ForBio – Course 2021
Phylogenomics

Torsten Struck – t.h.struck@nhm.uio.no
NHM UiO

Phylogenomics

- Original definition by Eisen (1998):
Improving Functional Predictions for Uncharacterized Genes by Evolutionary Analysis (Gene function prediction)

- Gene family evolution

- Most commonly used nowadays:
Reconstructing phylogenetic relationships using genomes or large part of the genome (e.g., transcriptomes)

- Studying horizontal gene transfer

The era of "-omics"

- Genomics
- Transcriptomics
- Proteomics

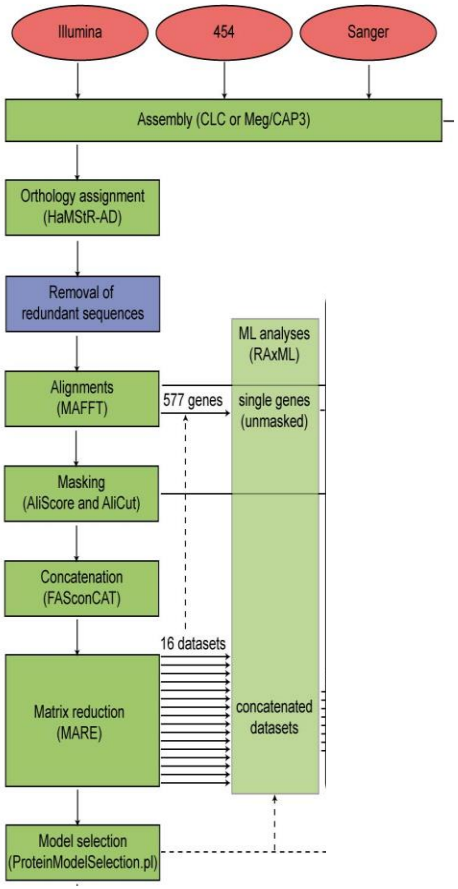
The era of "-omics"

- Genomics
- Transcriptomics
- Proteomics

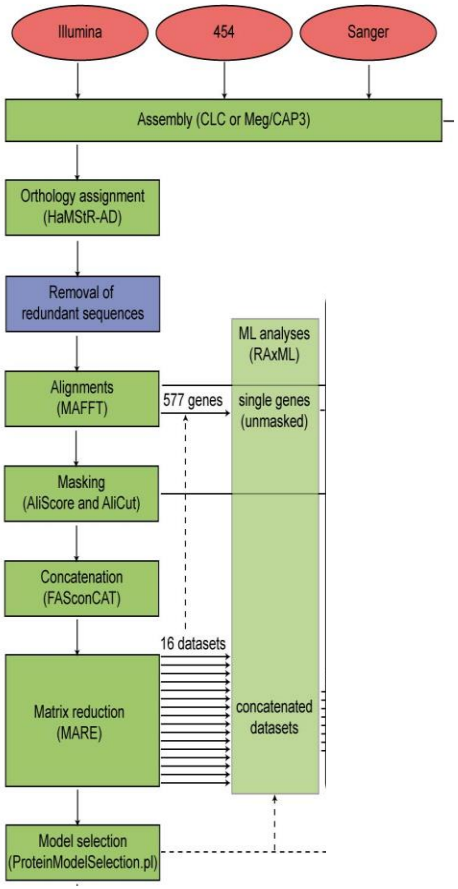
Next generation sequencing:

- 454 pyrosequencing
- Illumina (Solexa) sequencing
- SOLiD sequencing
- Ion semiconductor sequencing (Ion torrent)
- Single molecule real time sequencing (Pacific Biosciences)
- Nanopore DNA sequencing

The standard procedure

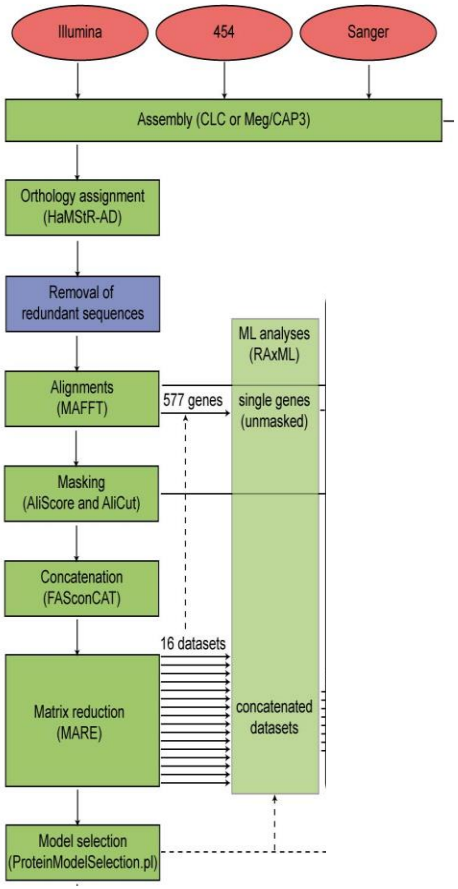


The standard procedure



Sequence data generation

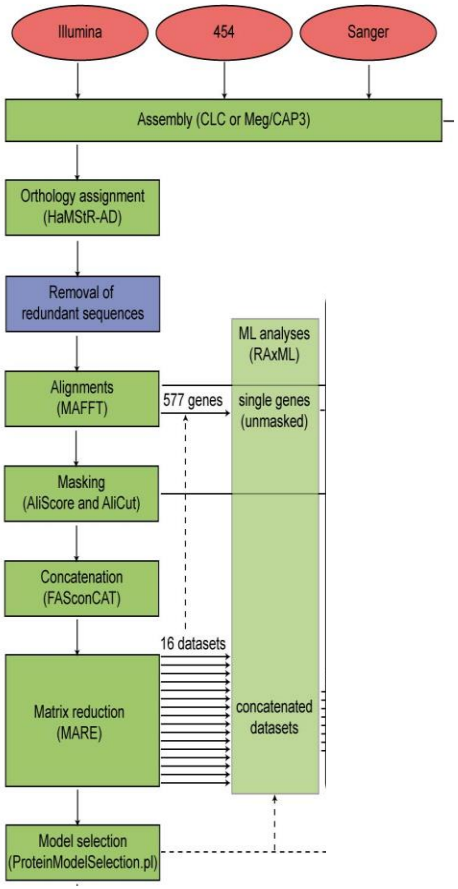
The standard procedure



Sequence data generation

Assembly of short reads into larger genes

The standard procedure

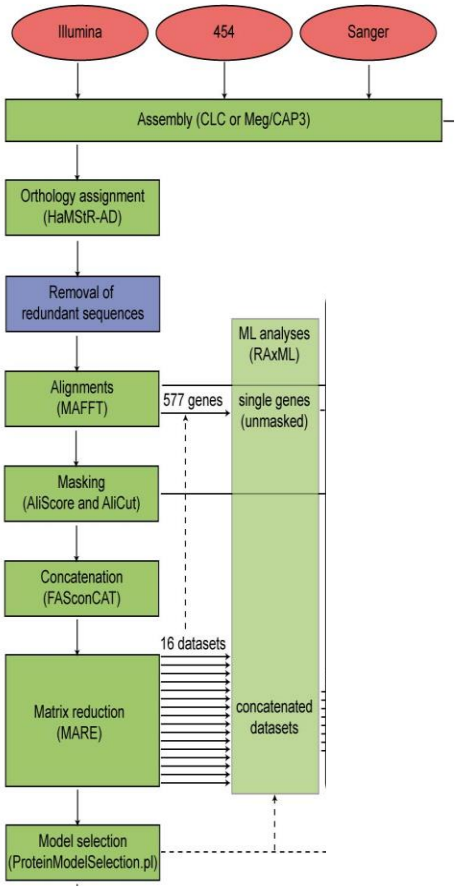


Sequence data generation

Assembly of short reads into larger genes

Determination of orthologous genes

The standard procedure



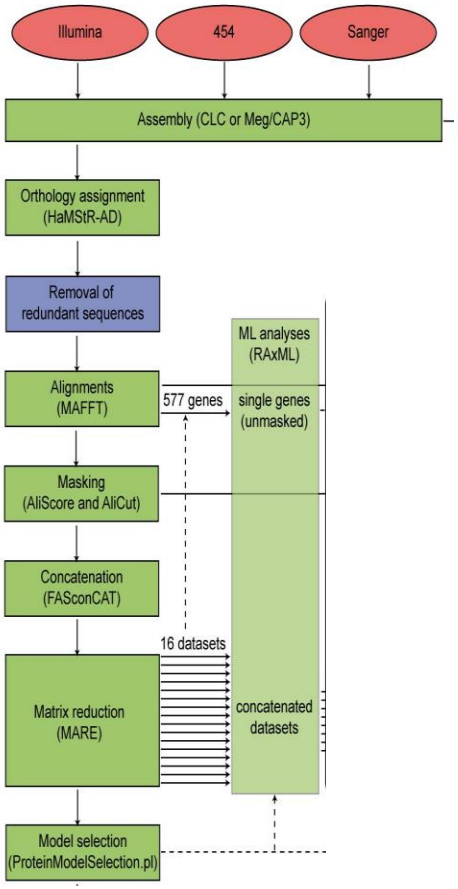
Sequence data generation

Assembly of short reads into larger genes

Determination of orthologous genes

Alignment of the genes

The standard procedure



Sequence data generation

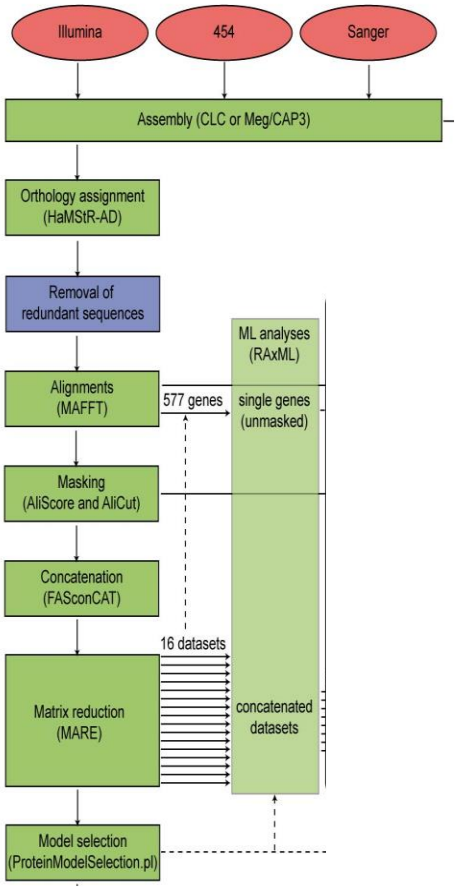
Assembly of short reads into larger genes

Determination of orthologous genes

Alignment of the genes

Masking of positions, which are not well aligned

The standard procedure



Sequence data generation

Assembly of short reads into larger genes

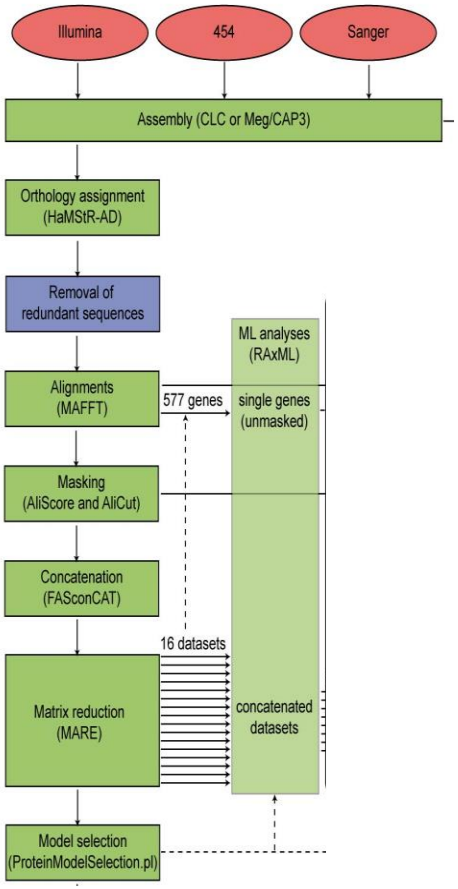
Determination of orthologous genes

Alignment of the genes

Masking of positions, which are not well aligned

Generation of supermatrix

The standard procedure



Sequence data generation

Assembly of short reads into larger genes

Determination of orthologous genes

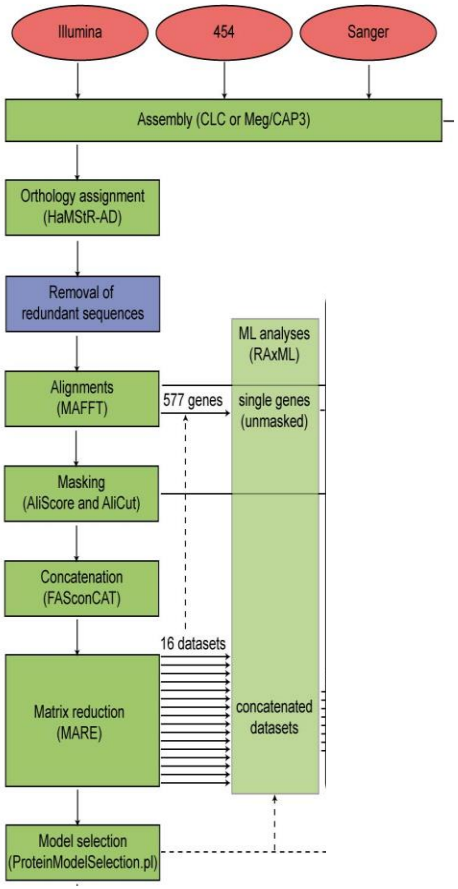
Alignment of the genes

Masking of positions, which are not well aligned

Generation of supermatrix

Condensation of supermatrix to increase data coverage

The standard procedure



Sequence data generation

Assembly of short reads into larger genes

Determination of orthologous genes

Alignment of the genes

Masking of positions, which are not well aligned

Generation of supermatrix

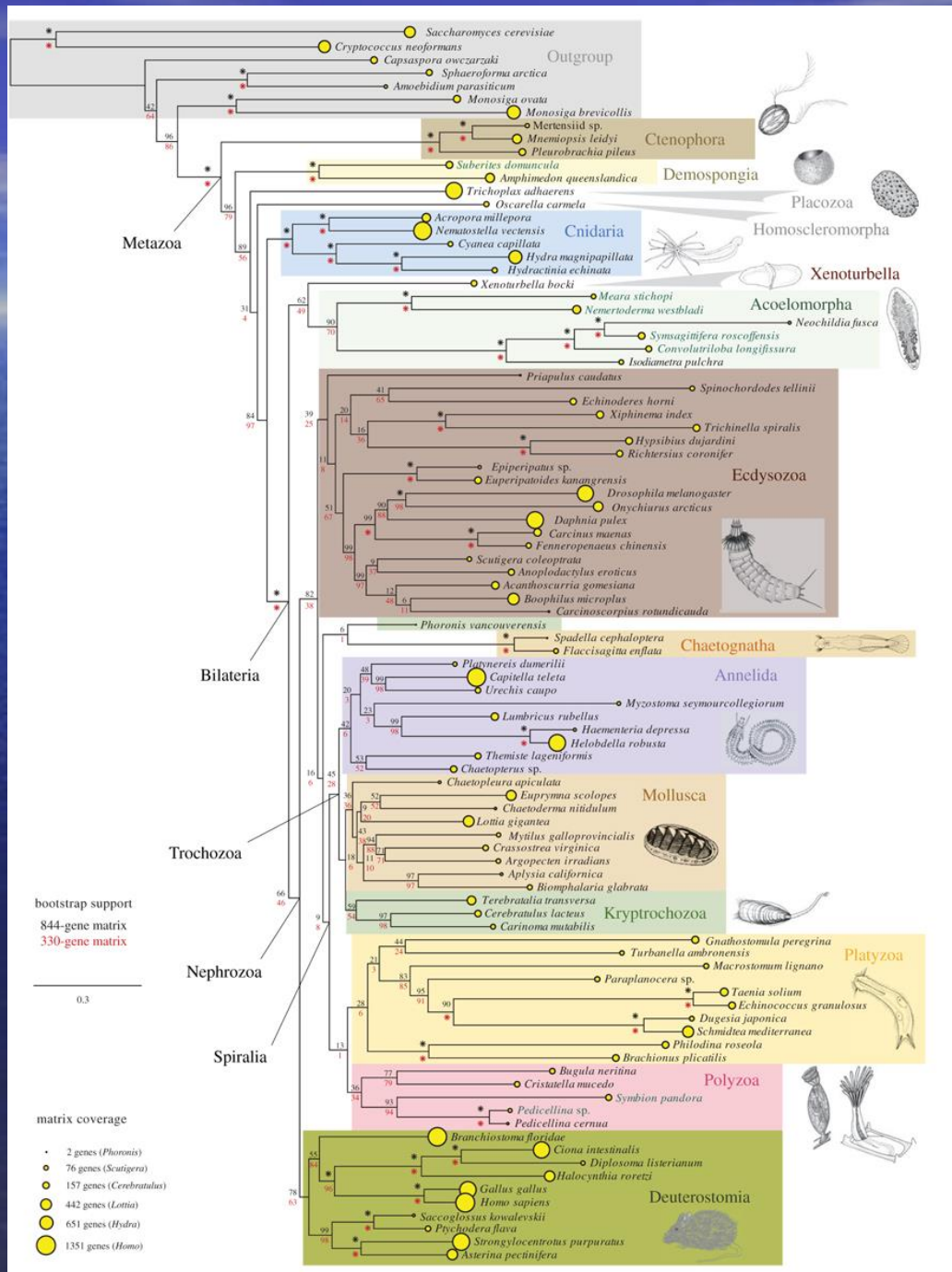
Condensation of supermatrix to increase data coverage

Reconstruction of the relationships of the taxa to each other

Phylogenomics

- Monophyly of Platyzoa

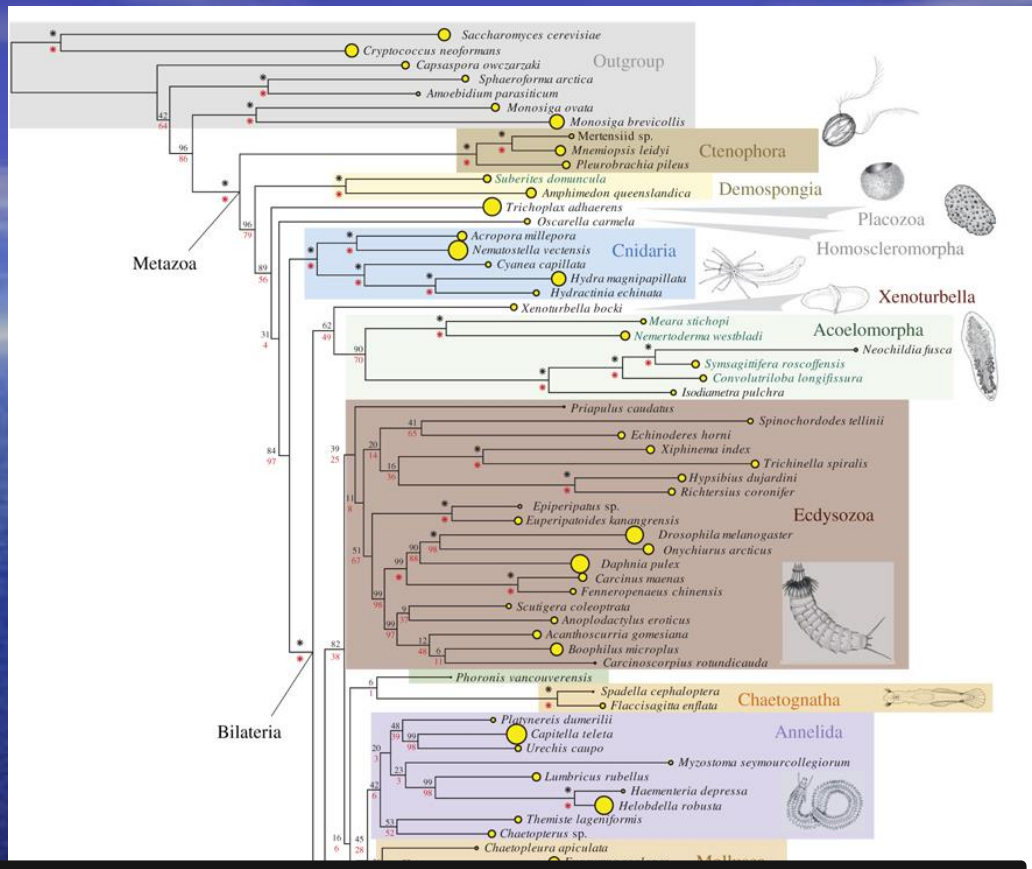
Hejnol et al. (2009). Proc. R. Soc. B: Biol. Sci.

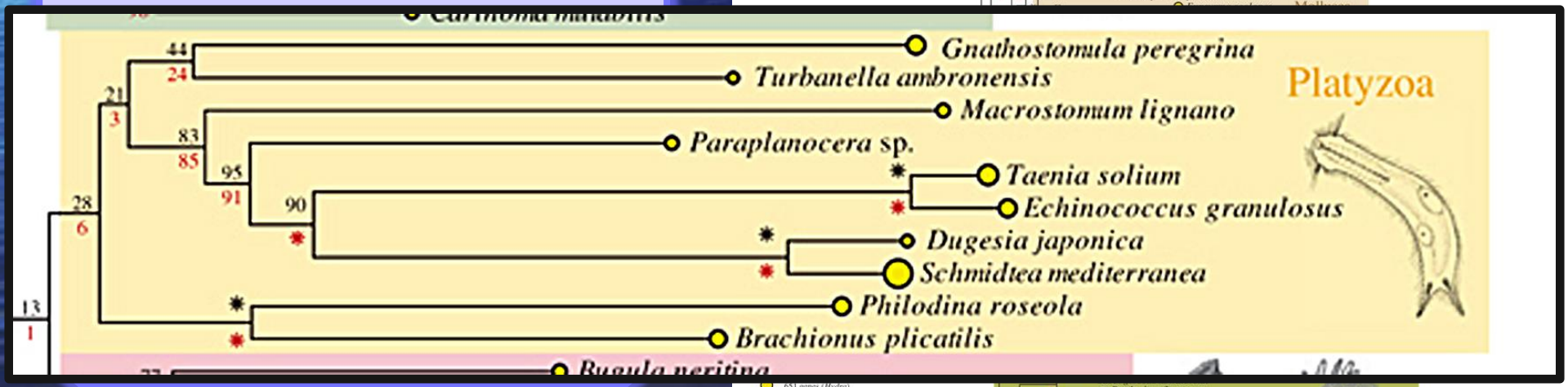


Phylogenomics

- Monophyly of Platyzoa
- Except for Platyhelminthes, only one or two species

Hejnol et al. (2009). Proc. R. Soc. B: Biol. Sci.

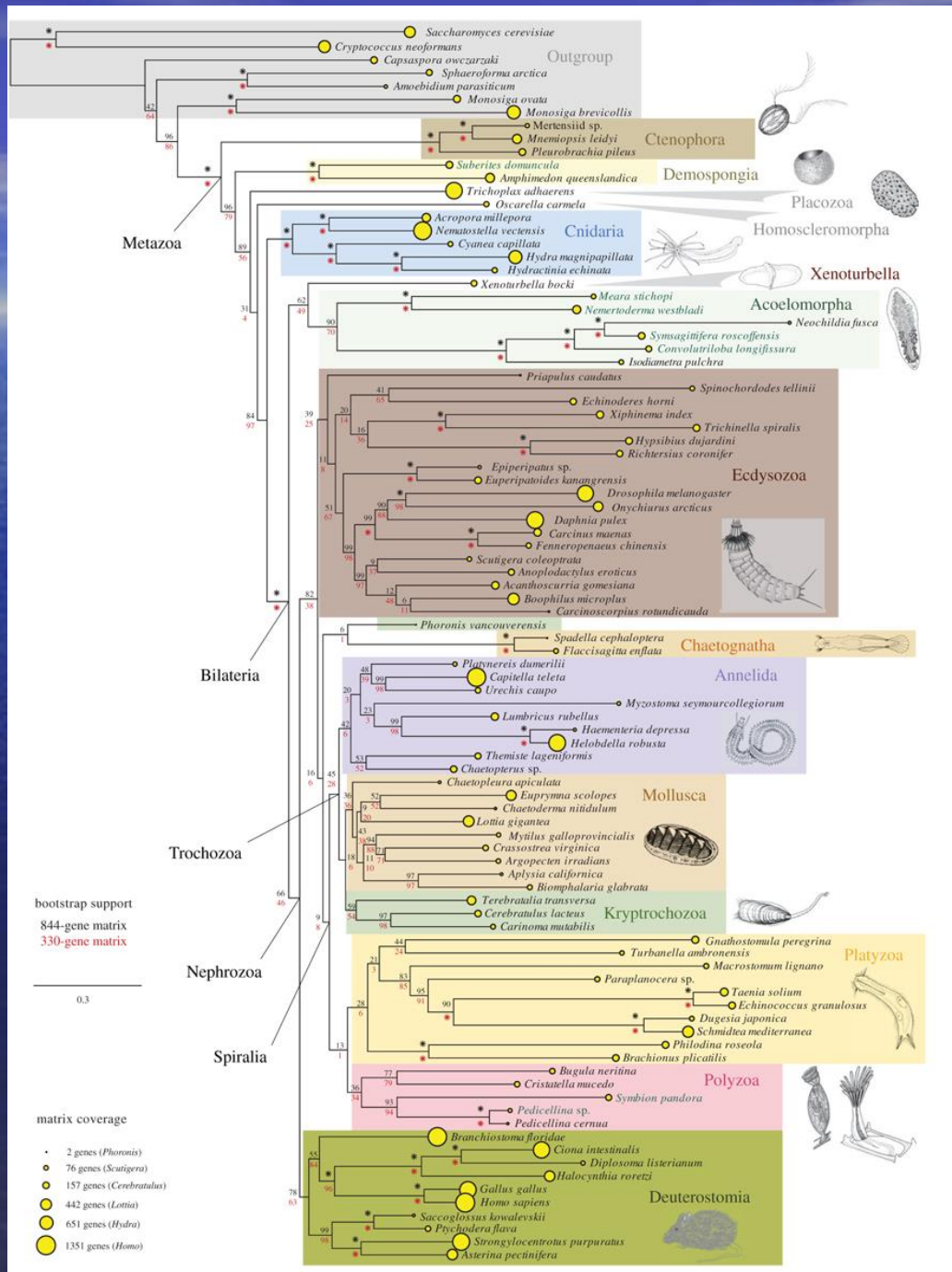




Phylogenomics

- Monophyly of Platyzoa
- Except for Platyhelminthes, only one or two species
- Small libraries

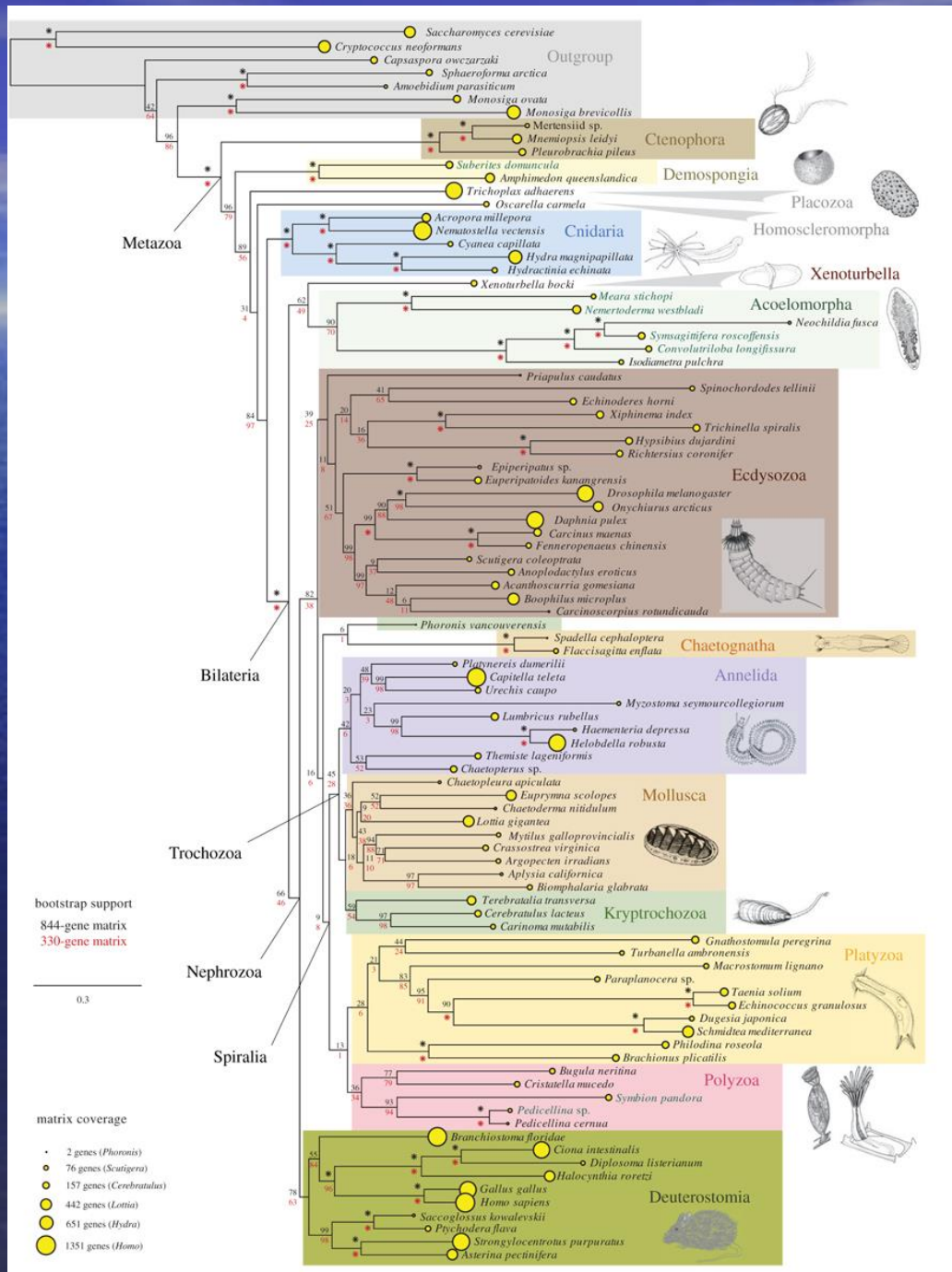
Hejnol et al. (2009). Proc. R. Soc. B: Biol. Sci.



Phylogenomics

- Monophyly of Platyzoa
- Except for Platyhelminthes, only one or two species
- Small libraries
- Low coverage (74 out of 1,487 genes)

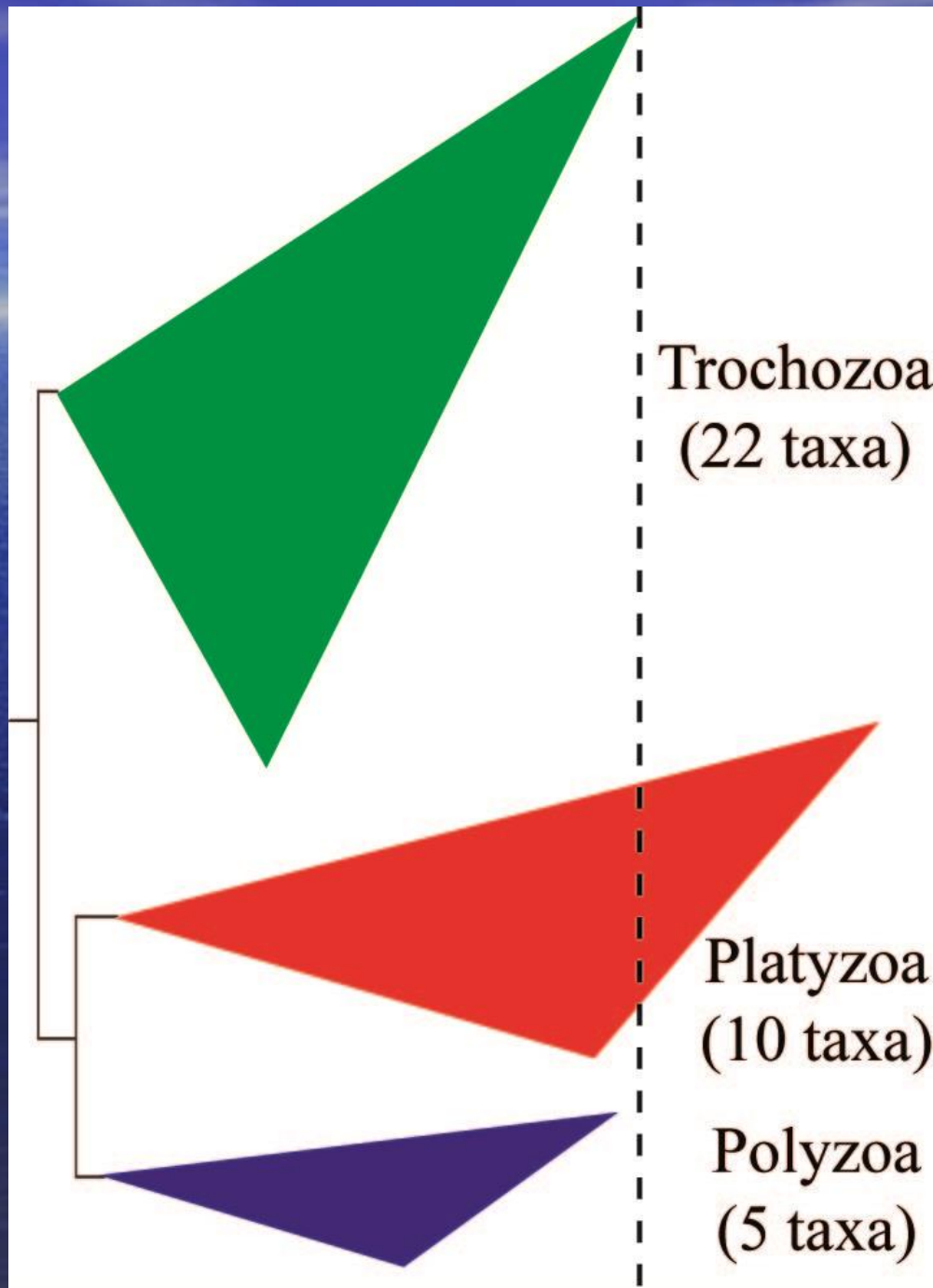
Hejnol et al. (2009). Proc. R. Soc. B: Biol. Sci.



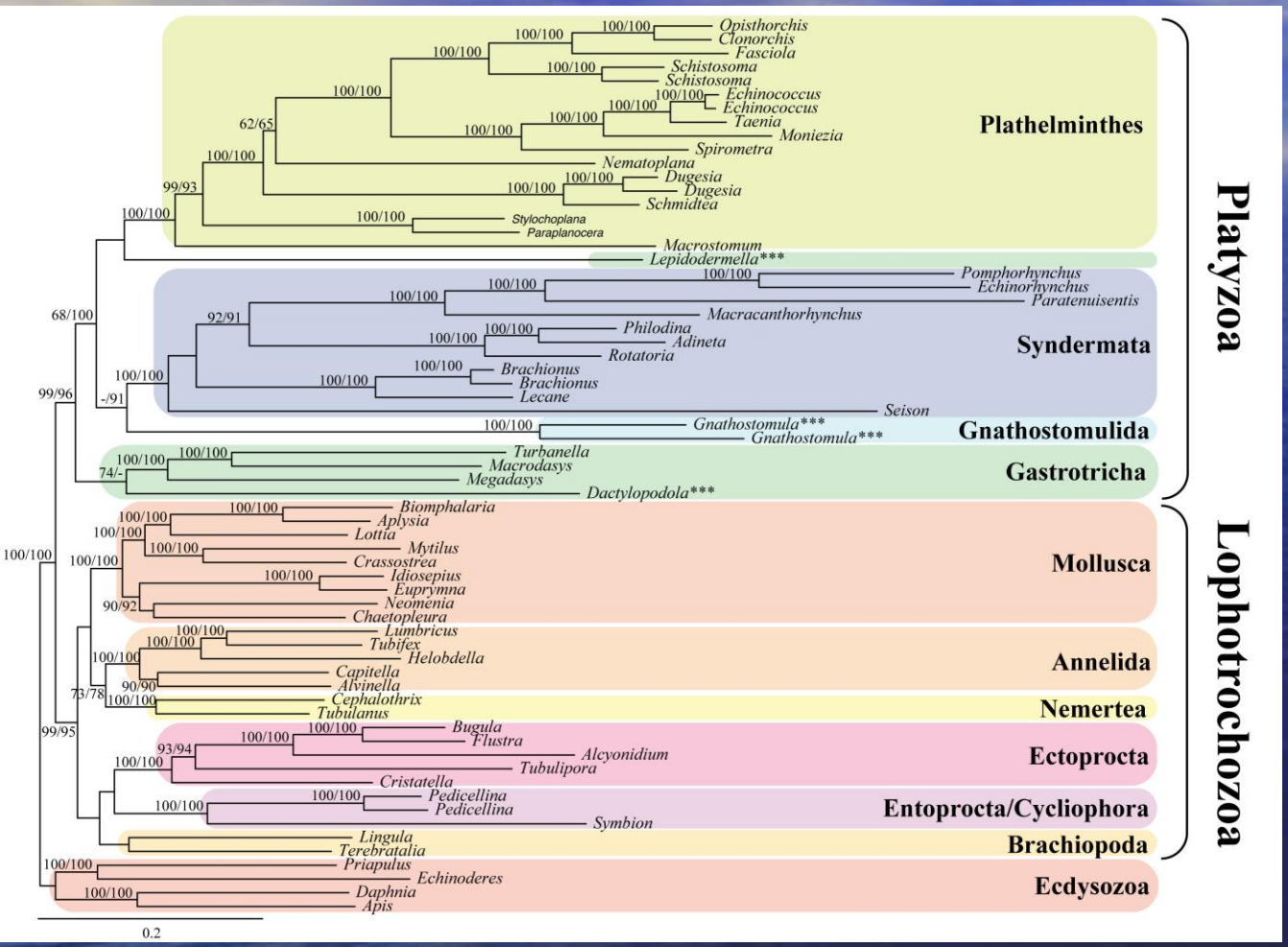
Phylogenomics

- Monophyly of Platyzoa
- Except for Platyhelminthes, only one or two species
- Small libraries
- Low coverage (74 out of 1,487 genes)
- Long branches

Hejnol et al. (2009). Proc. R. Soc. B: Biol. Sci.



Brute-force phylogenomics



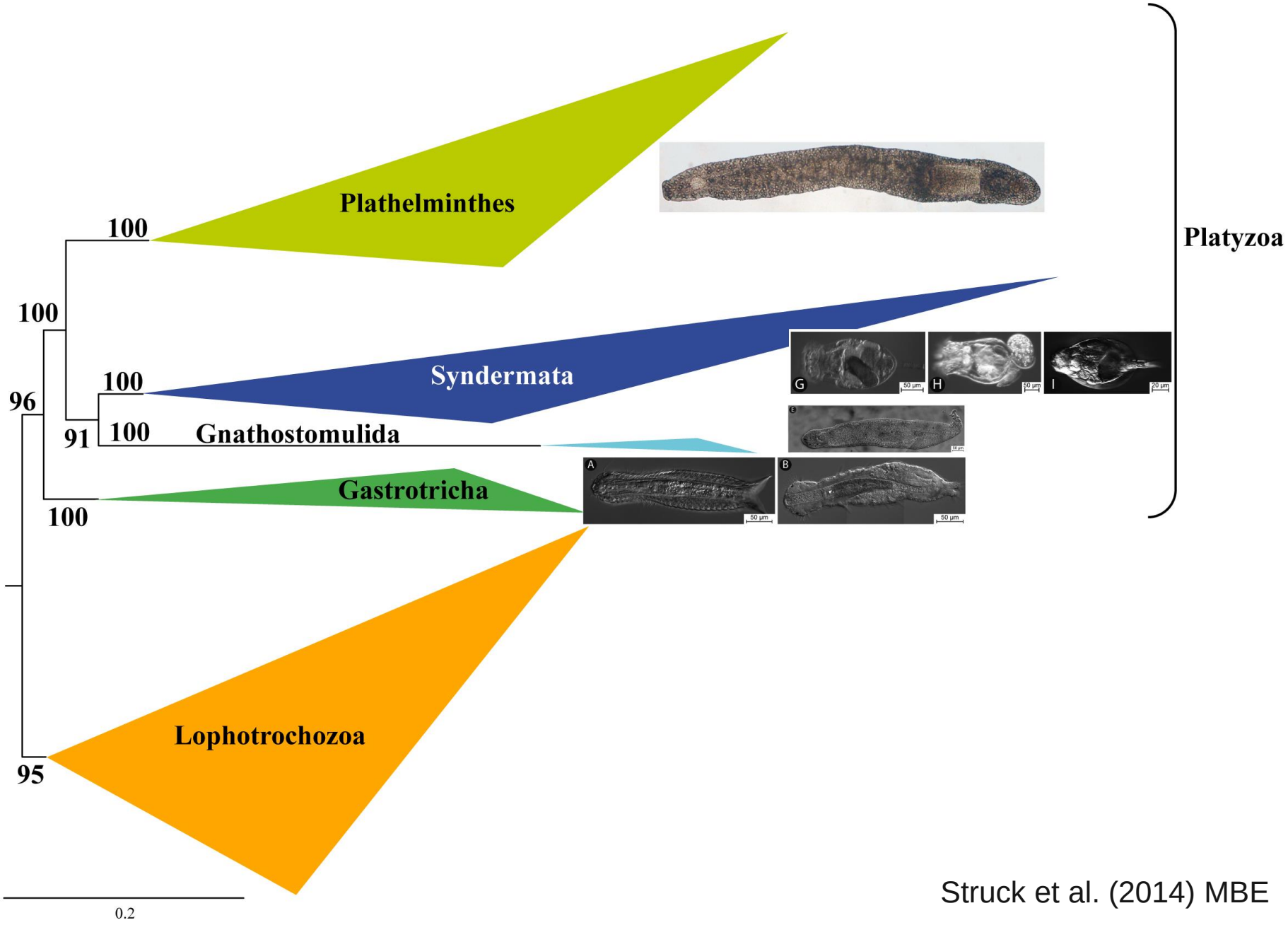
559 ortholog genes
82162 aa
35.7% coverage/taxon

RAxML
Only bootstrap values ≥ 70%
Sensitivity analyses
*** non-stable taxa

Latter BS values without *Lepidodermella* and *Dactylopodola* due to long-branch issues

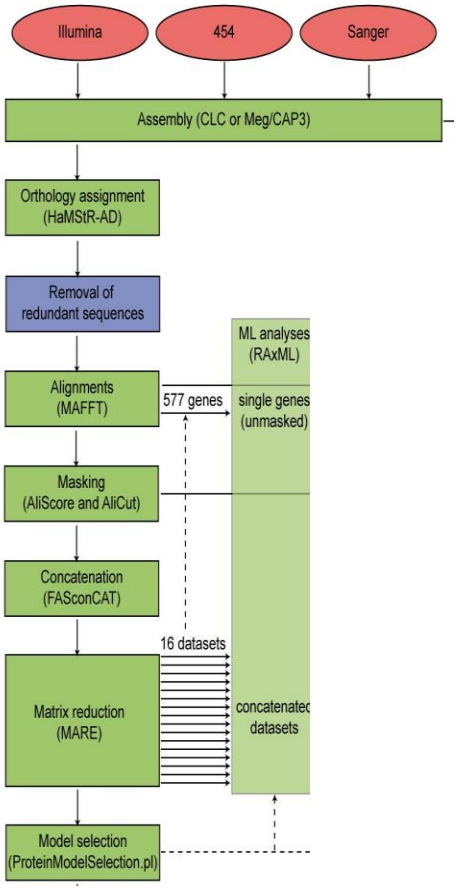
Struck et al. (2014) MBE

Brute-force phylogenomics

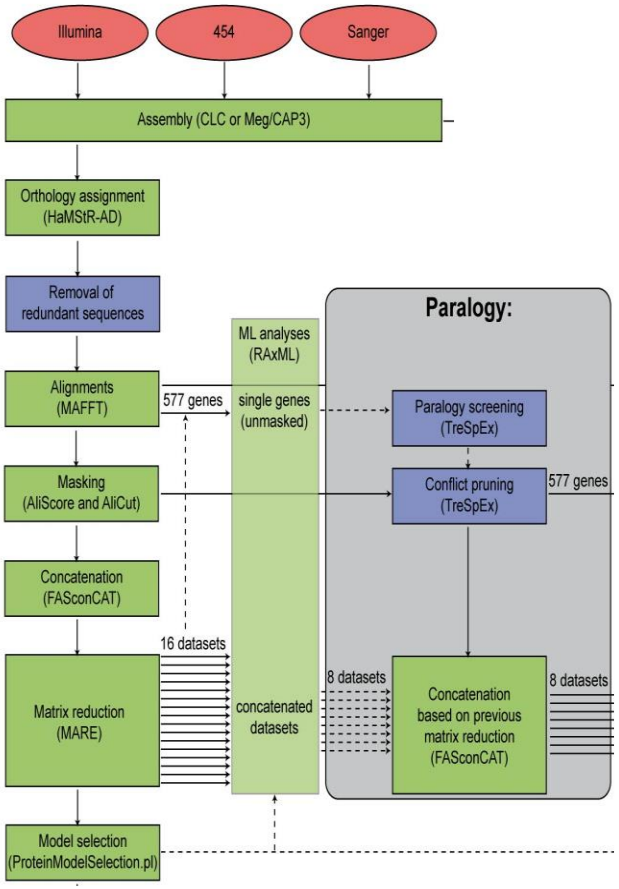


Struck et al. (2014) MBE

Going beyond the standard

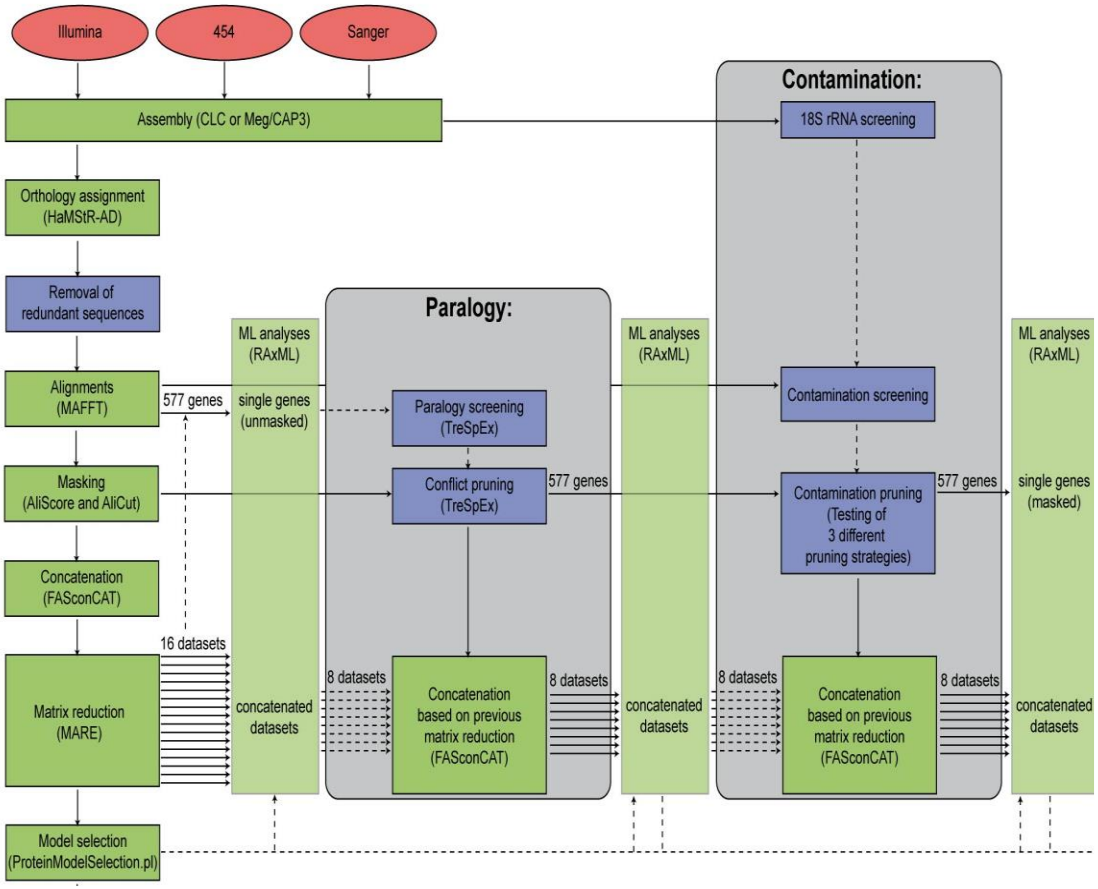


Going beyond the standard



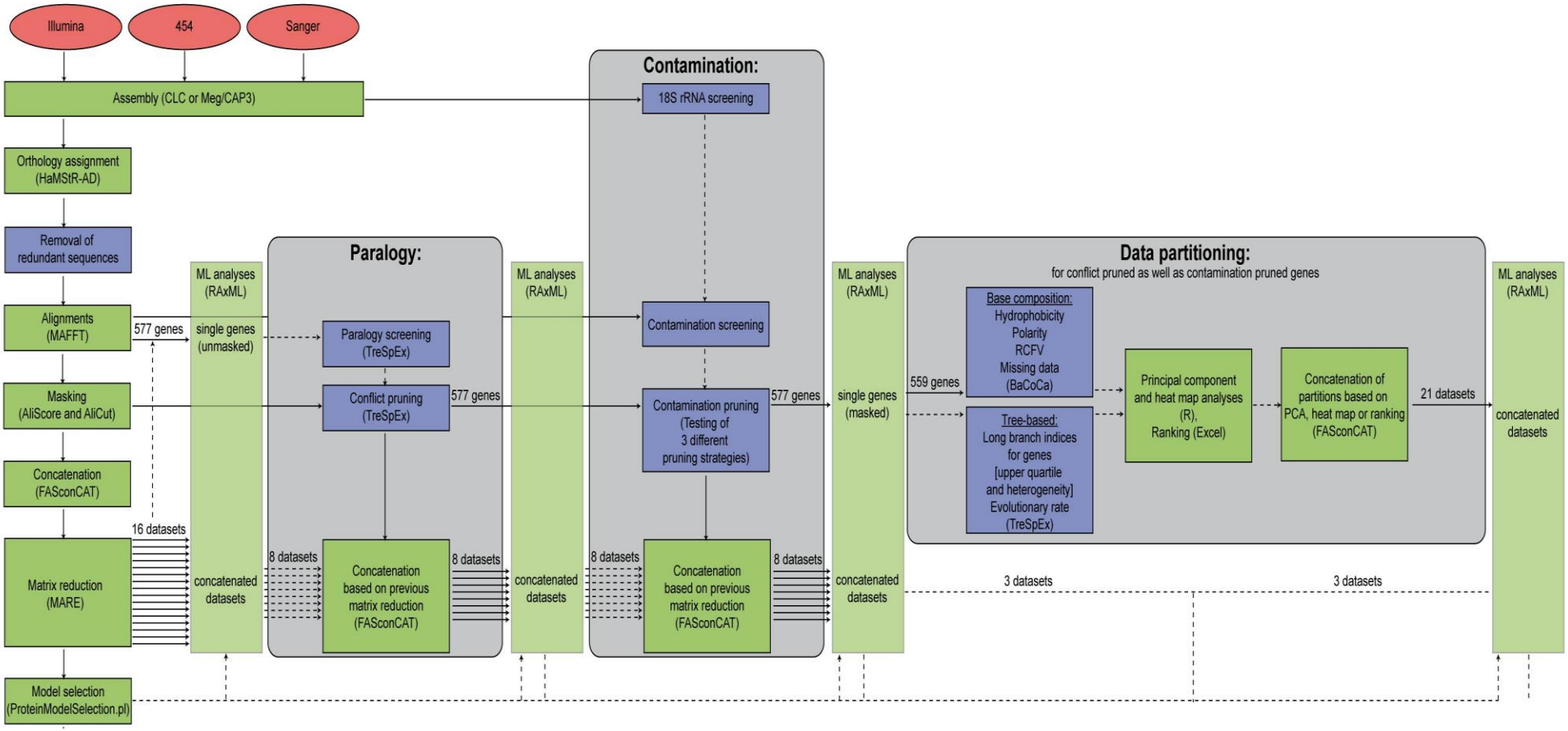
Struck (2013) PLoS one
Struck (2014) Evol. Bioinf.

Going beyond the standard



Struck et al. (2014) MBE

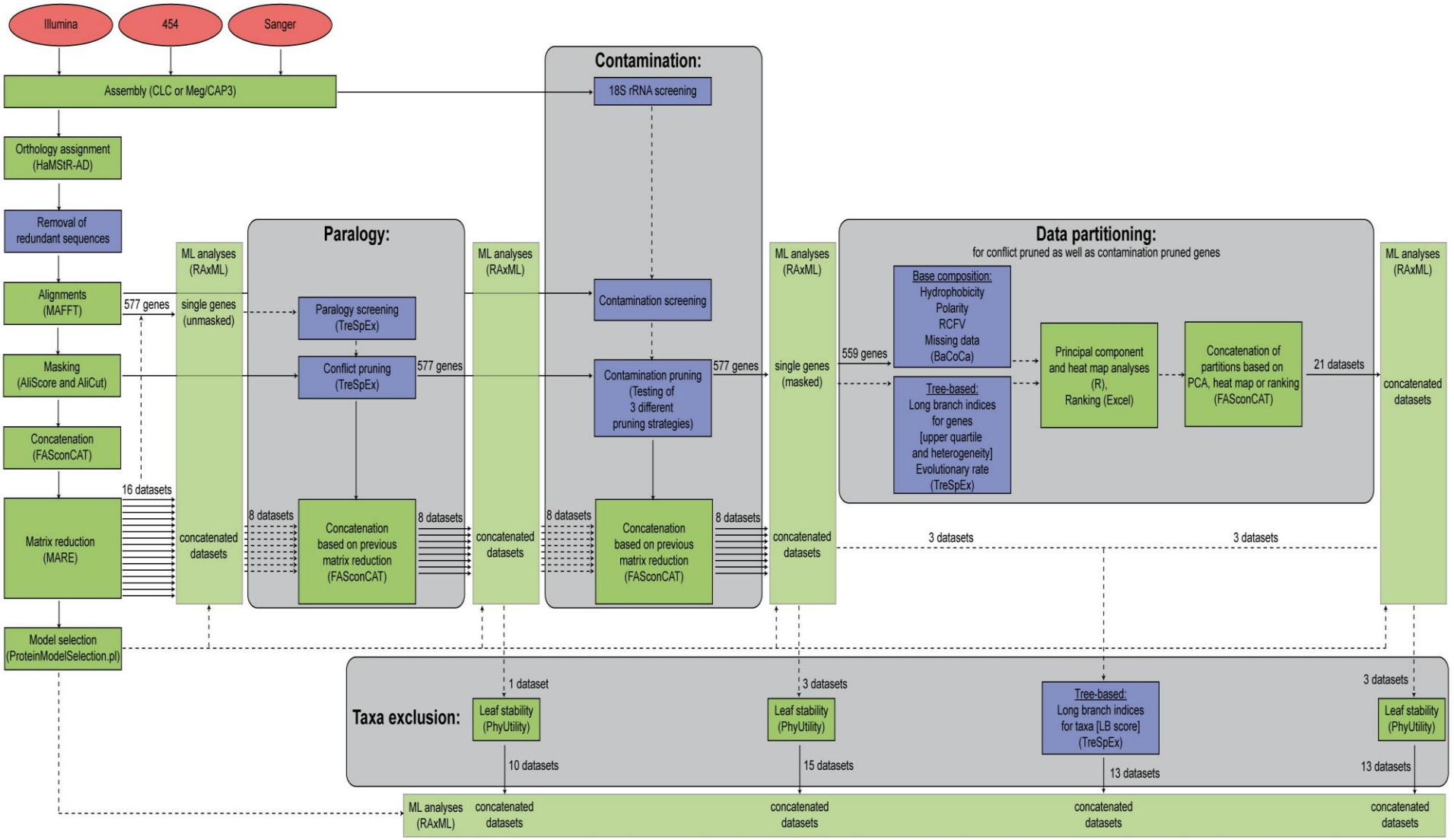
Going beyond the standard



Struck et al. (2014) MBE
Zhong et al. (2011) BMC Evol. Biol.

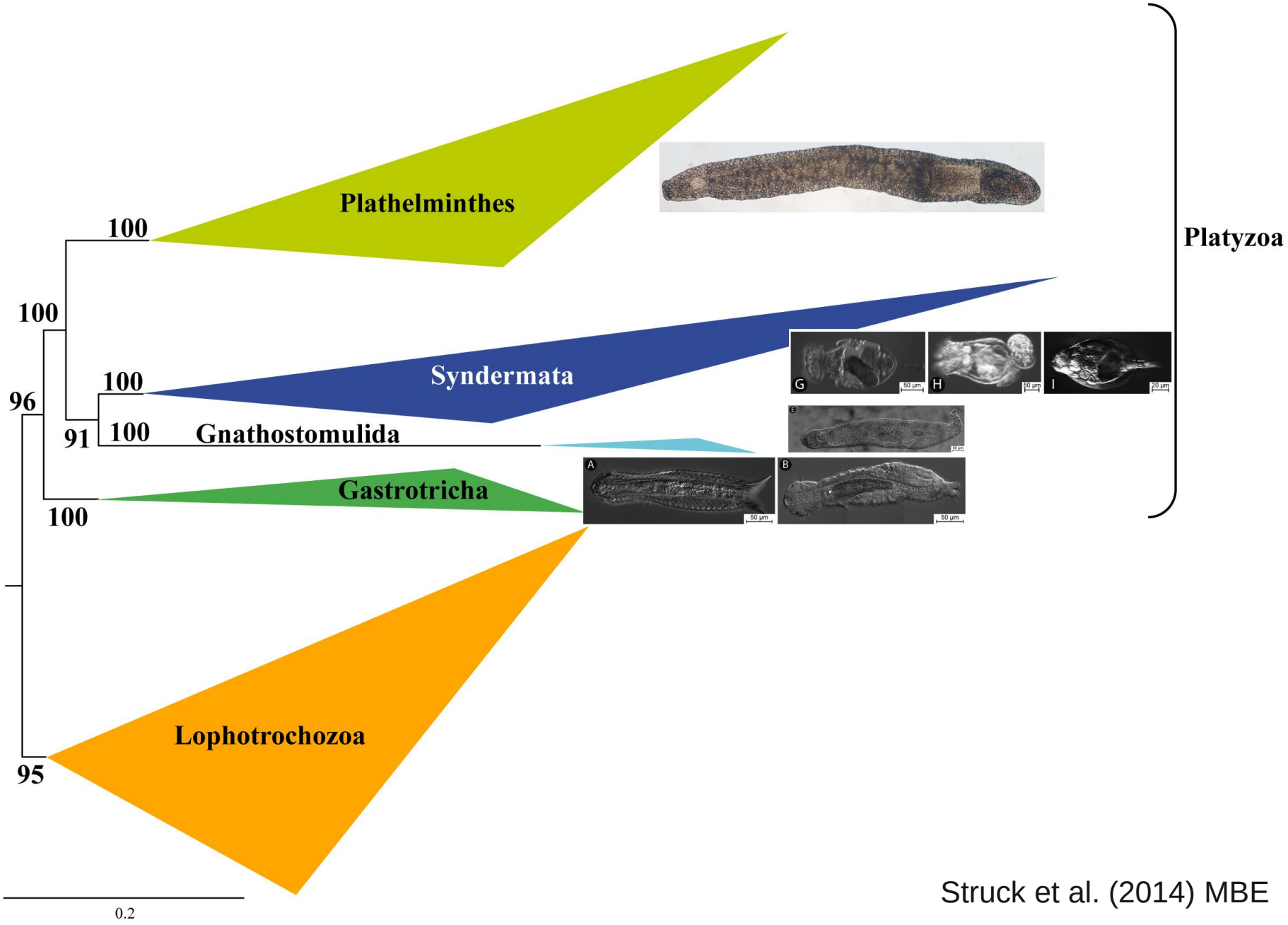
Kück & Struck (2014) MPE
Struck (2014) Evol. Bioinf.

Going beyond the standard



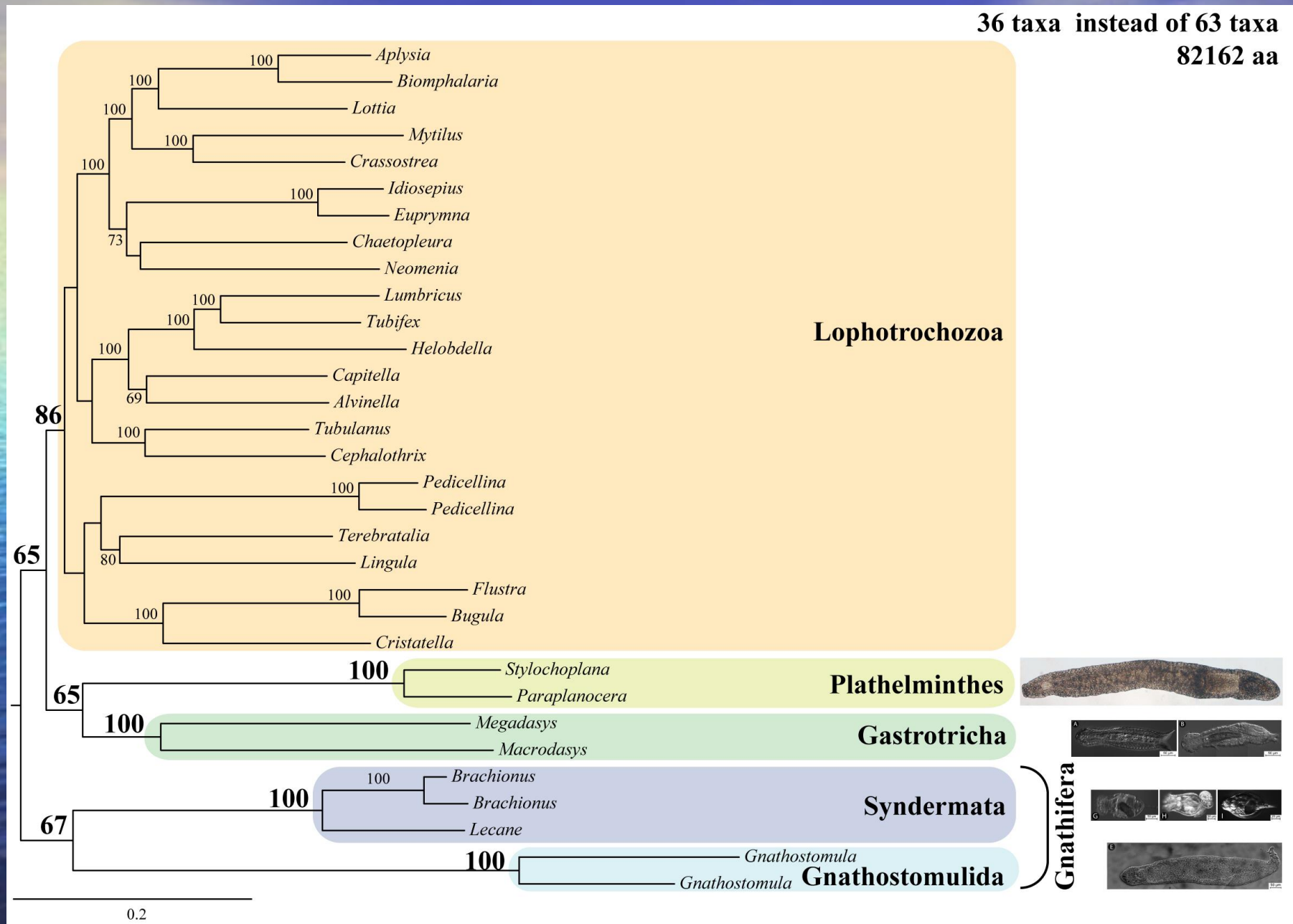
Struck et al. (2014) MBE

Brute-force phylogenomics



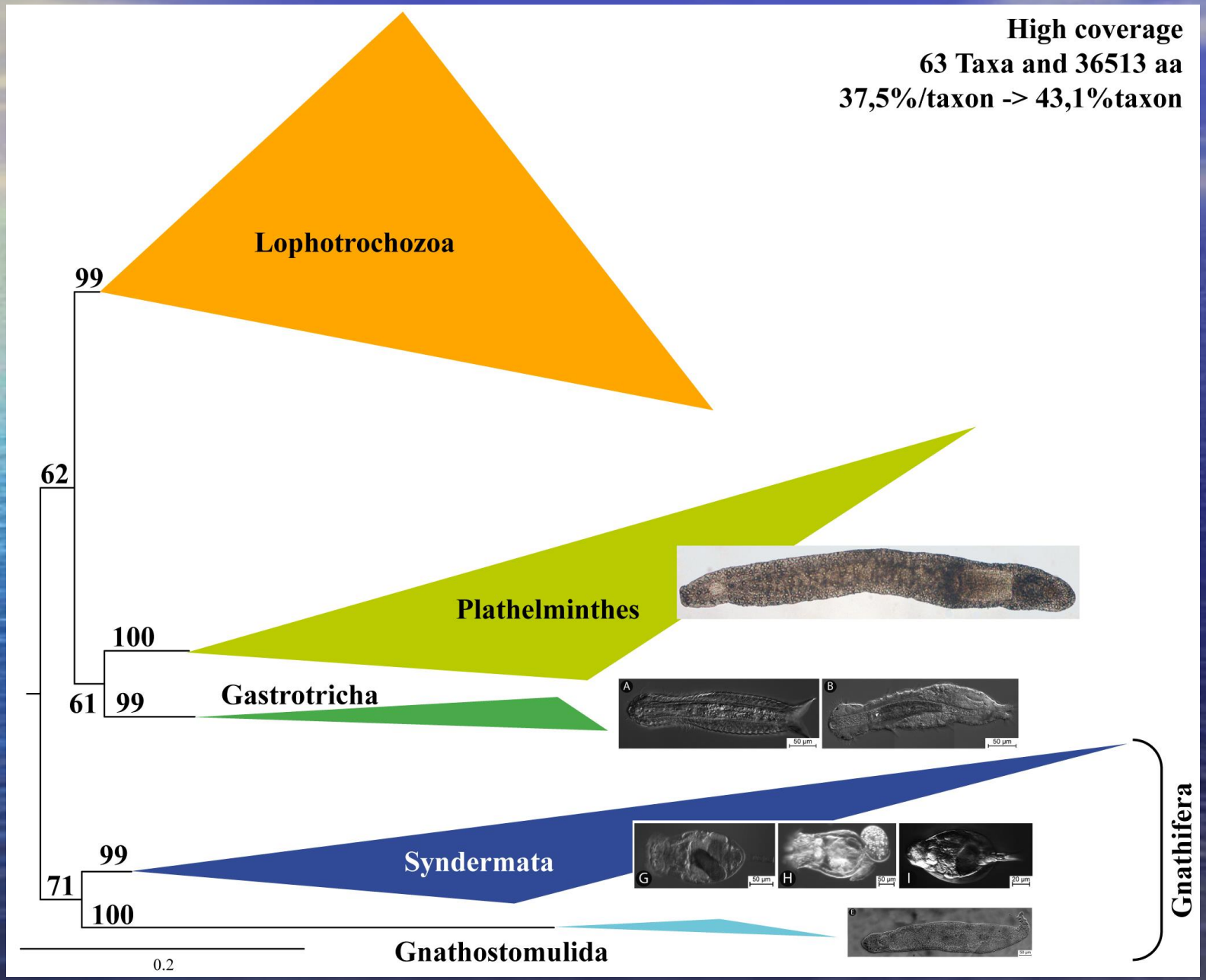
Struck et al. (2014) MBE

Exclusion of long-branched taxa

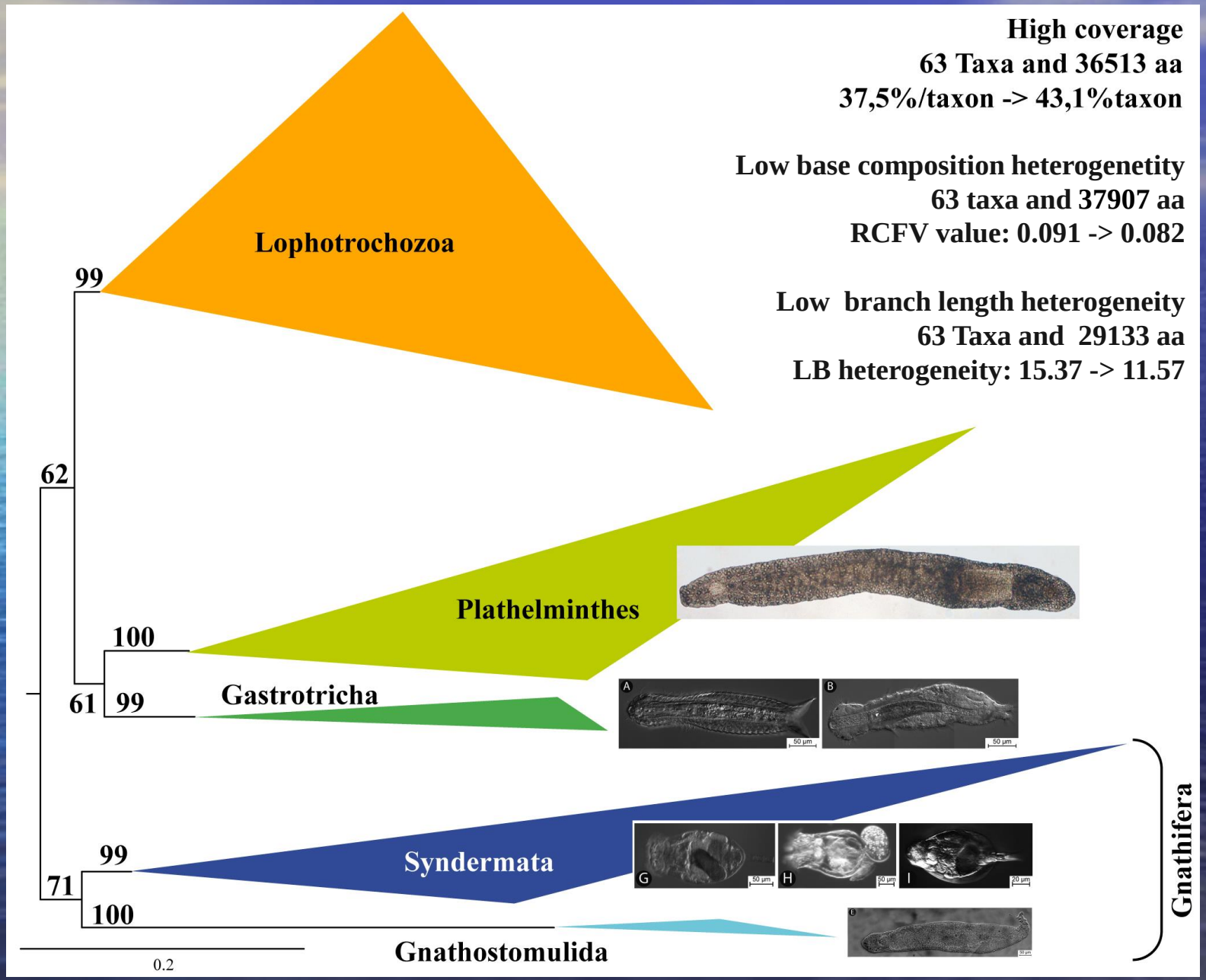


Struck et al. (2014) MBE

Exclusion of biased data

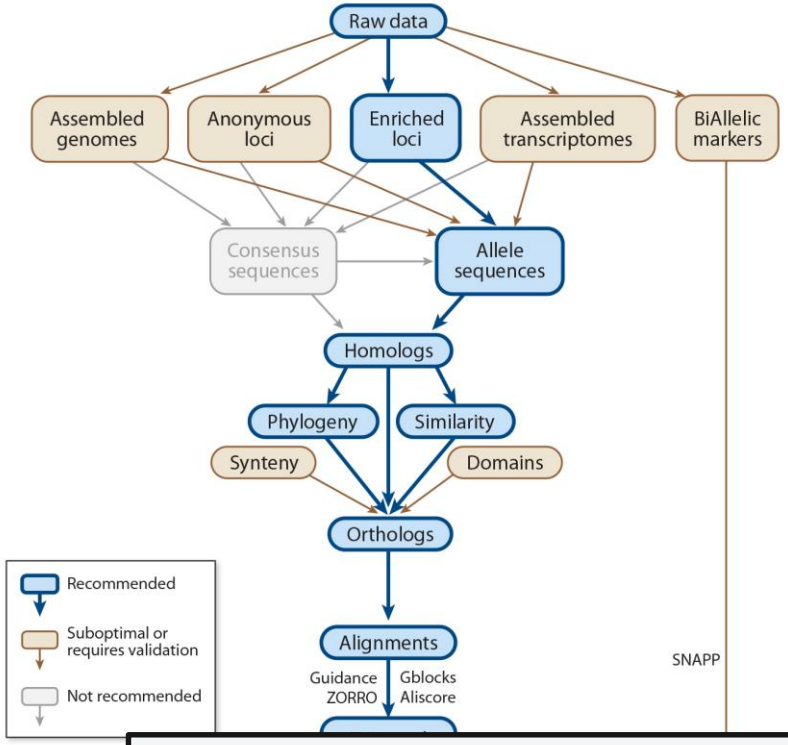


Exclusion of biased data

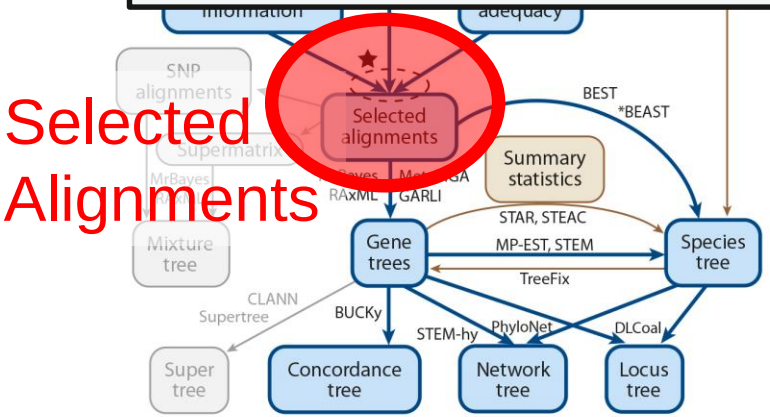


Struck et al. (2014) MBE

Data quality



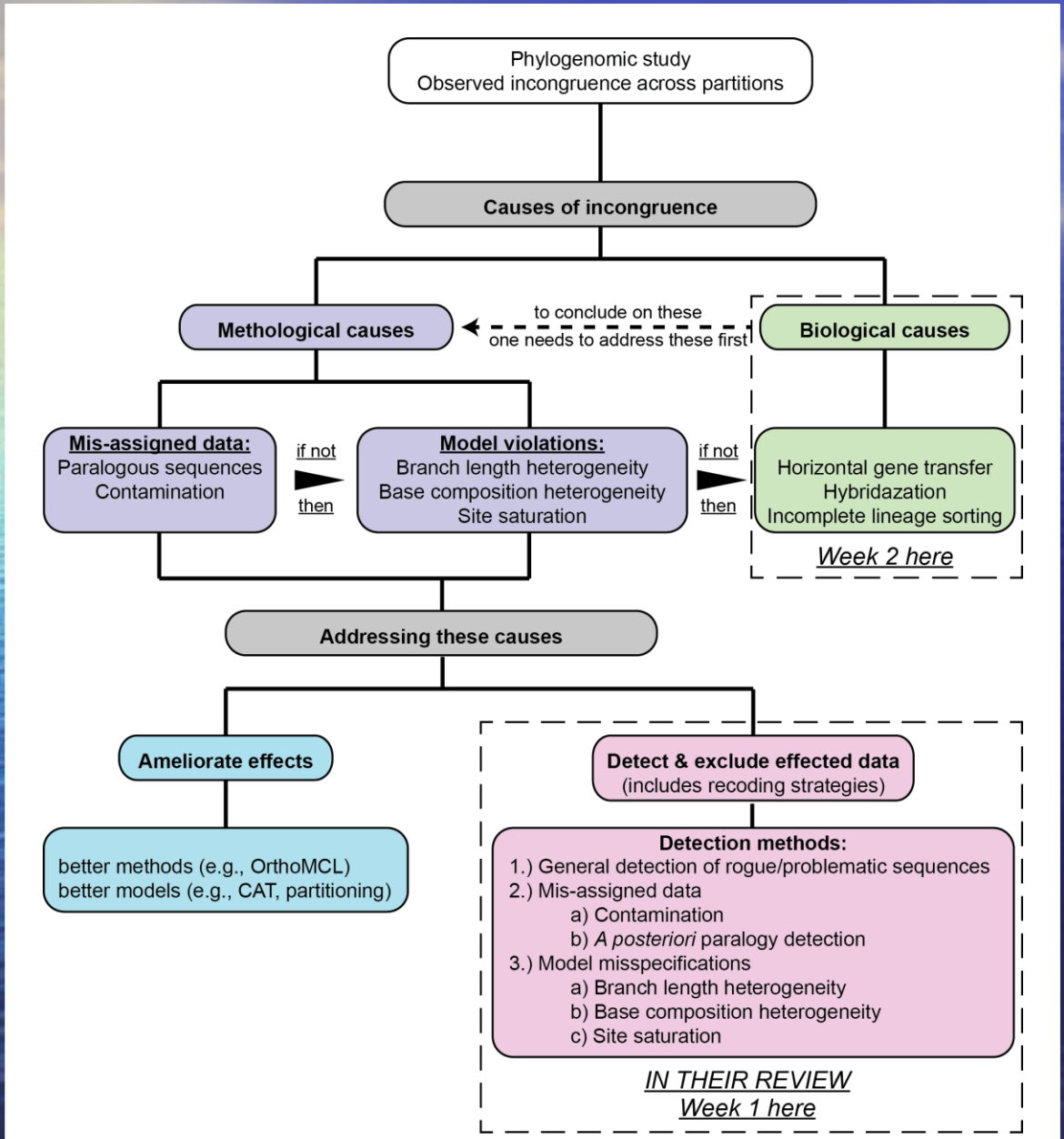
Given the rapidly increasing quantity of available data, systematists would benefit from an integrated system for data subsampling that would allow selection of a sufficient number of loci using objective criteria for minimizing phylogenetic error (indicated by a *star* in **Figure 4**).



Selected Alignments

Lemmon & Lemmon (2013) AREES

Data quality



Modified from Fleming et al. (in review)

Program week 1

Day	Task on GitHub	Topic
Monday	Day1_Morning	Introductory lectures
	Day1_Afternoon	Contamination
Tuesday	Day2_Morning	Paralogy
	Day2_Afternoon	Missingness & Low signal
Wednesday	Day3_Morning	Rogue taxa
	Day3_Afternoon	Model misspecification
Thursday	Day4_Morning	Evolutionary rate & saturation
	Day5_Morning	Branch length heterogeneity 1
Friday	Day4_Afternoon	Compositional heterogeneity
	Day5_Afternoon	Branch length heterogeneity 2

Your own computer & SAGA
https://documentation.sigma2.no/jobs/submitting.html

Login into SAGA
ssh *$USERNAME@saga.sigma2.no*
Password: *your password*
Project number: NN9458K

Information on GitHub (cookbooks for each task)
https://github.com/ForBioPhylogenomics/tutorials