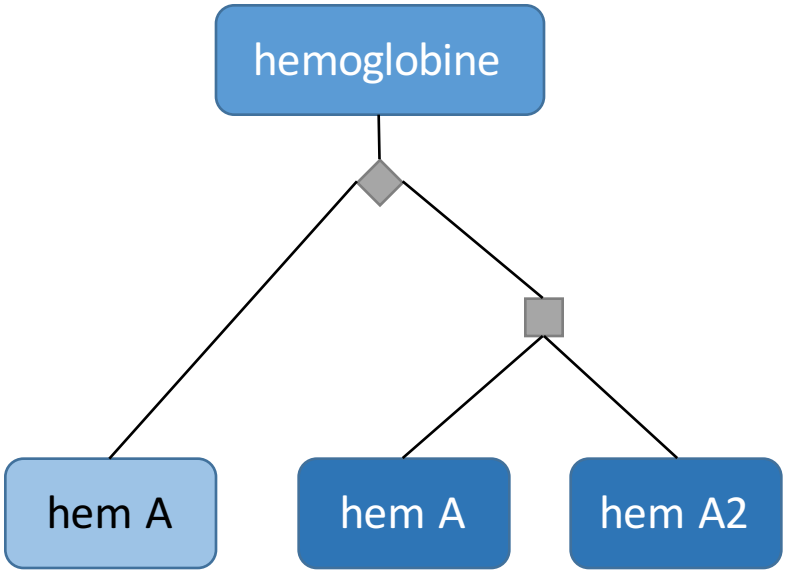
# Studying Paralogs in Linguistics

**31st TBI Winterseminar**

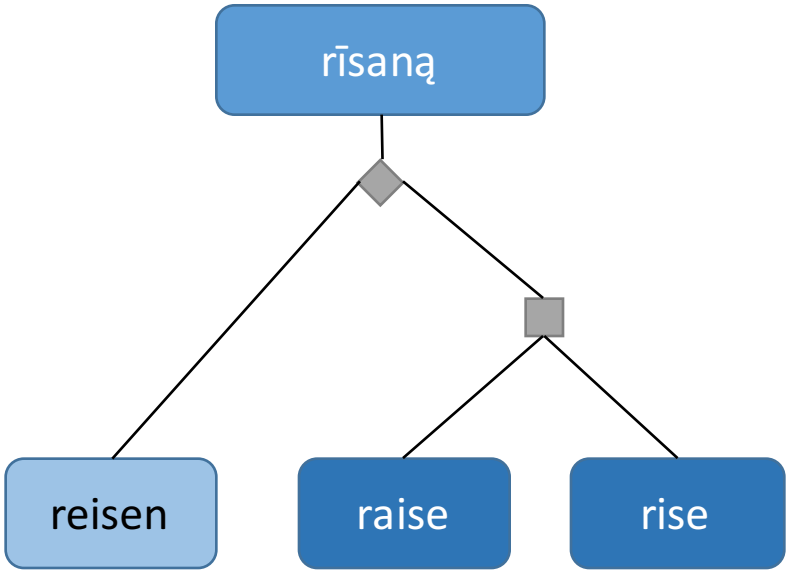
Nancy Retzlaff
Bioinformatics Group Leipzig
Max Planck Institute for Mathematics in the Sciences

# Paralogs in Biology and Linguistics

## Biology




## Linguistics




gene/ word in ancestral species/ language

gene/ word in species/ language 1

gene/ word in species/ language 2

speciation event

duplication event

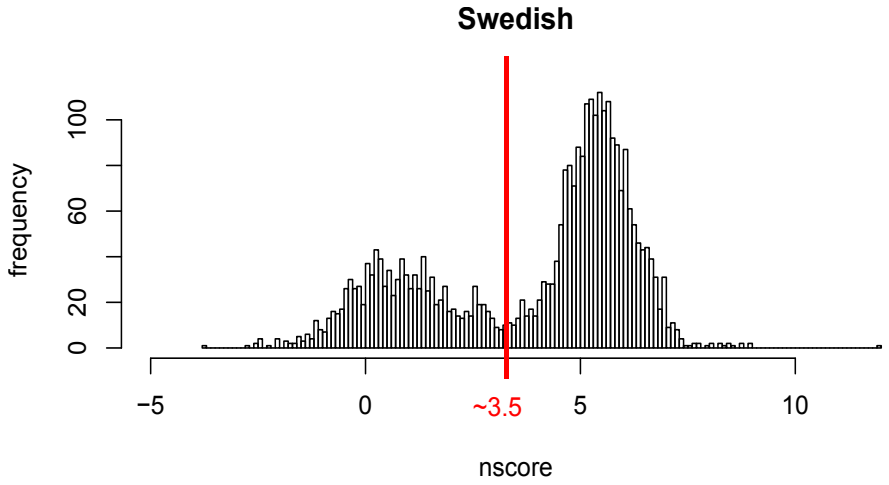
# Intercontinental Dictionary Series

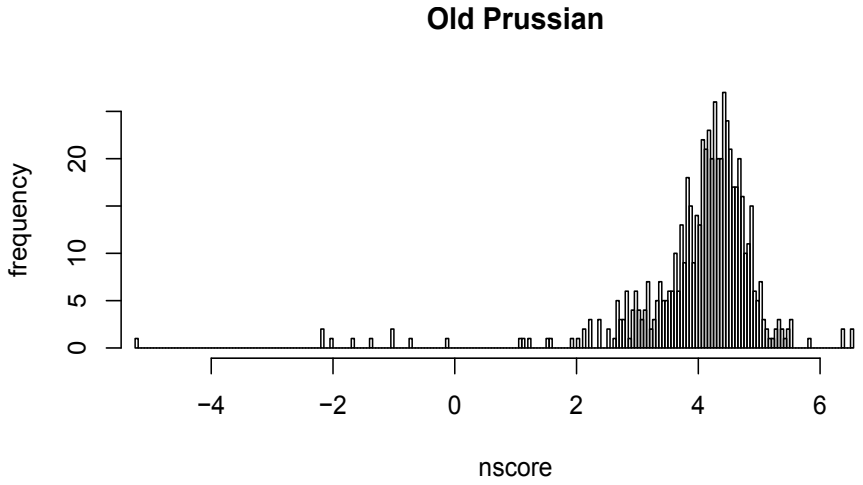
| Meaning entry | Corresponding word |
|---|---|
| 1.100 | world |
| 1.210 | earth |
| 1.210 | land |
| ... | .... |

- Words assigned to meanings, i.e., meanings stay consitent over all languages
- 155 languages in 9 language families

# Finding Paralog Candidates I

- Alignments for each language to itself
- Language-specific cut-off $\theta$



# Finding Paralog Candidates II

- Sampling of
  i. 1000 alignments with words of similar meaning $N_s$
  ii. 1000 alignments with words of different meaning $N_d$

- Count occurences of alignment scores $> \theta$

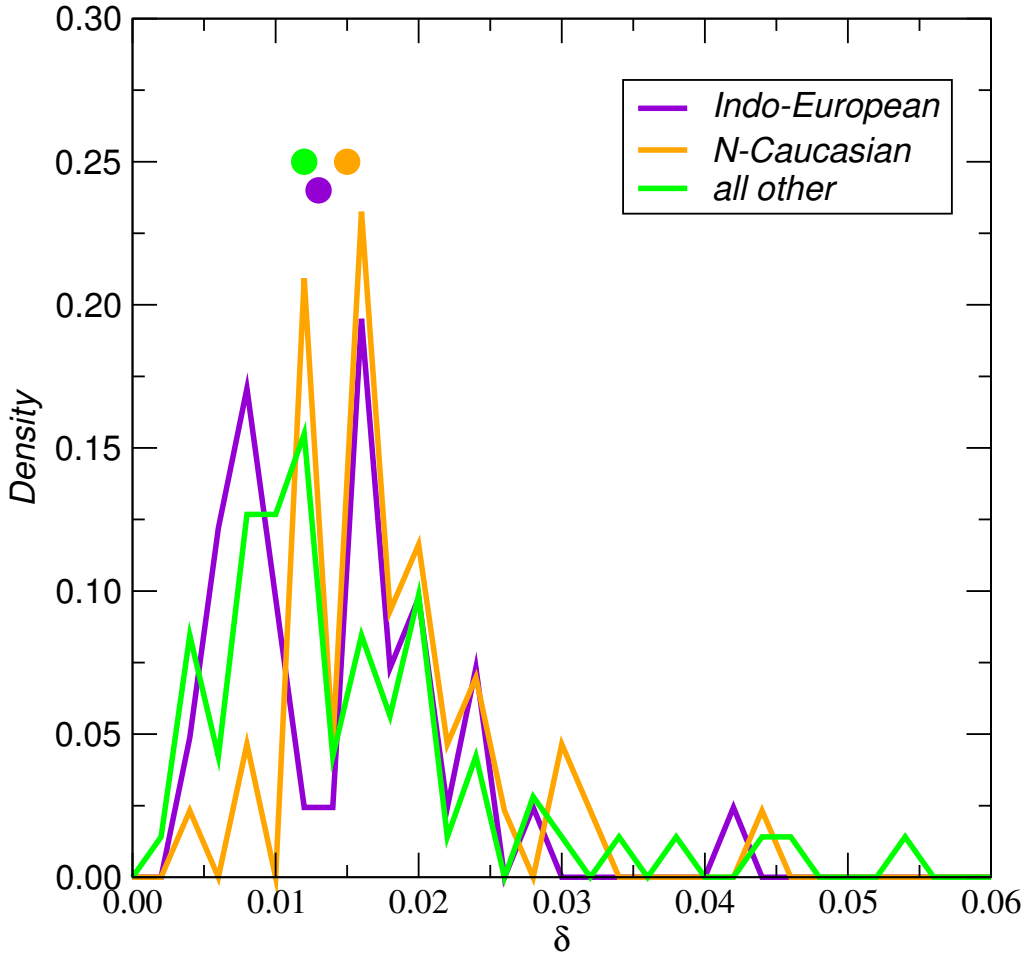
$$n_i = |\{n \in N_i : score(n) > \theta\}|;$$
$$\mathrm{i} \in \{s, d\}$$

# Permutation Test

- Calculate $\delta$ for each language:

$$\delta = \frac{n_s - n_d}{1000}$$

- Divide data into 3 sets of almost the same size



# Discussion

- $\forall$ languages : $\delta \geq 0$
- Signal over all languages/ language families

→Language specific (morphologic) feature responsible?

→Case study of reduplication

# World Atlas of Language Structures

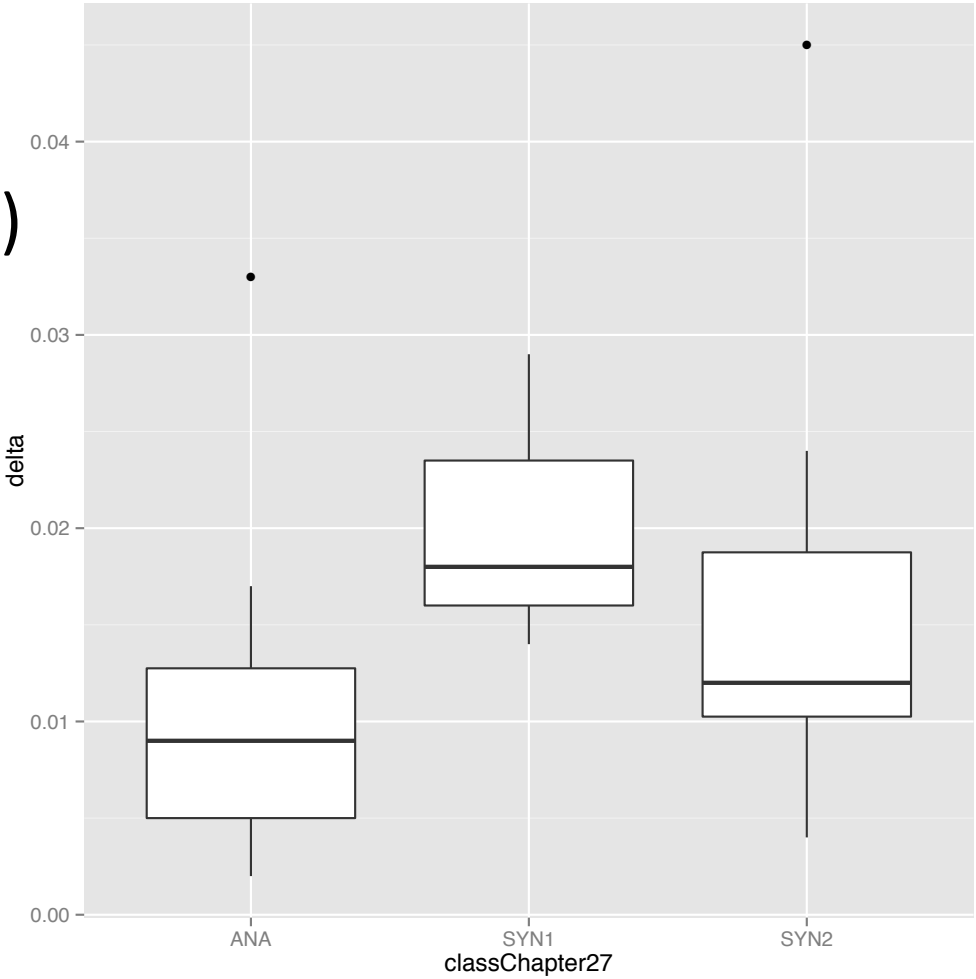
- Database for structural features of languages
- For each linguistic feature, feature values are assigned to the available languages

Reduplication

| | Value | Representation |
|---|---|---|
| ● | Productive full and partial reduplication | 278 |
| ● | Full reduplication only | 35 |
| ○ | No productive reduplication | 55 |
| | **Total:** | 368 |

# Evaluation and Discussion

- Mapping WALS features to linguistic features (analytical and synthetic)
- Not enough data (in total only 33 languages!!!) due to a lack of entries in WALS
- t-Test did not show significant results (p-Value = 0.06401)



# Closing remarks

- Studying language evolution bears challenges
  - Not enough data available
  - Small community doing computational linguistics
  - Working with people in the humanities as a problem in itself (Trudie)

- Jaro says:
  - ,mum,mmd g

# Thanks for your attention!

Questions?