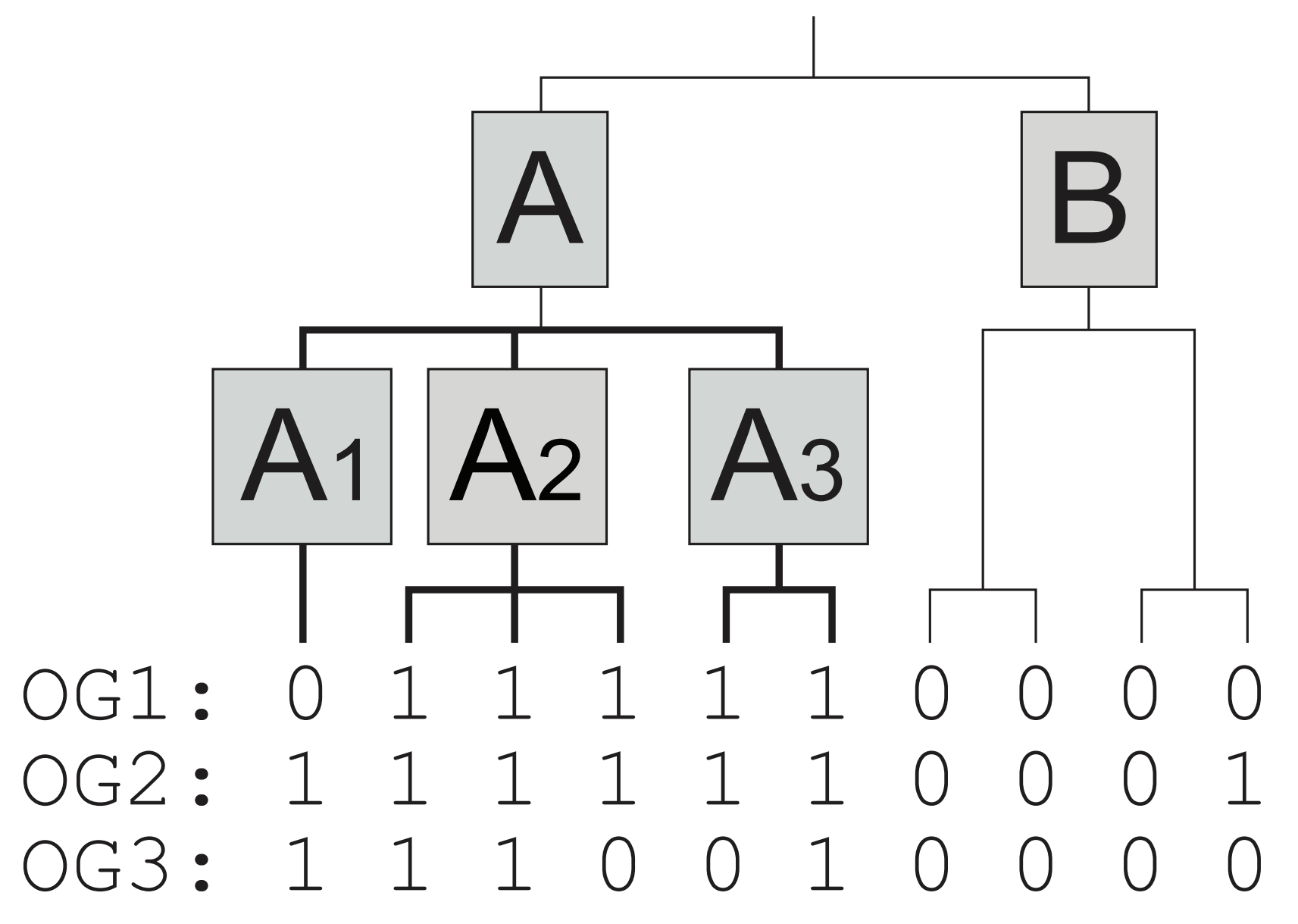
# Signature genes as a phylogenomic tool

Bas E. Dutilh, Berend Snel, Thijs J.G. Ettema, Ying He, Maarten L. Hekkelman and Martijn A. Huynen
CMBI / NCMLS, Radboud University Nijmegen Medical Centre, The Netherlands

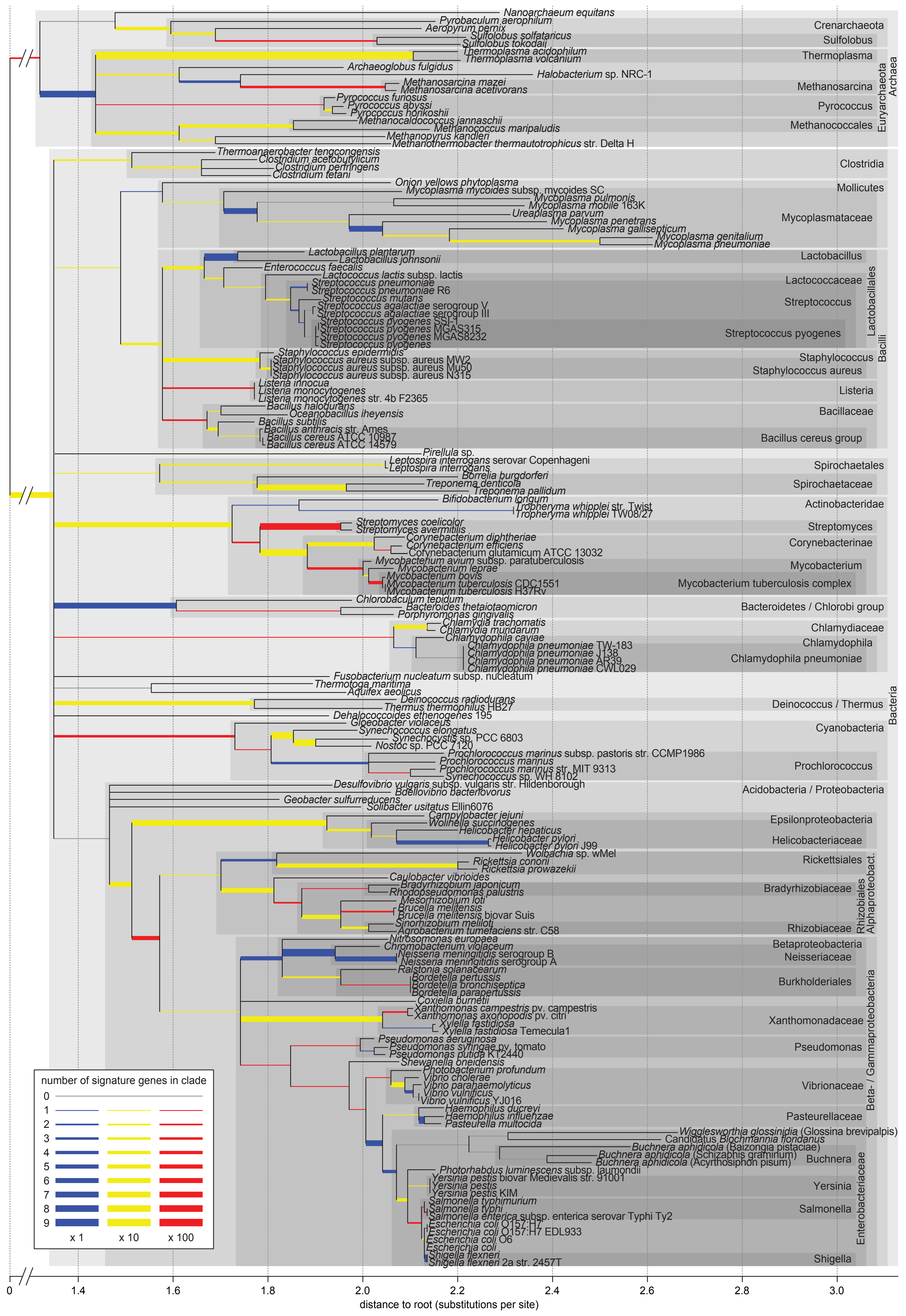
## introduction

Signature genes are unique to a taxonomic clade and are present in all daughter lineages. They can be used for the phylogenetic characterisation of sequence samples, including incomplete genomes and metagenomic samples. We have tested the reliability of signature genes as a phylogenomic tool, and implemented the method in a web server.



- OG1 is not a signature for clade A as it is not present in daughter A1.
- OG2 is not a signature for clade A because it is present in clade B.
- OG3 is a signature for clade A, as it is present in all the daughter lineages A1, A2 and A3, but not in clade B.

The coverage score of a signature is iteratively calculated as the average of the scores in the daughter clades. For OG3 in clade A: ((1/1)+(2/3)+(1/2))/3 = 0.72.



We found 8,362 signatures for 112 prokaryotic taxa, given the COGs and NOGs in STRING 7.0 (von Mering et al. 2007) and the reference phylogeny above. This tree is based on a superalignment (Ciccarelli et al. 2006), nodes with bootstrap <80% were collapsed.

## cross-validation

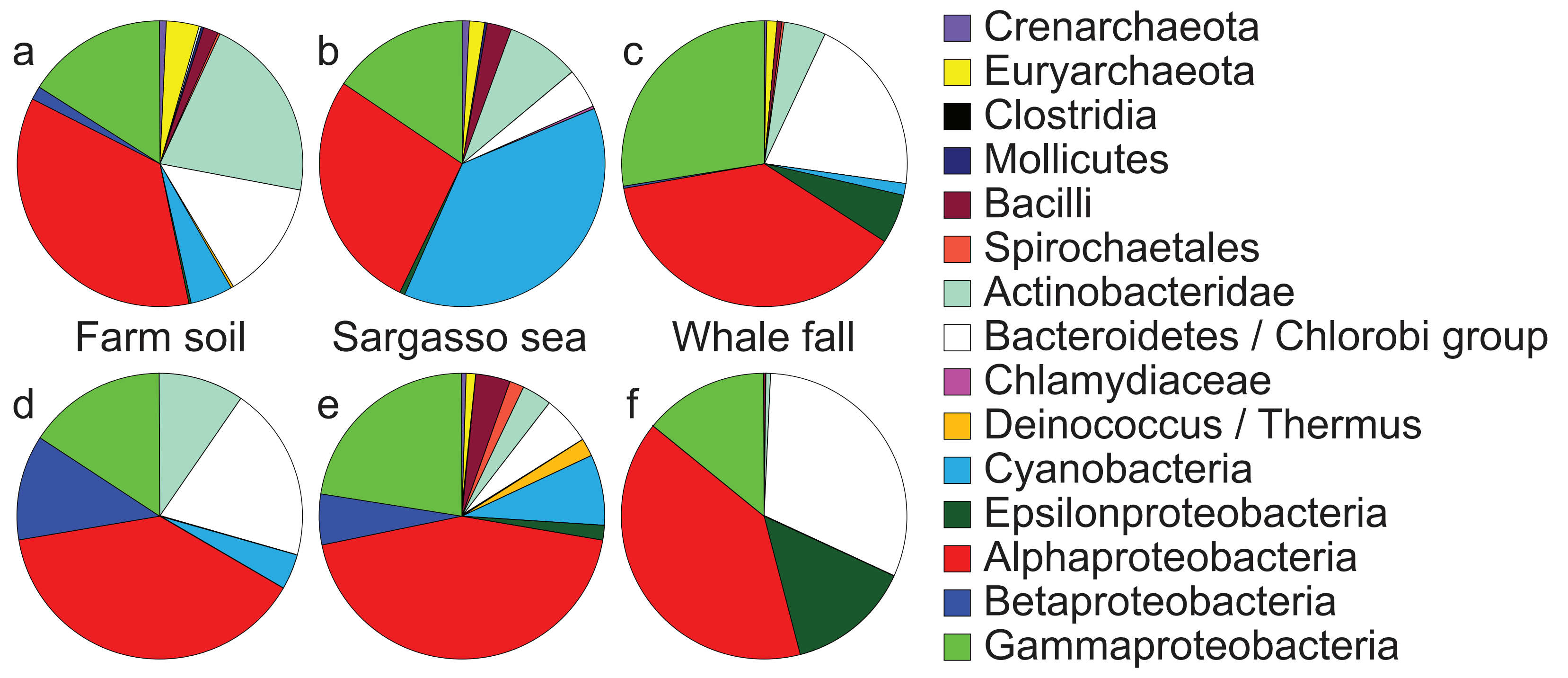
We removed up to 30% of the species from the data set. In the adjusted phylogeny, a removed protein could be:

- A signature for a correct ancestor (tp)
- A signature for another clade (not an ancestor; fp)
- Not a signature, but a signature for an ancestor in the original phylogeny with all species (fn)
- Not a signature, nor a signature originally (tn)

| | |
|---|---|
| sensitivity (tp/(tp+fn)) | *71.5%* |
| specificity (tn/(tn+fp)) | *98.7%* |
| precision (tp/(tp+fp)) | *91.7%* |
| accuracy (true/all) | *94.1%* |

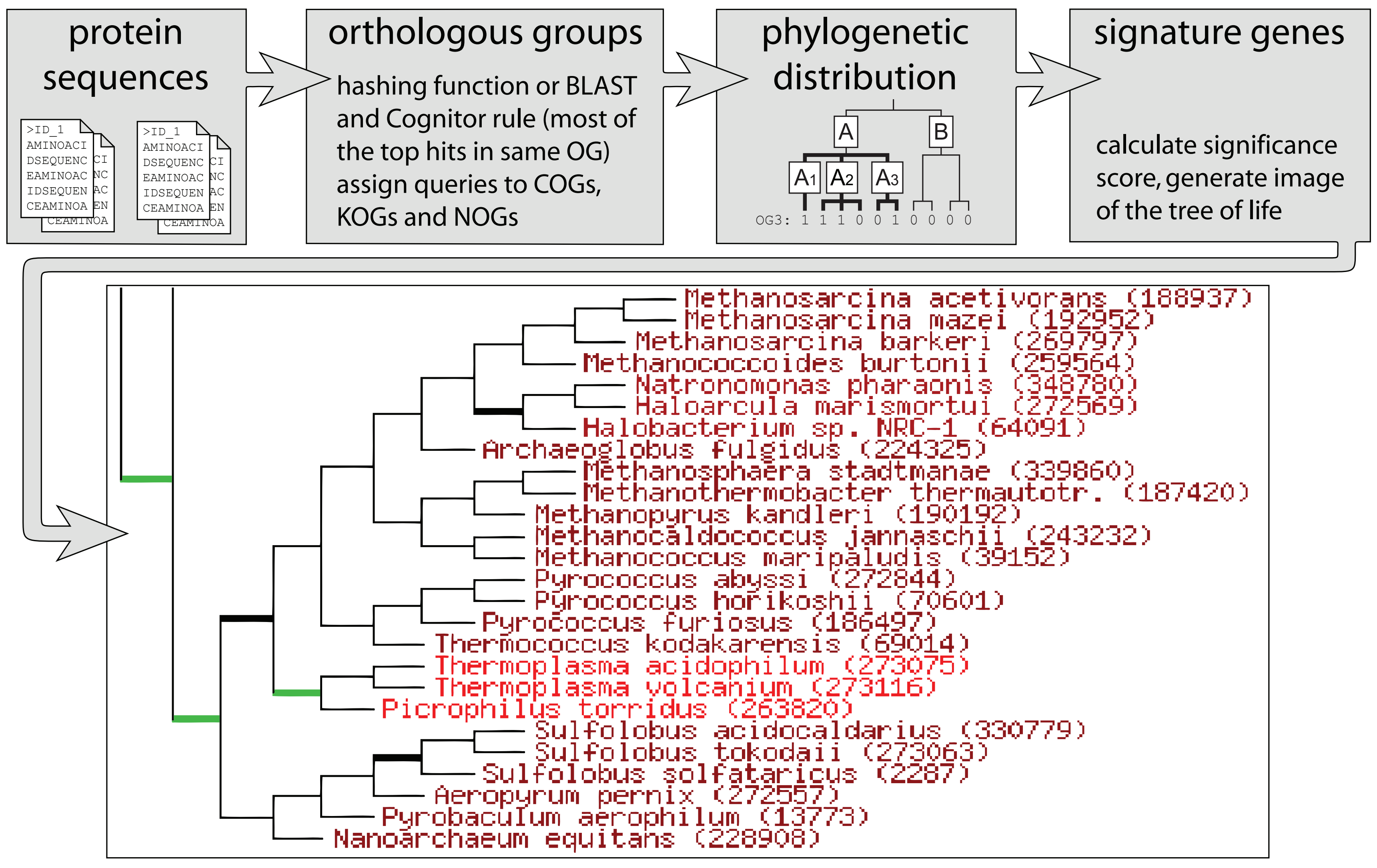
## metagenomics data

Signature genes qualitatively find the same clades in metagenomic samples as experiments based on phylogenetic markers, and provide an independent point of view.



a, b, and c: total numbers of signature genes found for each clade (including subclades); d, e and f: percentages of these clades found in the original analyses using several phylogenetic markers (Venter et al. 2004; Tringe et al. 2005). Signatures could not be identified for clades that were not in the reference tree, these are not shown.

## web server



protein sequences

orthologous groups: hashing function or BLAST and Cognitor rule (most of the top hits in same OG) assign queries to COGs, KOGs and NOGs

phylogenetic distribution

signature genes: calculate significance score, generate image of the tree of life

>ID_1
AMINOACI
DSEQUENC
EAMINOAC
IDSEQUEN
CEAMINOA
A
B
A1
A2
A3
OG3: 1 1 1 0 0 1 0 0 0 0
Methanosarcina acetivorans (188937)
Methanosarcina mazei (192952)
Methanosarcina barkeri (269797)
Methanococcoides burtonii (259564)
Natronomonas pharaonis (348780)
Haloarcula marismortui (272569)
Halobacterium sp. NRC-1 (64091)
Archaeoglobus fulgidus (224325)
Methanosphaera stadtmanae (339860)
Methanothermobacter thermautotr. (187420)
Methanopyrus kandleri (190192)
Methanocaldococcus jannaschii (243232)
Methanococcus maripaludis (39152)
Pyrococcus abyssi (272844)
Pyrococcus horikoshii (70601)
Pyrococcus furiosus (186497)
Thermococcus kodakarensis (69014)
Thermoplasma acidophilum (273075)
Thermoplasma volcanium (273116)
Picrophilus torridus (263820)
Sulfolobus acidocaldarius (330779)
Sulfolobus tokodaii (273063)
Sulfolobus solfataricus (2287)
Aeropyrum pernix (272557)
Pyrobaculum aerophilum (13773)
Nanoarchaeum equitans (228908)

Flow chart of the Signature web server: www.cmbi.ru.nl/signature.

- Input amino acid queries (FASTA, no maximum).
- Assign queries to orthologous groups (COG, KOG, NOG).
- Assess distribution of OG in given reference phylogeny, check signature definition.
- Calculate significance score (observed/expected ratio) and generate an insightful image of the tree of life, highlighting clades with signature genes in green and red.

Results based on 1,956 sequences in *Ferroplasma* scaffolds (Tyson et al. 2004), 974 unique OG assigned, 196 signature OGs (took <10 hours).

## references

- Ciccarelli, F.D., T. Doerks, C. von Mering, C.J. Creevey et al. (2006) "Toward automatic reconstruction of a highly resolved tree of life". Science 311:1283-1287.
- Dutilh, B.E., B. Snel, T.J.G. Ettema and M.A. Huynen (submitted) "Signature genes as a phylogenomic tool".
- Dutilh, B.E., Y. He, M.L. Hekkelman and M.A. Huynen (submitted) "Signature: a web server for taxonomic characterization of sequence samples using signature genes".
- Tringe, S.G., C. von Mering, A. Kobayashi, A.A. Salamov et al. 2005. "Comparative metagenomics of microbial communities". Science 308:554-557.
- Tyson, G.W., Chapman, J., Hugenholtz, P., Allenet al. (2004) "Community structure and metabolism through reconstruction of microbial genomes from the environment". Nature, 428, 37-43.
- Venter, J.C., K. Remington, J.F. Heidelberg, A.L. Halpern et al. 2004. "Environmental genome shotgun sequencing of the Sargasso Sea". Science 304:66-74.
- von Mering, C., L.J. Jensen, M. Kuhn, S. Chaffron et al. (2007) "STRING 7 recent developments in the integration and prediction of protein interactions". Nucleic Acids Res 35:D358-362.