Language models

Remco R. Bouckaert
remco@cs.auckland.ac.nz

Centre for Computational Evolution
Department of Computer Science, University of Auckland
Max Planck Institute for the Science of Human History

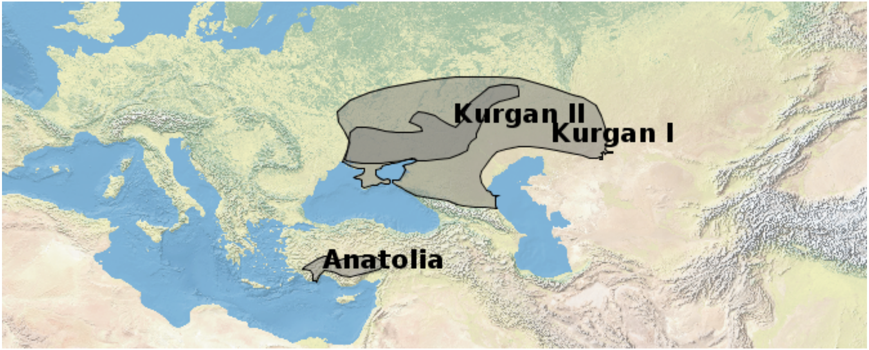
Seili, September 2017

Origin of Indo-European

"the most well-studied and yet still most recalcitrant problem in historical linguistics"

Diamond, Belwood, Science 2003

Two competing theories



Kurgan $<$ 6000BP, Anatolian $>$ 8000BP

Step 1. Building a database of cognates
word list

language	hand	mother	father	...
English	hand	mother	father	...
Dutch	hand	moeder	vader	...
German	hand	mutter	vater	...
French	main	mère	père	...
Spanish	mano	madre	padre	...
Dhudhuroa	?	papa	mama	...

Step 1. Building a database of cognates

word list

language	hand	mother	father	...
English	hand	mother	father	...
Dutch	hand	moeder	vader	...
German	hand	mutter	vater	...
French	main	mère	père	...
Spanish	mano	madre	padre	...
Dhudhuroa	?	papa	mama	...

cognate list

language	hand	mano	mother	papa	father	mama	...
English	1	0	1	0	1	0	...
Dutch	1	0	1	0	1	0	...
German	1	0	1	0	1	0	...
French	0	1	1	0	1	0	...
Spanish	0	1	1	0	1	0	...
Dhudhuroa	?	?	0	1	0	1	...

103 languages (of which 20 ancient), 207 meanings, 6280 cognates, 5362 patterns

Meanings: I all and animal ashes at back bad bark because belly big bird bite black blood blow bone breast breathe burn child cloud cold come count cut day die dig dirty dog drink dry dull dust ear earth eat egg eye fall far fat father fear feather few fight fingernail fire fish five float flow flower fly fog foot four freeze fruit full give good grass green guts hair hand he head hear heart heavy here hit hold horn how hunt husband ice if in kill knee know lake laugh leaf left leg lie live liver long louse man many meat moon mother mountain mouth name narrow near neck new night nose not old one other person play pull push rain red right rightside river road root rope rotten round rub salt sand say scratch sea see seed sew sharp short sing sit skin sky sleep small smell smoke smooth snake snow some spit split squeeze stab stand star stick stone straight suck

CTMC

Simples model: continuous time Markov chain model

$$\begin{array}{c} 0 \\ 1 \end{array} \left(\begin{array}{cc} - & 1 \\ 1 & - \end{array} \right) \times \left(\begin{array}{c} f_0 \\ f_1 \end{array} \right) = \left(\begin{array}{cc} - & f_1 \\ f_0 & - \end{array} \right) = Q$$

f_0, f_1 equilibrium frequency of a 0 or 1 respectively

$$P(x_i = j | x_{\pi_i} = j, t, \theta) = e^{tQ}_{j,k}$$

Covarion

0 in alignment is either slow 0 or fast 0
1 in alignment is either slow 1 or fast 1

$$\begin{array}{l} \text{fast}\left\{\begin{array}{l}0:\\1:\end{array}\right. \\ \text{slow}\left\{\begin{array}{l}0:\\1:\end{array}\right.\end{array} \begin{pmatrix} - & 1 & s & 0\\ 1 & - & 0 & s\\ s & 0 & - & \alpha\\ 0 & s & \alpha & -\end{pmatrix} \times \begin{pmatrix} f_0\\ f_1\\ f_0\\ f_1\end{pmatrix} = \begin{pmatrix} - & f_1 & sf_0 & 0\\ f_0 & - & 0 & sf_1\\ sf_0 & 0 & - & \alpha f_1\\ 0 & sf_1 & \alpha f_0 & -\end{pmatrix} = Q$$

- f_0, f_1 equilibrium frequency of a 0 or 1 respectively
- s switch rate between fast and slow
- α slow mutation rate

Stochastic Dollo

Dollo principle: every trait appears only once, but can die out many times
New features appear according to a Poisson process with rate r

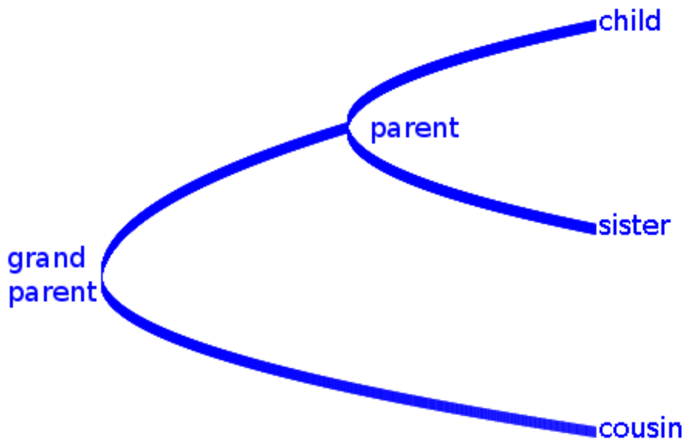
$$\begin{array}{rc} & \begin{array}{cc} 0 & 1 \end{array} \\ \begin{array}{r} 0: \\ 1: \end{array} & \left[\begin{array}{cc} - & 0 \\ \mu & - \end{array}\right] \end{array} = Q$$

μ rate of extinction

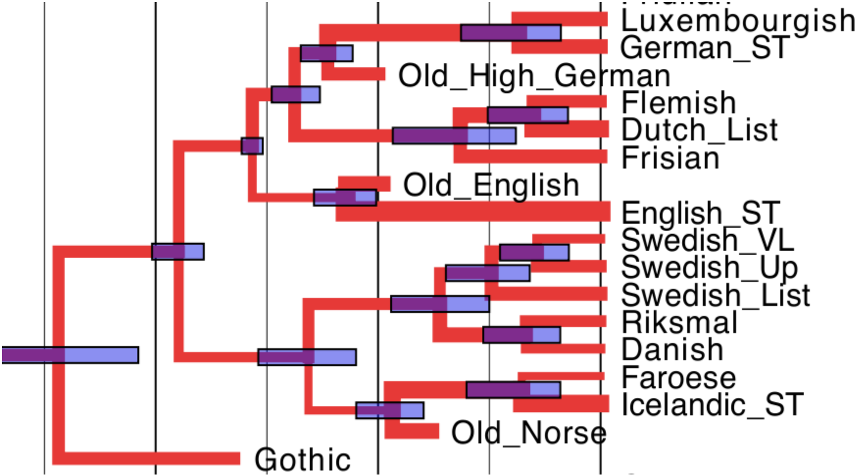
MK model

- for structural data (linguistics), morphological data (biology)
- generalisation of Jukes Cantor: for a trait with k traits, the $k \times k$ rate matrix has rates all equal
- MKv model: MK with ascertainment correction for the trait being present at least once

Step 3. Building family trees of languages



Step 3. Building family trees of languages



Step 3. Building family trees of languages

Phylogenetic model defined by the following components:

Substitution model: binary Covarion

- beats CTMC, Stochastic Dollo

Branch rate model: Relaxed clock

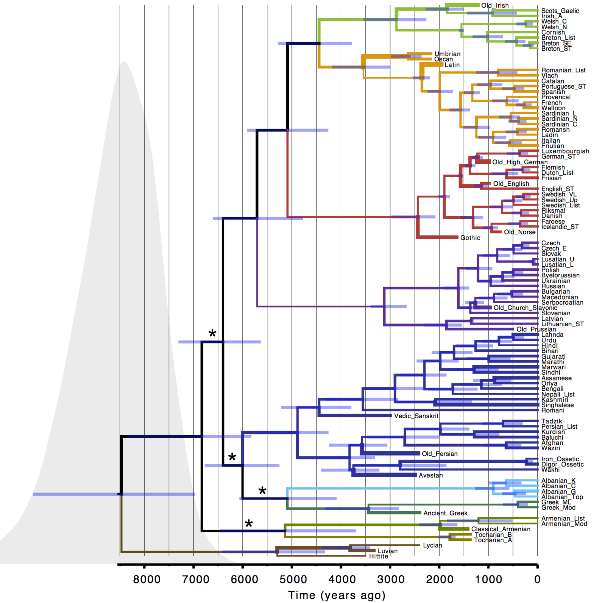
- beats strict clock

Tree prior: Bayesian Skyline plot

- non-parametric tree prior

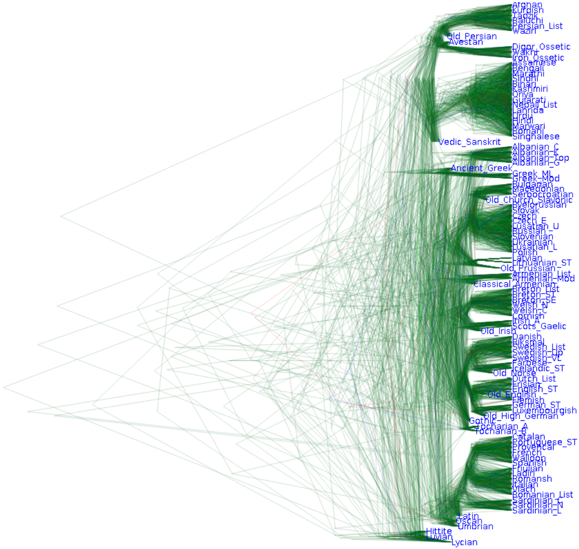
Ascertainment correction for cognates

Step 4. Calibrating the age of the tree



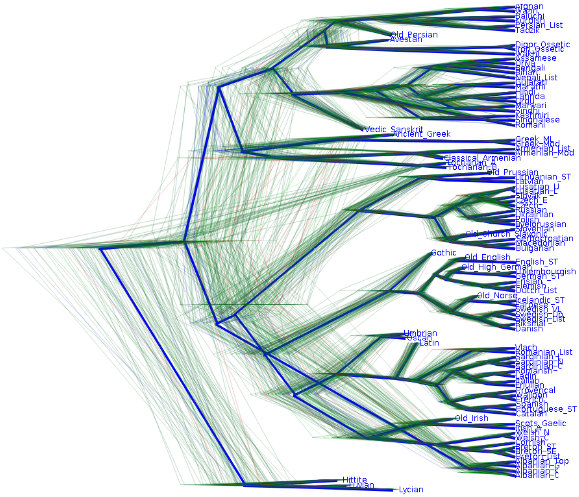
Step 4. Calibrating the age of the tree

Prior

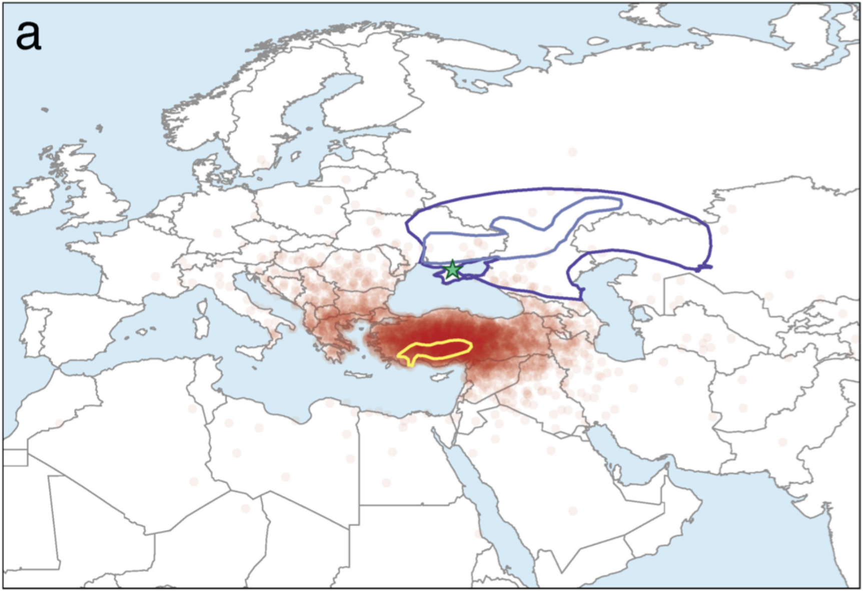


Step 4. Calibrating the age of the tree

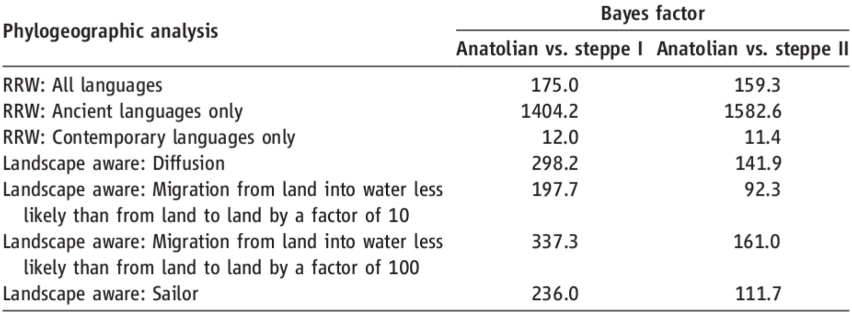
Posterior



Step 6. Testing between the two homeland hypotheses



Step 6. Testing between the two homeland hypotheses



Phylogeographic analysis	Bayes factor	
	Anatolian vs. steppe I	Anatolian vs. steppe II
RRW: All languages	175.0	159.3
RRW: Ancient languages only	1404.2	1582.6
RRW: Contemporary languages only	12.0	11.4
Landscape aware: Diffusion	298.2	141.9
Landscape aware: Migration from land into water less likely than from land to land by a factor of 10	197.7	92.3
Landscape aware: Migration from land into water less likely than from land to land by a factor of 100	337.3	161.0
Landscape aware: Sailor	236.0	111.7

How robust are these findings?

Choice of languages

- same results with ancient languages only
- same results with modern languages only

Differentiate between water and land

- same results with landscape aware model
- same results with different parameters

Only cognate information used, not phonology

- same results with tree constrained to the one based on phonology

Improved modelling of cognates (Chang et al. 2015)

word list

language	hand	mother	father	...
English	hand	mother	father	...
Dutch	hand	moeder	vader	...
German	hand	mutter	vater	...
French	main	mère	père	...
Spanish	mano	madre	padre	...
Dhudhuroa	?	papa	mama	...

cognate list

language	ascetainment	hand	mano	mother	papa	father	mama	...
English	0	1	0	1	0	1	0	...
Dutch	0	1	0	1	0	1	0	...
German	0	1	0	1	0	1	0	...
French	0	0	1	1	0	1	0	...
Spanish	0	0	1	1	0	1	0	...
Dhudhuroa	0	?	?	0	1	0	1	...

Improved modelling of cognates (Chang et al. 2015)

word list

language	hand	mother	father	...
English	hand	mother	father	...
Dutch	hand	moeder	vader	...
German	hand	mutter	vater	...
French	main	mère	père	...
Spanish	mano	madre	padre	...
Dhudhuroa	?	papa	mama	...

cognate list

language	ascetainment	hand	mano	mother	papa	father	mama	...
English	0	1	0	1	0	1	0	...
Dutch	0	1	0	1	0	1	0	...
German	0	1	0	1	0	1	0	...
French	0	0	1	1	0	1	0	...
Spanish	0	0	1	1	0	1	0	...
Dhudhuroa	0	?	?	0	1	0	1	...

cognate list

language	ascetainment	hand	mano	ascetainment	mother	papa	ascetainment	father	mama	...
English	0	1	0	0	1	0	0	1	0	...
Dutch	0	1	0	0	1	0	0	1	0	...
German	0	1	0	0	1	0	0	1	0	...
French	0	0	1	0	1	0	0	1	0	...
Spanish	0	0	1	0	1	0	0	1	0	...
Dhudhuroa	?	?	?	0	0	1	0	0	1	...

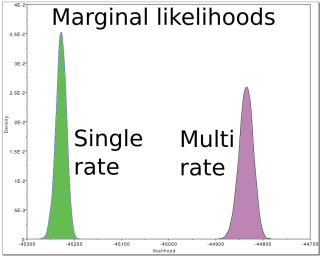
Improved modelling of cognates (Chang et al. 2015)

word list

language	hand	mother	father	...
English	hand	mother	father	...
Dutch	hand	moeder	vader	...
German	hand	mutter	vater	...
French	main	mère	père	...
Spanish	mano	madre	padre	...
Dhudhuroa	?	papa	mama	...

cognate list

language	ascetainment	hand	mano	mother	papa	father	mama	...
English	0	1	0	1	0	1	0	...
Dutch	0	1	0	1	0	1	0	...
German	0	1	0	1	0	1	0	...
French	0	0	1	1	0	1	0	...
Spanish	0	0	1	1	0	1	0	...
Dhudhuroa	0	?	?	0	1	0	1	...



cognate list

language	ascetainment	hand	mano	ascetainment	mother	papa	ascetainment	father	mama	...
English	0	1	0	0	1	0	0	1	0	...
Dutch	0	1	0	0	1	0	0	1	0	...
German	0	1	0	0	1	0	0	1	0	...
French	0	0	1	0	1	0	0	1	0	...
Spanish	0	0	1	0	1	0	0	1	0	...
Dhudhuroa	?	?	?	0	0	1	0	0	1	...

+ 1 relative rate for every meaning class = 207-1 extra parameters

Miscellaneous Chang et al. 2015

- Remove many languages
- Remove many calibrations
- Assume ancestrality of ancient languages
- Topology assumptions e.g. (Anatolian,(Tocharian, rest))
- IElex data fixes

Miscellaneous Chang et al. 2015

- Remove many languages
- Remove many calibrations
- Assume ancestrality of ancient languages
- Topology assumptions e.g. (Anatolian,(Tocharian, rest))
- IElex data fixes

Result:

- Much younger origin, supporting Kurgan hypothesis
- Much less credible dates
  - Latin expansion $<$ 2200 BP (Hispania) – 1900 BP (Dacia) not 763-1286 BP
  - Norse to Faroese and Icelandic $<$ 1100BP not 212-494 BP
  - Gaelic into Scotland $<$ 1500BP not 259-655 BP

Miscellaneous Chang et al. 2015

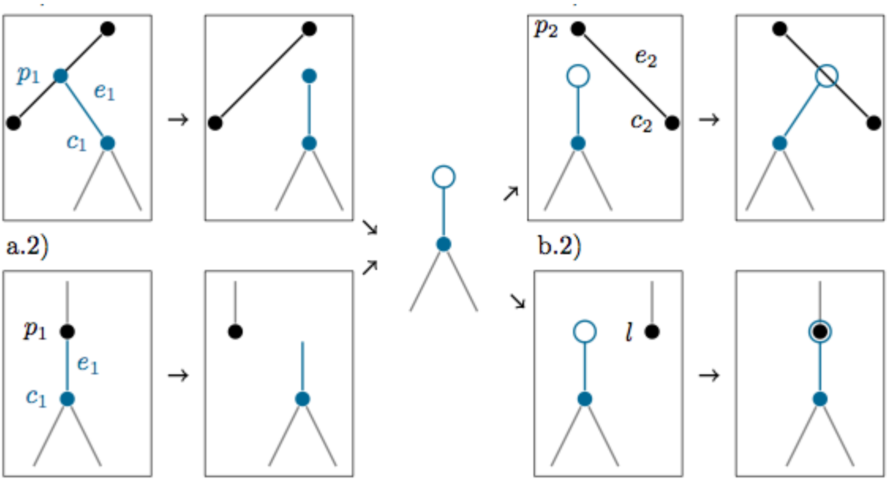
- Remove many languages $\leftarrow$ keep 103 languages
- Remove many calibrations $\leftarrow$ keep all calibrations
- Assume ancestrality of ancient languages $\leftarrow$ test assumption
- Topology assumptions e.g. (Anatolian,(Tocharian, rest))$\leftarrow$ relax
- IElex data fixes

Result:

- Much younger origin, supporting Kurgan hypothesis
- Much less credible dates
  - Latin expansion $<$ 2200 BP (Hispania) – 1900 BP (Dacia) not 763-1286 BP
  - Norse to Faroese and Icelandic $<$ 1100BP not 212-494 BP
  - Gaelic into Scotland $<$ 1500BP not 259-655 BP

Test for ancestrality

Sampled ancestors model (Gavryushkina et al. 2014) uses reversible jump



Issues with IELex data

Swadish word list

- modern languages – pick most commonly used words
  dog but not *hound* in English
- ancient languages – pick any documented words
  omnis & *totus* in Latin

Issues with IELex data

Swadish word list

- modern languages – pick most commonly used words
  dog but not *hound* in English
- ancient languages – pick any documented words
  omnis & *totus* in Latin

Observation:

- Many more cognates for ancient languages
- Extreme estimates for old-old Irish and old Ancient Greek (in Bouckaert et al. 2012)

Word lists revised for consistent definition of most commonly Celtic languages, Old Irish as well as Ancient Greek.

Tree prior

SA birth death skyline

What we know:

- Speciation process (not coalescent)
- Root height range = 5K,10K
- Nr of contemporary languages = 80 out of 400-600
- Allow 9 ancestral languages

Tree prior

SA birth death skyline

What we know:

- Speciation process (not coalescent)
- Root height range = 5K,10K
- Nr of contemporary languages = 80 out of 400-600
- Allow 9 ancestral languages

What to specify:

- birth rate U[0, 0.56...]
- death rate U[0, 100]
- sampling rate - non zero when sampling
- origin height $< 10K$
- sample proportion < 0.2

Tree prior

SA birth death skyline

What we know:

- Speciation process (not coalescent)
- Root height range = 5K,10K
- Nr of contemporary languages = 80 out of 400-600
- Allow 9 ancestral languages

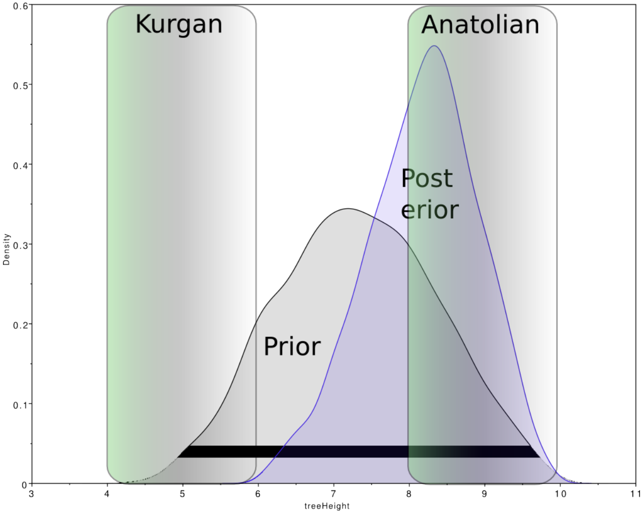
What to specify:

- birth rate U[0, 0.56...]
- death rate U[0, 100]
- sampling rate - non zero when sampling
- origin height $< 10K$
- sample proportion < 0.2

SACount Prior = 6.2 [2-9] Posterior = 0.5 [0, 1]– mostly Luvian ancestral to Lycian

Kurgan : Anatolian root height 2:1

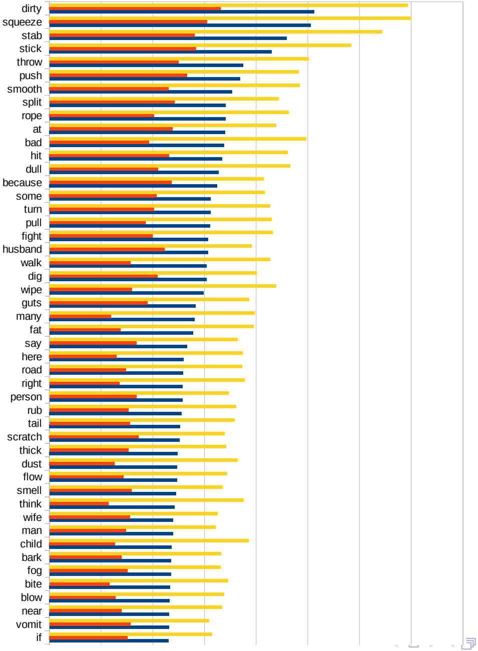
Origin estimates



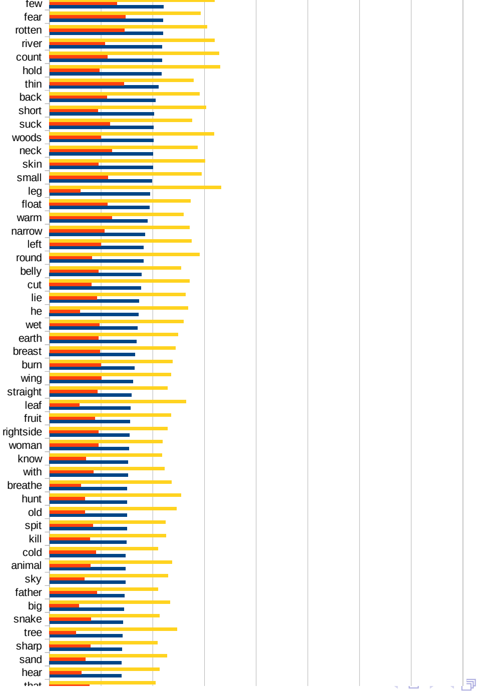
Root height Prior = 6.9 [5.0, 9.0] Posterior 7.9 [6.5,9.5]

Bayes Factor >> 100 in favour of Anatolian hypothesis

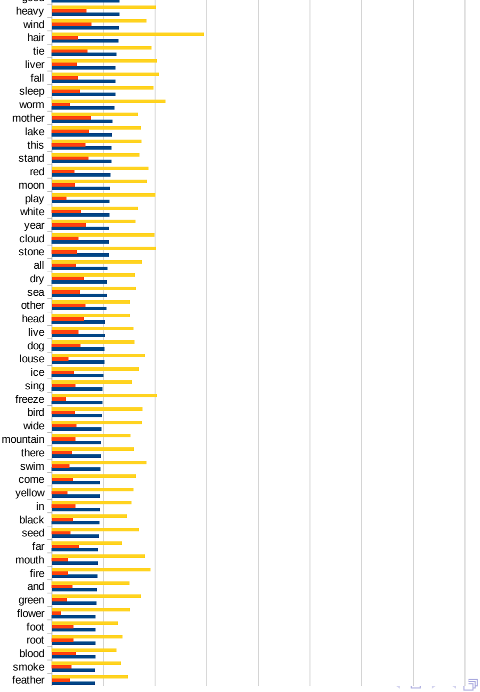
Rates


dirty
squeeze
stab
stick
throw
push
smooth
split
rope
at
bad
hit
dull
because
some
turn
pull
fight
husband
walk
dig
wipe
guts
many
fat
say
here
road
right
person
rub
tail
scratch
thick
dust
flow
smell
think
wife
man
child
bark
fog
bite
blow
near
vomit
if

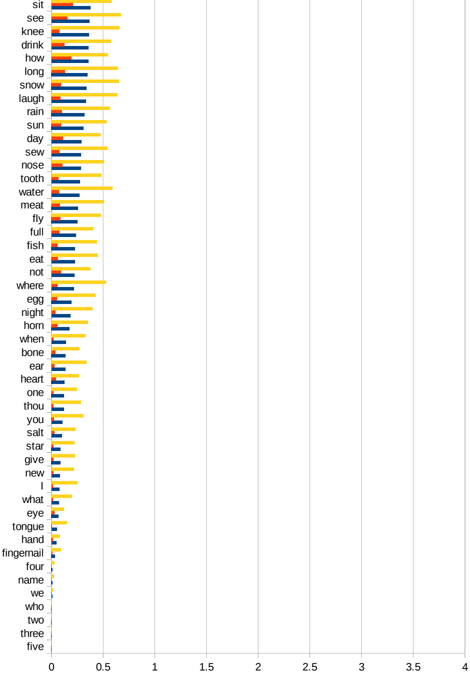
Rates


few
fear
rotten
river
count
hold
thin
back
short
suck
woods
neck
skin
small
leg
float
warm
narrow
left
round
belly
cut
lie
he
wet
earth
breast
burn
wing
straight
leaf
fruit
rightside
woman
know
with
breathe
hunt
old
spit
kill
cold
animal
sky
father
big
snake
tree
sharp
sand
hear
that

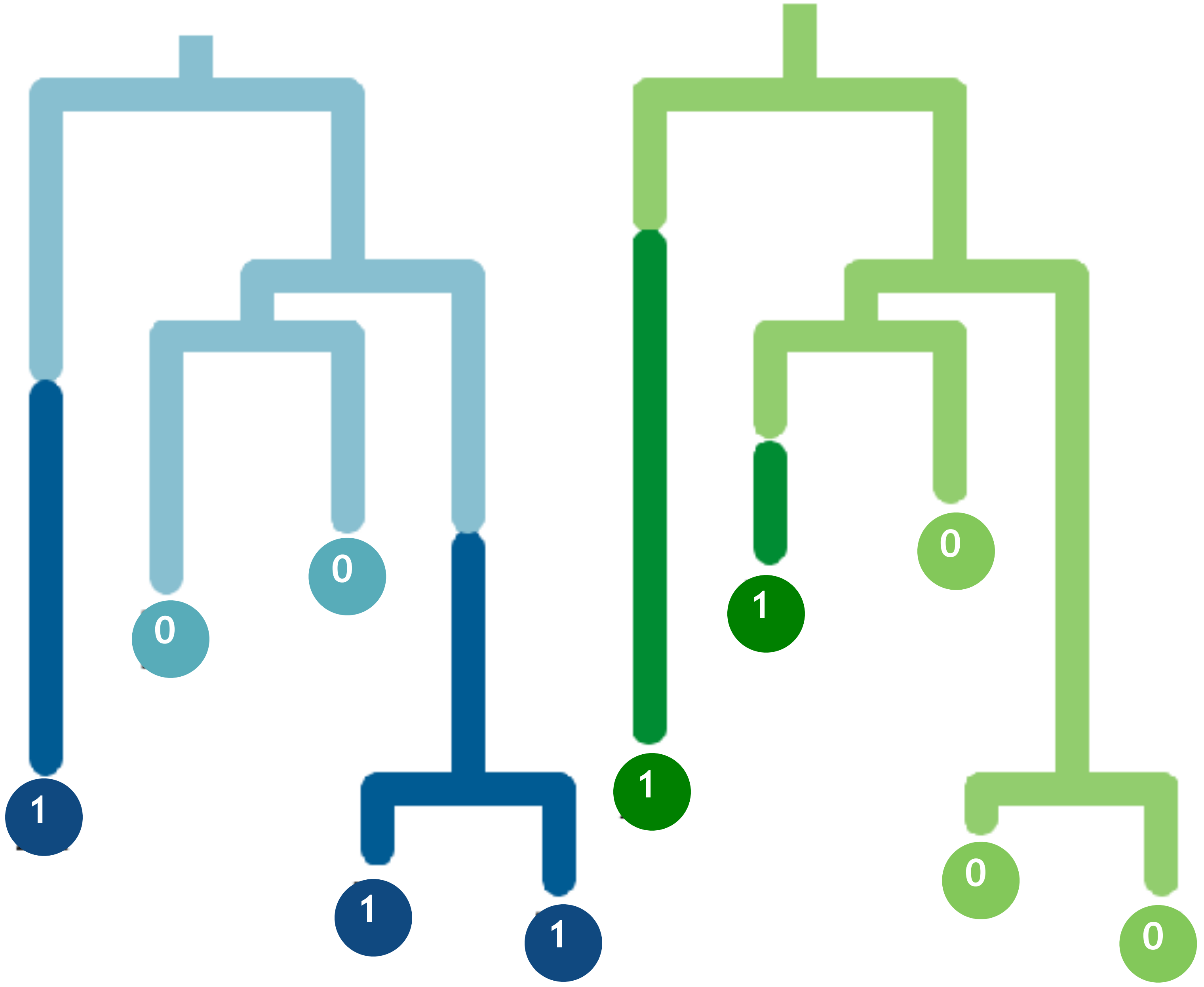
Rates

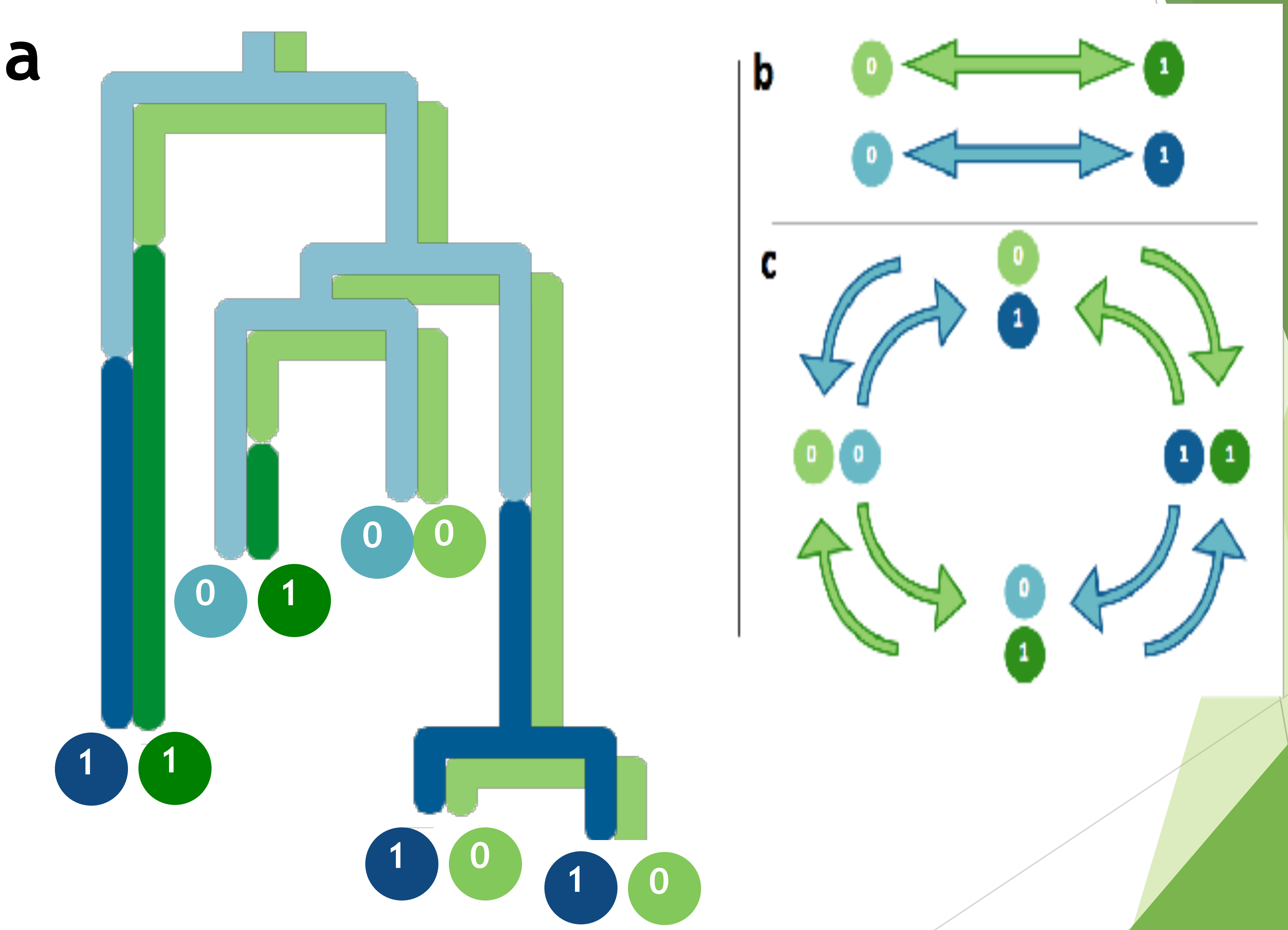

good
heavy
wind
hair
tie
liver
fall
sleep
worm
mother
lake
this
stand
red
moon
play
white
year
cloud
stone
all
dry
sea
other
head
live
dog
louse
ice
sing
freeze
bird
wide
mountain
there
swim
come
yellow
in
black
seed
far
mouth
fire
and
green
flower
foot
root
blood
smoke
feather

Rates

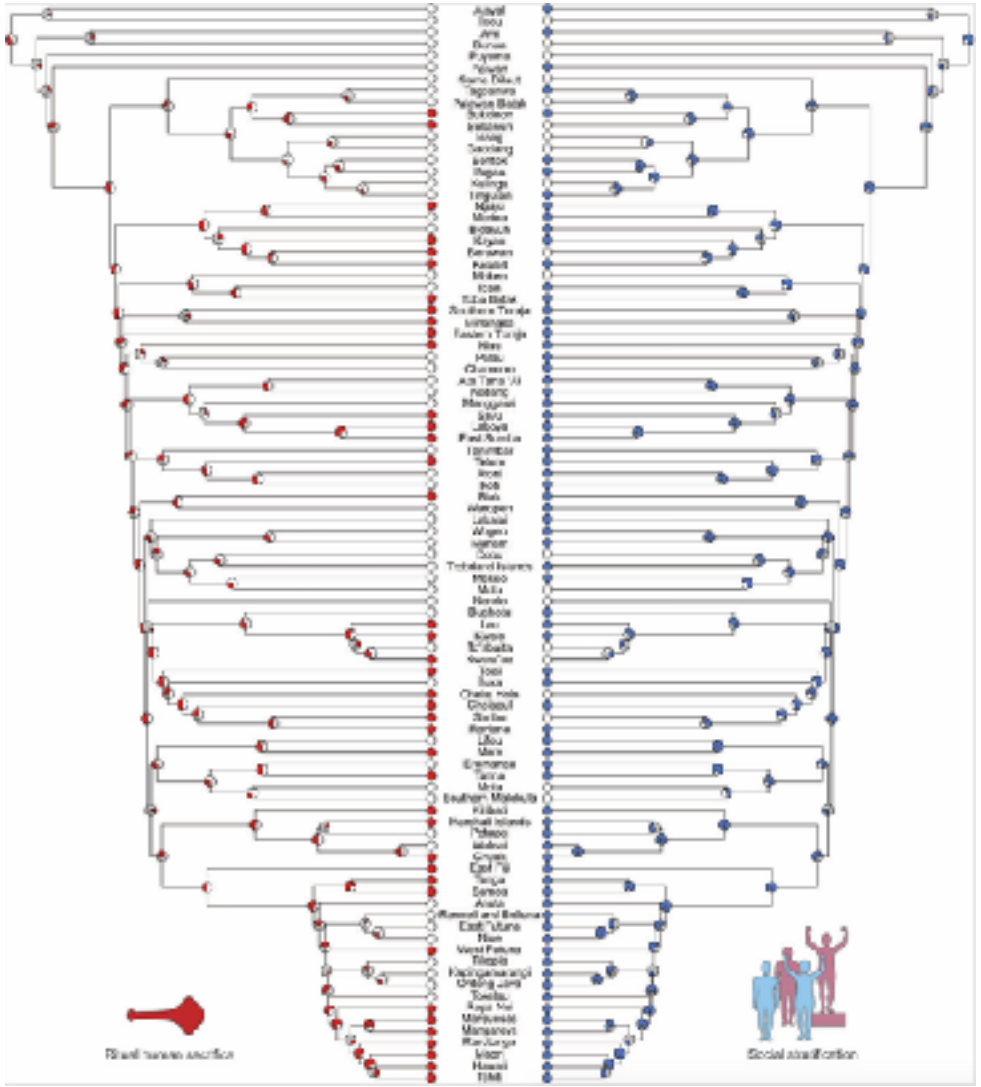

sit
see
knee
drink
how
long
snow
laugh
rain
sun
day
sew
nose
tooth
water
meat
fly
full
fish
eat
not
where
egg
night
horn
when
bone
ear
heart
one
thou
you
salt
star
give
new
I
what
eye
tongue
hand
fingernail
four
name
we
who
two
three
five
0
0.5
1
1.5
2
2.5
3
3.5
4

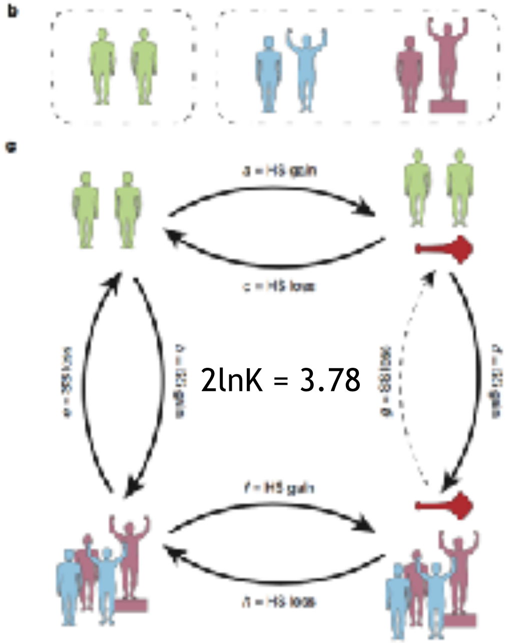
a




a
0
1
0
0
1
1
1
0
1
0
b
0
1
0
1
c
0
1
0
0
1
1
0
1

Human sacrifice sustained social stratification





2lnK = 3.78

Watts et al. (2016). *Nature*

Summary

- Introduce phylogenetical/phylogeographical methods adapted to linguistics
- Since 2012:
  - Improved model – better fit to data
  - SA Test – ancestrality not supported by data
  - Refined data – somewhat more credible dating of internal nodes
- Homeland of Indo-European languages is identified as Anatolia

What's next:

- Develop more realistic models using input from linguists

Keeping up to date: `http://language.cs.auckland.ac.nz`