of adaptation of micro-organisms (mutations, domain shuffling and gene exchanges).

-Different species living in the same environmental niche will most definitely present common phenotypic features reflected in genomes similarities, blurring the line between species boundaries.

Acknowledgments

J. Peter Gogarten

Former Lab members :

Lorraine Olendzenski
Olga Zhaxybayeva
Reshma Shial
Daniel Shock

Current Members :

Maria Poptsova
Greg Fournier
Ali Senejani
Kristen Swithers
Tim Harlow

My wife Nathalie

The Benson's Lab :

Derek Bickhard
Juliana Mastronunzio

The Noll's Lab :

Dhaval Nanavati
Tu Nguyen
John Dipippo

Ph.D. Thesis Committee members

All Faculty and Students of the MCB department

Funding Agencies