A new hybrid hypothesis for the origin and spread of the Indo-European languages

Russell Gray, Max Planck Institute for the Science of Human History, Jena

Theories of Indo-European Origin

The origin of Indo-European languages

"the most intensively studied, yet still most recalcitrant, problem of historical linguistics"

Diamond and Bellwood, *Science*, 2003

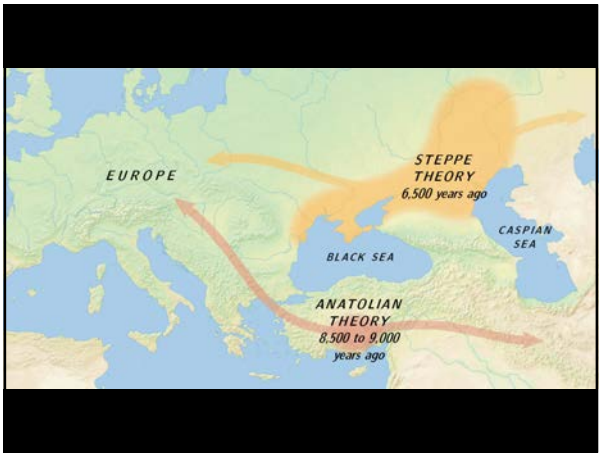
Talk structure

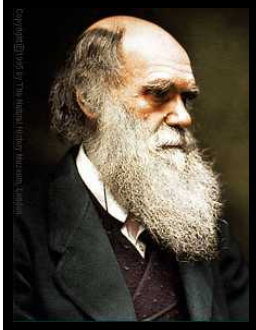
1. The challenge(s)
2. Bayesian phylolinguistics made easy
3. The Indo-European debate goes Bayesian
4. Archaeogenetics to the rescue?
5. A new hypothesis for the origin of PIE

"the quest for the origins of the Indo-Europeans has all the fascination of an electric light in the open air on a summer night: it tends to attract every species of scholar or would-be savant who can take pen to hand."

Mallory, 1989.

So why do I care?

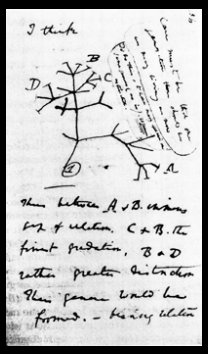




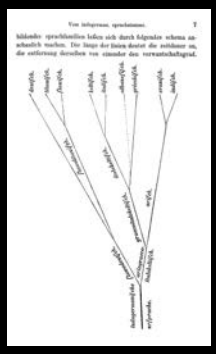
"The formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are curiously parallel. … We find in distinct languages striking homologies due to community of descent, and analogies due to a similar process of formation"

The Descent of Man, 1871

Tree of life



Tree of languages



Retentions vs innovations

Karl Brugmann
1884

Sympliesiomorphies vs synapomorphies

Willi Hennig
1950/1966

Glottoclock

Morris Swadesh
1952

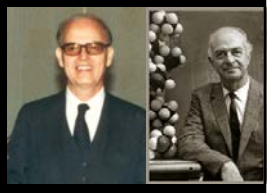
t = log c/2 log r

c = % shared cognates

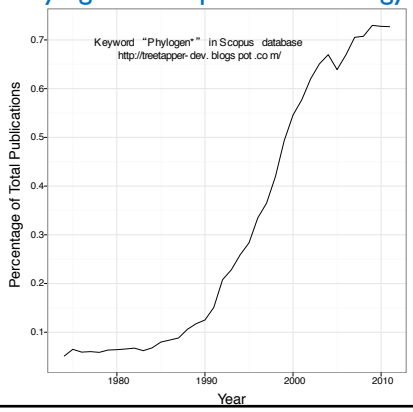
r ≈ 81% (200 word list)

Molecular clock

Zuckerkandl & Pauling 1962

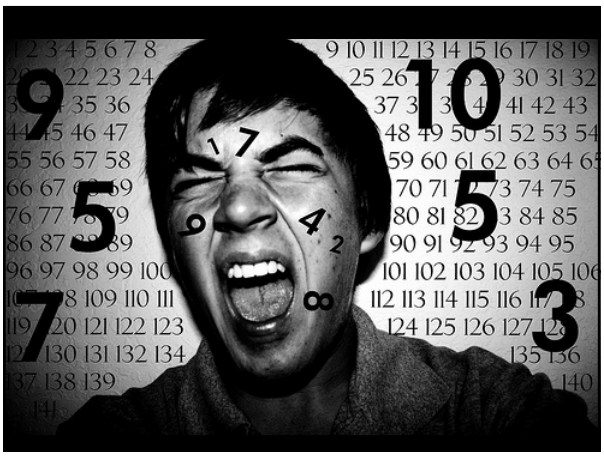


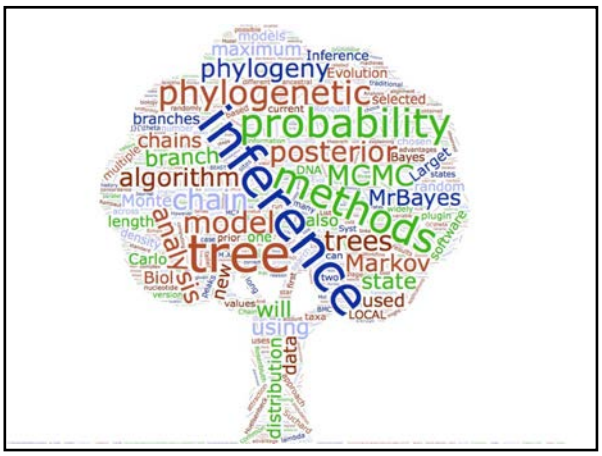
Phylogenetic explosion in biology



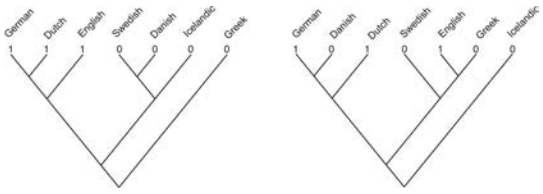
Talk structure

1. The challenge(s)
2. Bayesian phylolinguistics made easy
3. The Indo-European debate goes Bayesian
4. Archaeogenetics to the rescue?
5. A new hypothesis for the origin of PIE





Which tree is more likely



Why use computers?

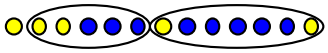
Languages	# rooted trees
3	3
4	15
5	105
6	945
7	10395
8	135135
9	2027025
10	34459425
20	8.2×10^{21}
50	2.7×10^{76}
100	3.4×10^{184}

(2n-5)!!

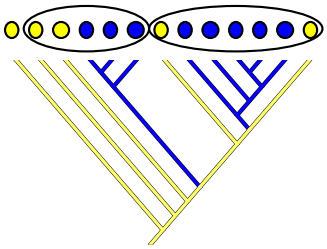
Modern Bayesian Phylogenetic Inference

1. Data
2. Model
3. Priors
4. Tree search

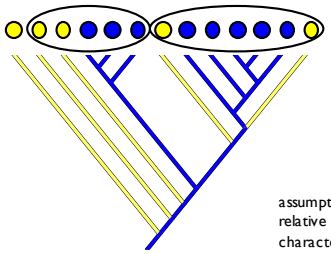
What is the ancestral state?



Depends on the tree…

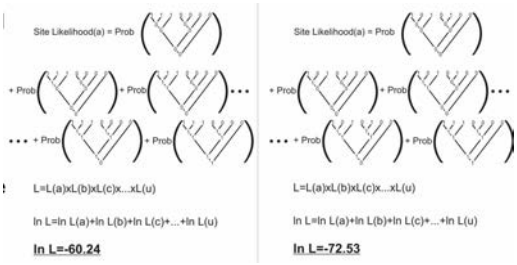


And the model…



assumptions about the relative probabilities of character state changes matter

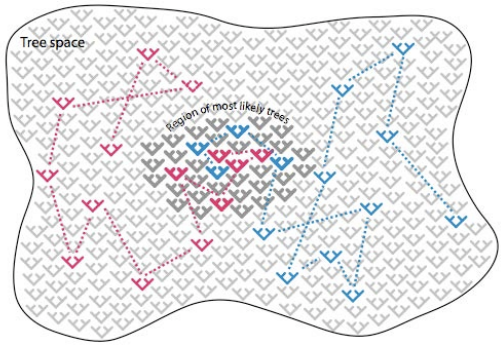
Likelihood calculation 2



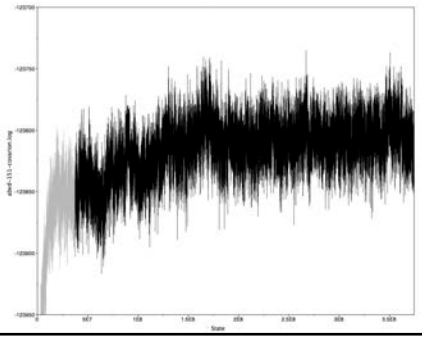
Three models of lexical evolution

1. Equal probability of cognate gains and losses
2. Dollo (gains can only occur once, assumes stochastic clock))
3. Covarion

MCMC search (Markov chain Monte Carlo)

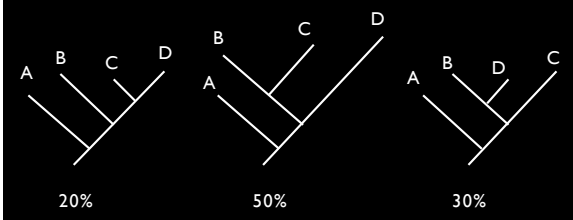


Convergence?



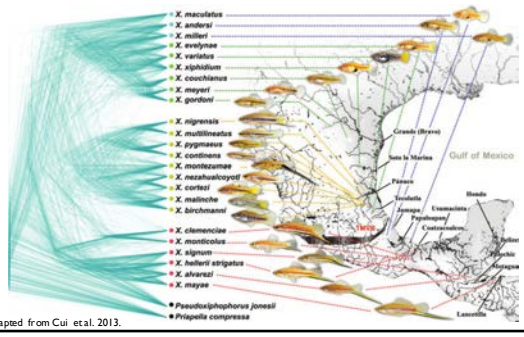
Bayesian MCMC inference

posterior probability of the trees given the priors, the data and the model



No one tree to rule them all
Posterior sample of trees reveals uncertainty

Densitree visualisation



Adapted from Cui et al. 2013.

What we are *NOT* doing…

1. Counting cognates to get % shared cognates
2. Pairwise comparisons of languages
3. Assuming constant rates

What is to be gained?

- Level playing field – no cherry picking, explicit optimality criterion to evaluate subgrouping hypotheses
- Quantify uncertainty
- Estimate dates

Talk structure

1. The challenge(s)
2. Bayesian phylolinguistics made easy
3. The Indo-European debate goes Bayesian

Data

- Dyen et al. (1992) - 84 languages, 2449 cognate sets
- Added three extinct languages
- Swadesh list - 200 items of basic vocabulary
  - numerals, kinship terms, terms for body parts, and basic verbs
  - Relatively resistant to borrowing
- Recognized borrowing removed - e.g. English "*mountain*" borrowed from French "*montagne*" was not coded as cognate.

Binary coding

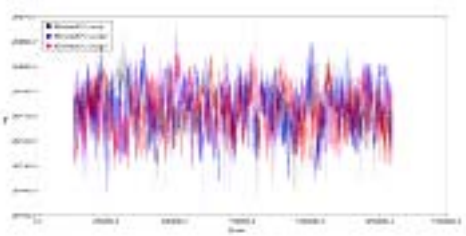
English	And[1]	Big[1]	Fire[1]	Meat[1]	Rub[1]	Water[1]
German	Und[1]	Gross[2]	Feuer[1]	Fleisch[2]	Reiben[1]	Wasser[1]
Dutch	En[1]	Groot[2]	Vuur[1]	Vleesch[2]	Wrijven[1]	Water[1]
Swedish	Och[2]	Stor[3]	Eld[2]	Kott[3]	Gnida[2]	Vatten[1]
Icelandic	Og[2]	Stor[3]	Eldr[2]	Hold[3]	Nua[3]	Vatn[1]
Danish	Og[2]	Stor[3]	Ild[2]	Kod[4]	Gnide[2]	Vand[1]
Greek	Ke[3]	Meghalos[4]	Fotia[3]	Kreas[5]	Trivo[4]	Nero[2]

a

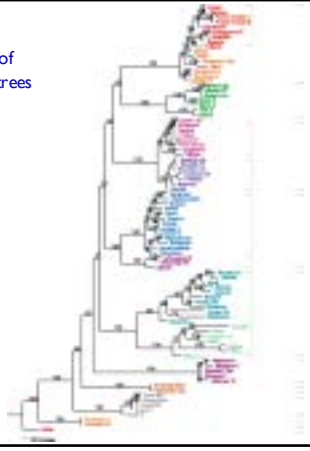
	A	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u
English	1	0	0	1	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	1	0
German	1	0	0	0	1	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0
Dutch	1	0	0	0	1	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0
Swedish	0	1	0	0	0	1	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0
Danish	0	1	0	0	0	1	0	0	1	0	0	0	0	1	0	0	1	0	0	1	0
Icelandic	0	1	0	0	0	1	0	0	1	0	0	0	1	0	0	0	0	1	0	1	0

Bayesian MCMC tree-estimation

- MrBayes (v.2 & v.3) - Huelserbeck & Ronquist
- 10 independent runs with 4 chains
- 1,300,000 generations
- First 300,000 discarded as "burnin"
- Sample every 10,000

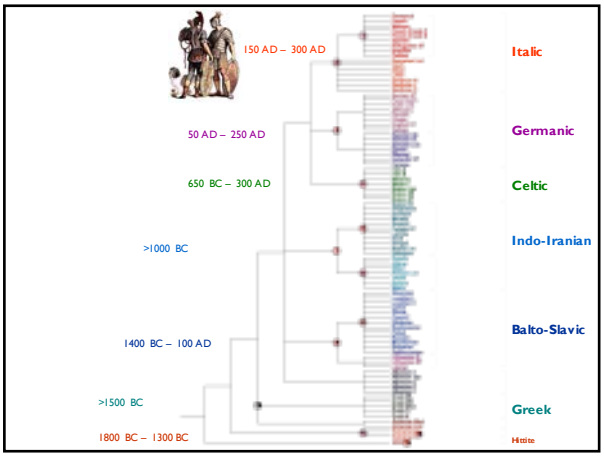


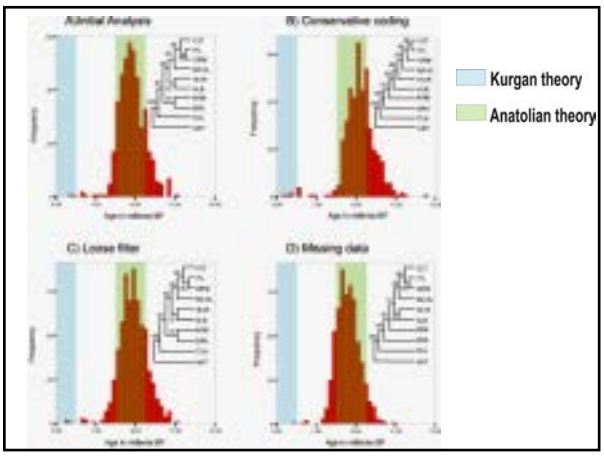
Majority rule consensus tree of 10,000 MCMC trees



"linguists don't do dates"

April & Robert McMahon (2006)





Responses to Gray & Atkinson 2003…

"This kind of study is exactly what linguistics needs."

"…simply reconfirms the unreliability of any glottochronological model, no matter what improvements are made."

"Utter bollocks"

"Junk science"

"X#@$?&! (unrepeatable)"

Criticisms of our method

1. Wrong answer
   - linguistic paleontology rejects this time date – Larry Trask, Sussex Univ.
2. Reliance on lexical data
   - *"Lexical data is the least reliable type of data"* - Don Ringe, Univ. of Pennsylvania
3. Model misspecification
   - *"used techniques that are not appropriate for their data."* - Tandy Warnow, Univ. of Texas
4. Cognate sets not independent

The argument from the wheel

English	*wheel*	Hom. Greek	*kuklos*
OHGerman	*wel*	Sanskrit	*cakra*
Olcelandic	*hjol*		

Proto-Indo-European word **kwekwlo-* "wheel"

"How could a language that was last spoken around 10,000 years ago have words for things that were not invented until 4000 years later?"

Larry Trask

The argument from the wheel

Other explanations…

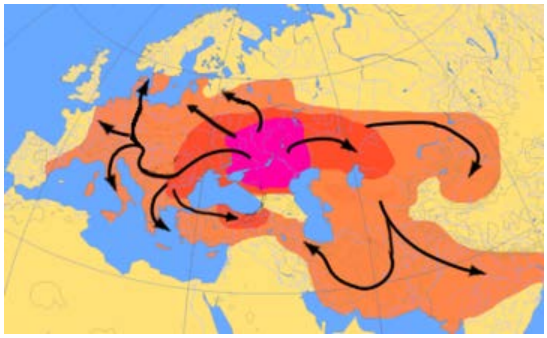
- Borrowing
  - new technology
- Semantic shift

*kwel- (to turn, rotate) ➚ *kola* "wheel" – Old Russian
➔ *kuklos* "wheel" - Greek
➘ *cakra* "wheel" - Sanskrit

Should we believe the lexical data?

- Constraining the tree to Ringe et al. 2002 topology gives similar date estimates

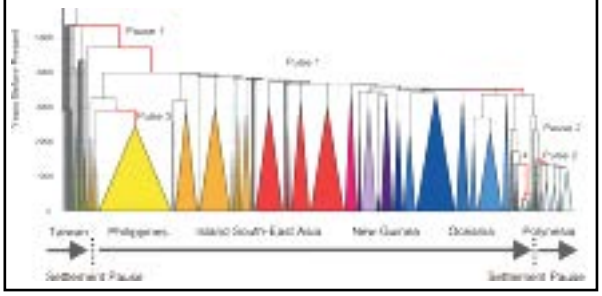
Center of gravity/center of diversity

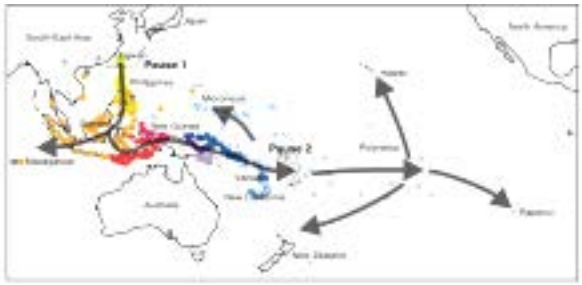


Austronesian expansion

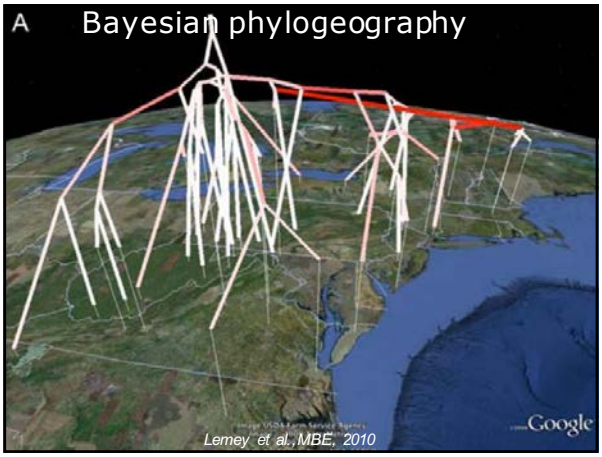
Sequence, timing, pulses and pauses

Gray, Drummond & Greenhill. 2009. *Science*, 323, 479-483.





Gray, Drummond, & Greenhill (2009) Science, 323, 479-483.



The team…

- Alex Alekseyenko
- Quentin Atkinson
- Remco Bouckaert
- Alexei Drummond
- Michael Dunn
- Russell Gray
- Simon Greenhill
- Philippe Lemey
- Marc Suchard

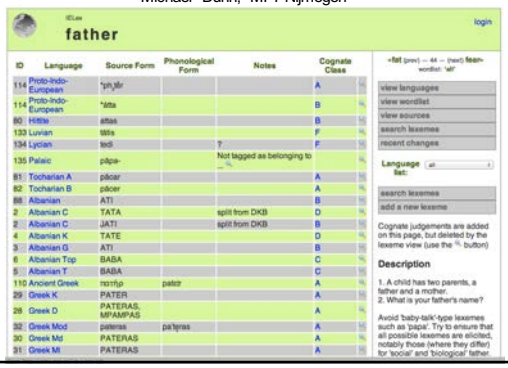
Bayesian phylogeography

1. Data (basic vocab)
2. Location (language ranges)
3. Diffusion model
4. Calibration data to date language divergences
5. Bayesian MCMC inference of phylogeny in BEAST

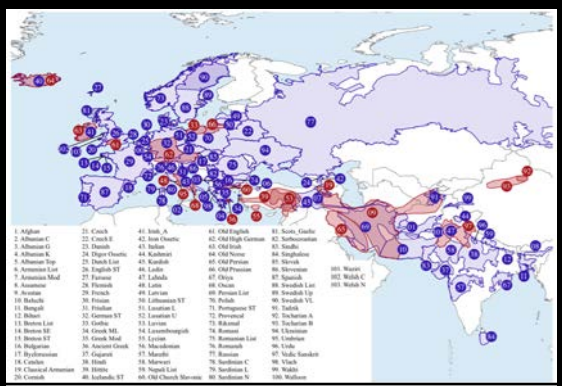
New (improved) data

http://ielex.mpi.nl/

Michael Dunn, MPI Nijmegen



+language location data

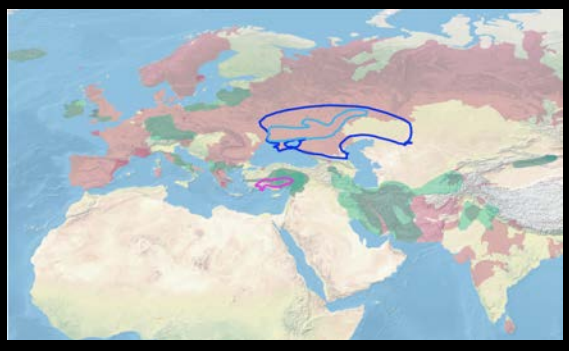


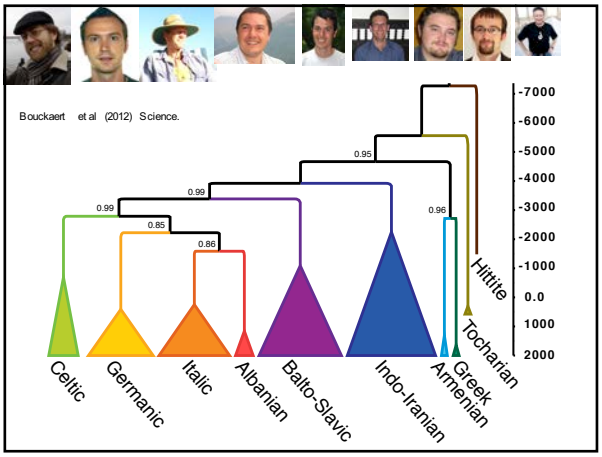
+ model of spatial diffusion

+ Bayesian inference of phylogeny in BEAST

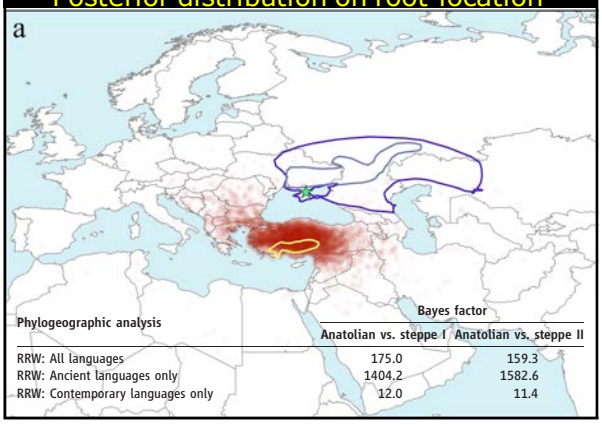
- Substitution models
  - Simple binary reversible model (0 ↔ 1)
  - Binary covarion model slow(0 ↔ 1) fast (0 ↔ 1)
  - Stochastic dollo model (0 → 1 one gain, many loses)
- Clock
  - Uncorrelated lognormal relaxed clock (Drummond et al., 2006)

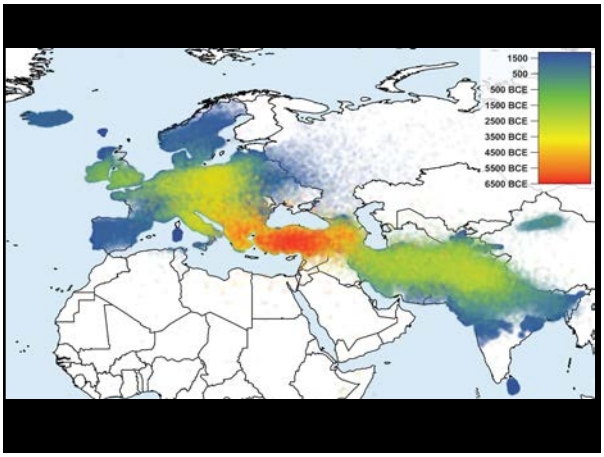
= test origin hypotheses

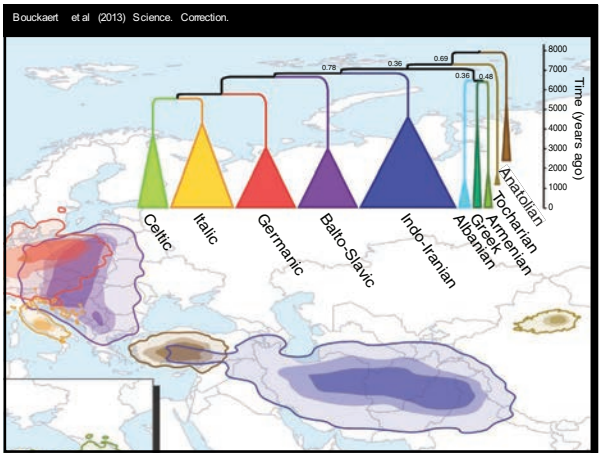




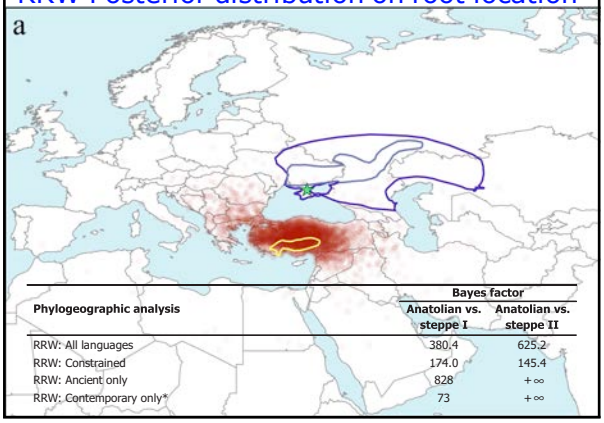
Posterior distribution on root location



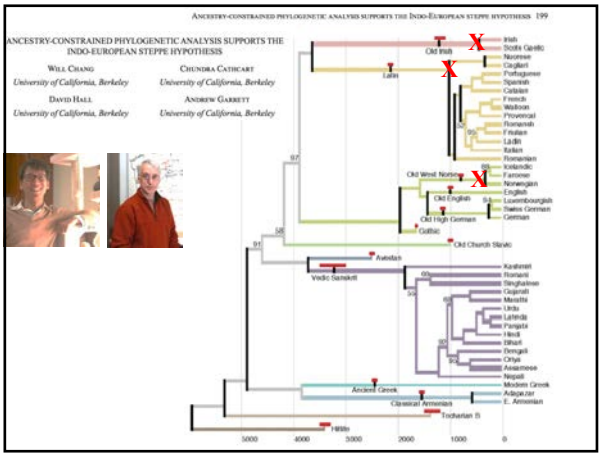




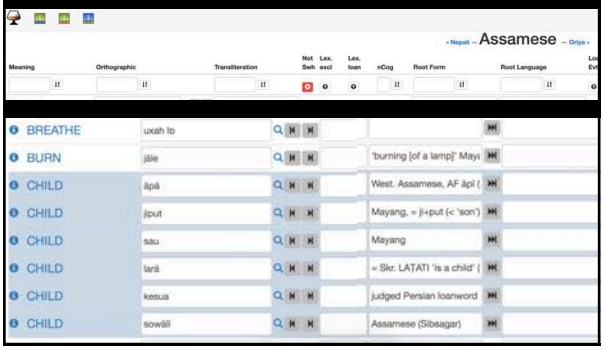
RRW Posterior distribution on root location

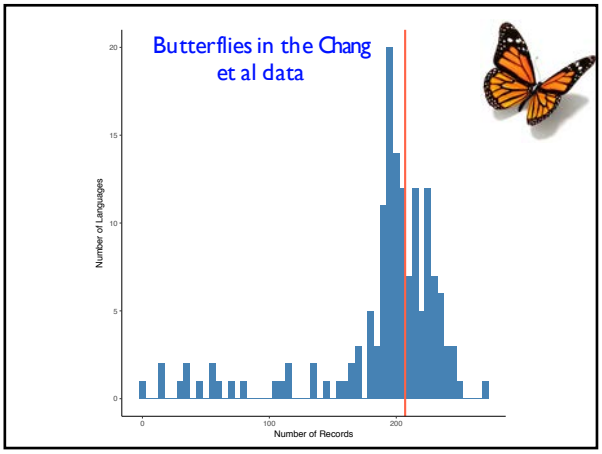


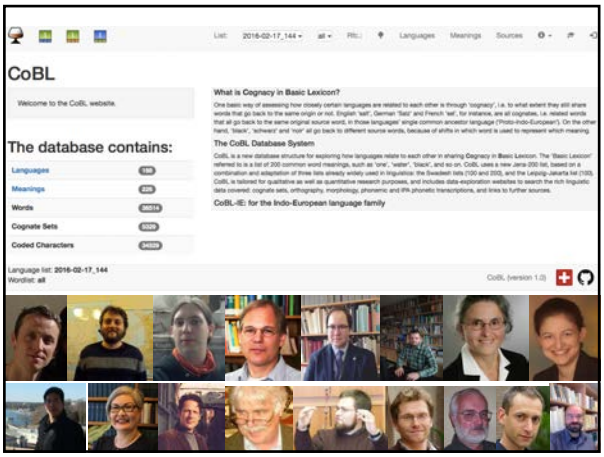
Phylogeographic analysis	Bayes factor: Anatolian vs. steppe I	Bayes factor: Anatolian vs. steppe II
RRW: All languages	380.4	625.2
RRW: Constrained	174.0	145.4
RRW: Ancient only	828	+∞
RRW: Contemporary only*	73	+∞



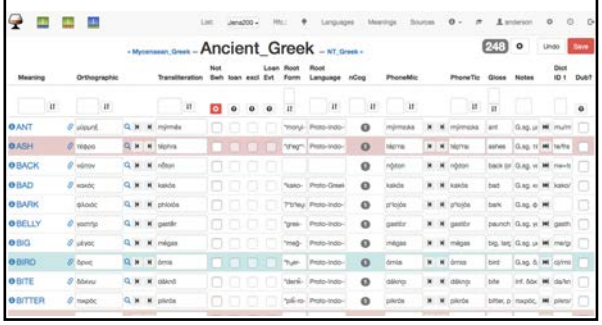
Big problem – inconsistent data coding
(Swadesh policy of most commonly used lexeme not consistently followed)





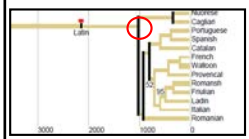


COBL – Ancient Greek

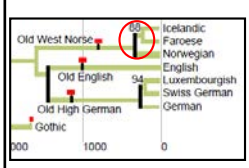


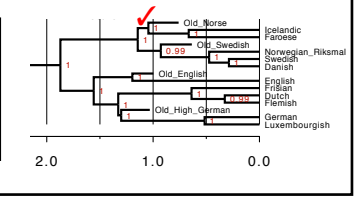
Better data, multi-state, better model that allows but does not enforce direct ancestry, new results….

Chang et al 2014



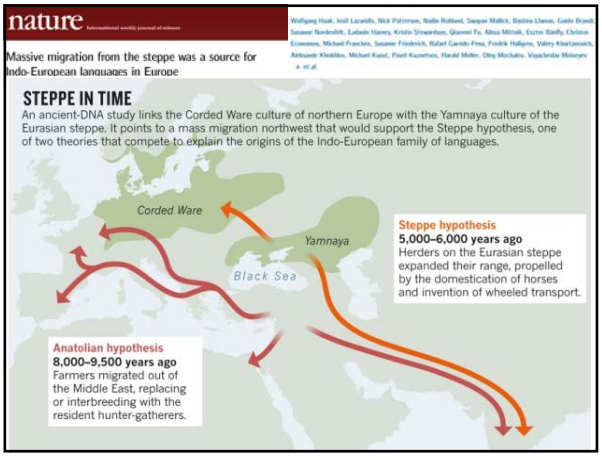
New Jena results

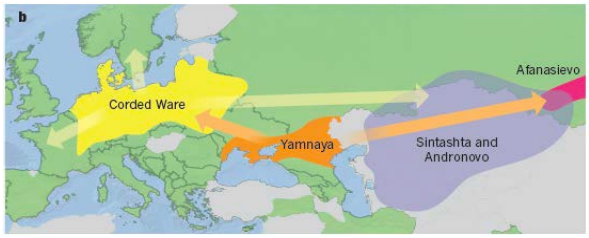


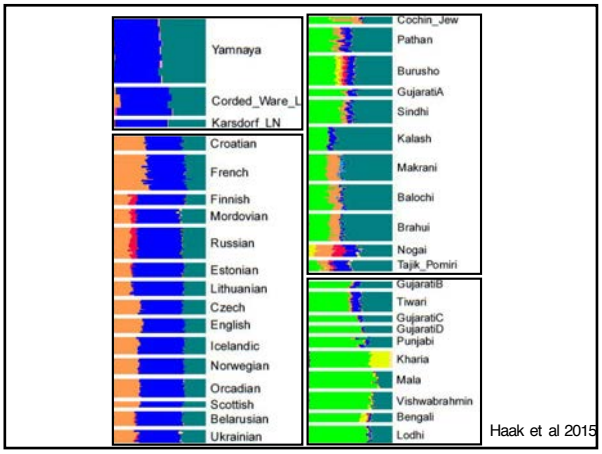


Talk structure

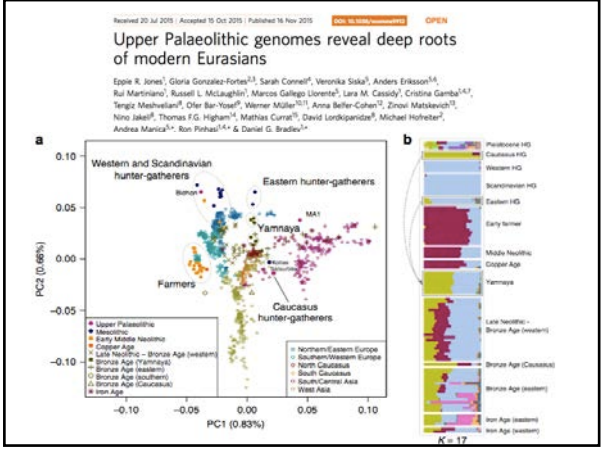
1. The challenge(s)
2. Bayesian phylolinguistics made easy
3. The Indo-European debate goes Bayesian
4. Archaeogenetics to the rescue?







Haak et al 2015



Talk structure

1. The challenge(s)
2. Bayesian phylolinguistics made easy
3. The Indo-European debate goes Bayesian
4. Archaeogenetics to the rescue?
5. A new hypothesis for the origin of PIE

A hybrid model for the origin and spread of Indo-European
Haak, Heggarty, Krause & Gray

A hybrid model for the origin and spread of Indo-European
Haak, Heggarty, Krause & Gray