A stochastic Dollo-model for phylogenetic inference:

model elaboration and checking.

Q Atkinson, GN, D Welch, and R Gray 'From words to dates: water into wine, mathemagic or phylogenetic inference', *Trans. Philological Soc.* 103, 193-219 (2005)

GN, RD Gray, 'Dated ancestral trees from binary trait data and its application to the diversification of languages', *J. Roy. Statist. Soc. B, 70:545-566* (2008)

RJ Ryder, GN, 'Missing data in a stochastic Dollo model for cognate data, and its application to the dating of Proto-Indo-European', *arXiv:0908.1735* (2009 and JRSSC, to appear).

Geoff Nicholls, Statistics, Oxford *Model Uncertainty*, CRISM, Warwick, 30/5/2010

Data

1. *Ringe et al. 2002* 24 Indo-European languages 4000 distinct traits (*schematically*)

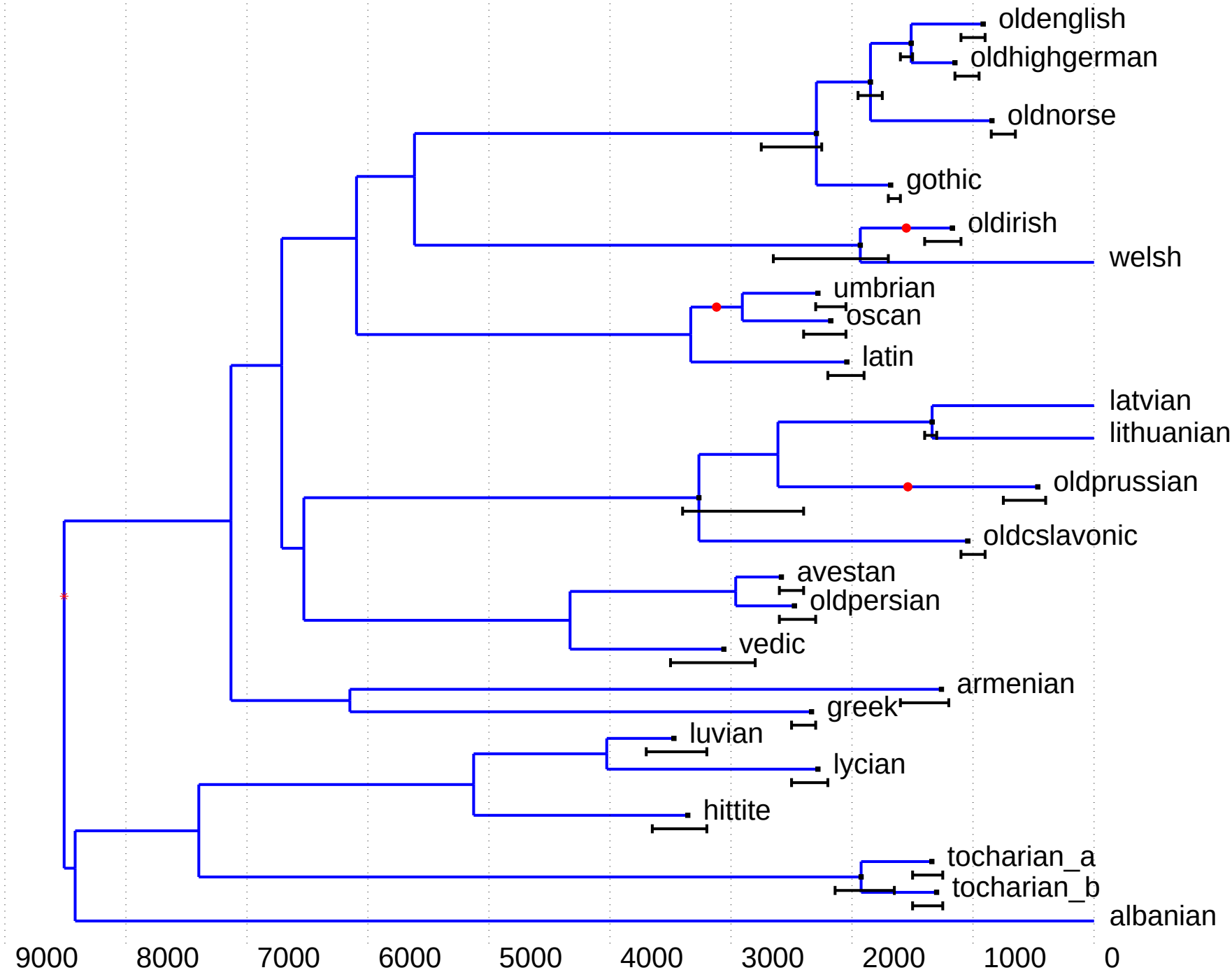
	ALL	AND	ANIMAL
oldenglish	1000000	100000000000000	00010000 . . .
oldhighgerman	1000000	100000000000000	00010000
gothic	1000000	????????????????	00001000
latin	0100000	00000001000000	00000100
umbrian	0100000	00000000000010	????????. . .
...	...	...	...

New Data

2. *M Rother 2008*, from Sprachatlas von Bayerisch-Schwaben
Includes dialects Ries, Ostfranken, Nordbaiern, Ostlech(Sued), Ammersee(Ost),Lechrain,...

3. *S Greenhill et al 2008*
400 Austronesian languages (Maori, Indonesian, Malagasy,...), 30000 lexical traits

4. *A Kitchen et al 2009*
25 Semitic languages (Hebrew, Akkadian, Amharic,...), 674 lexical traits


oldenglish
oldhighgerman
oldnorse
gothic
oldirish
welsh
umbrian
oscan
latin
latvian
lithuanian
oldprussian
oldcslavonic
avestan
oldpersian
vedic
armenian
greek
luvian
lycian
hittite
tocharian_a
tocharian_b
albanian
9000
8000
7000
6000
5000
4000
3000
2000
1000
0

Generic trait-observation model

$g = (E, t, k) \quad t = (t_1, \ldots t_{2L-1})$

$z = (z_1, \ldots z_N) \quad z_a = (\tau_a, i_a) \in [g]$

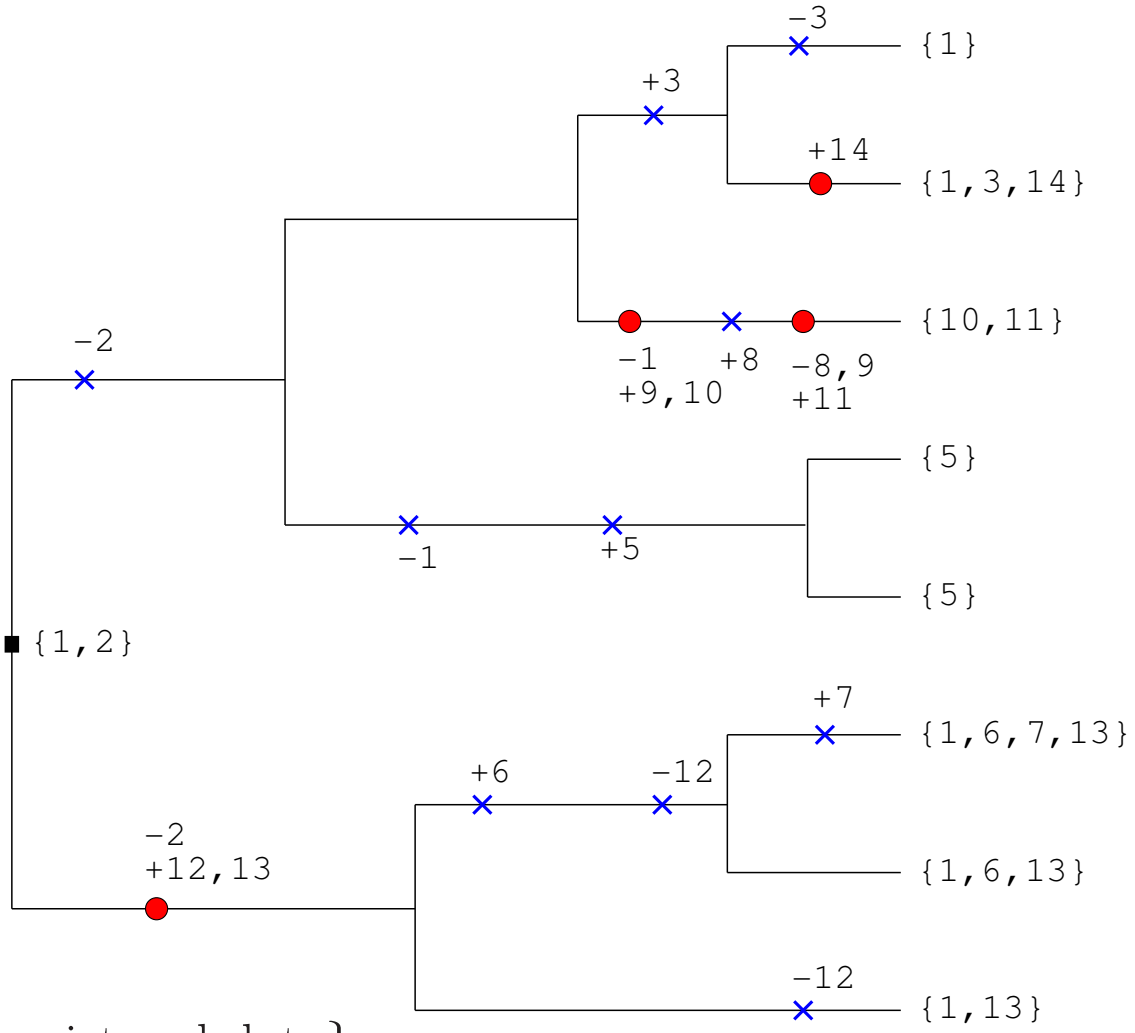
λ: set-element birth rate;

μ: set element death rate;

ρ: catastrophe rate, $k = (k_1, \ldots, k_{2L-2})$

κ, death prob, $\nu = \lambda\kappa/\mu$ mean births

ξ_i: probability element a visible at leaf i



$$
\begin{aligned}
Z &\sim \Pi(\lambda; [g]), \\
\mathcal{E}_{Z_0} &= \{\text{birth at } Z_0 \in [g] \text{ yields registered data}\} \\
\lambda(z_0) &= \lambda \Pr(\mathcal{E}_{Z_0} | g, \mu, \lambda, \xi, Z_0 = z_0) \\
Z|\mathcal{E}_Z &\sim \Pi(\lambda(z); [g]), \\
\Pr(dz|g, \mu, \lambda, \xi) &= f_Z(z)dz \\
&= \frac{1}{N!} e^{-\Lambda([g])} \prod_{a=1}^{N} \lambda(z_a) dz_a
\end{aligned}
$$

[D*]	[I*]	[D~]
100000000000000	111111111111111	100000000000000
101000000000001	010111111111111	?0?00000000001
000000000011000	101110011101l0	0?000??001?00?
000010000000000	111111111111111	000010000000000
000010000000000	111101111111111	0000?000000000
100001100000l0	101111111111111	1?000110000010
100001000000l0	111111111111111	100001000000l0
100000000000l0	111111111111100	100000000000??

[I]	[D]
1 111 1 11	1 000 0 00
0 111 1 11	? 000 0 01
1 100 1 10	0 0?? 1 0?
1 111 1 11	0 100 0 00
1 011 1 11	0 ?00 0 00
1 111 1 11	1 011 0 10
1 111 1 11	1 010 0 10
1 111 1 00	1 000 0 ??

$$
\begin{aligned}
Y_a &= \sum_{i=1}^{L} \mathbf{I}^*_{i,a} \mathbf{D}^*_{i,a} \\
R_2(\tilde{D}) &= (\tilde{D}_a : Y_a > 1) \\
\mathcal{E}_{Z_a} &= \{\mathbf{D}_a = R_2(\tilde{\mathbf{D}}_a)\}
\end{aligned}
$$

Likelihood

$$P[\mathbf{D}=D \;\mid\; g,\mu,\lambda,\xi,\mathbf{D}=R(\tilde{\mathbf{D}})] = \int P[\mathbf{D}=D|g,\mu,\xi,Z=z,\mathbf{D}=R(\tilde{\mathbf{D}})]\, f_Z(z)dz,$$

$$= \frac{1}{N!}\left(\frac{\lambda}{\mu}\right)^N \exp\left(-\frac{\lambda}{\mu}\sum_{\langle i,j\rangle\in E} P[\mathcal{E}_Z|Z=(t_i,i),g,\mu,\kappa,\xi](1-e^{-\mu(t_j-t_i+k_iT_C)})\right)$$

$$\times\prod_{a=1}^{N}\left(\sum_{\langle i,j\rangle\in E_a}\sum_{d^*\in\mathcal{D}_a} P[\mathbf{D}^*_a=d^*|Z_a=(t_i,i),g,\mu,\kappa](1-e^{-\mu(t_j-t_i+k_iT_C)})\right)$$

$$\times\prod_{a=1}^{N}\prod_{i=1}^{L}\xi_i^{I_{i,a}}(1-\xi_i)^{1-I_{i,a}}$$

Posterior

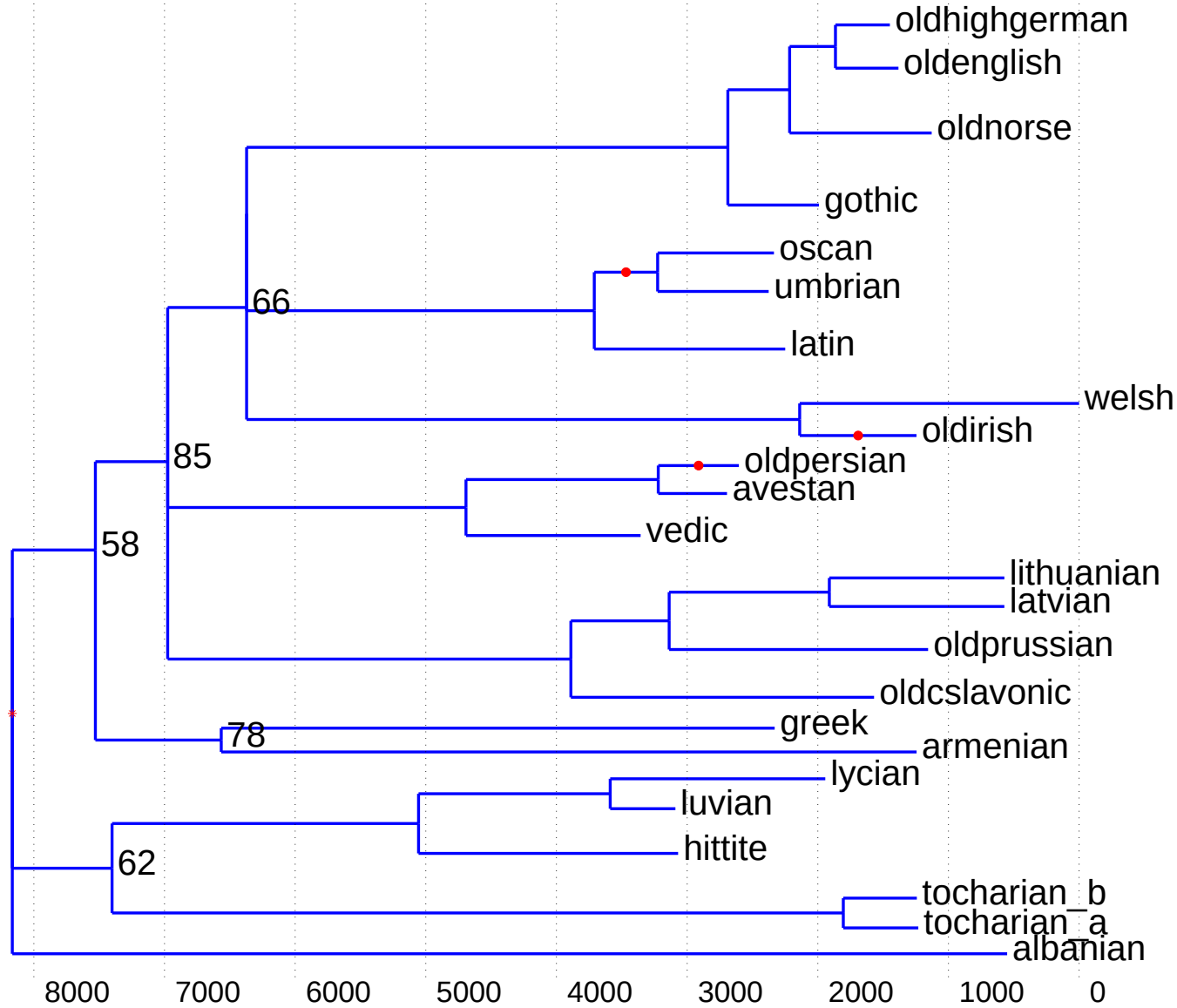
$$p(g,\mu,\lambda,\kappa,\rho,\xi|\mathbf{D}=D) = \frac{1}{\mu\lambda}f_G(g|T)p_R(\rho)\frac{e^{-\rho|g|}(\rho|g|)^{k_T}}{k_T!} \times P[\mathbf{D}=D|g,\mu,\lambda,\xi,\mathbf{D}=R(\tilde{\mathbf{D}})]$$

Notation

$$\mathcal{D}_a = \left\{\text{possible } D^*_{:,a}\text{-vectors given } D_{:,a}\right\}$$
$$E_a = \text{edges where birth at } Z_a \text{ could evolve to } D_{:,a}.$$

Quantity	*95% HPD*	*units*
root age	(7110,9750)	yr BP
death rate μ	(1.7,2)	# per 10^4yr
cat. rate ρ	(0.31,0.97)	# per 10^4yr
death prob. κ	(0.3,0.42)	

A catastrophe every 15000yr
each $\simeq$ 2400yr.



Model Error: Predicting Calibrations

Predict ages of nodes for which calibration data is available.
Quantify goodness-of-fit using Bayes factors - would we reject the truth?

$\Gamma^{(c)}$ space of all trees respecting c'th constraint, and $\Gamma = \bigcap_{c=0}^{C} \Gamma^{(c)}$.

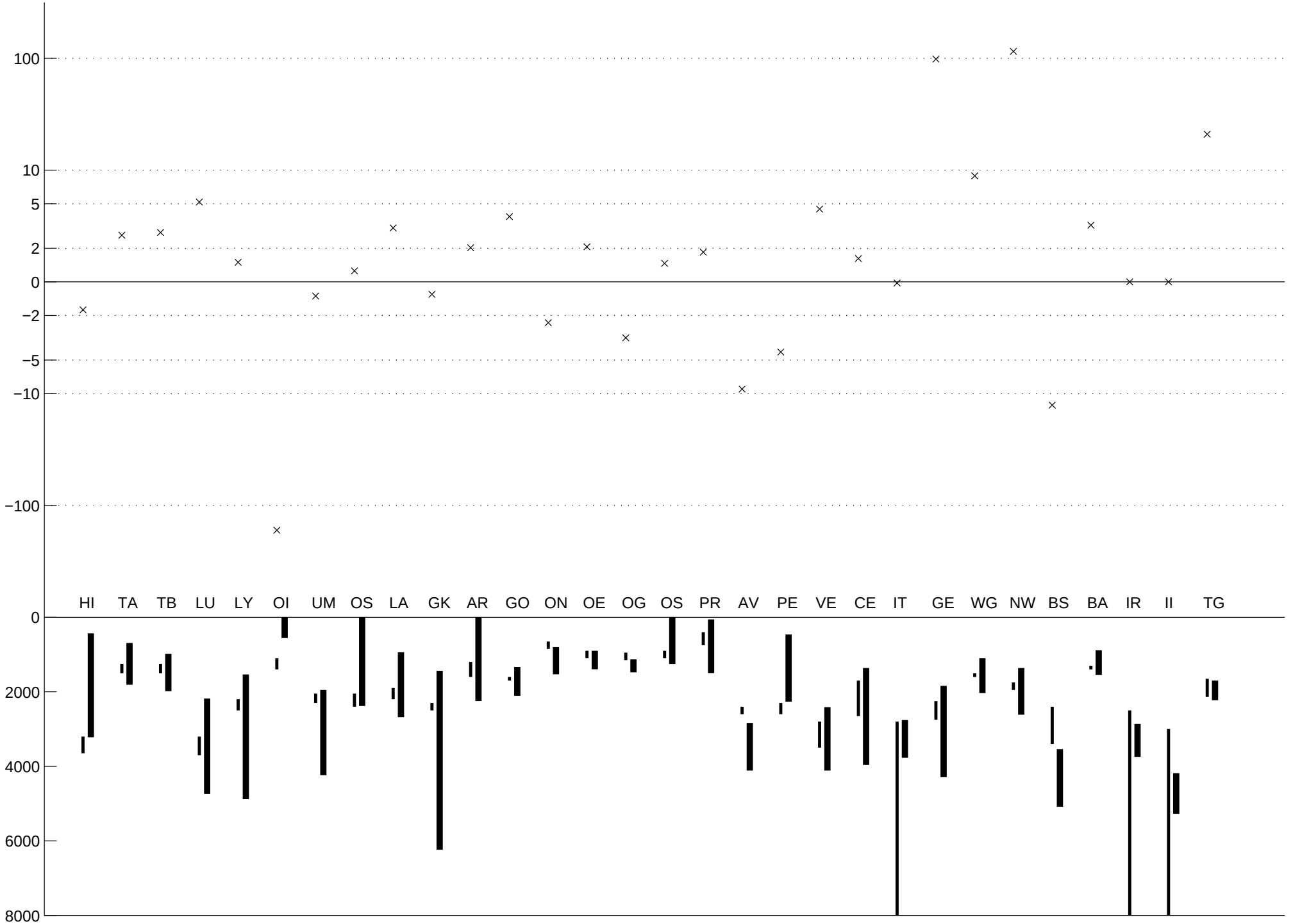
For $c = 1, 2, ..., C$ let

$$\Gamma^{-c} = \bigcap_{\substack{c'=0 \\ c' \neq c}}^{C} \Gamma^{(c')}.$$

$M_0 : g \in \Gamma$, $M_1 : g \in \Gamma^{-c}$. Evidence for c'th constraint

$$\begin{aligned} B_c &= \frac{P(D|g \in \Gamma)}{P(D|g \in \Gamma^{-c})} \\ &= \frac{P(D|g \in \Gamma \cap \Gamma^{-c})}{P(D|g \in \Gamma^{-c})} \\ &= \frac{P(g \in \Gamma|D, g \in \Gamma^{-c})}{P(g \in \Gamma|g \in \Gamma^{-c})}. \end{aligned}$$

Constraints with $2\log(B_c) \lesssim -5$ conflict other model and data. Reject 3 of 30.


100
10
5
2
0
−2
−5
−10
−100
HI
TA
TB
LU
LY
OI
UM
OS
LA
GK
AR
GO
ON
OE
OG
OS
PR
AV
PE
VE
CE
IT
GE
WG
NW
BS
BA
IR
II
TG
0
2000
4000
6000
8000

Model error: targeted checks

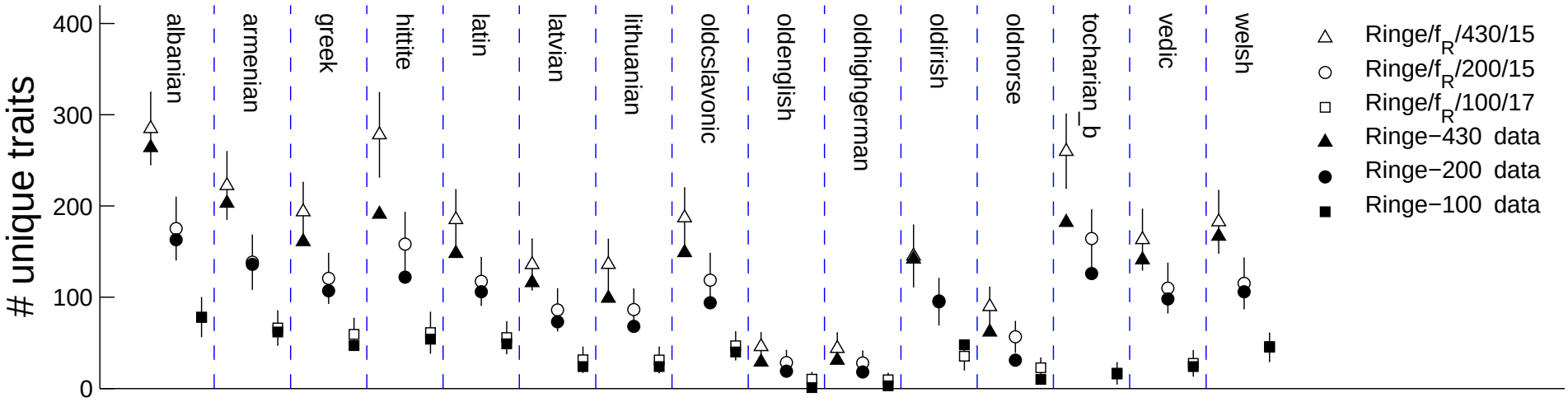
- rate heterogeneity (space-time)
  - explicit (catastrophe) model shows some STRH
  - fit simulated data Gamma rate heterogeneity
- rate heterogeneity (across traits)
  - split data by columns into groups SW100/200, and refit (core 1-100 slower than 101-200)
  - exploratory check based on predictive replicates (tail of very slow traits)
  - fit simulated data
- borrowing/lateral transfer
  - fit simulated data (successful)
  - split data into two groups of language, refit (evidence for some LT)
- several others (*eg* meaning-category non-empty, data missing in blocks)

Model error: generic checks for misfit

External Data: remove singletons and predict # for each leaf.
D^* reserved data, $f(D)$ function of data, ψ parameters, predictive DBN

$$P(D'|D) = \int P(D'|\psi)p(\psi|D)d\psi$$

and estimate $P(f(D^*) > f(D')|D)$ (relevant here).



Posterior predictive distributions for the number of unique states, from the Ringe02 data: △ 328 meaning categories; ∘ Swadesh-200 list; □ Swadesh 100 list.

Predictive replicates: estimate $P(f(D) > f(D')|D)$ for $f(D)$ function of data.
$Y^{(n)} = \#$ traits displayed at exactly n leaves.

y-axis $\mathsf{E}(\tilde{Y}^{(n)}|D) - Y^{(n)}(D)$ and $\pm 2\text{std}(\tilde{Y}^{(n)}|D)$. x-axis $n = 2, 3, ..., L$.

(TL) $S/MH50/U200$ (TR) $S/MH25/U200$ (LL) Ringe/f_R/328/15 (LR) Ringe/f_R/100/17.