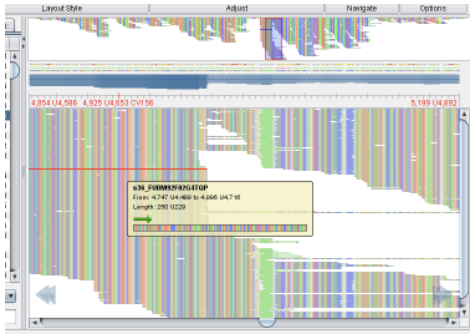
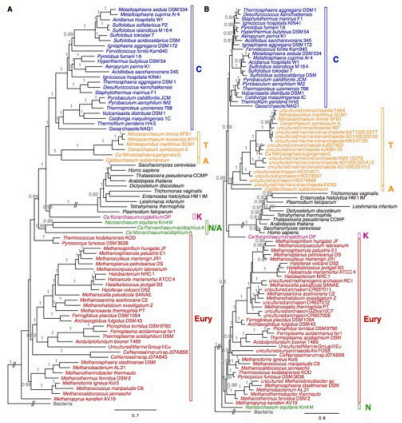
UPPSALA
UNIVERSITET

Single-cell genomics pipeline: from raw reads to phylogenomics using Galaxy





Lionel Guy and Thijs Ettema, Department of Cella and Molecular Biology and SciLifeLag, Uppsala University, Sweden

Galaxy Community Conference 2013

Oslo, Norway, 2013-07-01

SciLifeLab

UPPSALA
UNIVERSITET

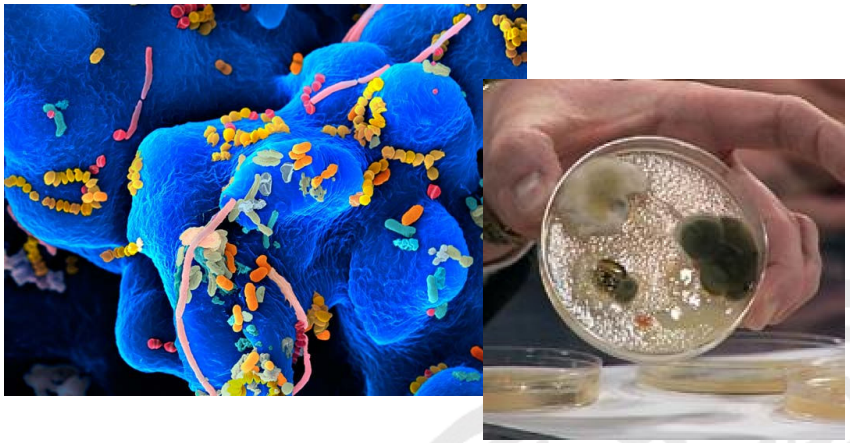
- Introduction
  - Microbes
  - Single-cell genomics
  - Underlying scientifice question
- Implementation in our lab
  - Sample pipeline
  - Galaxy pipelines
- Conclusions

UPPSALA UNIVERSITET

“Microbes rule the world”

Table 5. Number and biomass of prokaryotes in the world

Environment	No. of prokaryotic cells, $\times 10^{28}$	Pg of C in prokaryotes*
Aquatic habitats	12	2.2
Oceanic subsurface	355	303
Soil	26	26
Terrestrial subsurface	25–250	22–215
Total	415–640	353–546



- Microbes are abundant: ~5’000’000’000’000’000’000’000’000’000’000’000 prokaryotic cells ($5x10^{30}$, five million trillion trillion)
- All cells in all humans: $7x10^{9} \times 10^{13} = 7x10^{22}$, i.e. ~70’000’000’000’000’000’000’000 (a mere 70 billion trillion)

Table 6. Relationship of plant and prokaryotic biomass to primary productivity

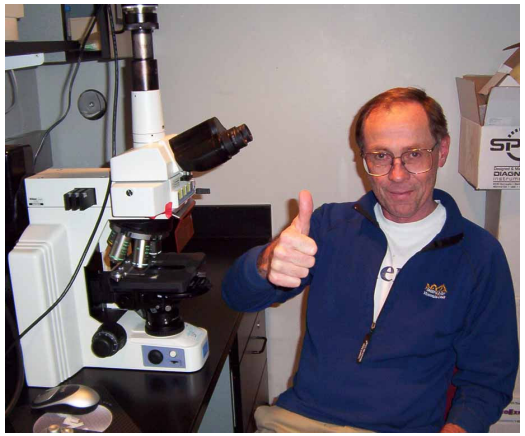
Ecosystem	Net primary productivity,* Pg of C/yr	Total carbon content, Pg of C		
		Plant*	Soil and aquatic prokaryotes	Subsurface prokaryotes
Terrestrial	48	560	26	22–215
Marine	51	1.8	2.2	303

- Roughly the same biomass as plants (~500 Pg C)

Source: Whitman WB et al, PNAS 1998
Baserga 1985

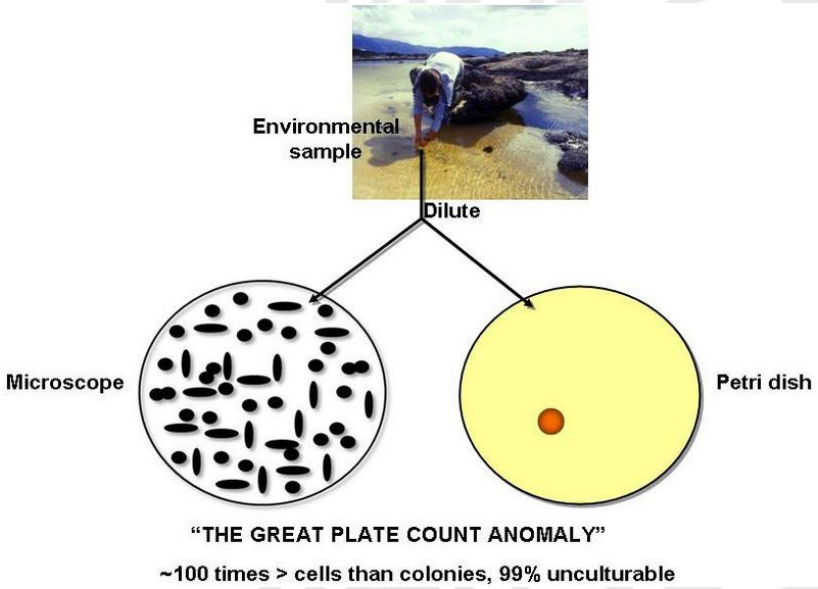
UPPSALA UNIVERSITET

Studying microbes



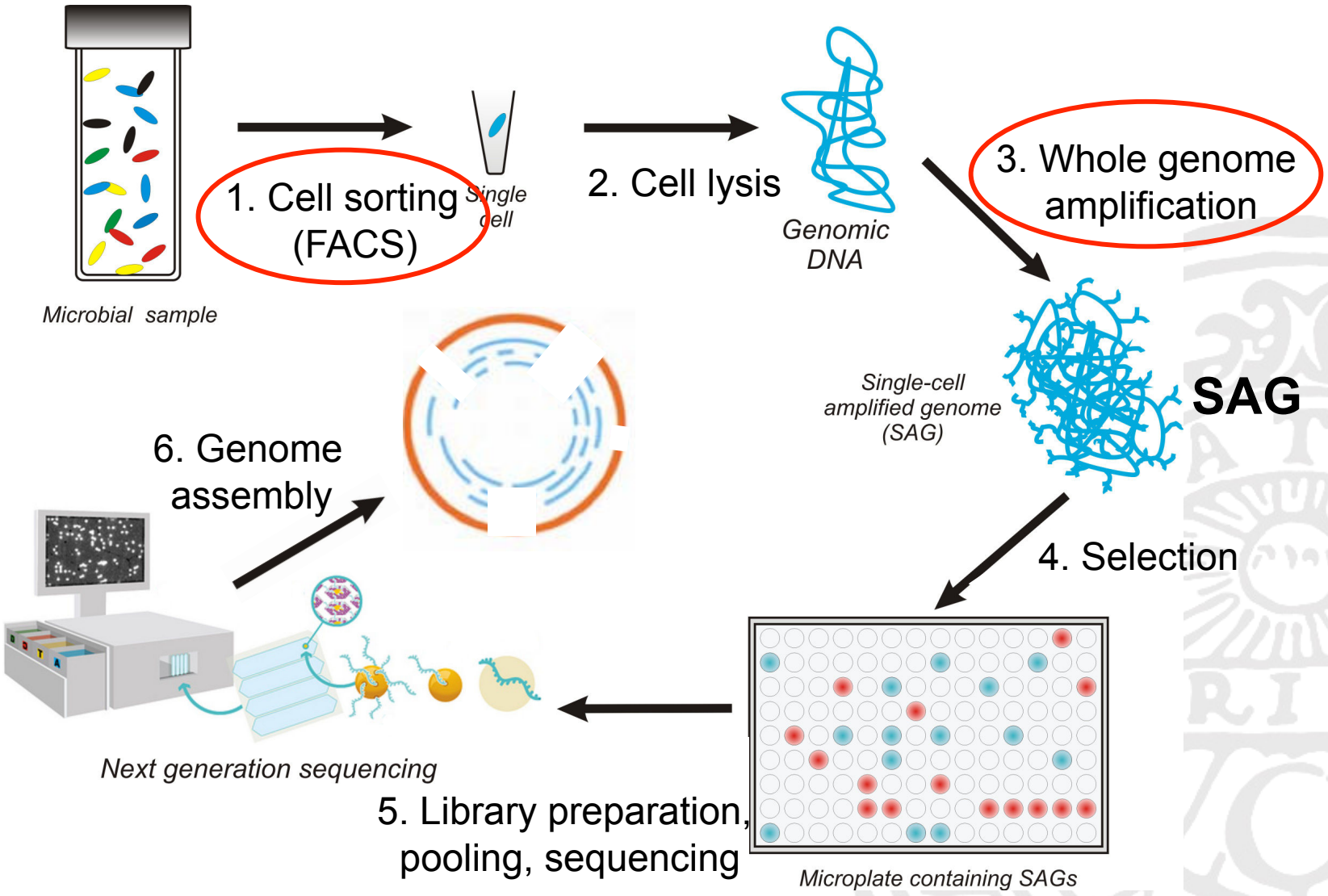
- Norman Pace (Univ. Colorado, Boulder), 2001:
- "*Imagine if our entire understanding of biology were based on a visit to the **zoo**... And that's exactly the situation we've been at in the **microbial** world until really quite recently*"

- ~1 % of prokaryotes grow on plate ("zoo", aka "The Great Plate Count Anomaly")
- ~99% to be studied!!!



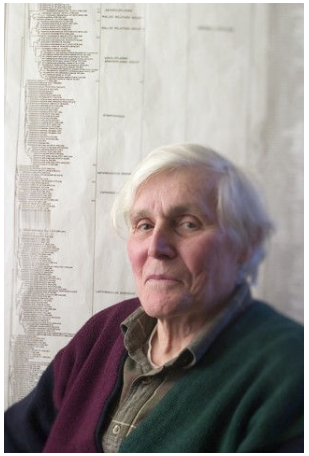
UPPSALA UNIVERSITET

How? Single-cell genomics



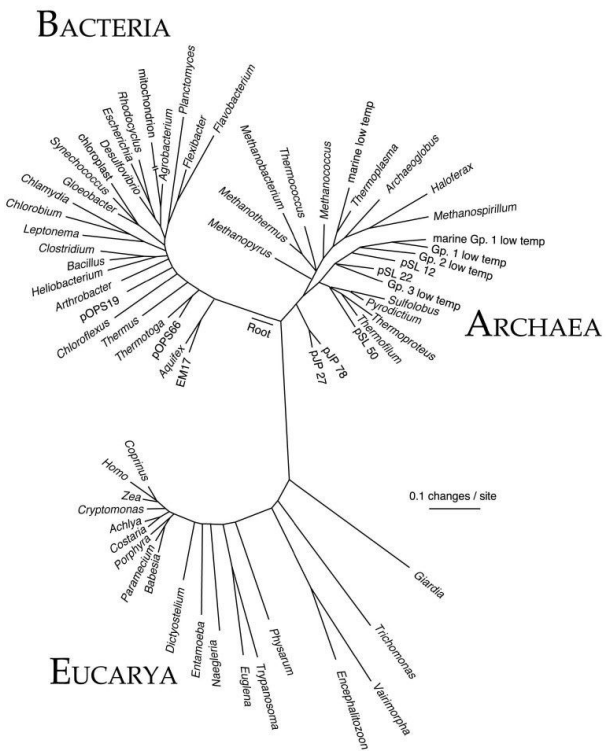
UPPSALA UNIVERSITET

Why? Origin of eukaryotes

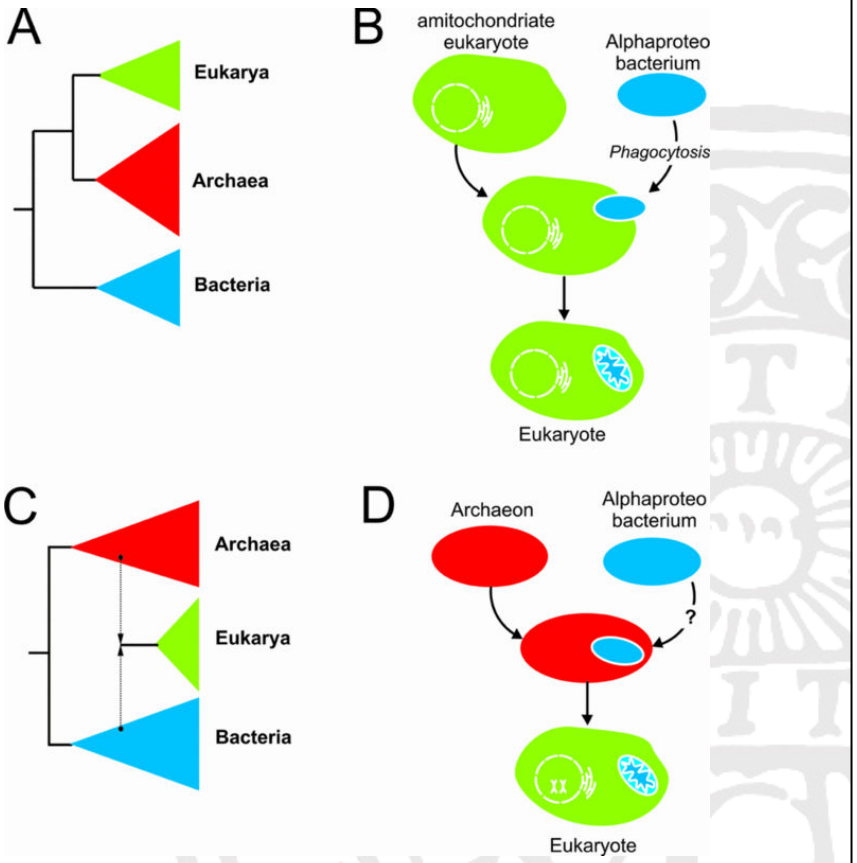


Carl Woese

- Three Domains of Life?



Pace, Science 1997

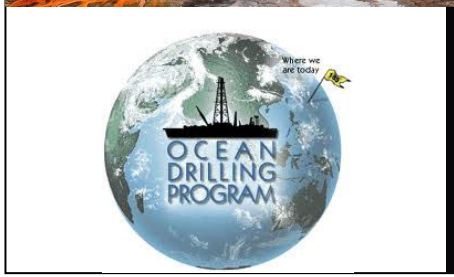


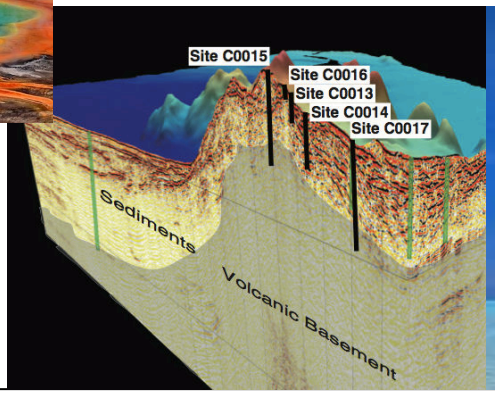
Martijn & Ettema, Biochem Soc Transac 2013

UPPSALA UNIVERSITET

What? Sampling

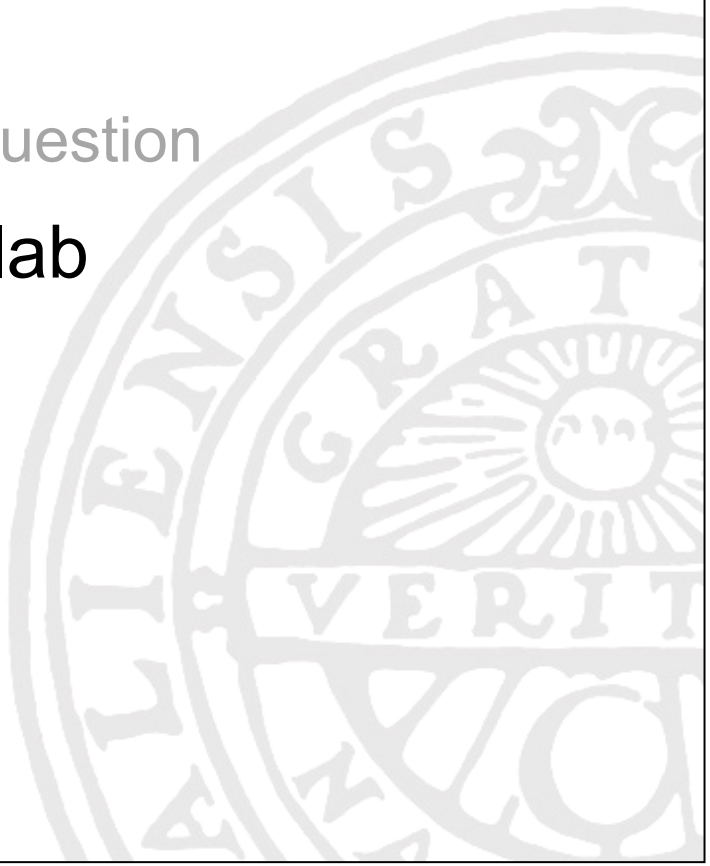
- Diverse and unexplored environments:
  - New Zealand, 1000 hot springs project (Matthew Scott)
  - Yellowstone hot springs
  - Ocean Drilling Program expeditions (expedition 331, Deep Hot Biosphere)
  - Hawaii oceanic water
  - Sala silver mine, Sweden





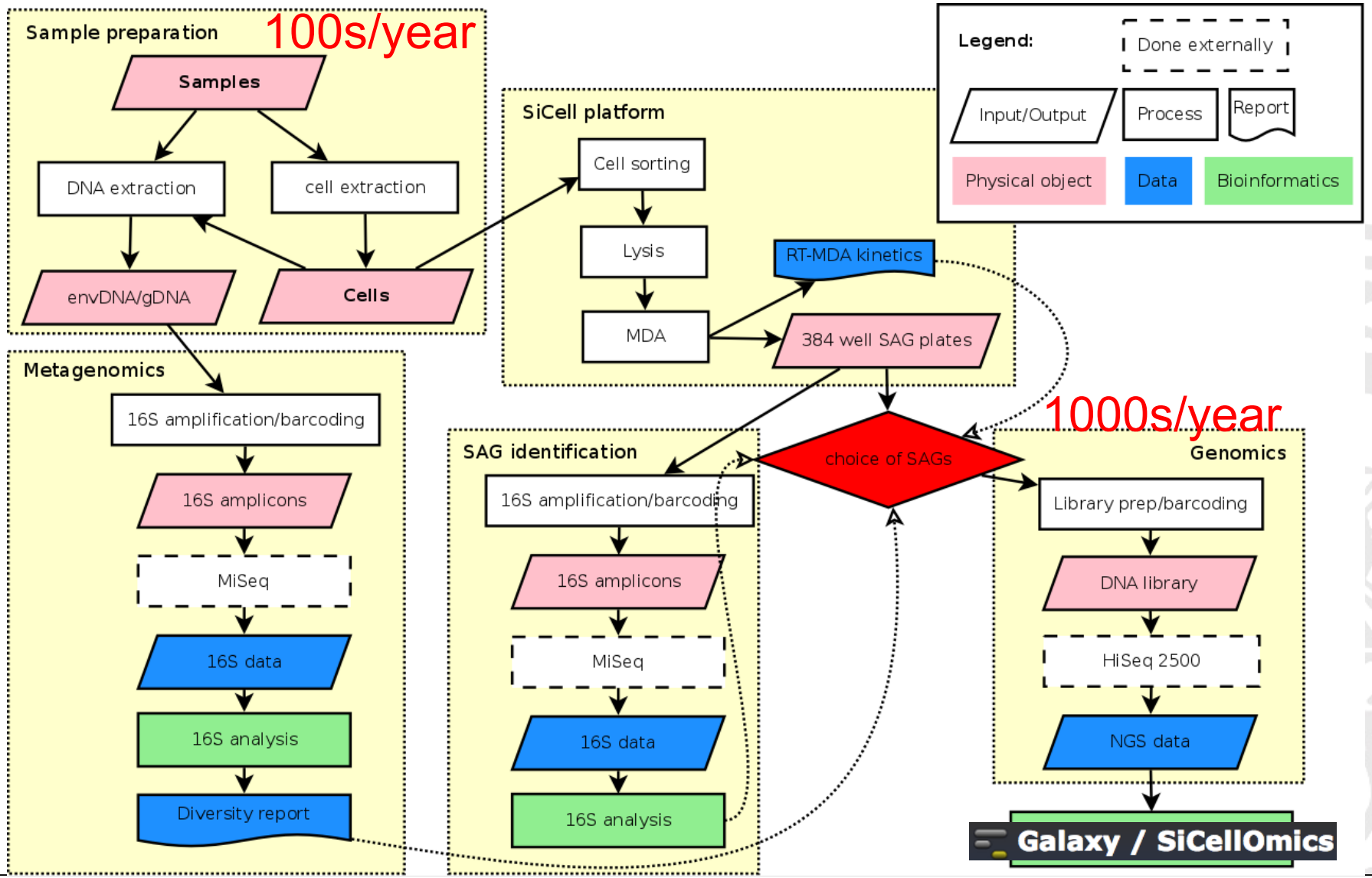
UPPSALA
UNIVERSITET

- Introduction
  - Microbes
  - Single-cell genomics
  - Underlying scientifice question
- Implementation in our lab
  - Sample pipeline
  - Galaxy pipelines
- Conclusions



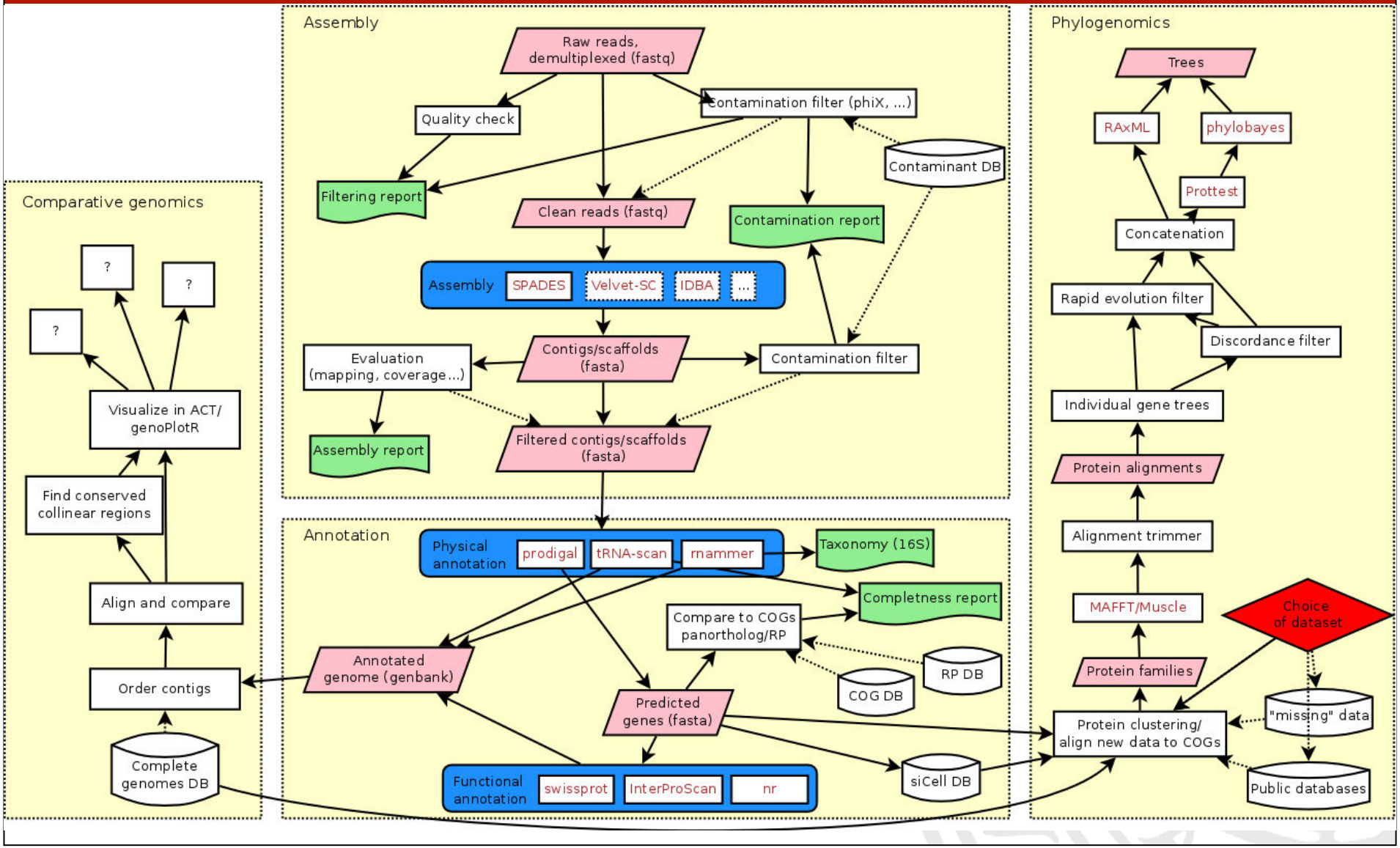
UPPSALA UNIVERSITET

Sample pipelines



UPPSALA UNIVERSITET

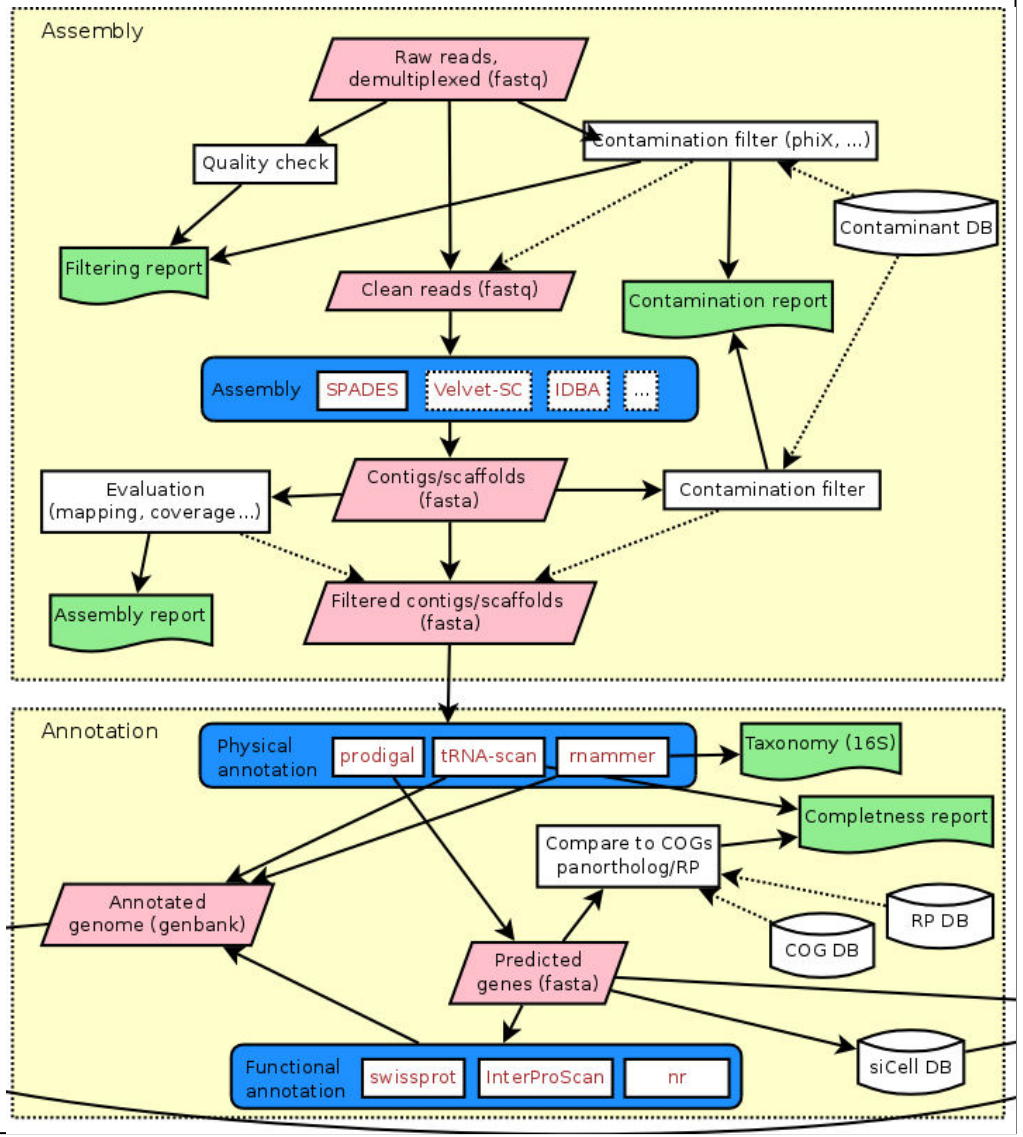
Galaxy pipelines



UPPSALA UNIVERSITET

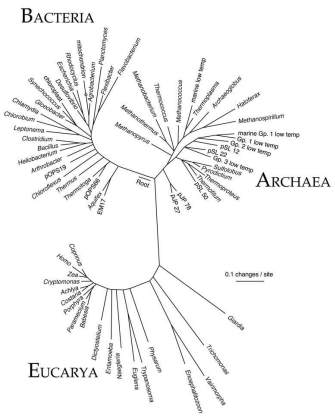
Assembly/annotation

- Input: raw reads from sequencing center
- Output: annotated genome
- Reports:
  - Reads quality
  - Contaminations
  - Closest sequenced organism
  - Completeness

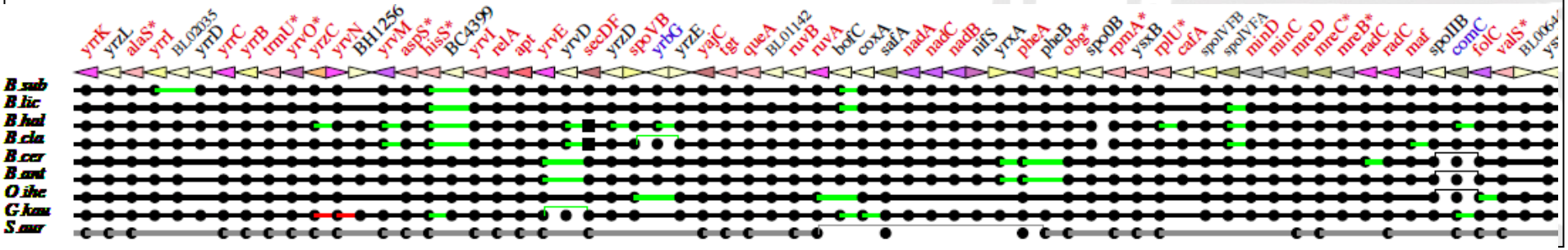


UPPSALA UNIVERSITET

Phylogenomics

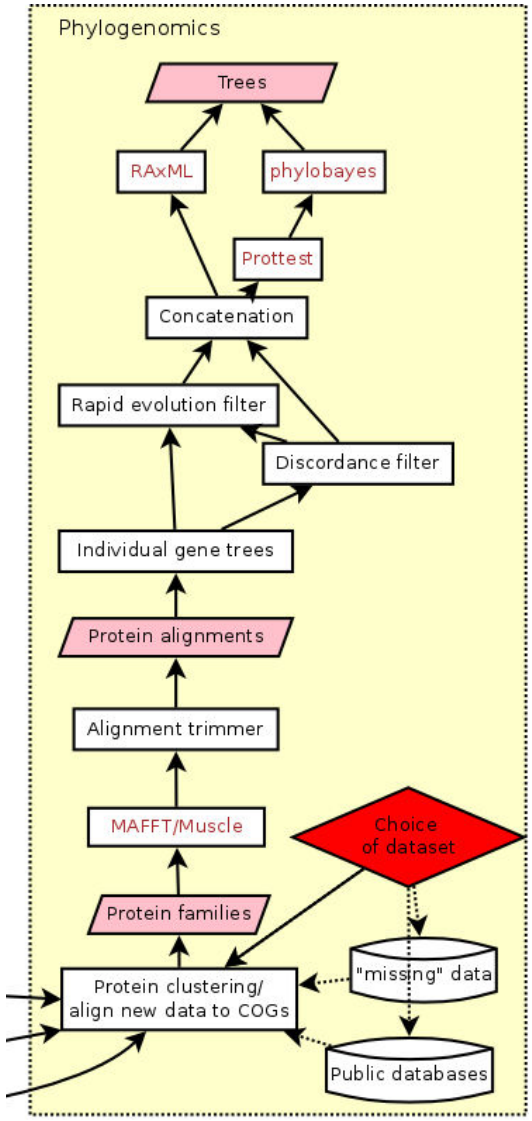


- Ribosomal RNA: 1.5 kb (SSU/16S/18S), 3 kb (LSU)
- Use proteome: in average ~2000 genes per genome (range 150-12000):
  - Find orthologous genes, align, concatenate, run phylogeny
- Issues:
  - Tree of species ≠ Tree of genes: paralogs, horizontal gene transfers (HGT)
  - Distant homologies hard to assess
  - Very few genes conserved in all genomes small (~30)
  - Phylogenies computationally costly
  - SAGs are incomplete (20-90% of the genome)



UPPSALA UNIVERSITET

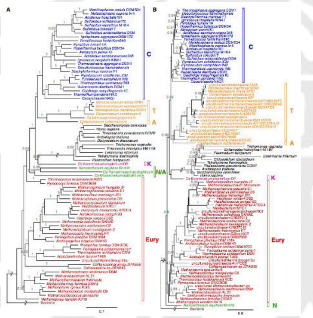
Phylogenomics pipeline



- Task: find a set of orthologous genes and organisms with the following properties:
  - Few HGTs
  - Little (and evenly spread) missing data
  - Representative genomes (<100)

Genomes

Genes	**A**	**B**	**C**
x	1	1	1
y	1	0	1
z	1	1	1



UPPSALA UNIVERSITET

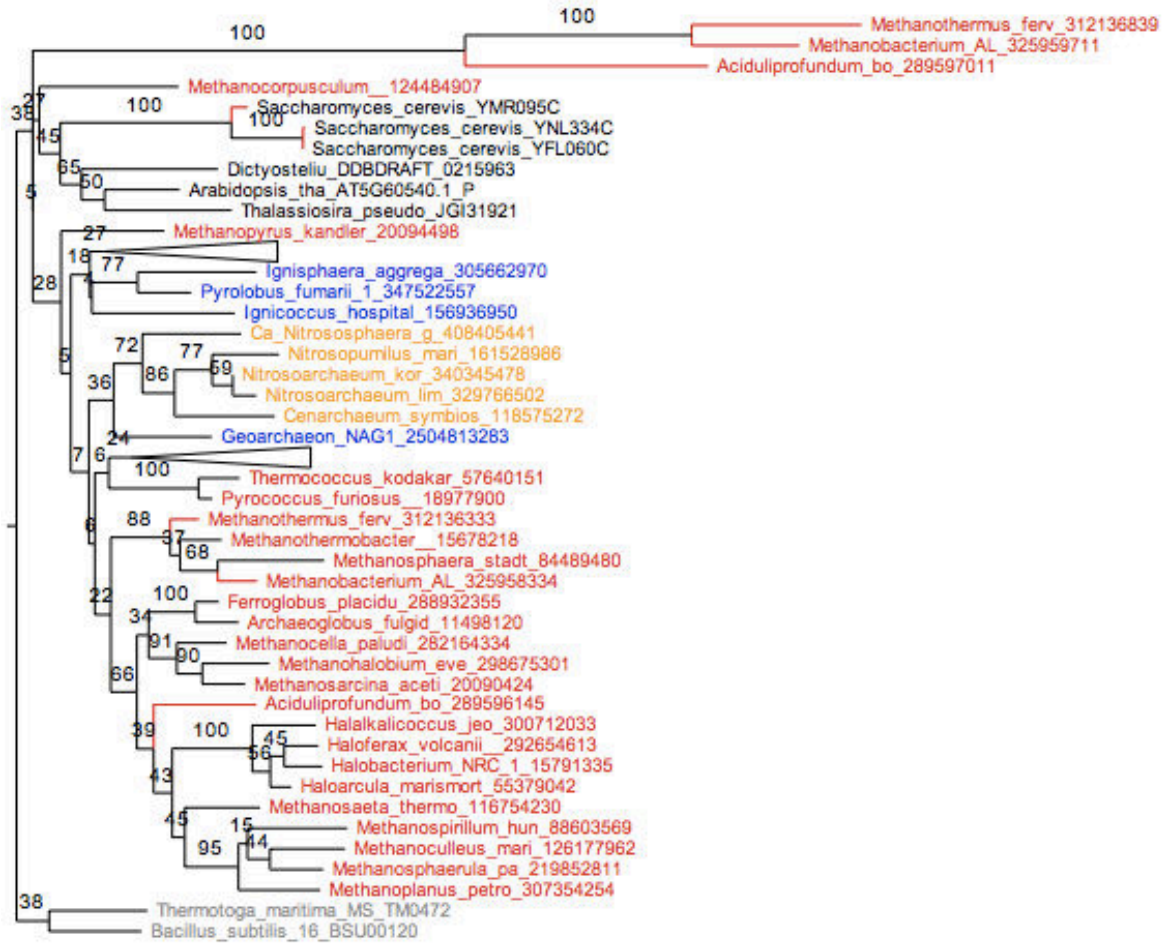
Phylogenomics – data seletcion

Genomes

Genes	A	B	C
x	1	1	3
y	0	1	1
z	1	2	0

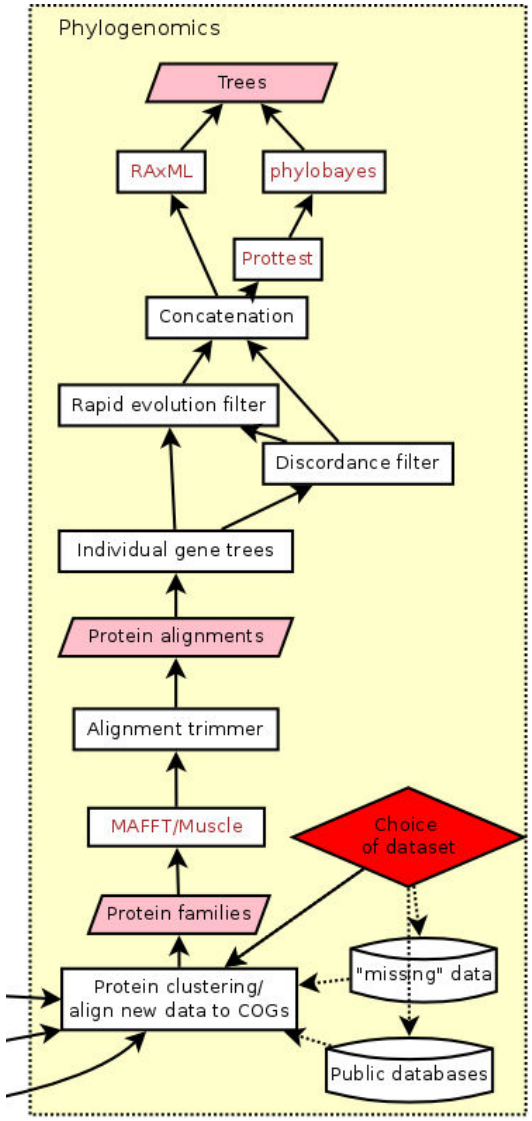
Genomes

Genes	A	B	C
x	1	1	1
y	0	1	1
~~z~~	~~1~~	~~2~~	~~0~~

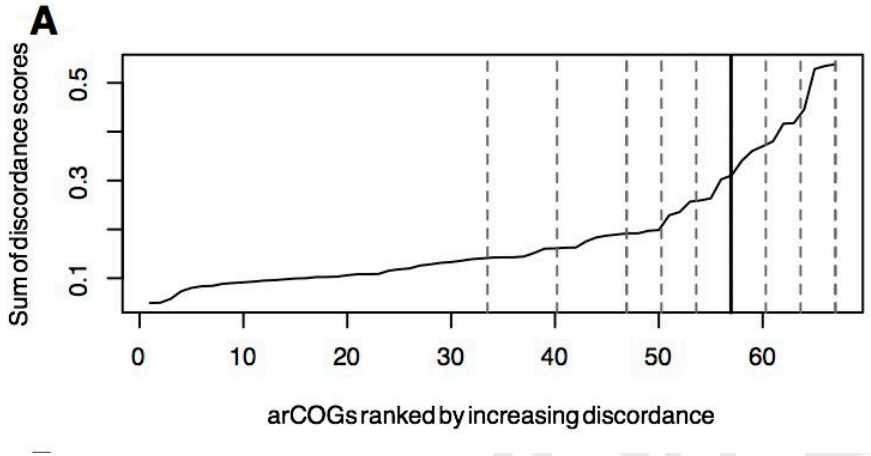


UPPSALA UNIVERSITET

Phylogenomics – filtering

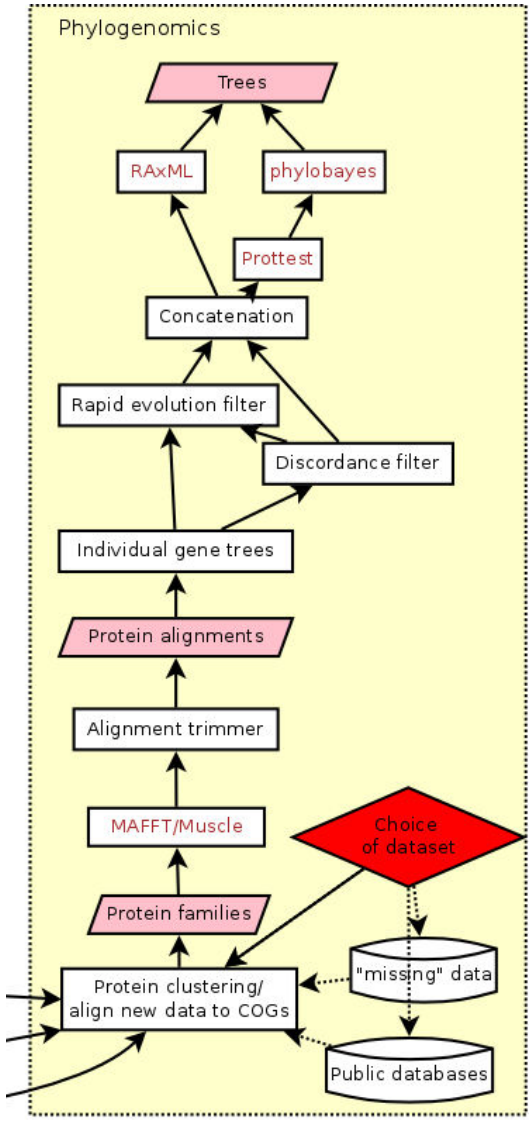


- Per gene: discordance filter
  - Compare all gene trees to all, count conflicts
  - Rank genes by sum of conflicts

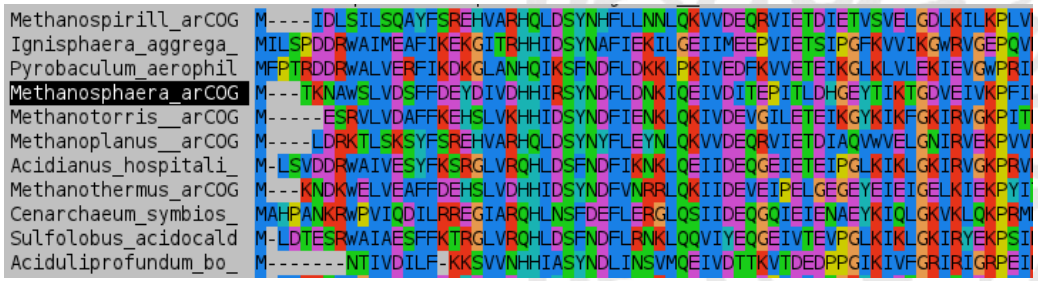


UPPSALA UNIVERSITET

Phylogenomics – filtering



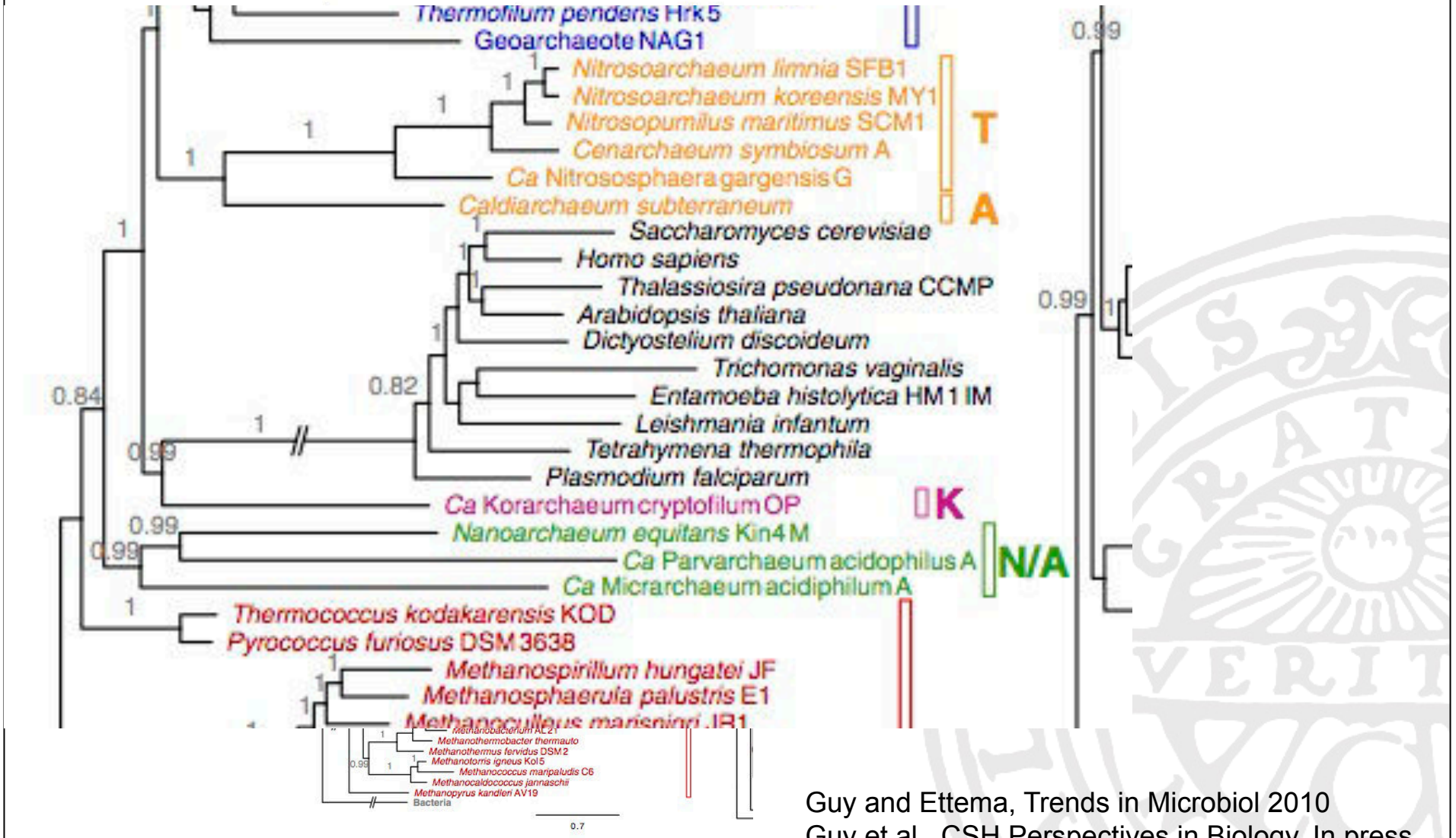
- Per site:
  - Calculate the global amino-acid biases in a protein alignment
  - Remove one site (column), recalculate
  - Rank sites by how much they contribute to the global bias



Viklund et al, MBE 2012

UPPSALA UNIVERSITET

A robust Tree of Life



Guy and Ettema, Trends in Microbiol 2010
Guy et al., CSH Perspectives in Biology, In press

UPPSALA
UNIVERSITET

- Introduction
  - Microbes
  - Single-cell genomics
  - Underlying scientifice question
- Implementation in our lab
  - Sample pipeline
  - Galaxy pipelines
- Conclusions

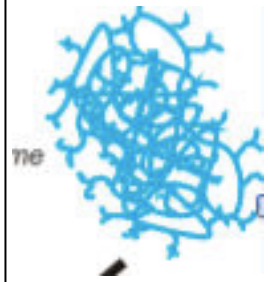
UPPSALA UNIVERSITET

Future work

- Automatic generation of clusters (OrthoMCL, TribeMCL, …)
- Automated trimming of paralogs (hard)
- Complete automated pipeline, incremental database

UPPSALA UNIVERSITET

Summary



Galaxy / SiCellOmics

- Sequencing genomes from single-cell in a high-througput manner
- Exploring genomics outside the 1% “zoo” is now possible
- Galaxy is the tool of choice!

UPPSALA UNIVERSITET

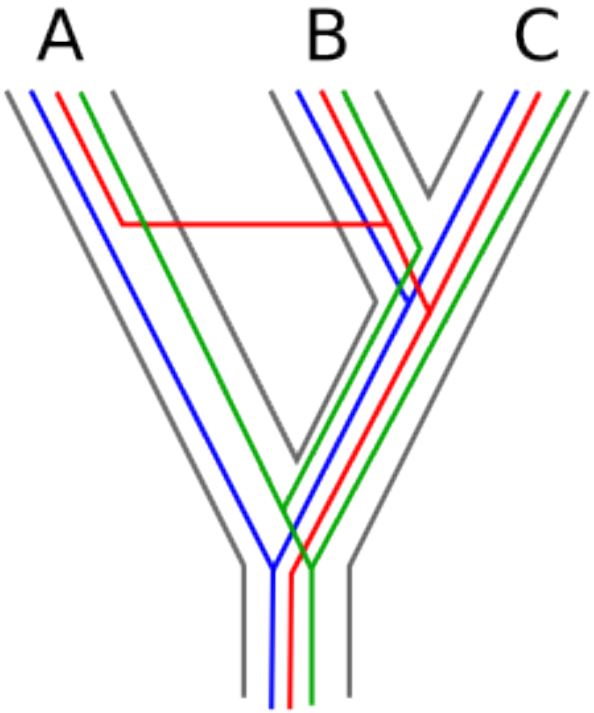
Acknowledgments

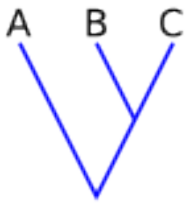
- SiCell:
  - Single-cell genomics (test) platform from SciLifeLab, Uppsala Unviersity.
  - First in Europe!
  - Thijs Ettema & Stefan Bertilsson, co-directors
  - Claudia Bergin & Anna-Maria Divne
- Ettema group (ICM, Uppsala University)
  - Thijs Ettema (PI)
  - Jimmy Saw
  - Anja Spang
  - Anders Lind
  - Joran Martijn
  - Janko Tackmann
  - Santhanam Kulasekara

UPPSALA
UNIVERSITET

UPPSALA UNIVERSITET

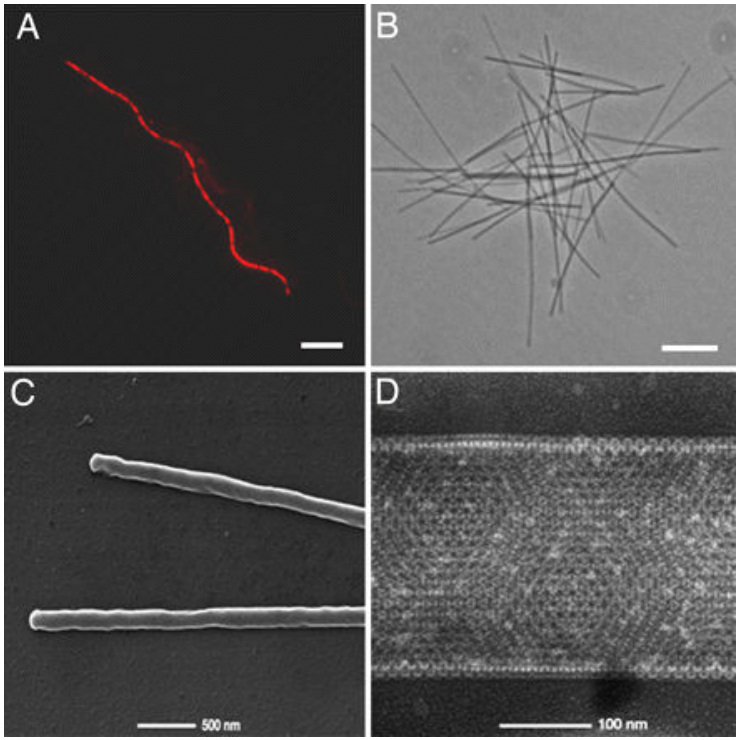
Gene tree =/= Species tree





UPPSALA UNIVERSITET

Korarchaeota

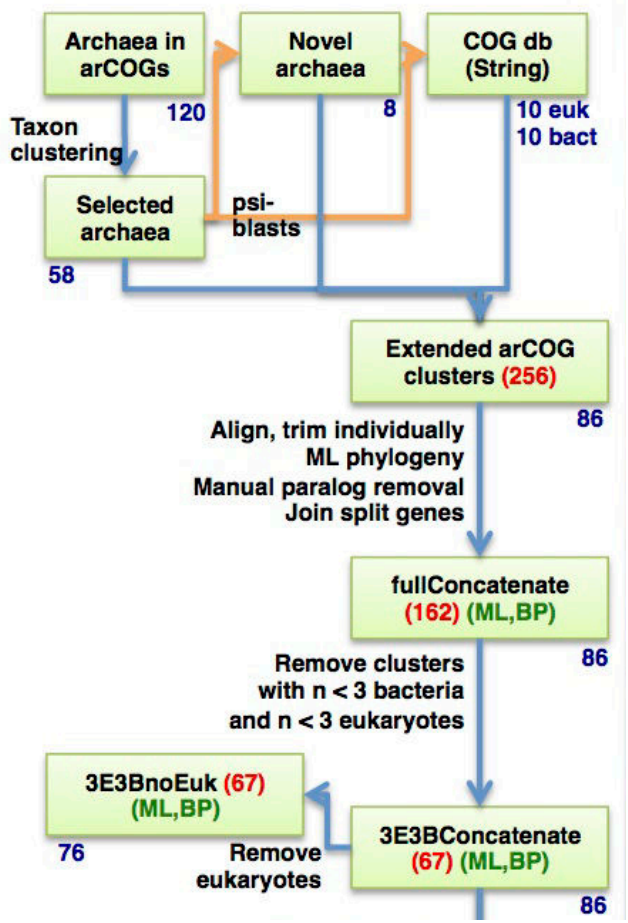


- “Ultrathin filamentous morphology”
- Obligate anaerobes
- Heterotrophes
- Peptides as principal carbon and energy source
- Our ancestor?

Elkins et al, PNAS 2008

UPPSALA UNIVERSITET

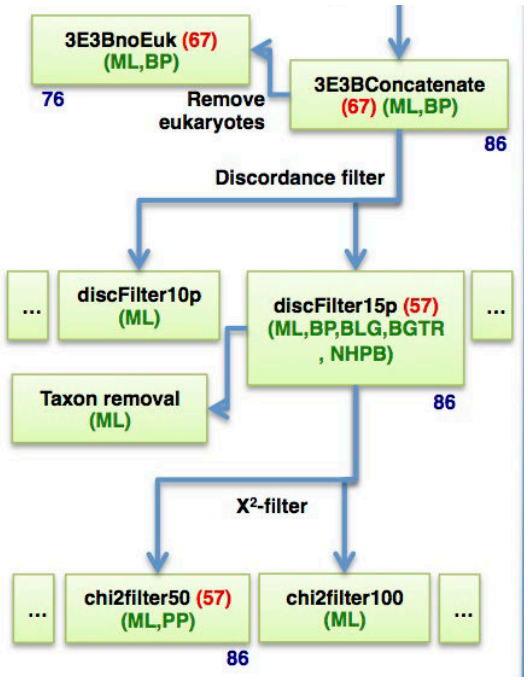
Pipeline – Data selection



- Start from arCOGs (Wolf et al, Biology direct 2012)
- Cluster taxa (CD-HIT, threshold = 70%)
- Add bacteria/eukaryotes (psi-blasts to restricted databases)
- Draw trees, manual check and remove paralogs
- Remove clusters with little or no eukaryotes

UPPSALA UNIVERSITET

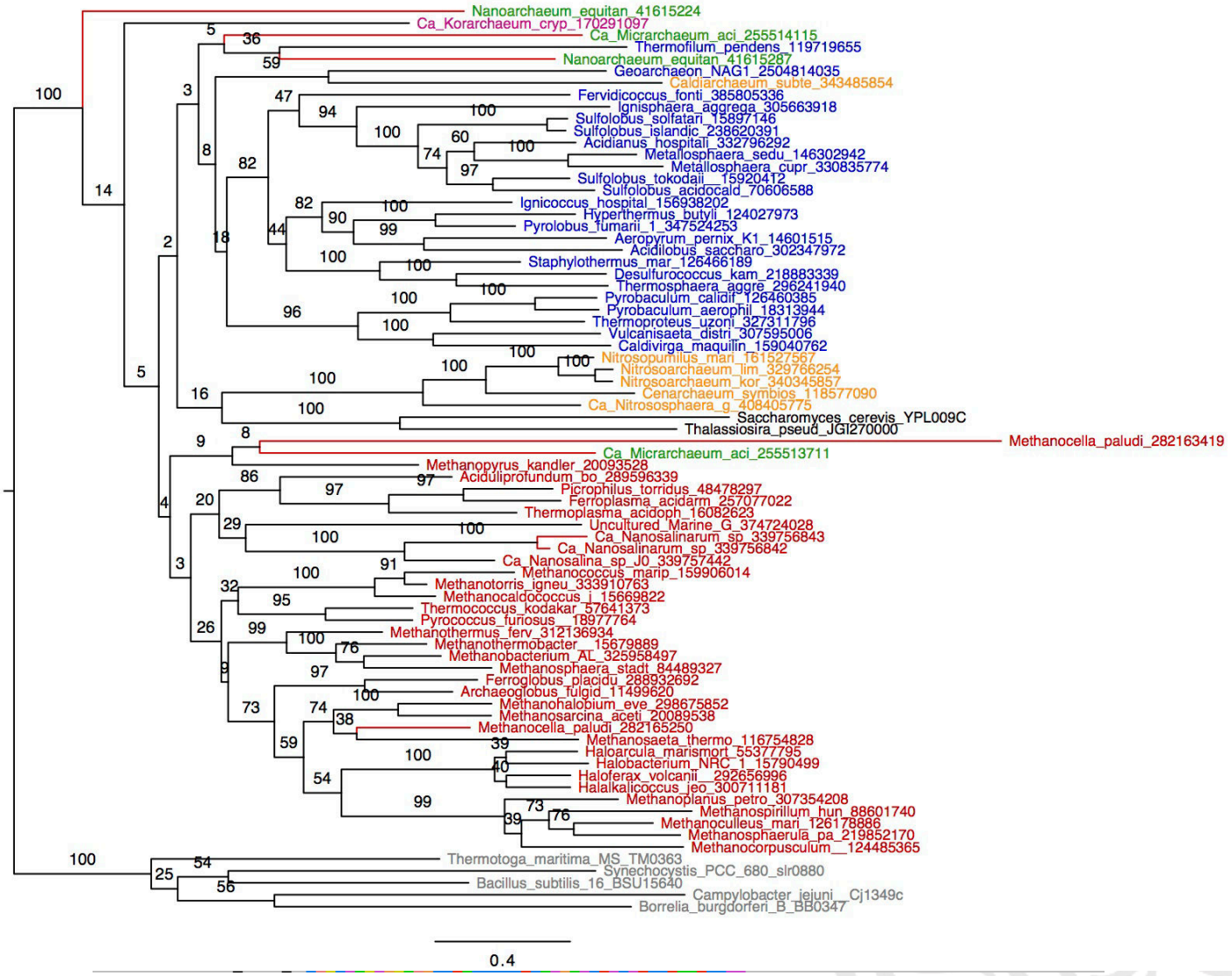
Position of eukaryotes



- Run ML and Bayesian trees with all eukaryotes
- Test different filters
  - Remove HGT
  - Remove sites with strong amino-acid composition bias
- Test the effect of removing taxa on the position of eukaryotes

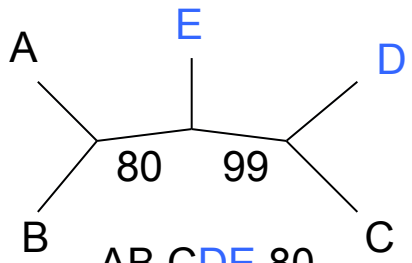
UPPSALA UNIVERSITET

Manual curation of trees



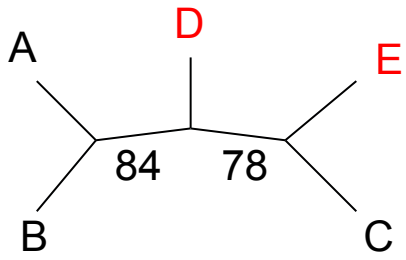
UPPSALA UNIVERSITET

Discordance filter



AB,CDE 80
ABE,CD 99
ABD,CE 12

AB,CDE 84
ABD,CE 80
AE,BCD 2

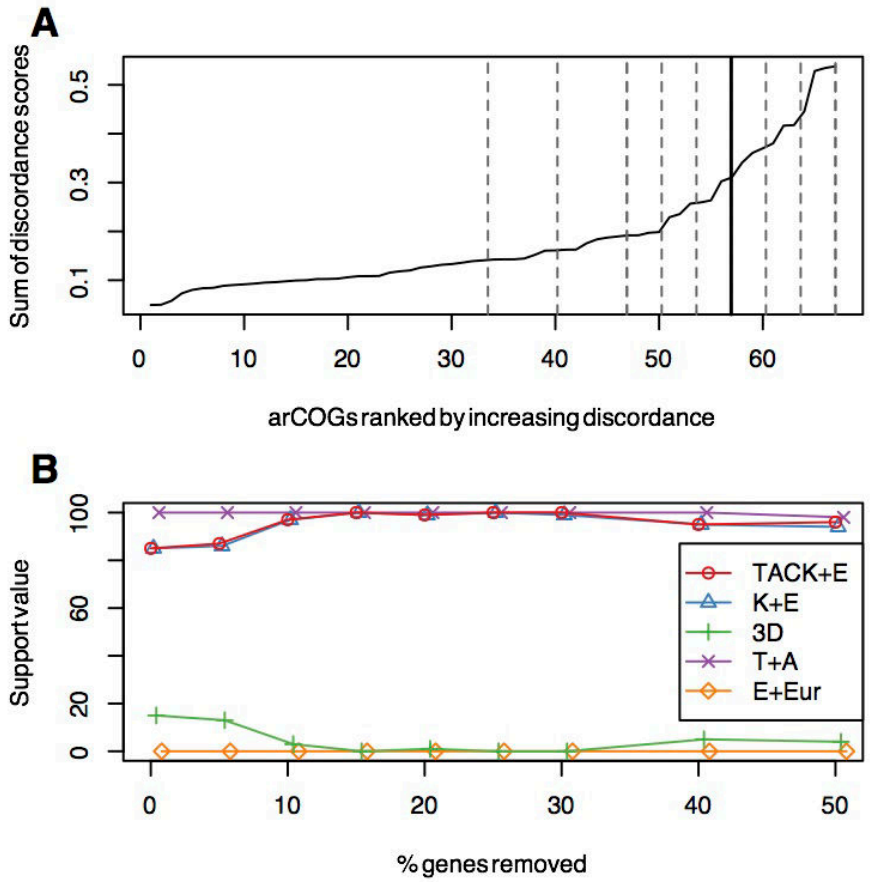


Disc. score = 2 / (2 * 2) = 0.5

- Goal: remove clusters with significant HGT
- Starting point: 67 clusters (3E3B dataset)
- For each possible pairwise comparison:
  - Get bipartitions for bootstrap trees in both trees
  - Keep only high-support bipartitions (HSB), >75
  - Sum the number of incompatible HSB
  - Divide that by the product of HSBs in each of the two trees
- The discordance score of one tree is the sum of all its pairwise discordance scores

UPPSALA UNIVERSITET

Discordance filter



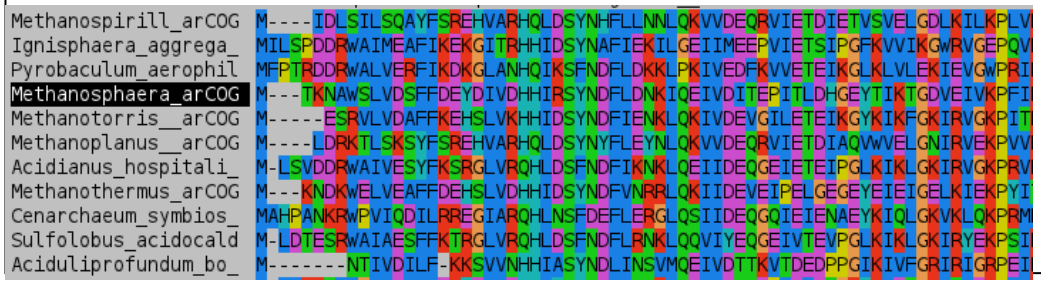
- Rank clusters by increasing discordance score
- Remove increasing fractions of the most discordant genes
- Assess the effect on key splits
- After removing 15% of the data, most nodes are supported either with 0 or 100

UPPSALA UNIVERSITET

Chi2 filter

$$\chi^2 = \sum_t \sum_{i=1}^{20} \frac{(O_i(t) - E_i)^2}{E_i},$$

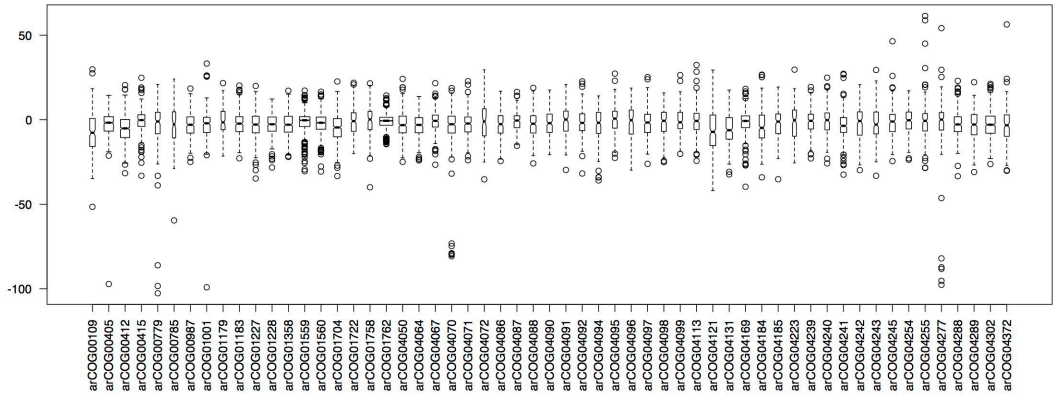
- Goal: remove sites with high amino-acid bias
- Principle:
  - Calculate χ^2-score for the whole alignment: aa composition difference between a row and the whole alignment, summed over aa and rows.
  - Trim one position, redo the calculation, get the difference between the trimmed and the complete alignment
  - Repeat for each site
- Estimates how much each site contributes to aa composition bias

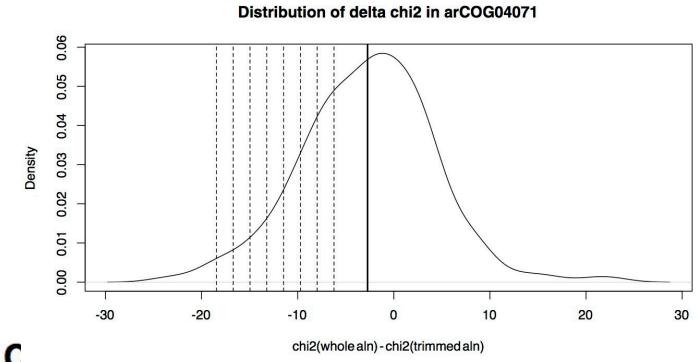


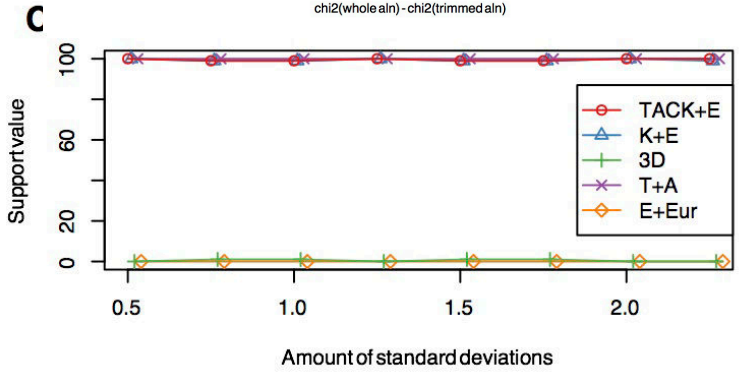
Viklund et al, MBE 2012

UPPSALA UNIVERSITET

Chi2 filter



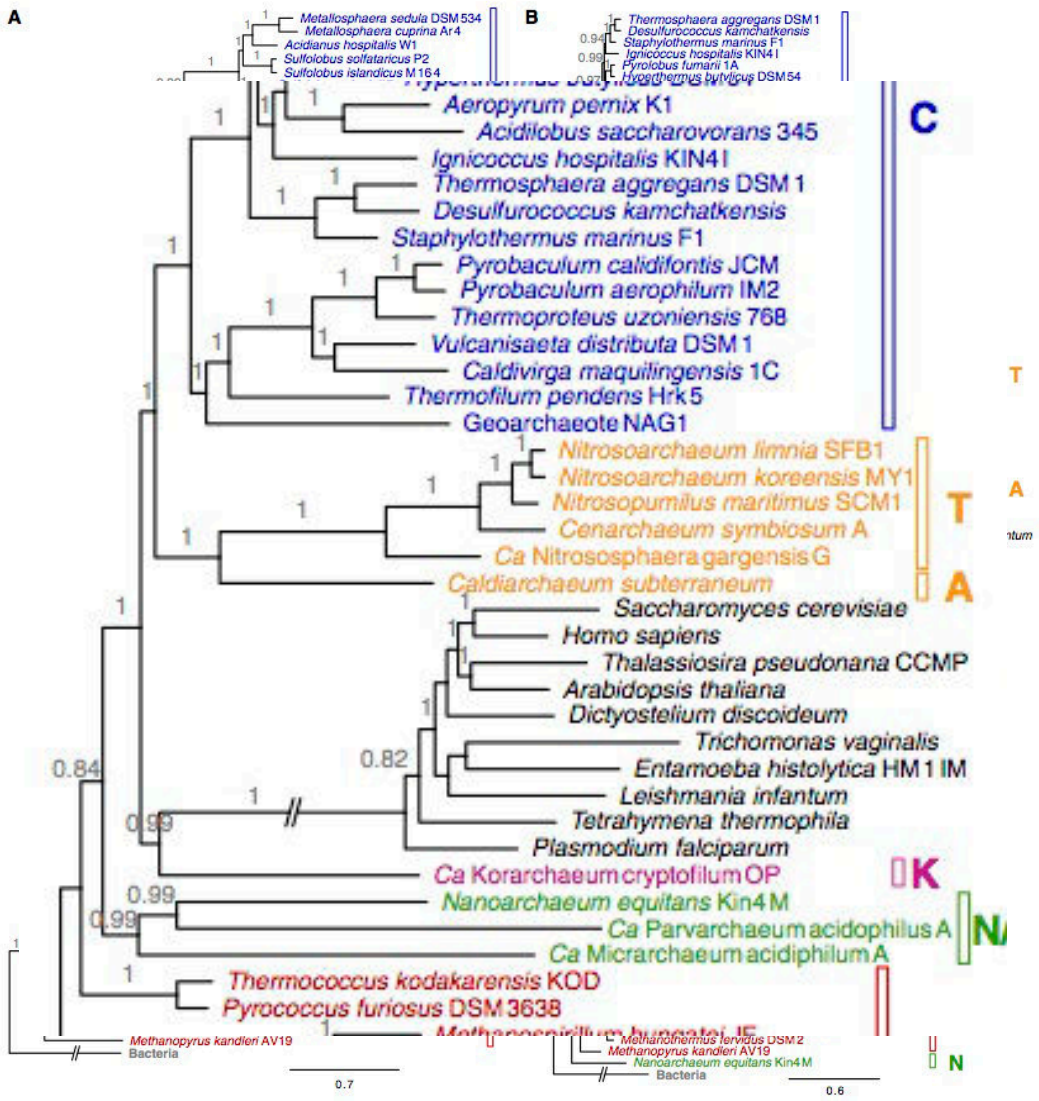




- Removing sites in a gene-specific manner: removing the most discordant sites, in terms of mean – fractions of sd
- Assess the effect on critical nodes
- No noticeable effect…

UPPSALA UNIVERSITET

“Best” tree



- A:
  - disc15p dataset
  - Phylobayes
  - CAT-Poisson
  - Kind of converged
- B: (for comparison)
  - Concatenate SSU/ LSU rDNA
  - Same program/ methods
  - Made by Jimmy Saw