How often are we wrong? A Bayesian assessment of taxonomic identifications for the National Wadeable Streams Assessment

Lester L. Yuan[1]

Jeroen Gerritsen[2]

James B. Stribling[2]

[1]Office of Research and Development, US EPA, Washington, DC

[2]TetraTech, Inc., Owings Mill, MD

Basic Question

- How consistent are observations of the presence and absence of different invertebrate taxa between different taxonomists?

  - Why focus on presence/absence?

    Presence/absence information is inherent to richness metrics and predictive models (e.g., RIVPACS).

More refined question

- If a taxonomist tells you that a certain taxon is in a sample, what is the probability that the taxon is truly in the sample?

- We would like to estimate a value for:

$$P(\textit{present} \mid \textit{observed})$$

Data

- National Wadeable Stream Assessment (NWSA) benthic samples were counted by 26 taxonomists in 11 different labs.

- 72 samples were randomly selected and re-identified by a single, QC taxonomist.

Estimating identification error rates

True Positive Rate: proportion of times that a taxon was observed, and was actually present.

False Positive Rate: proportion of times that a taxon was observed, but was actually absent.

Major Assumption

- QC taxonomist identifications are "truth" (for this analysis).
- All identification errors are committed by primary taxonomists.

Consistent identifications are the goal.

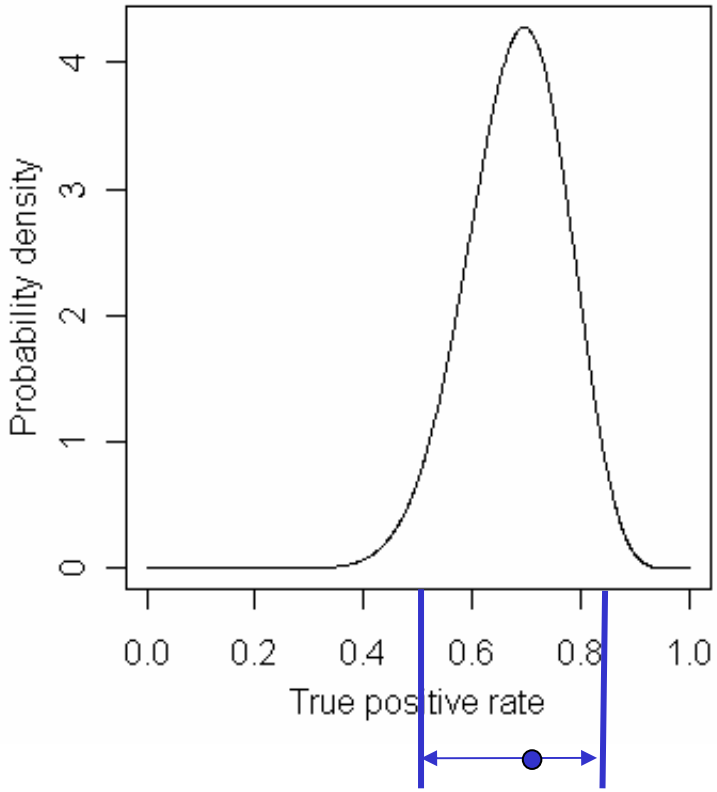
Example: Hydropsyche

		QC taxonomist	
		Present	Absent
Primary taxonomist	Observed	**16**	**6**
	Not observed	**7**	**43**

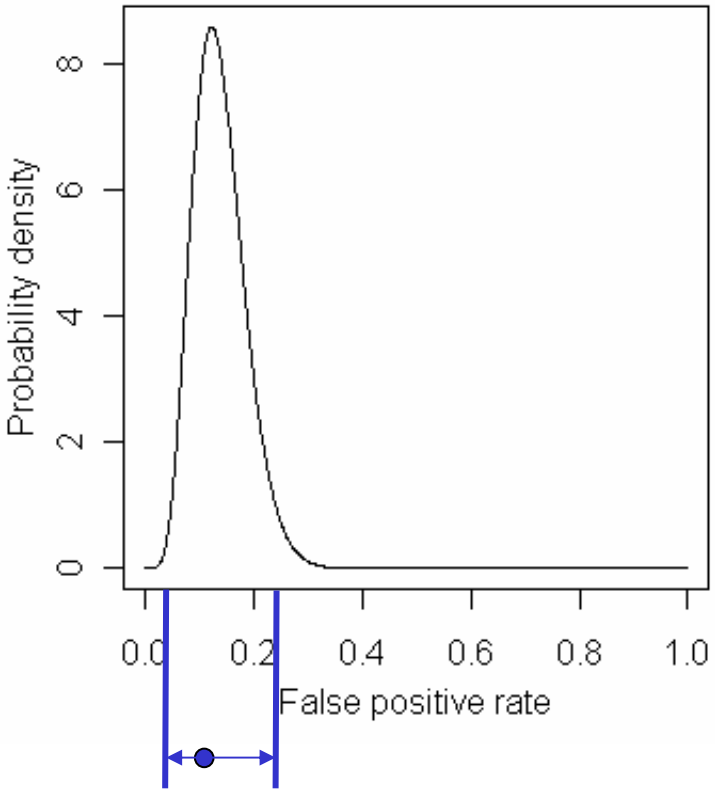
True positive rate: 16/(16 + 7) = 0.70

False positive rate: 6/(6 + 43) = 0.12

Uncertainty in error rates (Hydropsyche)



0.52 0.82



0.07 0.22

What do identification error rates tell us?

- True positive rate can be restated:
  - Given that a taxon is *present*, what is the probability that the taxon is *observed* in the sample?

- True positive rate = P(*observed* | *present*)

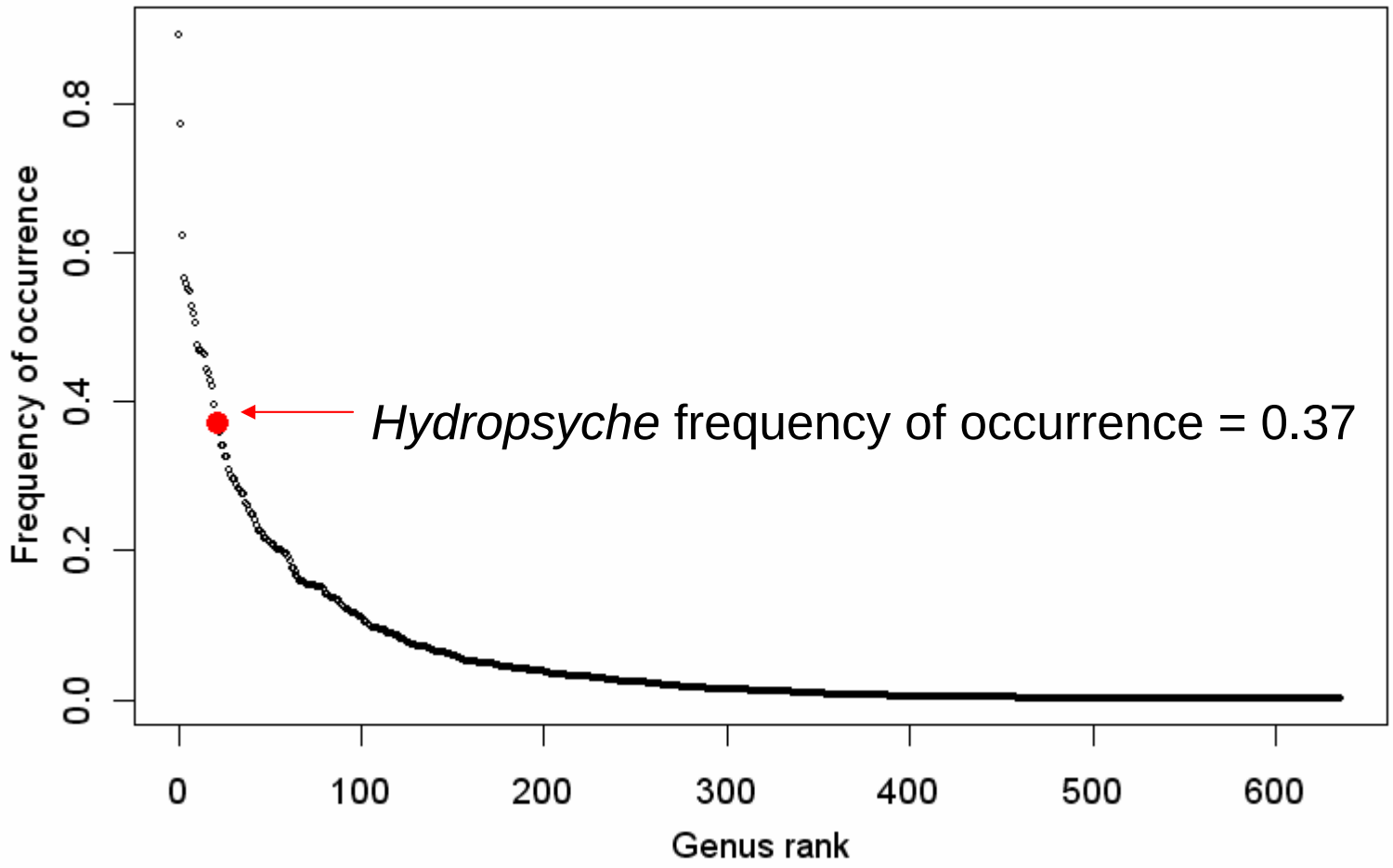
- We want to know P(*present* | *observed*).

NWSA data

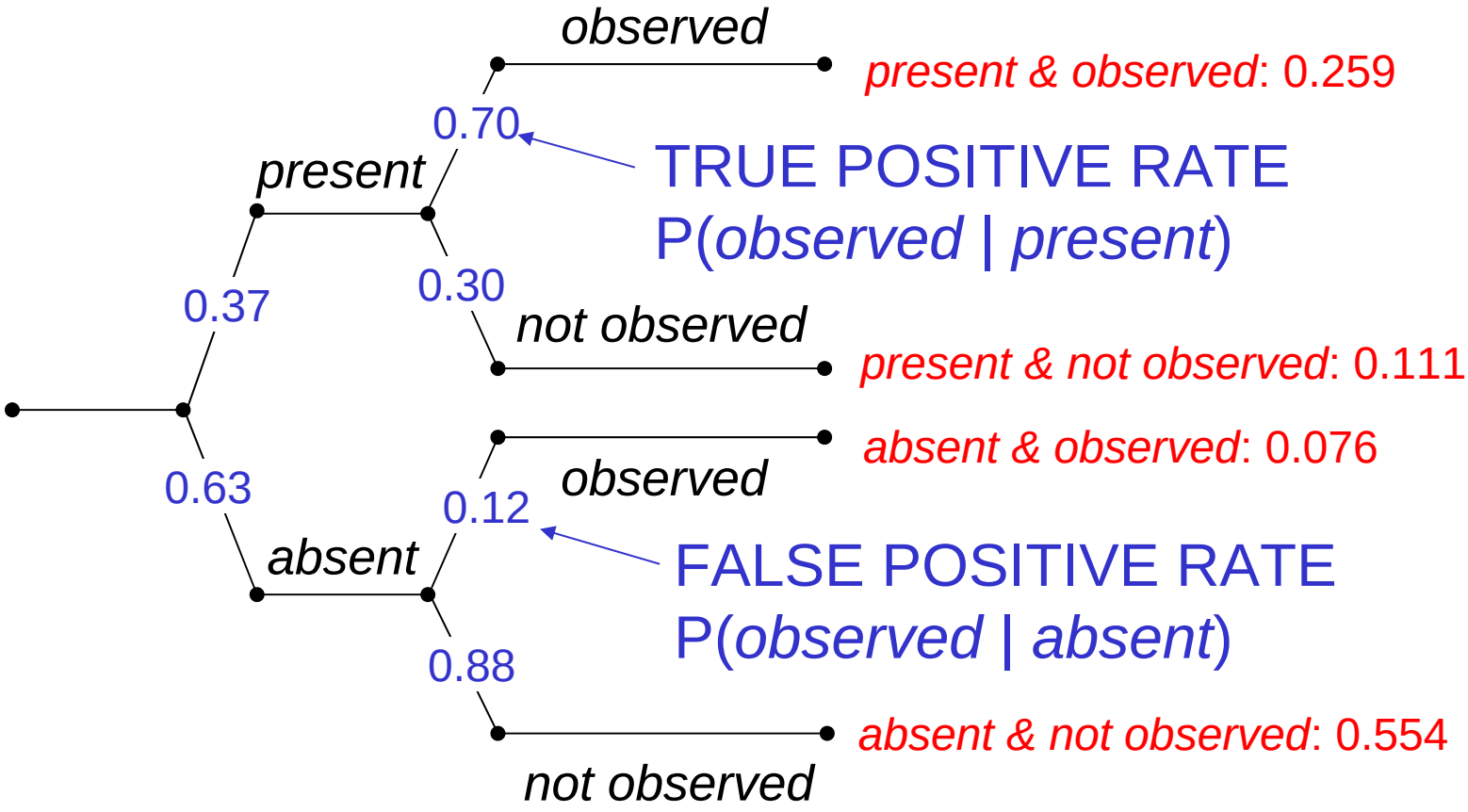
- Estimate frequencies of occurrence of different taxa from 750 NWSA benthic samples.

- Major assumption:
  - Assume primary taxonomist identifications match those of the QC taxonomist.

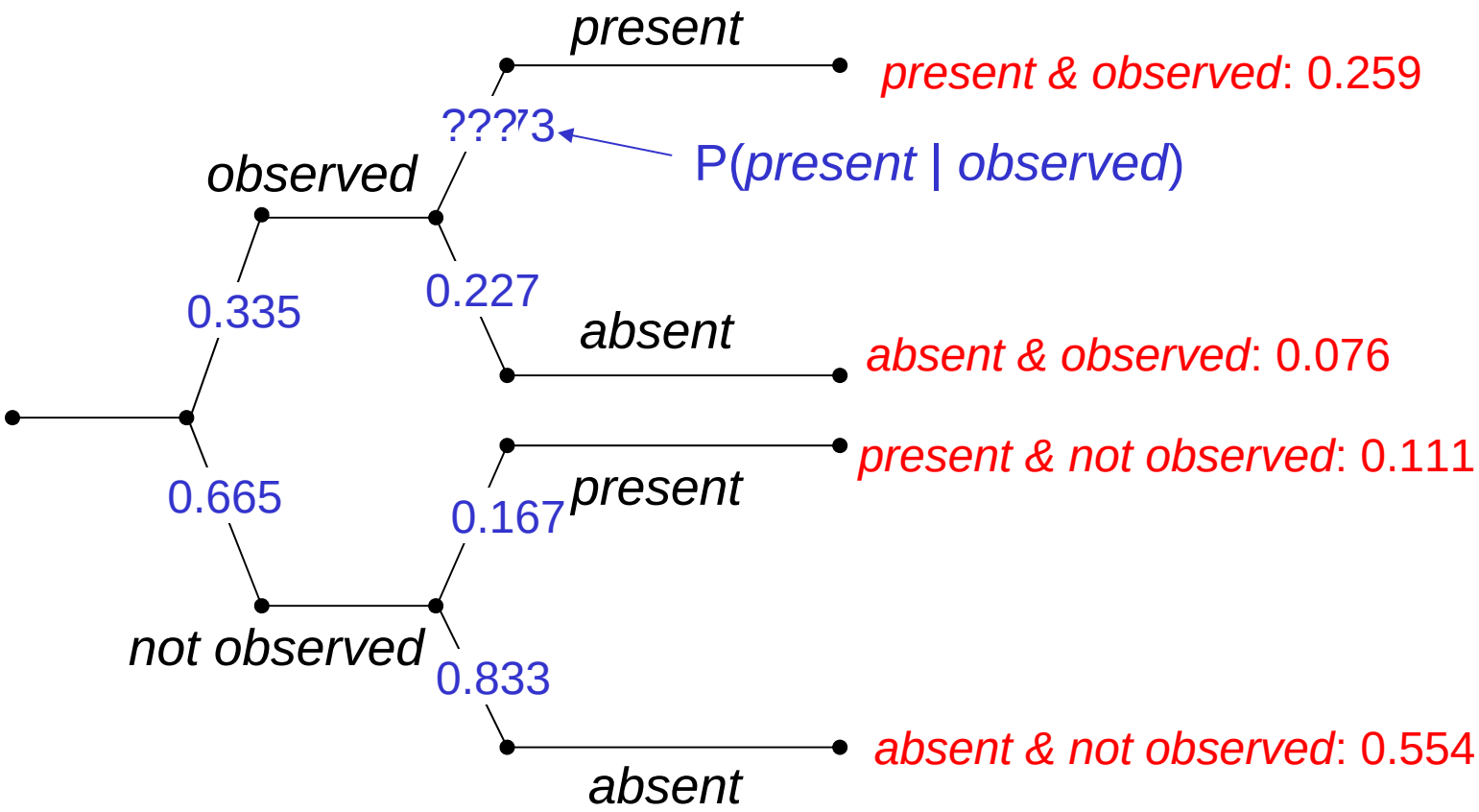
Frequencies of occurrence in WSA



Computing the posterior probability



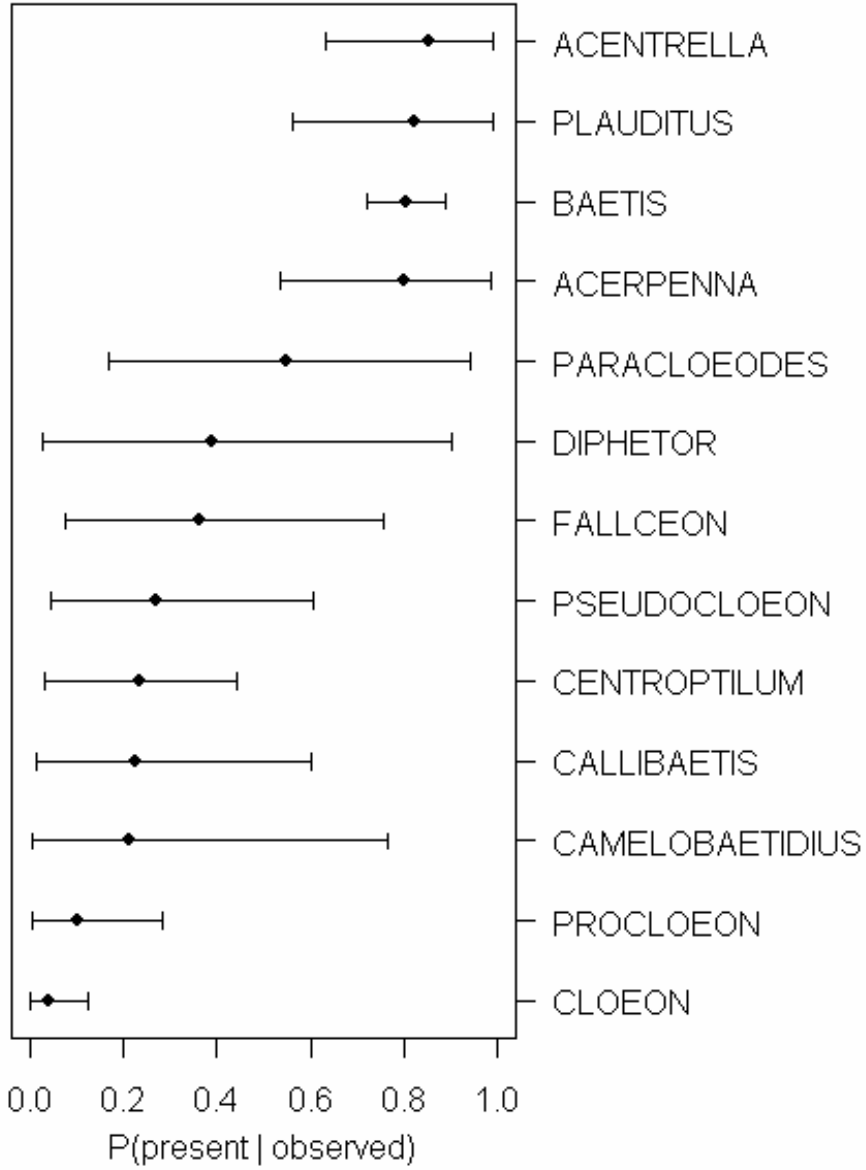
Computing the posterior probability (2)



Posterior probabilities: Baetidae

Acentrella, *Plauditus*, *Baetis*, and *Acerpenna* are consistently identified.

Centroptilum, *Procloeon*, and *Cloeon* are not.

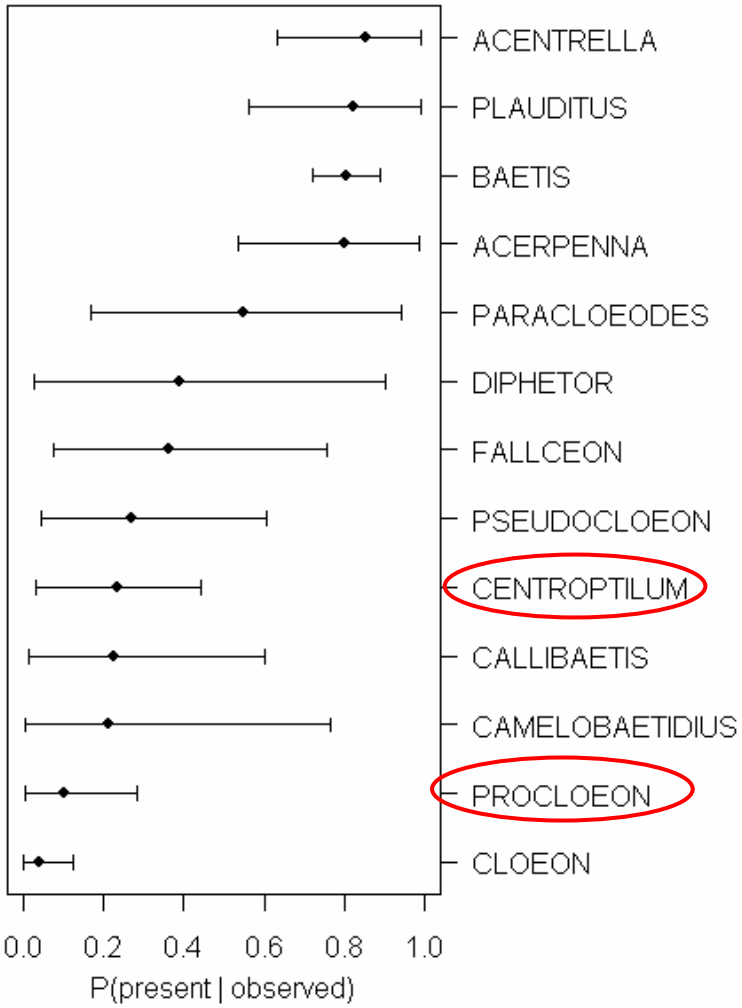


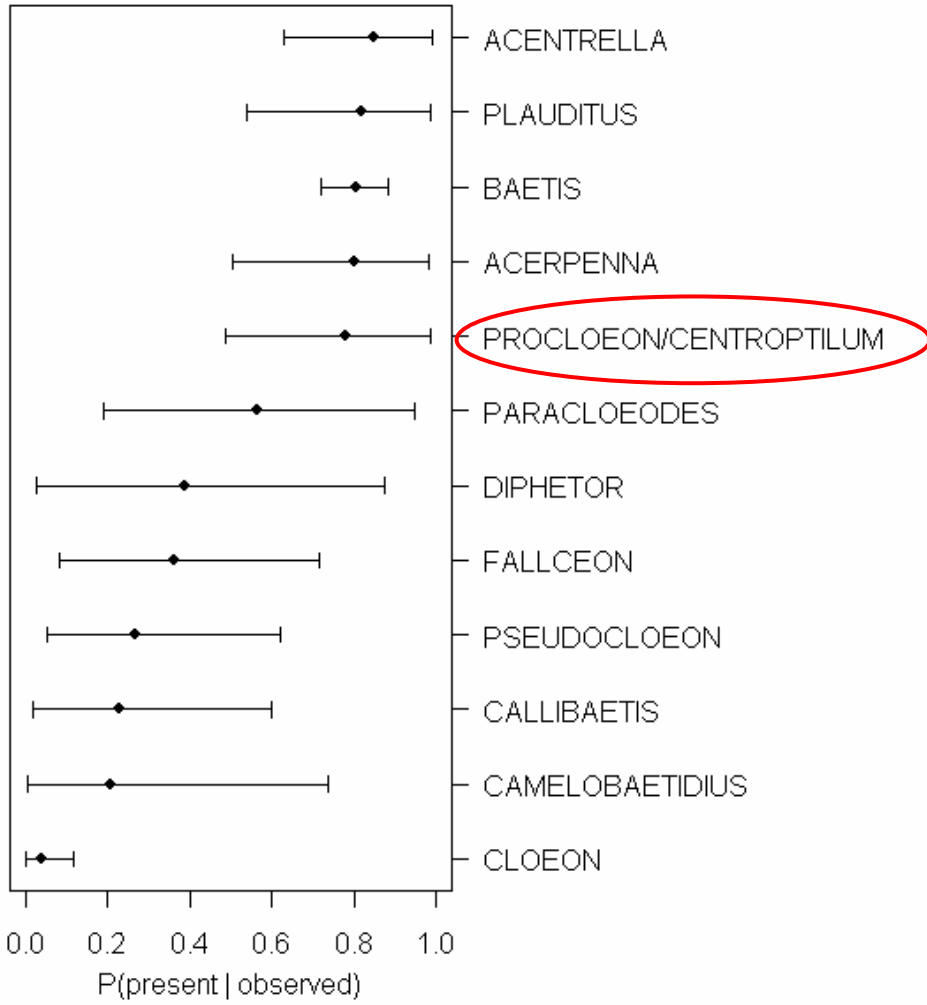
Summary Statistics

- Over 600 distinct genera in the NWSA
- 427 total genera analyzed from the QC
- 132 genera identified consistently.
- Identification of 62 genera differed significantly between primary and QC taxonomists.

- Many of the poorly identified genera could be handled by increasing the target taxonomic level.
  - 11 worms/leeches
  - 7 ceratopogonids
  - 5 mites
  - 5 mollusks

Combining genera…





Summary

- QC data and analysis provide useful insights into errors originating from taxonomic identification.

- Many errors can be mitigated by specifying more coarse target taxonomic levels (“lumping”).

Summary

- Identification consistency could only be determined for relatively common genera.
  - ~200 genera were observed in WSA data but not in the QC data.
  - ~230 genera were observed in the QC data so infrequently that the uncertainty in estimated error rates was greater than 50%.

- Identification of “problem” genera can help inform future assessments and training.