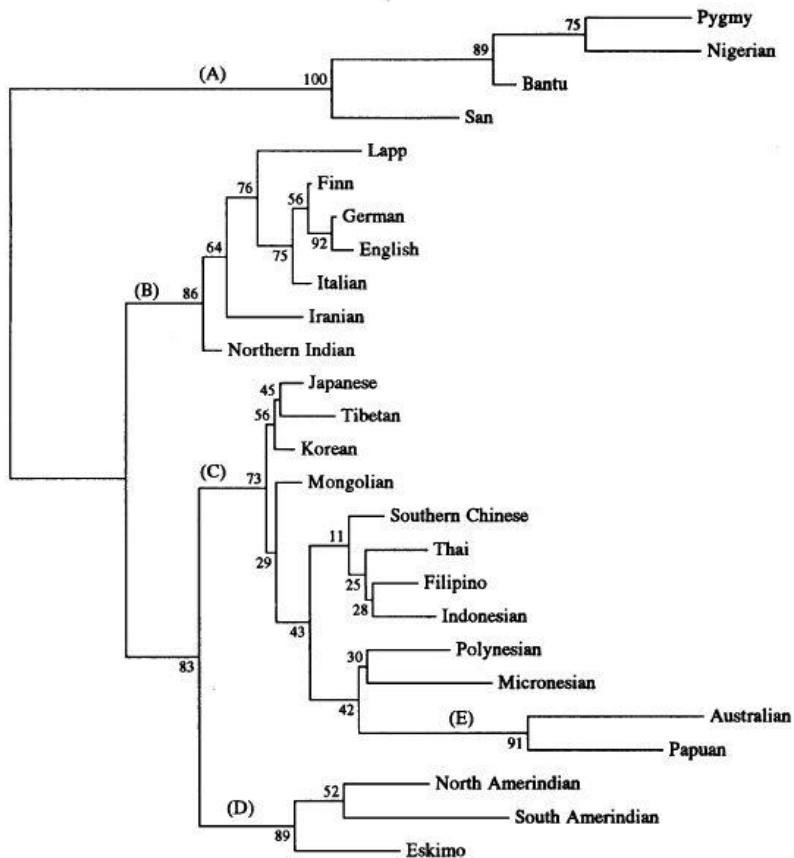


# 1,2,3, language!

---



Andrew Berdahl, Lucas Lacasa  
SFI – CSSS '09

# The idea

**DATASET: numbers 1 to 10 in over 5,000 languages**

***<http://zompist.com/numbers.shtml>***

- **Mapping between alphabet and DNA-like codons**
  - **Alphabet-based (Hamming-like)**
  - **Phonetic-based**
  - **Feature-based**
  
- **Use of evolutionary biology measures to quantify the distance between 2 strings**
  - **Global sequence alignment methods**
  
- **Generate the distance matrix between languages and its associated phylogenetic tree**
  
- **Compare with state of the art results**
  - **Are we able to deduce the language phylogenetic tree from “so few data”?**
  - **Importance of numbers within a language: fundamental quantity?**
  - **Relation to culture-based concepts (trading, etc)**
  
- **Relate with invasion-like spreading of culture**

# Alphabet mapping: criteria

- Each letter maps into a 3-nucleotide string from {A,T,C,G}
- Phonetic and feature-based properties are encoded in the mapping
- We finally have a new alphabet: each of the 26 letters is a 3-nucleotide string
- We concatenate the numbers in a single string
- We make global sequence alignment

THE INTERNATIONAL PHONETIC ALPHABET (2005)

CONSONANTS (PULMONIC)

	Bilabial	Labio-dental	Dental	Alveolar	Post-alveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Epi-glottal	Glottal
Nasal	m	ɱ		n	ɳ	ɲ	ɲ	ŋ	ɴ			
Plosive	p b	ɸ β	t d	ʈ ɖ	ʈ̪ ɖ̪	ʈ̡ ɖ̡	c ɟ	k ɡ	q ɢ		ʔ	ʔ̚
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ	ʕ	h ɦ
Approximant		ʋ	ɹ	ɻ		ɻ	j	ɰ				
Trill			r						R			
Tap, Flap		ɸ	ɾ	ɽ								
Lateral fricative				ɬ ɮ	ɬ̪ ɮ̪	ɬ̡ ɮ̡						
Lateral approximant				l	ɭ		ʎ	ʟ				
Lateral flap				ɭ								

Where symbols appear in pairs, the one to the right represents a modally voiced consonant, except for murmured ʕ. Shaded areas denote articulations judged to be impossible. Light grey letters are unorthodox extensions of the IPA.

## The mapping recipe

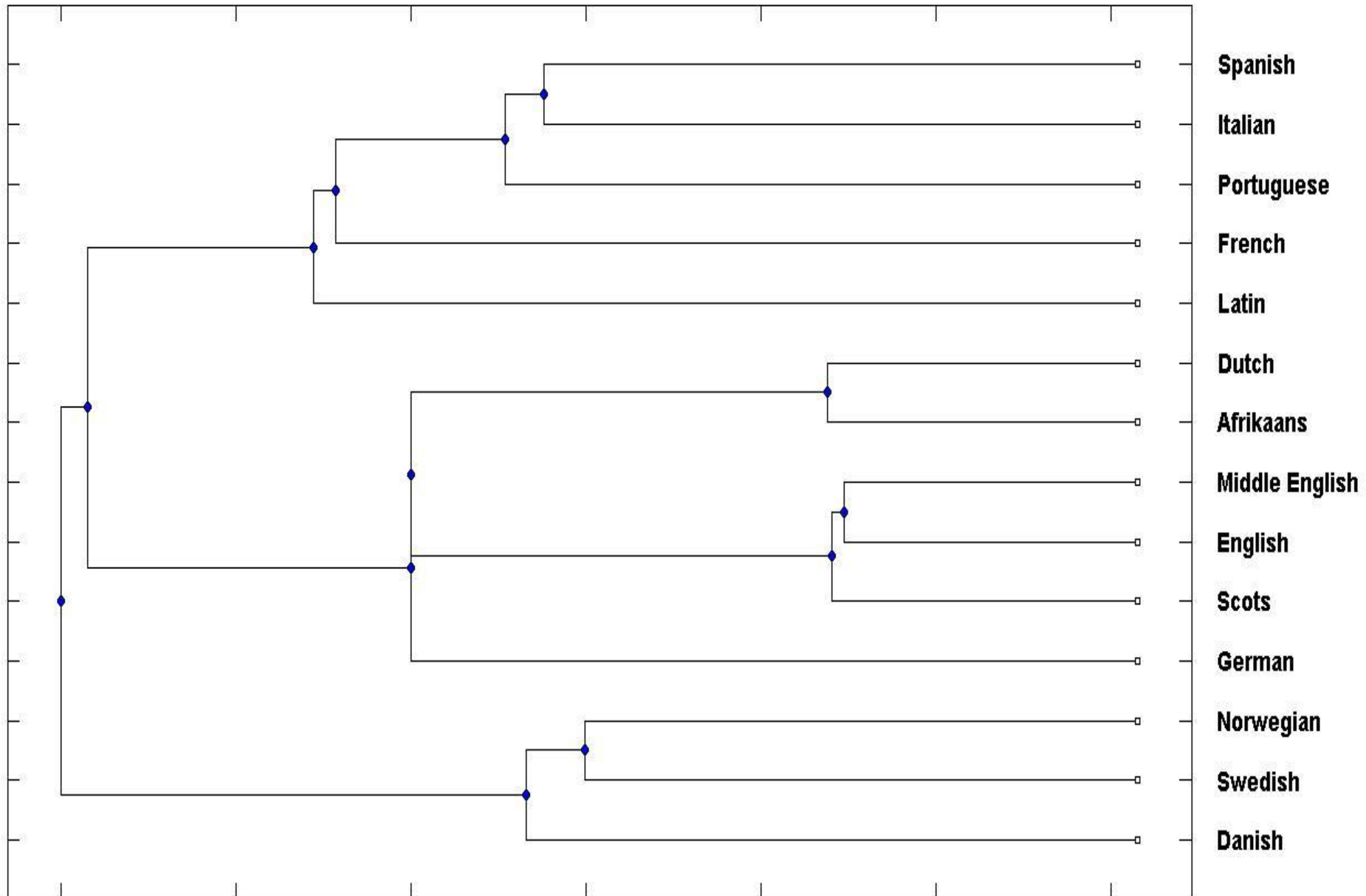
A	AAA
E	ACA
I	AGA
O	ATA
U	ATC
Y	AGG

B	CGT
P	CGA
V	CGG
F	CAG
W	AGG

C	GTA
D	GAA
T	GTT
Z	GTG

S	ATG
X	TTG
H	TTT
L	CCC
M	CTA
N	CTC
R	AGC

# Results (I): subset of 'familiar' languages: ***WORKS!***

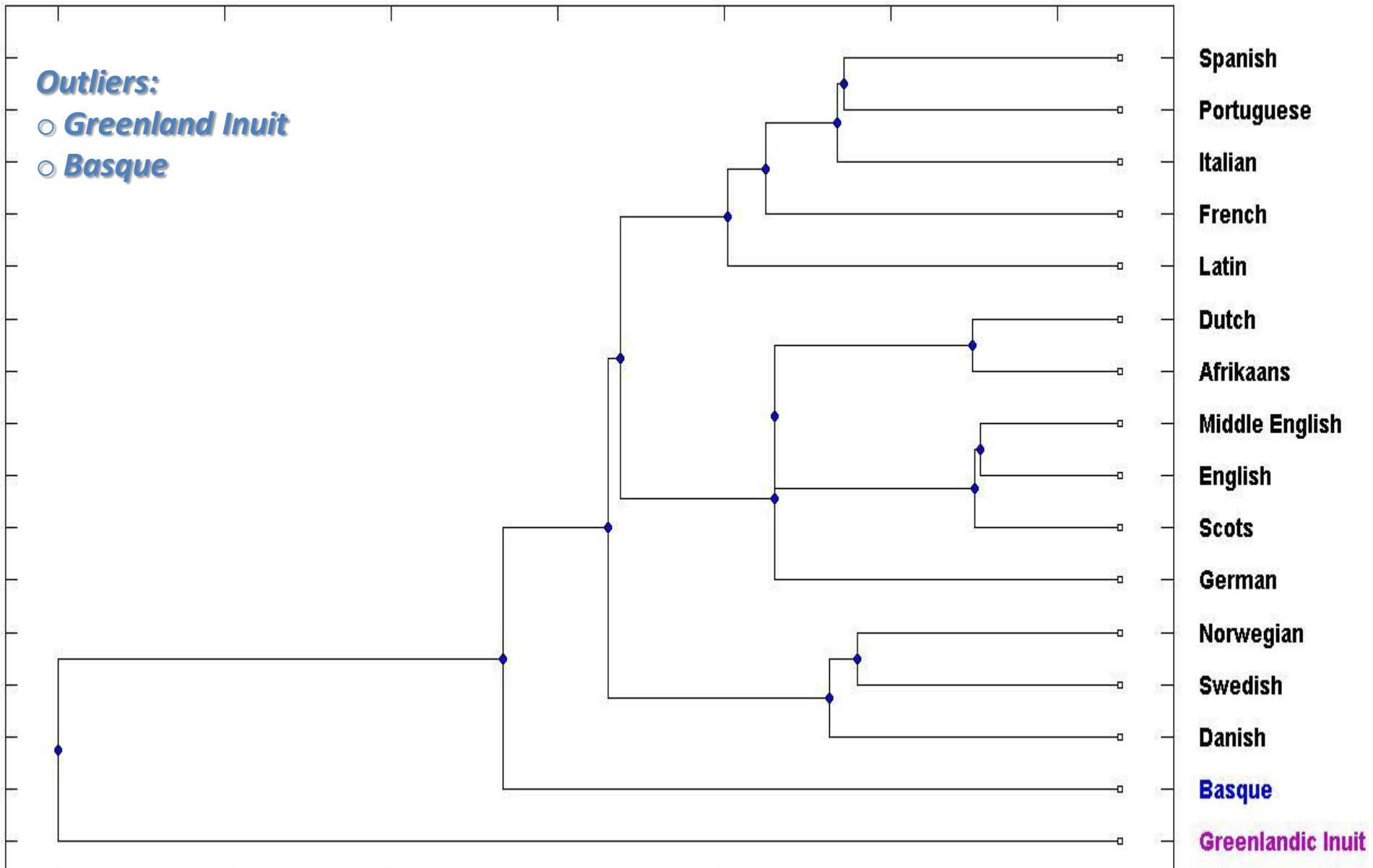


## Results (II): detecting the outliers: *WORKS!*

**Outliers:**

○ *Greenland Inuit*

○ *Basque*



# Conclusions and future directions

- ✓ Even with such small amount of data (10 numbers!!), results are promising
- ✓ Refine mapping criteria to take into account different linguistic properties
- ✓ Extend to different alphabets

***ACKNOWLEDGMENTS:*** D.E. Smith (SFI)