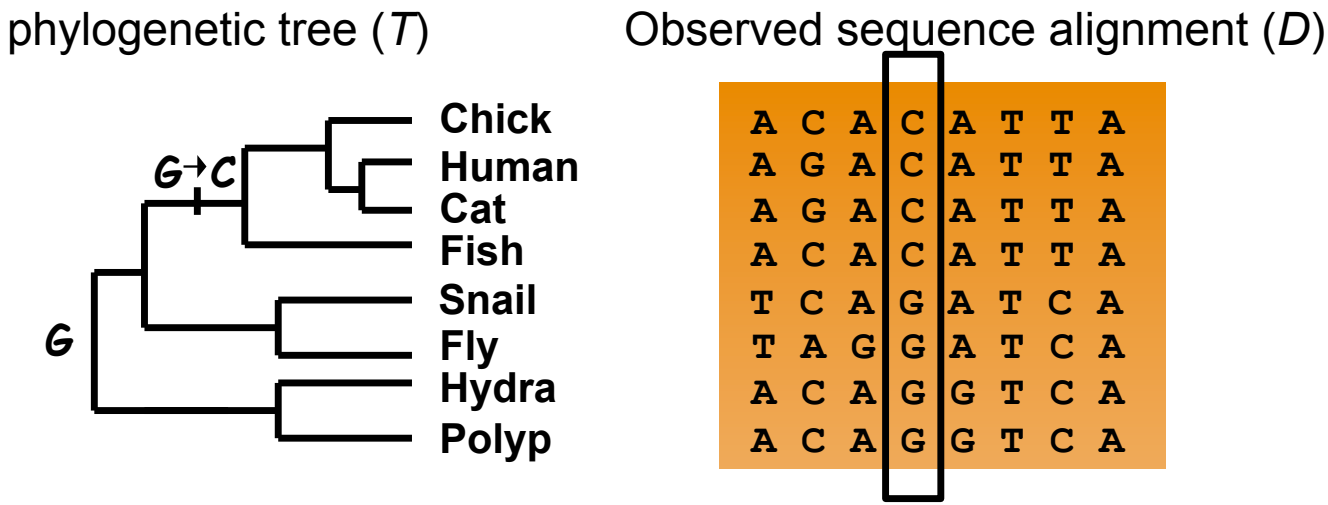
Bayesian models in evolutionary studies and their frequentist properties

Nicolas Lartillot

June 24, 2016

Molecules as documents of evolutionary history

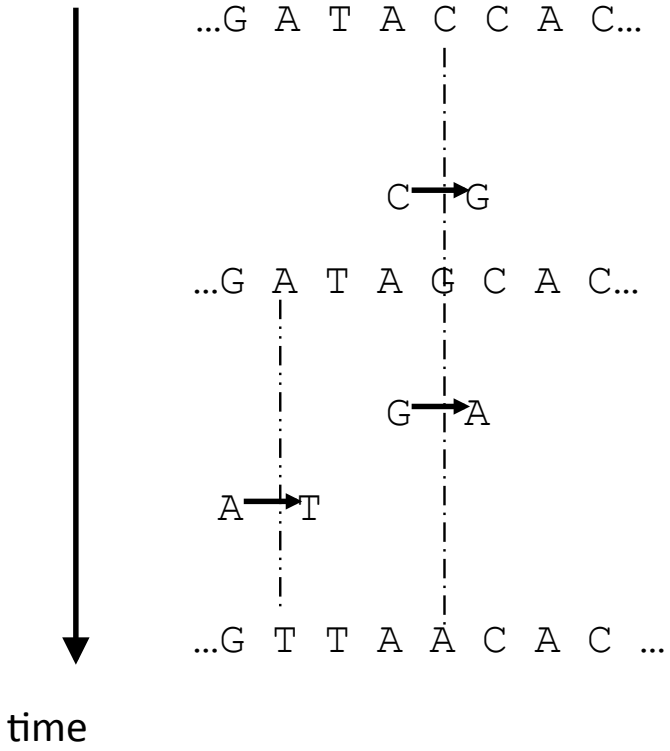


General aim

using aligned DNA sequences for:

- reconstructing phylogenies
- estimating divergence times
- inferring macro-evolutionary patterns
- characterizing molecular evolutionary processes

Probabilistic model of substitution: nucleotides

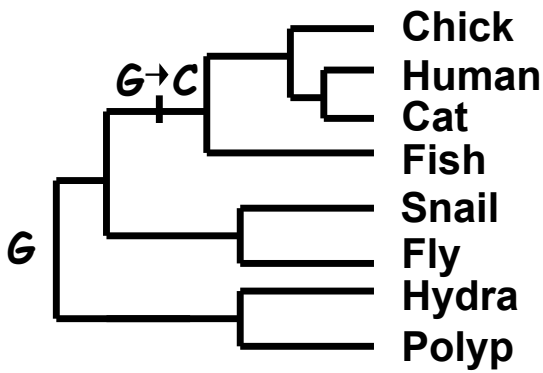


$$Q = \begin{array}{c|cccc} & A & C & G & T \\ \hline A & - & r\frac{\gamma}{2} & r\kappa\frac{\gamma}{2} & r\frac{1-\gamma}{2} \\ C & r\frac{1-\gamma}{2} & - & r\frac{\gamma}{2} & r\kappa\frac{1-\gamma}{2} \\ G & r\kappa\frac{1-\gamma}{2} & r\frac{\gamma}{2} & - & r\frac{1-\gamma}{2} \\ T & r\frac{1-\gamma}{2} & r\kappa\frac{\gamma}{2} & r\frac{\gamma}{2} & - \end{array}$$

- $r > 0$: substitution rate ($\sim 10^{-2}$ per million years in mammals)
- $\kappa > 0$: relative transition-transversion rate (~ 3).
- $0 < \gamma < 1$: equilibrium GC content (GC^*)

The likelihood

phylogenetic tree (T)



Observed sequence alignment (D)

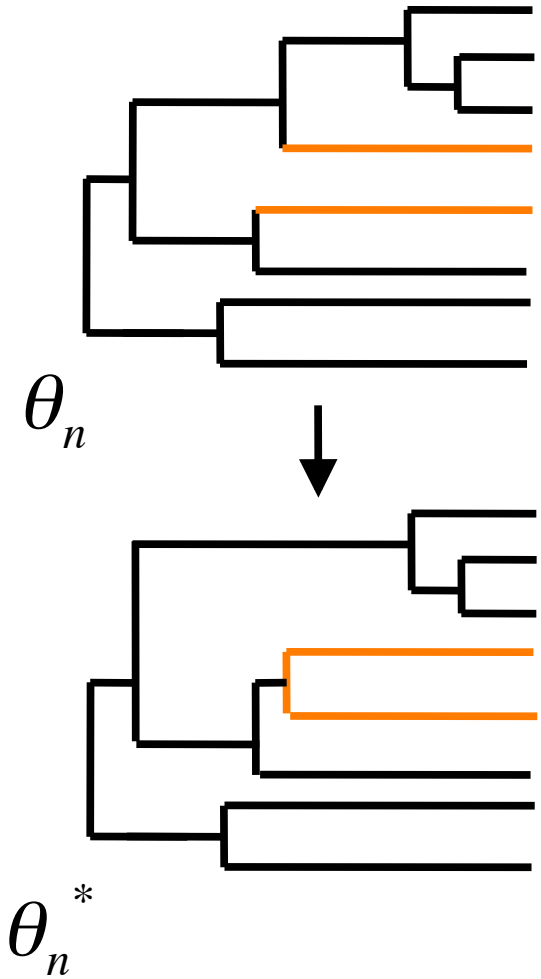
A	C	A	C	A	T	T	A
A	G	A	C	A	T	T	A
A	G	A	C	A	T	T	A
A	C	A	C	A	T	T	A
T	C	A	G	A	T	C	A
T	A	G	G	A	T	C	A
A	C	A	G	G	T	C	A
A	C	A	G	G	T	C	A

- D: data (columns X_i, $i = 1..N$, assumed to be i.i.d.)
- $\theta = (T, r, \kappa, \gamma)$: parameters of the model
- The likelihood:

$$p(D \mid \theta) = \prod_i p(X_i \mid \theta)$$

- most often, vague priors are used

Markov chain Monte Carlo



1. Propose a move $\theta_n \rightarrow \theta_n^*$

   According to kernel $q(\theta, \theta^*)$

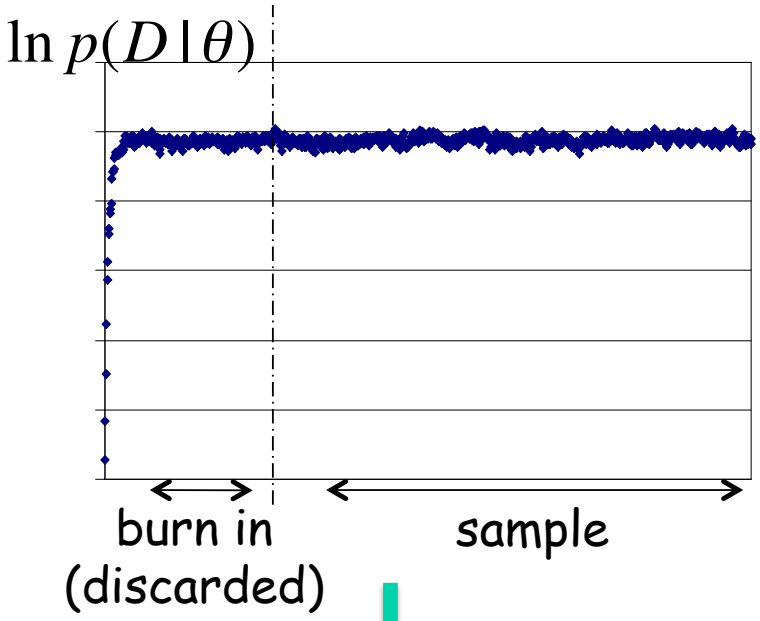
2. Accept with probability

$$p = Min\left\{1, \frac{p(\theta_n^* \mid D) q(\theta_n^*, \theta_n)}{p(\theta_n \mid D) q(\theta_n, \theta_n^*)}\right\}$$

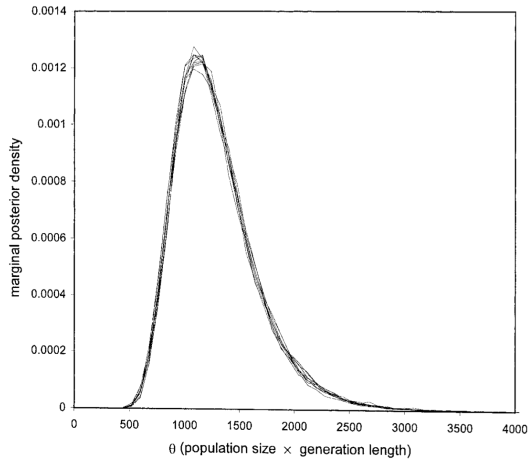
3. Iterate

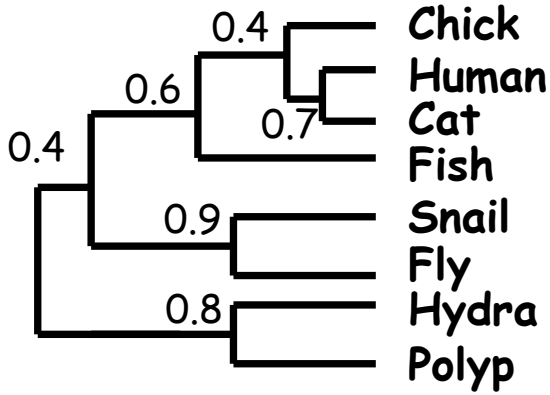
- alternate with Metropolis-Hastings on rates and branch lengths

Inference by marginalization of the posterior

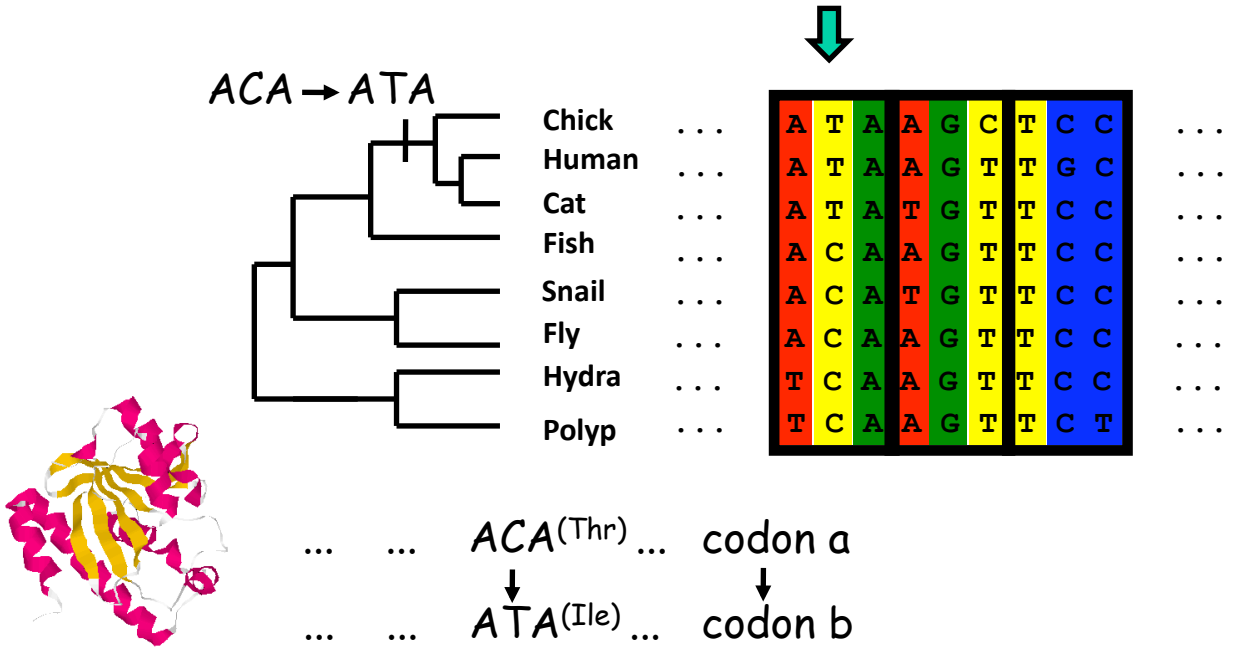


$$(\theta_k)_{k=1..K} \sim p(\theta \mid D)$$





Codon model with global effect



Given 4×4 nucleotide rate matrix Q, define 61×61 codon matrix R:

$$
\begin{aligned}
R_{\mathrm{ACA}\to\mathrm{ACC}} &= Q_{\mathrm{A}\to\mathrm{C}} \\
R_{\mathrm{ACA}\to\mathrm{ATA}} &= Q_{\mathrm{C}\to\mathrm{T}} \cdot \omega \\
R_{\mathrm{ACA}\to\mathrm{AGC}} &= 0 \\
&\cdots
\end{aligned}
$$

$\omega = dN/dS$: relative non-synonymous / synonymous rate

Codon model with global effect

Parameters

- phylogenetic tree (fixed tree or uniform prior over tree topologies)
- branch lengths (hierarchical exponential)
- parameters of the 4×4 nucleotide rate matrix Q (vague priors)
- $\omega = dN/dS$ (vague prior: e.g. half-Cauchy distribution)

Application: characterizing the selective regime

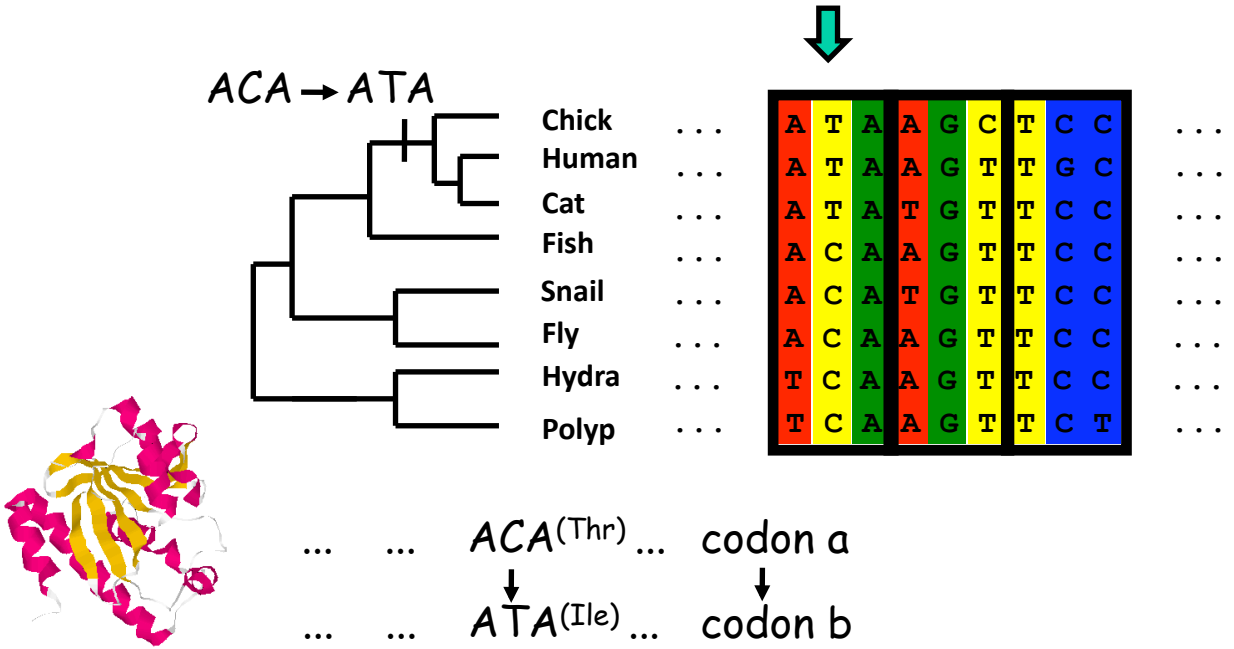
- estimation of ω: median and 95% credible interval
- $\omega > 1$: signature of positive selection
- apply method successively over all protein-coding genes
- find genes such that $p(\omega > 1 \mid D)$ is high

Posterior distribution on ω^*

Gene	post mean	95% CI	$p(\omega^* > 1 \mid D)$
S1PR1-*67-325*	0.681	(0.538, 0.857)	0.001
RBP3-*54-412*	0.726	(0.654, 0.806)	0.000
VWF-*62-392*	0.960	(0.865, 1.063)	0.220
SAMHD1-*67-543*	**1.731**	**(1.542, 1.935)**	$>$ 0.99
TRIM5α-*68-363*	**1.240**	**(1.128, 1.355)**	$>$ 0.99
BRCA1-*64-941*	**1.188**	**(1.123, 1.257)**	$>$ 0.99

Rodrigue and Lartillot, 2016 – based on a mechanistic codon model

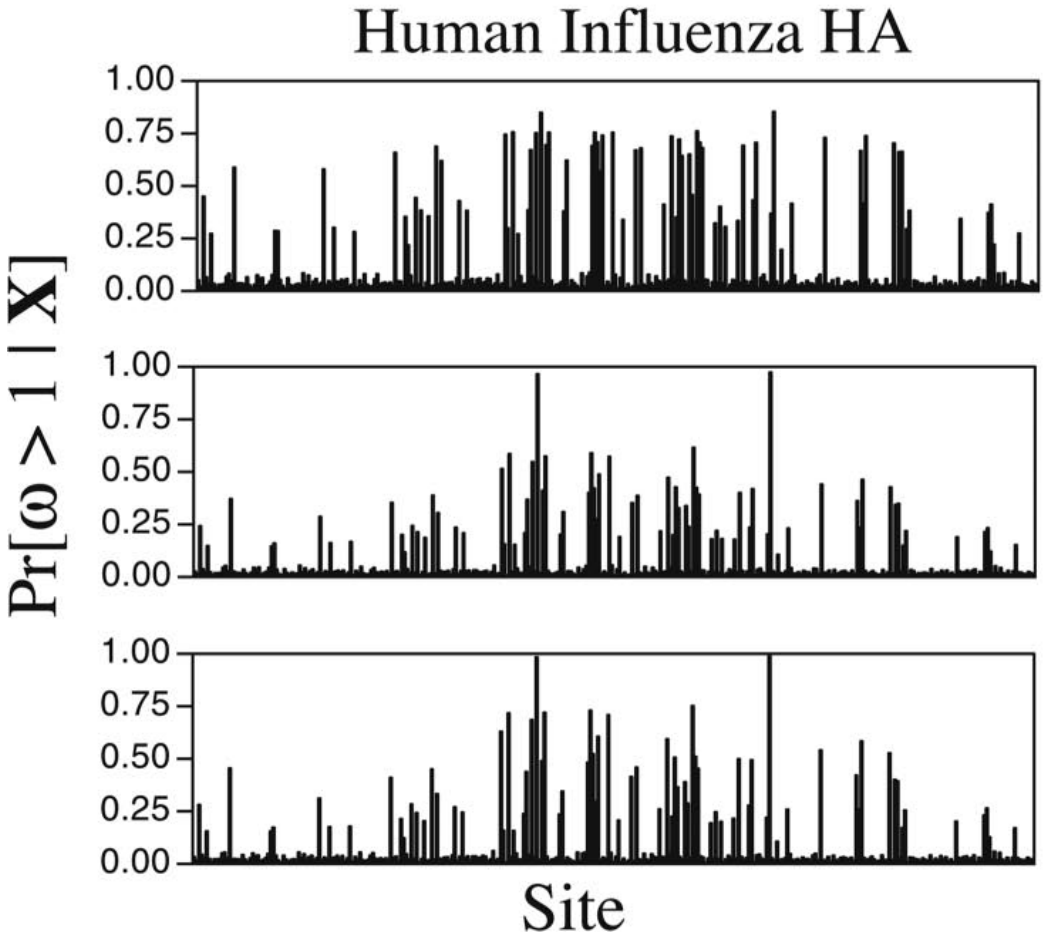
Codon model with site-specific effects



At coding position $i = 1..N$, define 61×61 codon matrix R^i:

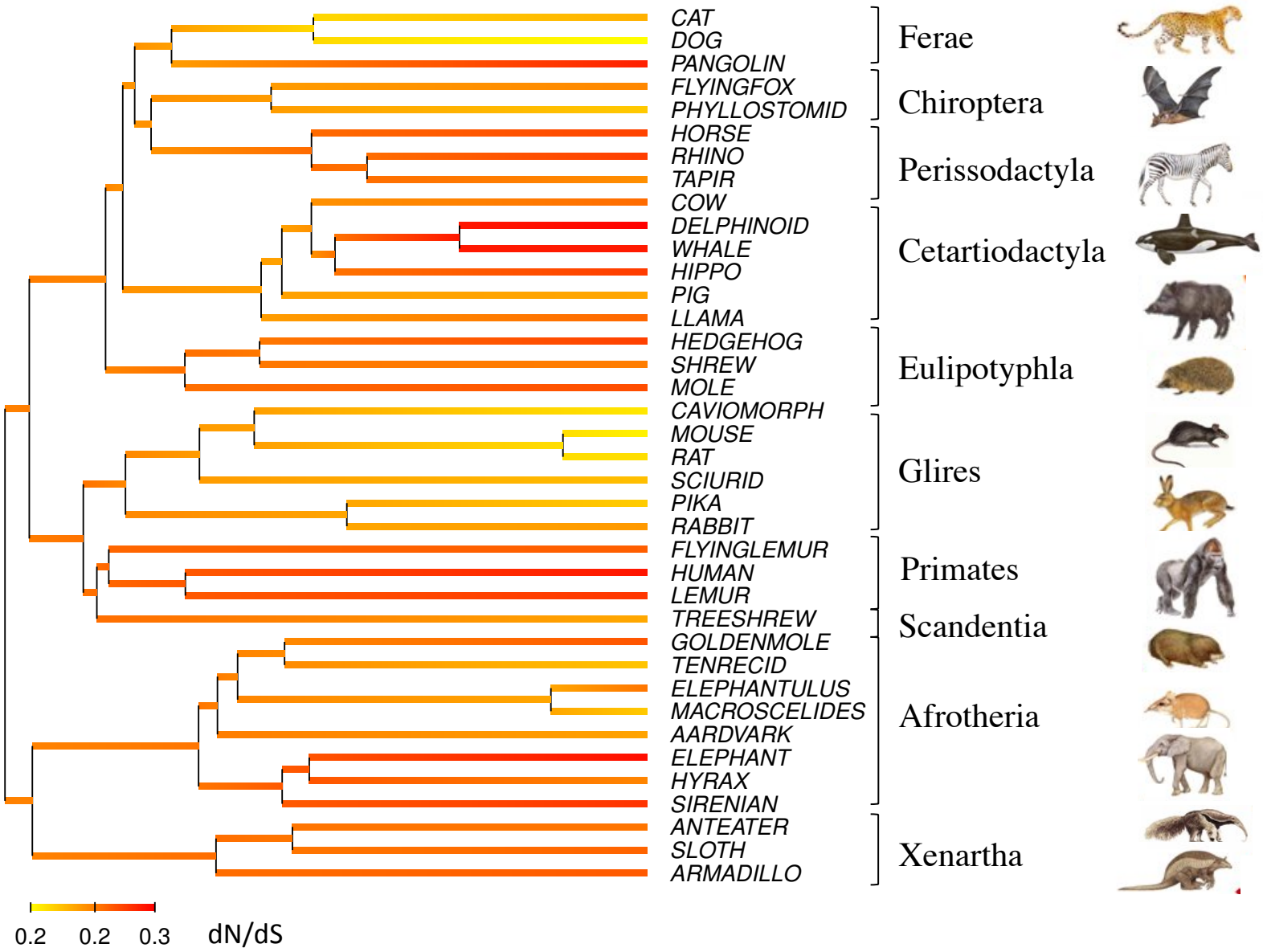
$$
\begin{aligned}
R^i_{\mathrm{ACA}\to\mathrm{ACC}} &= Q_{\mathrm{A}\to\mathrm{C}} \\
R^i_{\mathrm{ACA}\to\mathrm{ATA}} &= Q_{\mathrm{C}\to\mathrm{T}} \cdot \omega_i \\
R^i_{\mathrm{ACA}\to\mathrm{AGC}} &= 0 \\
&\dots
\end{aligned}
$$

Typical results with non-parameteric codon site-model

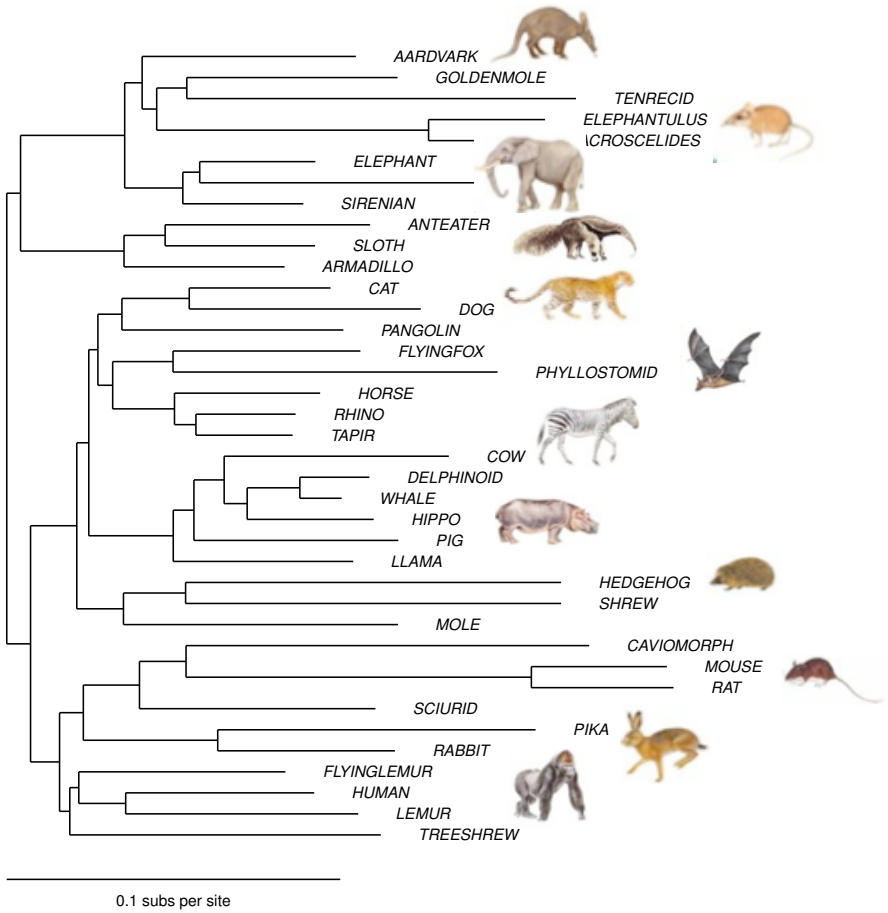


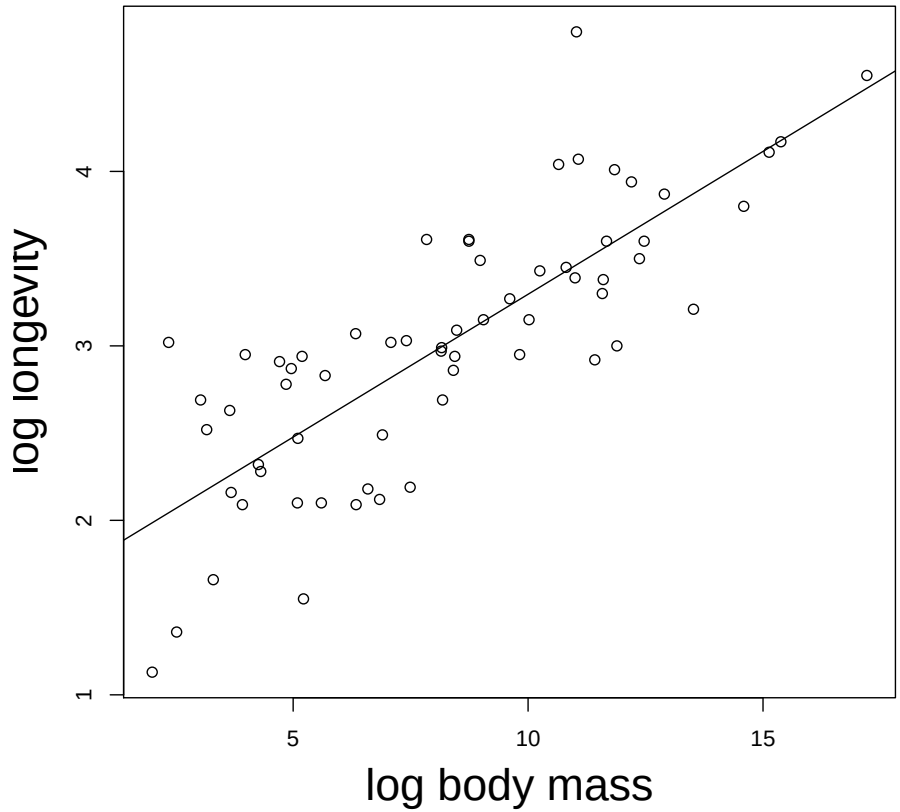
Huelsenbeck et al, 2006, PNAS 103:6263

Variation in $\omega = dN/dS$ over time

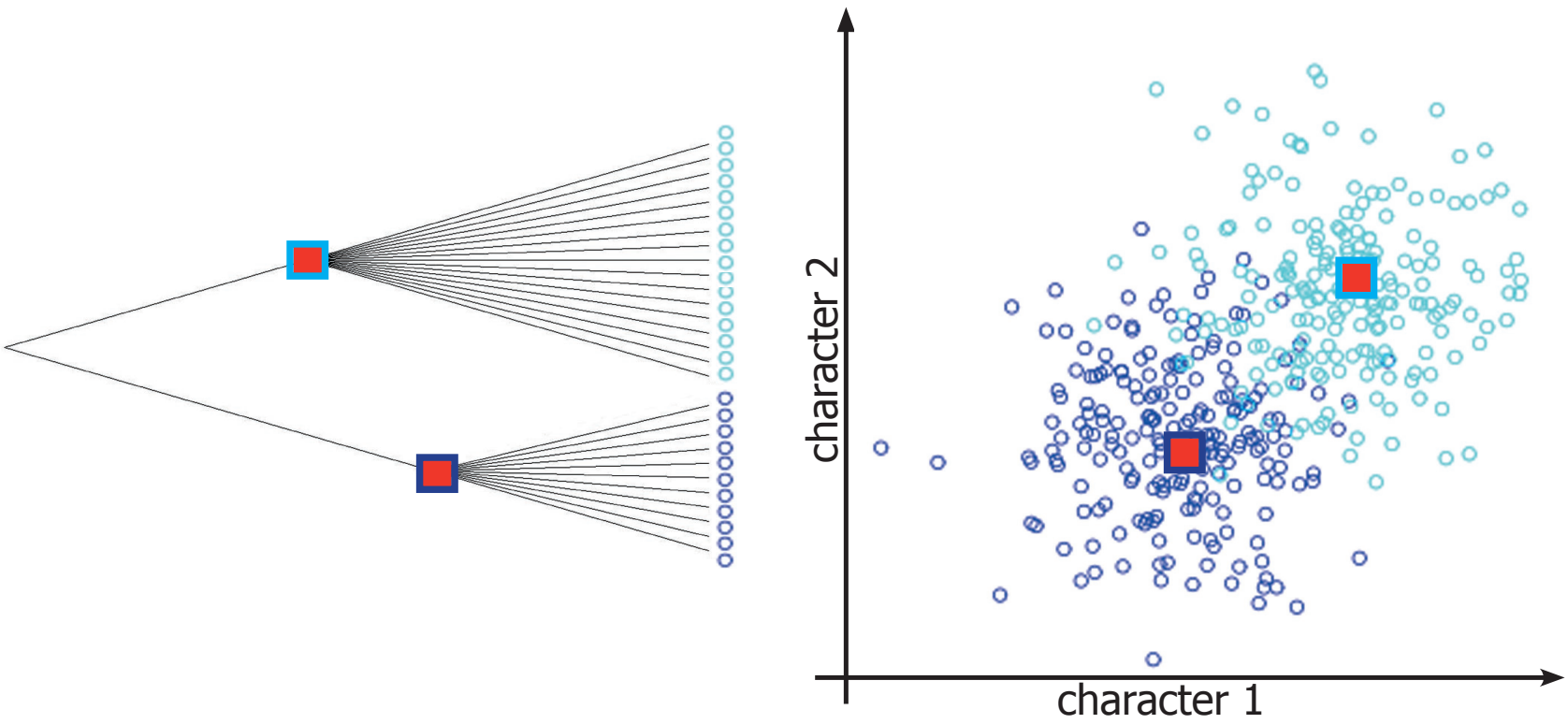


Multiple traits – correlated evolution



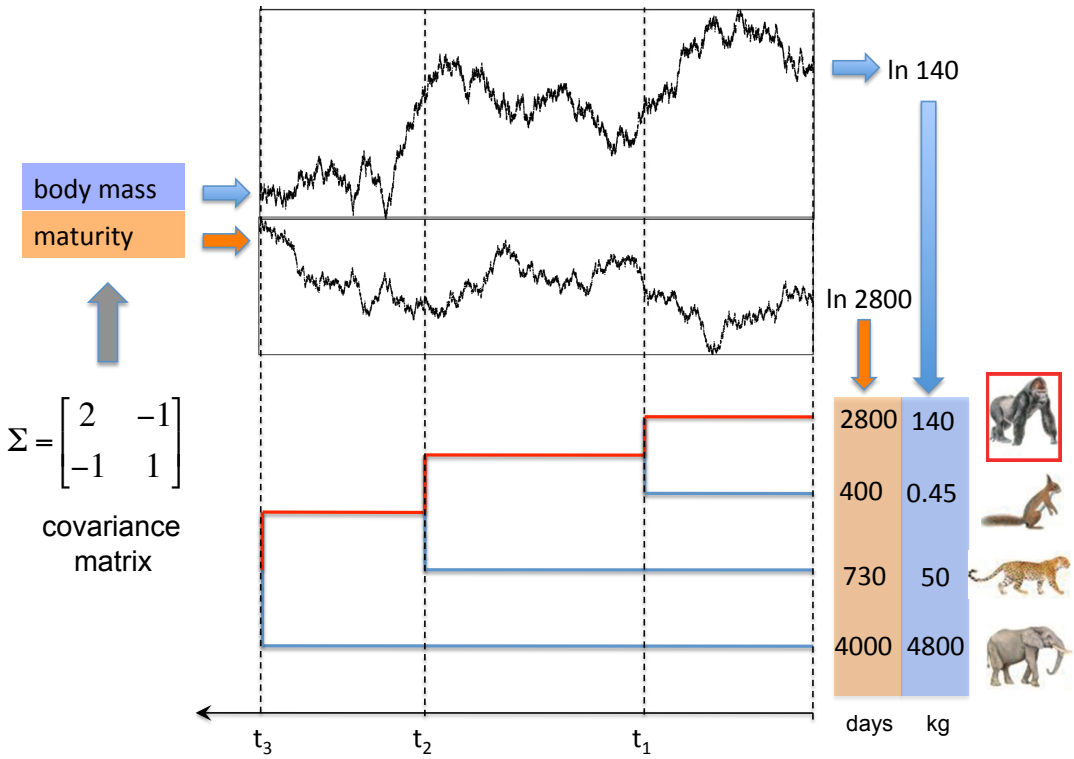


The problem of phylogenetic inertia



Felsenstein, 1985, Am Nat 125:1

Multivariate Brownian process along phylogeny



- Assume 2 traits follow bivariate Brownian motion
- vague prior on covariance matrix Σ
- (inv-Wish centered on diagonal matrix, with few d.f.)
- estimate Σ, assess whether correlation is positive/negative

Inferred correlations in placental mammals

Correlation	λ_S	ω	Maturity	Mass	Longevity
λ_S	—	−0.24*	−0.05	−0.20*	−0.16*
ω	—	—	−0.04	0.28*	0.25*
Maturity	—	—	—	0.40*	0.36*
Mass	—	—	—	—	0.48*
Posterior Prob.[b]	λ_S	ω	Maturity	Mass	Longevity
λ_S	—	0.01*	0.27	<0.01*	0.01*
ω	—	—	0.33	>0.99*	0.99*
Maturity	—	—	—	>0.99*	>0.99*
Mass	—	—	—	—	>0.99*

[a]Covariances estimated using the geodesic averaging procedure, and $\kappa = 10$. Asterisks indicate a posterior probability of a positive covariance smaller than 0.025 or greater than 0.975.
[b]Posterior probability of a positive covariance.
*Posterior probability >0.975 or <0.025.

Lartillot and Poujol, 2011, Mol Biol Evol, 28:729

Bayesian models in macro-evolutionary studies

Why Bayesian?

- integrating uncertainty over high-dimensional nuisances
- integrating multiple levels of macro-evolutionary processes
- complex models requiring sophisticated MCMC
- the RevBayes project (Hoehna et al, 2016, Syst Biol, in press)

Which Bayesian paradigm?

- mostly uninformative priors on top-level parameters
- meant for 'automatic' application to various problems
- increasingly large datasets available: effectively asymptotic
- Objective / Hierarchical / Empirical Bayes – not Subjective Bayes

Codon model with global $\omega = dN/dS$

- applied independently across many genes
- for each gene, point estimate and 95% CI for ω
- selecting genes for which $p(\omega > 1 \mid D) > c$

Codon model with site-specific effects

- for each site within a gene, point estimate and 95% CI for ω_i
- selecting sites for which $p(\omega_i > 1 \mid D) > c$

Comparative multivariate Brownian model

- over time, applied to a variey of problems
- point estimate and 95% CI for correlation between traits r
- usually, focus on whether $p(r > 0 \mid D)$ or $p(r < 0 \mid D) > 1 - \alpha$

A simple toy-example

Expression data transcriptome-wide

N genes. For gene $i = .1..N$:

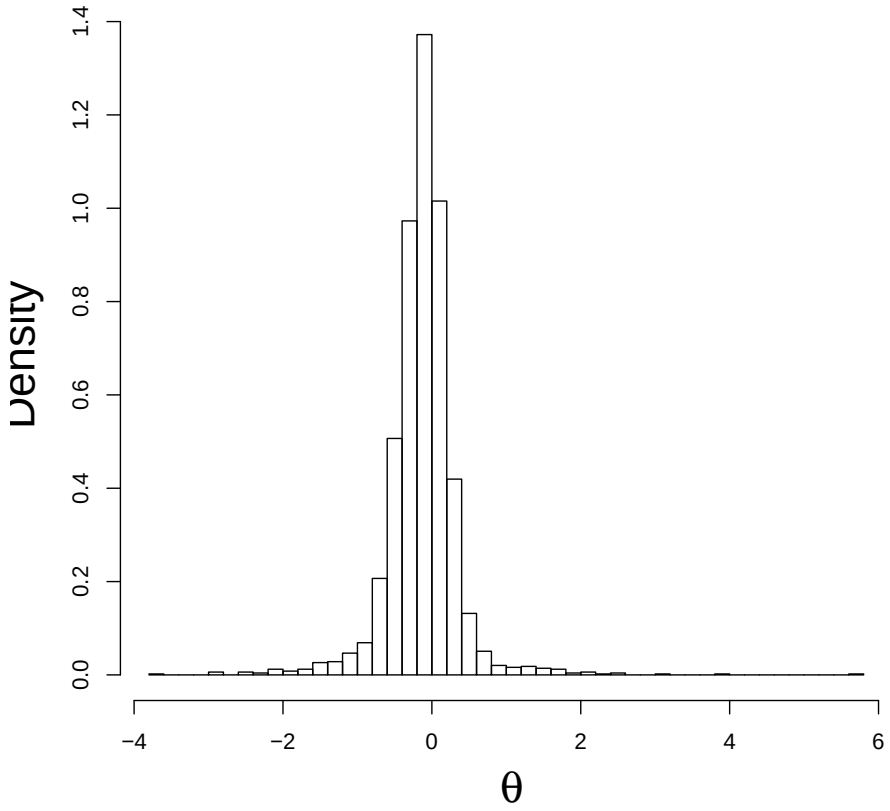
- x_i: measured differential expression (log ratio)
- θ_i^*: true differential expression

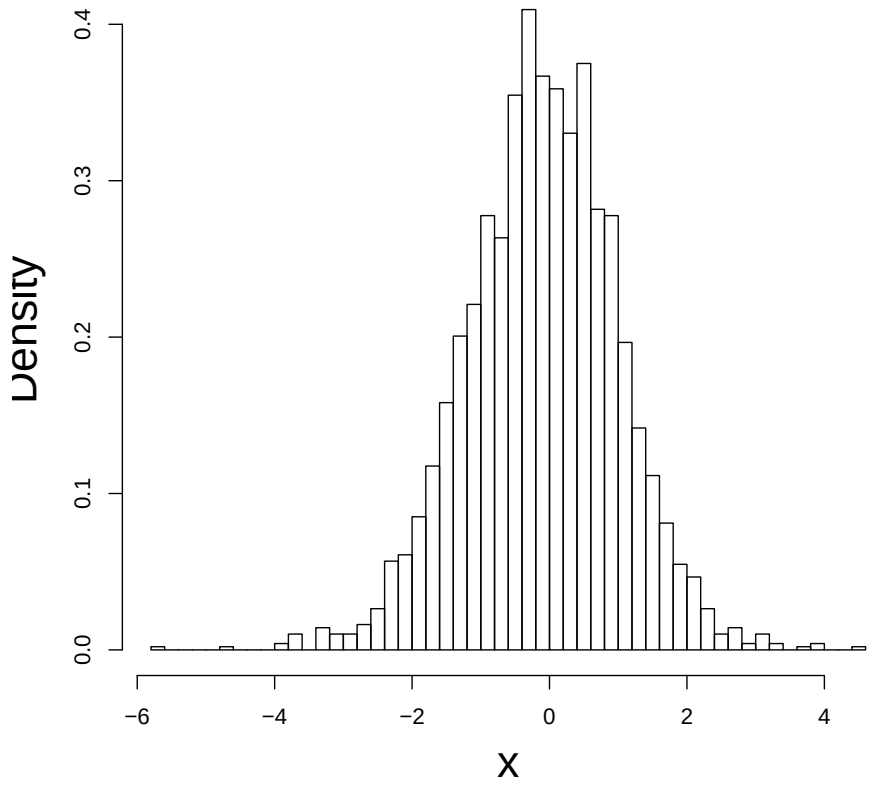
$$x_i \sim Normal(\theta_i^*, 1)$$

Two alternative inference schemes

- separate inference: each item (gene) considered individually
- joint inference: all items jointly analyzed (hierarchical model)
- frequentist properties of our inference and our selection ?

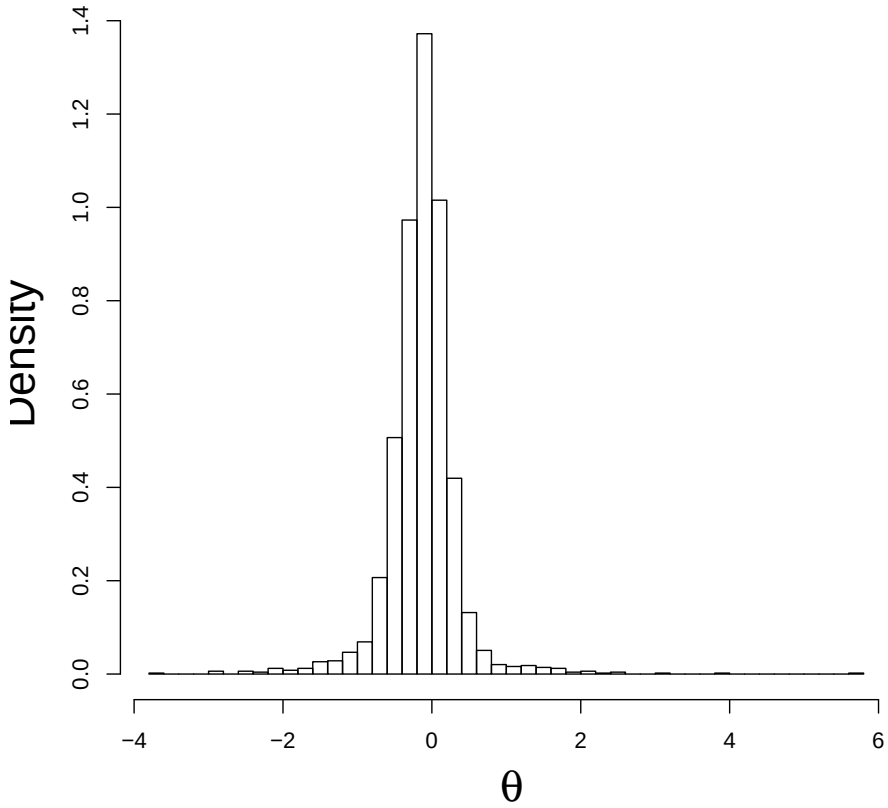
Toy example using empirical gene expression data

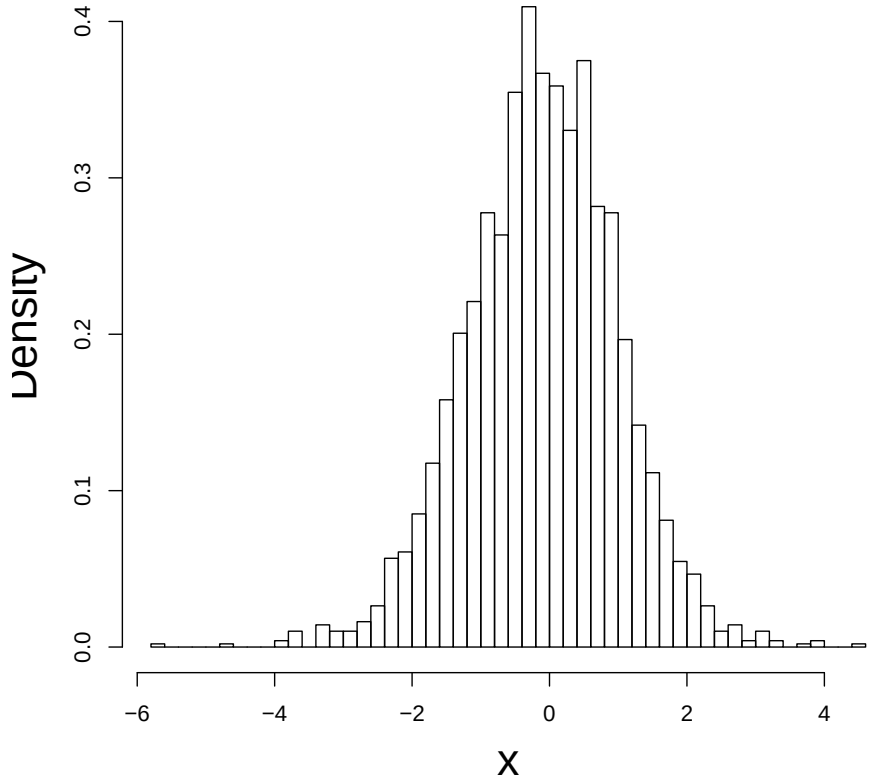




- data (right) simulated using empirical collection of θ_i^*'s (left)
- obtained from experimental gene expression data

Separate inference with uninformative prior





- true value is covered by 95% CI in 2272 cases out of 2393 (94%)
- 13 out of 2393 cases such that $p(\theta_i > 1.1 \mid X_i) > 0.95$
- 7 of them are such that true $\theta_i^* > 1.1$

Coverage versus calibration

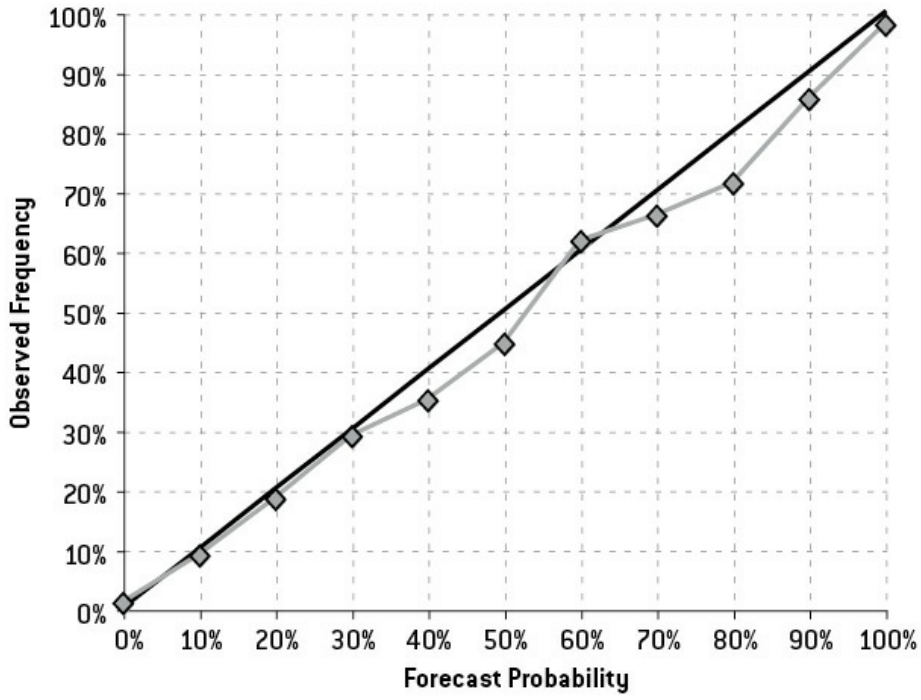
Coverage

- given: a confidence level $1 - \alpha$
- x is observed
- make a statement about θ (e.g. $3.90 < \theta < 6.10$)
- coverage: your statements are indeed true at a frequency $1 - \alpha$
- honest account of uncertainty in pure inference

Calibration

- given: a question about θ (e.g. is $\theta > 1.1$?)
- x is observed
- give your probability that answer to question is yes
- calibration: advertised probabilities = frequency of being correct
- more useful than coverage in a decision making context

Bayesian calibration



Nate Silver, The Signal and the Noise

Bayesian calibration

- advertised posterior probabilities = frequency of being correct
- more generally: implies posterior expected loss = true loss
- implies good control of true/false discovery rate

Empirically assessing calibration

for a given interval A (e.g. $A = (1.1, +\infty)$)

- define selected subset: $S_A(\alpha) = \{i,\ p(\theta_i \in A \mid X) > 1 - \alpha\}$
- compute nominal (or advertised) true discovery rate:

$$q_A(\alpha) = \frac{1}{|S_A(\alpha)|} \sum_{i \in S_A(\alpha)} p(\theta_i \in A \mid X)$$

- compute true discovery rate:

$$r_A(\alpha) = \frac{1}{|S_A(\alpha)|} \sum_{i \in S_A(\alpha)} \mathbb{1}[\theta_i^* \in A]$$

- calibration: $q_A(\alpha) = r_A(\alpha)$

Example based on simulations

- $N = 10000$ simulated genes
- $\theta_i^* \sim Normal(0, 3)$
- $x_i \sim Normal(\hat{\theta}_i, 1)$
- TDR cutoff: $1 - \alpha = 0.70$

prior variance	m.s. error	coverage (95% CI)	advertised TDR	TDR
$\sigma = 1$	2.78	0.58	-	-
$\sigma = 3$	0.94	0.95	0.86	0.86
$\sigma = 100$	1.04	0.96	0.88	0.81

Minimaxity

Worst-case risk

given a prior π:

- for any θ, define *frequentist* risk associated to π: $R(\pi, \theta)$
- find the worst-case risk (over θ)

$$R_{max}(\pi) \quad = \quad Max_{\theta}\, R(\pi, \theta)$$

Minimax prior

- find π^* which minimizes worst-case risk

$$\pi^* \quad = \quad ArgMin_{\pi}\, R_{max}(\pi)$$

- in many simple situations, leads to classical uninformative priors
- minimax, maximin, and maximum entropy priors

Simple normal model on θ

prior	$p(\theta)$	$\sim$	$Normal(0, \sigma^2)$
likelihood	$p(x \mid \theta)$	$\sim$	$Normal(\theta, 1)$
posterior	$p(\theta \mid x)$	$\sim$	$Normal\left(\frac{\sigma^2}{1+\sigma^2}x, \frac{\sigma^2}{1+\sigma^2}\right)$

Minimax: $\sigma \to \infty$

prior	$p(\theta)$	$\sim$	$Uniform(-\infty, +\infty)$
likelihood	$p(x \mid \theta)$	$\sim$	$Normal(\theta, 1)$
posterior	$p(\theta \mid x)$	$\sim$	$Normal\,(x, 1)$

- posterior credible interval: (x - 1.96, x + 1.96)
- identical to classical frequentist confidence interval

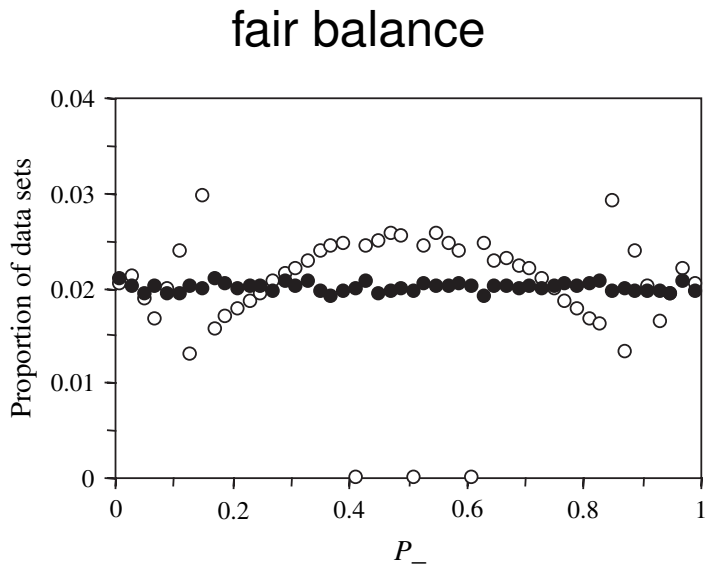
Objective Bayes controls for type I error

Selecting over-expressed genes

- H_0: $\theta_i \leq 1.1$ versus H_1: $\theta_i > 1.1$
- rejection of H_0 whenever one-sided 95% CI does not cover 1.1

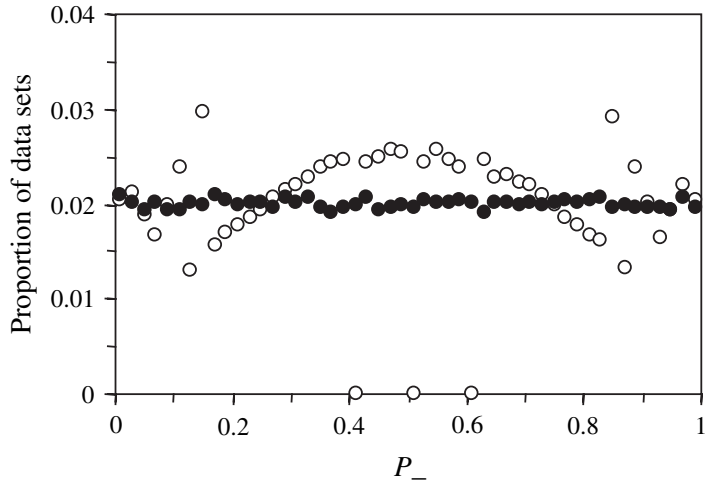
- imagine that, $\forall i = 1..N$, $\theta_i^* = 1.1$.
- H_0 rejected 5% of the times
- under objective Bayes, $p(H_0 \mid x_i)$ is in fact a p-value

The Fair-balance and the Star-tree 'paradoxes'

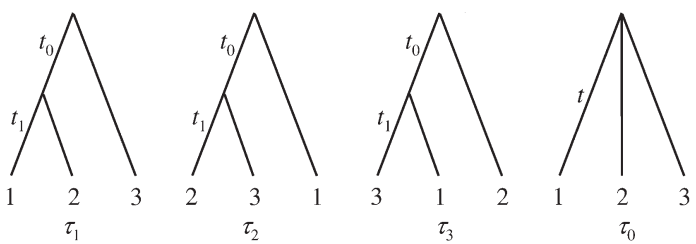


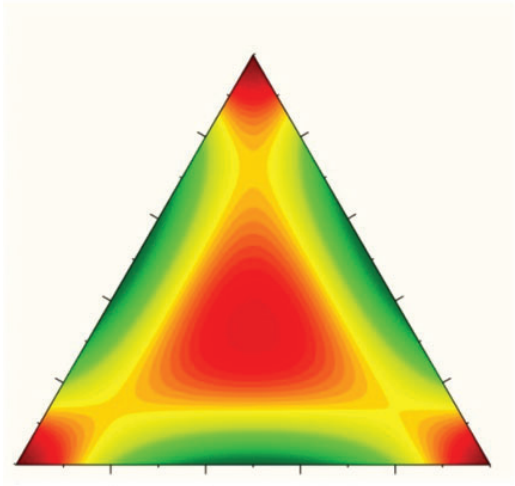
The Fair-balance and the Star-tree 'paradoxes'

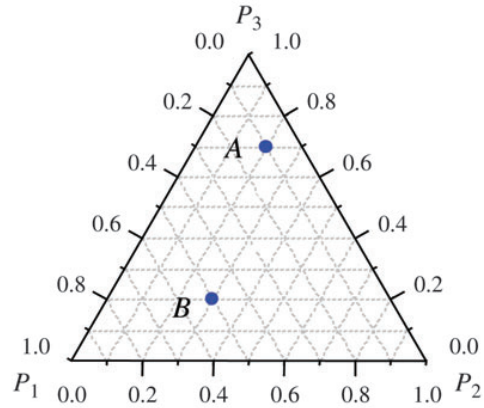
fair balance



star tree







Ziheng Yang, 2007, Mol Biol Evol, 24:1639

Objective Bayes

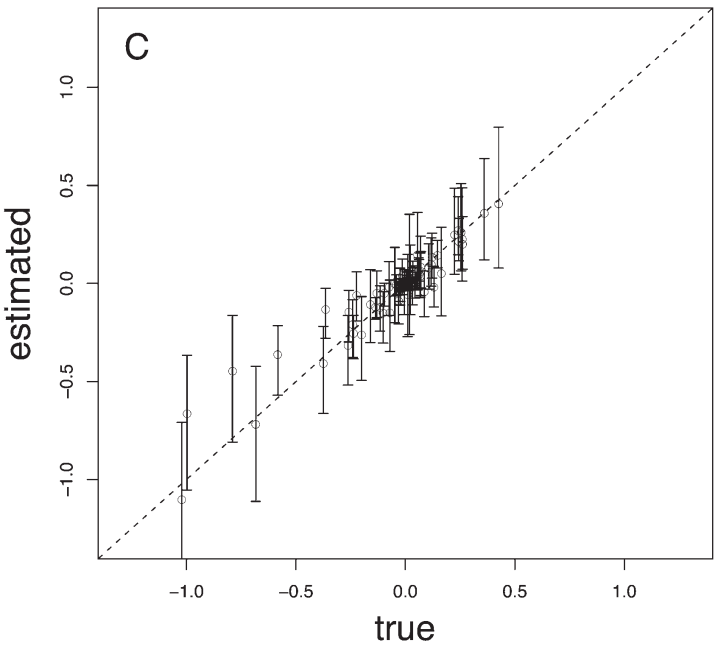
- non-informative priors are minimax
- Objective Bayes is closer to classical frequentism
- controls for type I error
- not well-calibrated

More general asymptotic results

- von Mises theorem: asymptotic normality of posterior
- credible intervals are asymptotic confidence intervals ($O(1/\sqrt{N})$)
- with objective priors: asymptotic convergence at least in $O(1/N)$

Empirical assessment of comparative model

coverage



type I error

Table 1. Rate of False Positives.[a]

	α				
Averaging Method	**0.100**	**0.050**	**0.010**	**0.001**	**0.0001**
Arithmetic	**0.050**	**0.022**	**0.002**	**0.001**	**0.000**
Geodesic	**0.049**	**0.021**	**0.000**	**0.000**	**0.000**

[a]Frequency, over 100 simulations under the diagonal model at which the posterior probability of a positive covariance is less than $\alpha/2$ or greater than $1 - \alpha/2$ (see text for details).

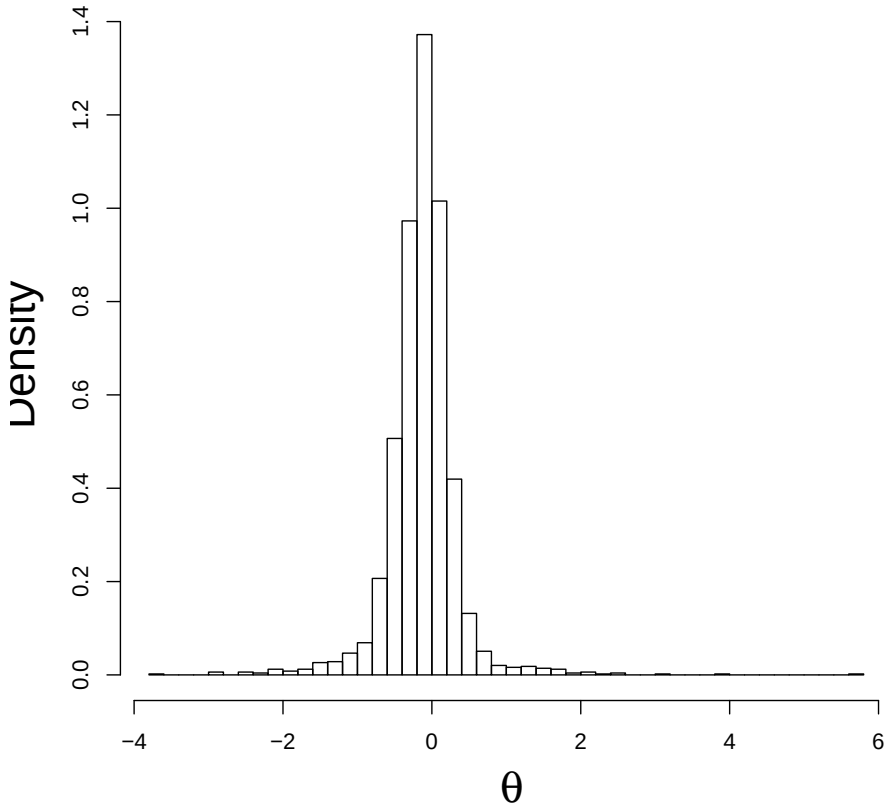
Lartillot and Poujol, 2011, Mol Biol Evol, 28:729

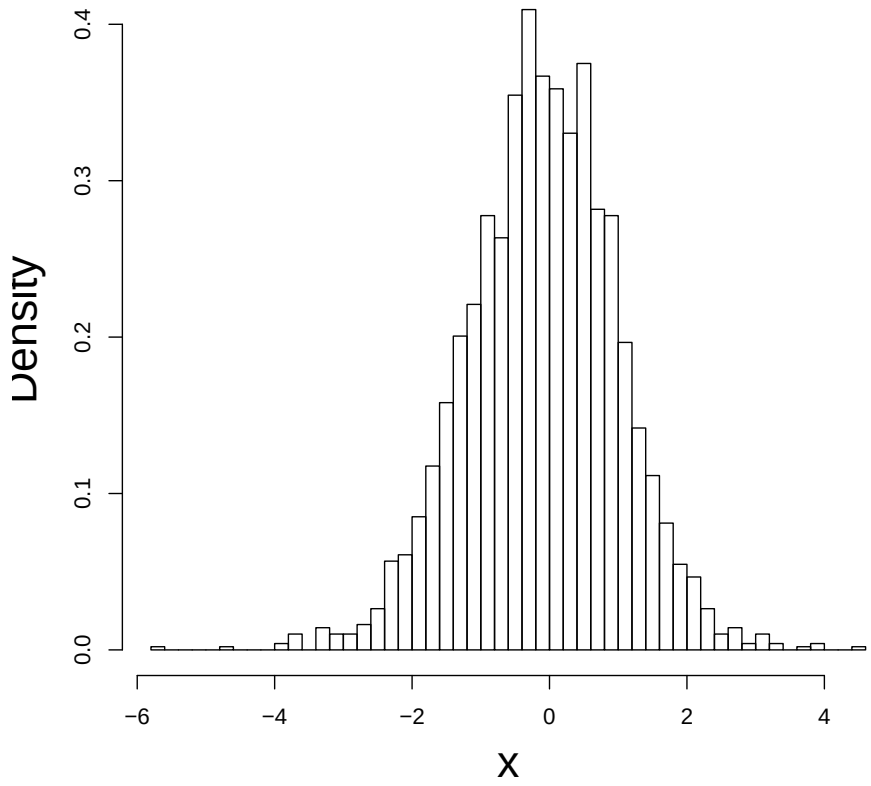
Example based on simulations

- $N = 10000$ simulated genes
- $\theta_i^* \sim Normal(0, 3)$
- $x_i \sim Normal(\theta_i^*, 1)$
- TDR cutoff: $1 - \alpha = 0.70$

prior variance	m.s. error	coverage (95% CI)	advertised TDR	TDR
$\sigma = 1$	2.78	0.58	-	-
$\sigma = 3$	0.94	0.95	0.86	0.86
$\sigma = 100$	1.04	0.96	0.88	0.81
$\bar{\sigma} = 2.99$	0.95	0.94	0.86	0.87

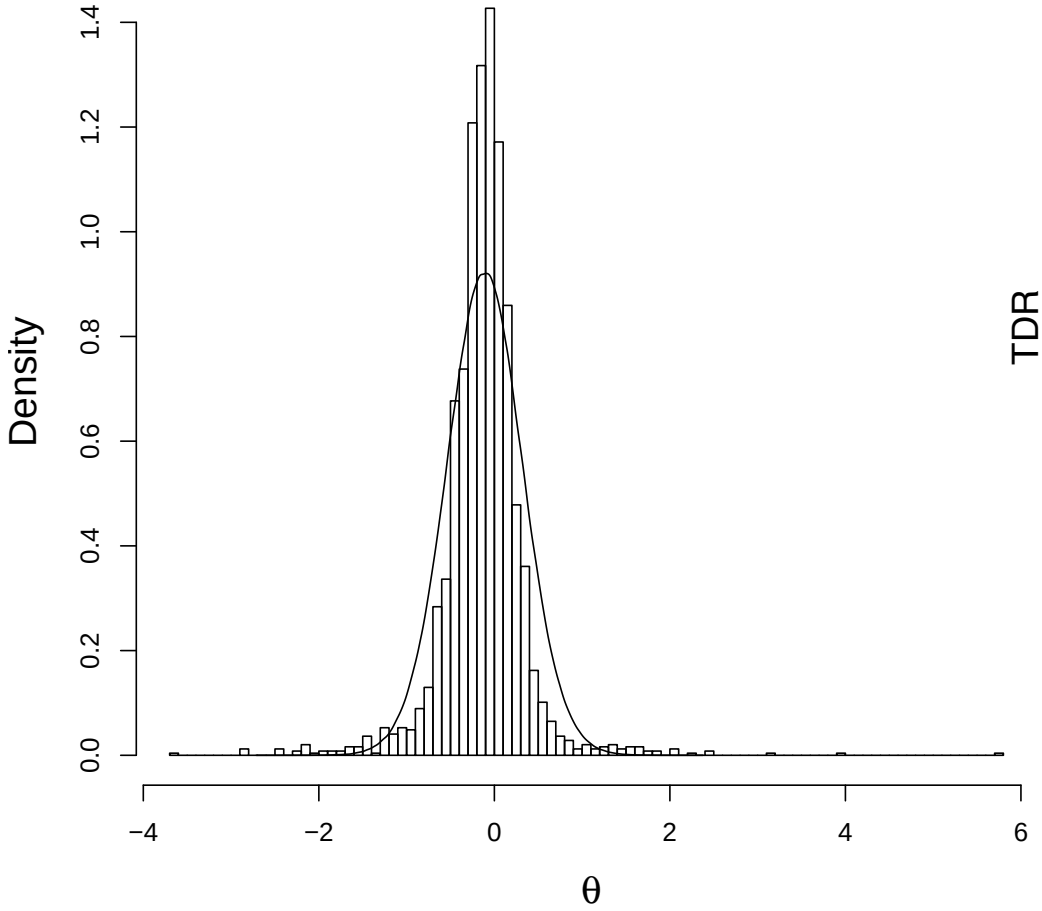
Example. Empirical gene expression data

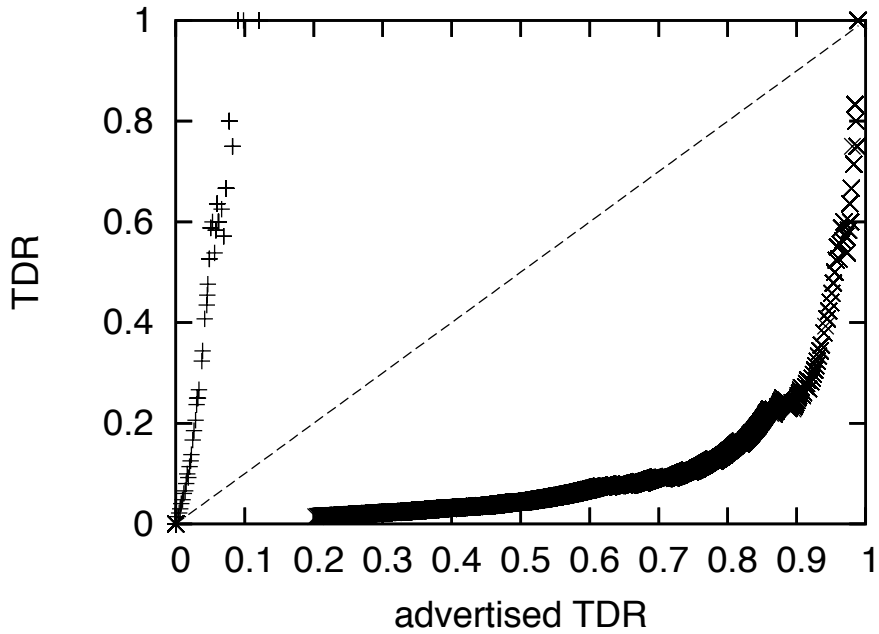




- data (right) simulated using empirical collection of θ_i^*'s (left)
- obtained from experimental gene expression data

Calibration under parametric (normal) model





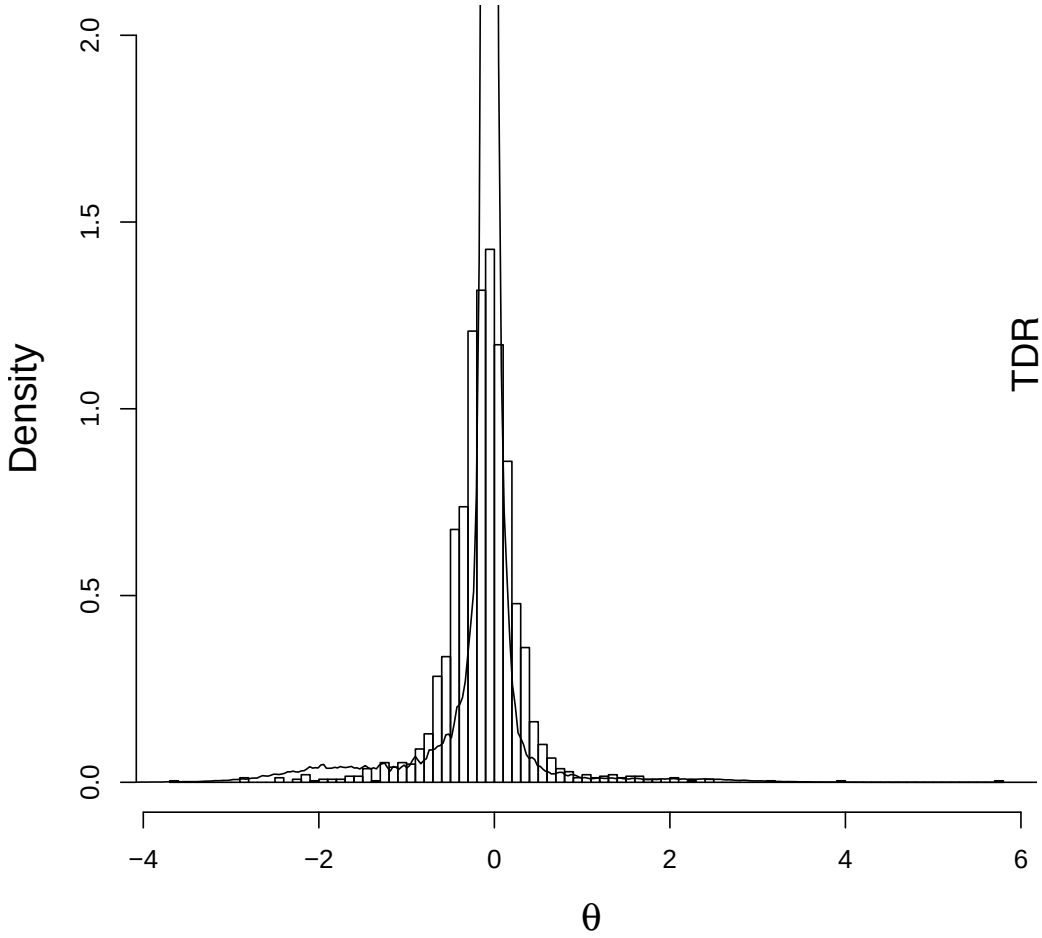
Stick-breaking representation (Sethuraman)

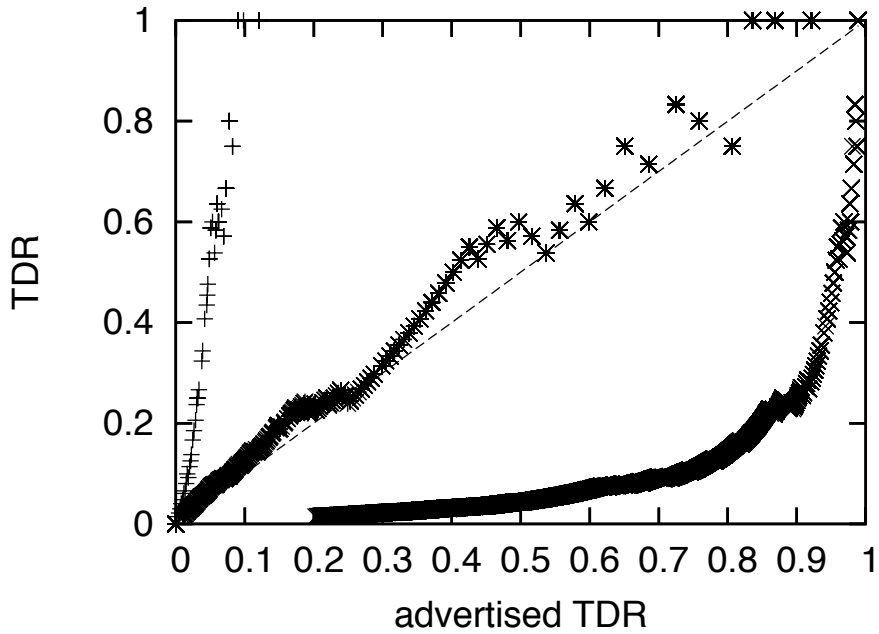
$$
\begin{aligned}
j = 1, 2, \ldots \qquad & Y_j \sim Beta(1, \alpha) \\
& p_j = \prod_{k<j} (1 - Y_k)\, Y_j \\
& \theta_j \sim G_0
\end{aligned}
$$

$$
G = \sum_j p_j \delta_{\theta_j}
$$

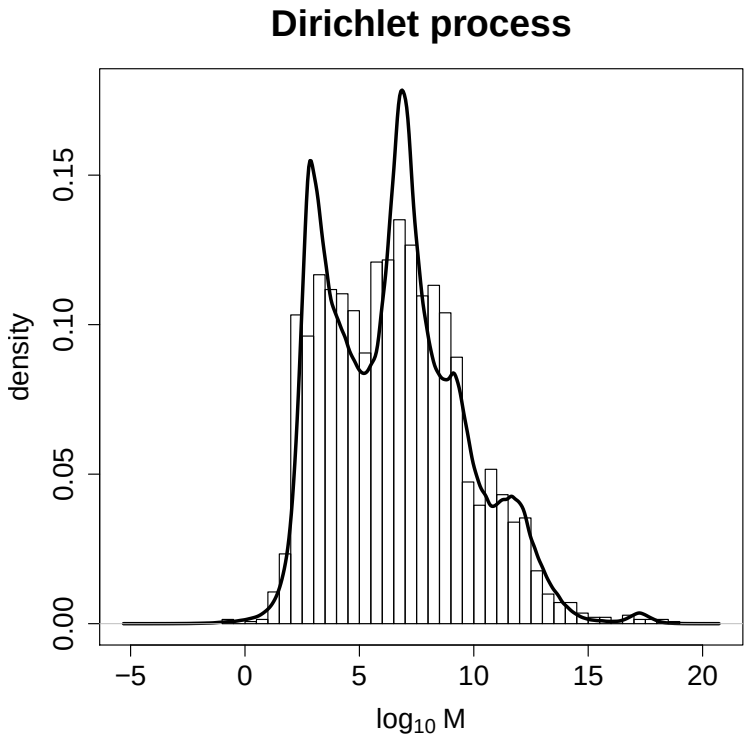
- $G \sim DP(\alpha G_0)$: infinite mixture
- infinite mixtures dense in space of distributions
- defines a non-parametric prior over distribution space
- MCMC over components represented in the data sample

Calibration – non-parametric model (Dirichlet process)



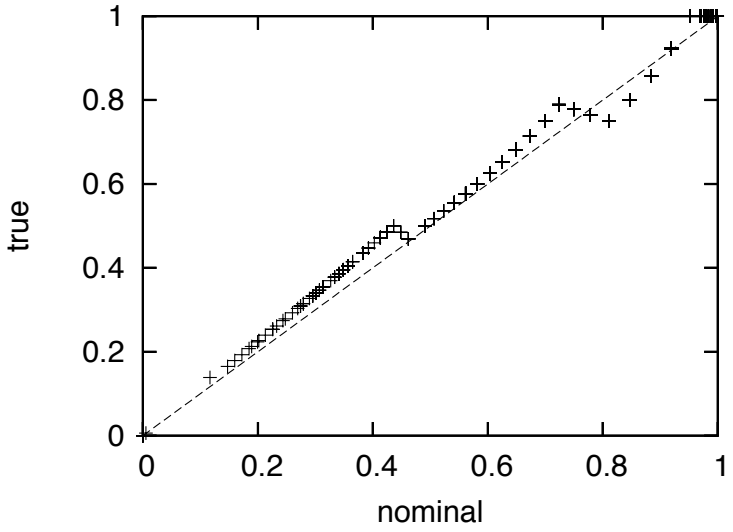


Calibration: log body size in mammals

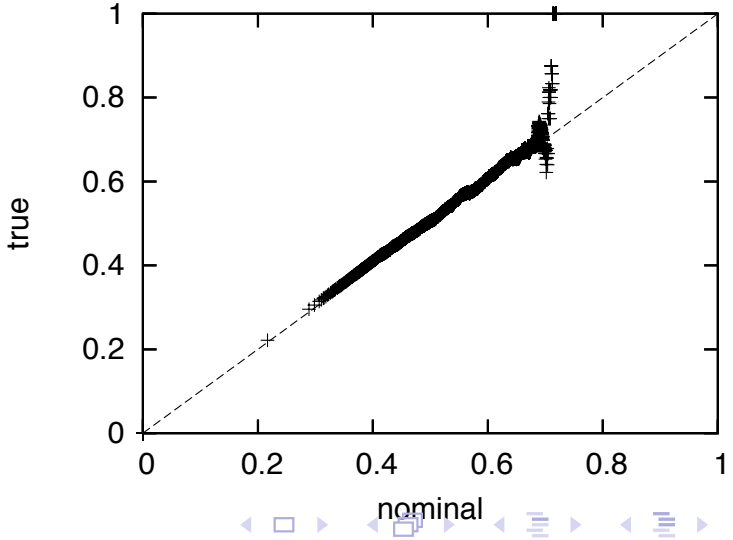


- $X_i \sim Normal(\theta_i^*, 1)$
- $\theta_i^* = \log_{10} M_i$

$A = (15, 20)$



$A = (3, 5)$



The dual frequentist meaning of posterior probabilities

Objective and simple (non-hierarchical) Bayes

- objective Bayes: fundamentally a *classical* frequentist meaning
- can be formalized in terms of minimaxity
- asymptotic coverage and control for type-I error – *not* calibration
- posterior probability semantics misleading here

Hierarchical or empirical Bayes

- borrow information across X_i's to estimate true distribution of θ_i's
- calibration (FDR control) on θ
- calibration fundamentally requires *shrinkage*
- big data, genomics: promising domains for using empirical Bayes
- non-parametric approach: general, but fragile and intensive

A short history of Bayesian inference (1)

Original goal (Bayes and Laplace)

- develop a language of *probabilistic inference*
- formulated in terms of prob. of hypotheses given observations
- Bayes theorem:

$$p(\theta \mid D) \propto p(D \mid \theta) p(\theta)$$

- turns out to depend on a prior – want it or not

Frequentist critique

- Fisher: uninformative priors ill-defined
- Neyman: only thing that can be controlled is type I error
- led to the classical frequentist paradigm

A short history of Bayesian inference (2)

Subjective Bayes (Savage and de Finetti)

- logical formalisation of personal beliefs
- making use of prior information
- don't claim to have any objective frequentist guarantees

Objective Bayes

- good formal definition of uninformative priors (minimaxity)
- best Bayesian proxy of classical frequentism

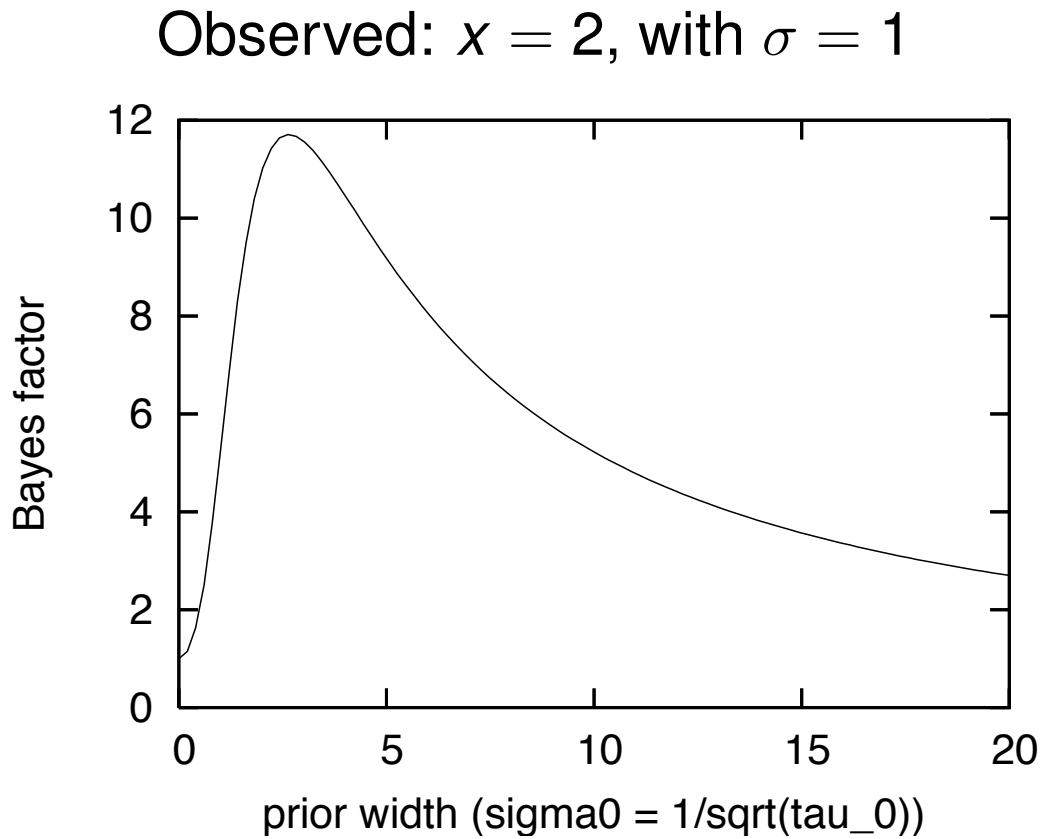
Empirical Bayes (Robbins, James, Stein)

- 1995: Benjamini and Hochberg (BH): false discovery rate
- Efron: BH method implicitly based on empirical Bayes argument
- realization that multiple settings carry with them their own prior

Bayes factor

Testing a point null under normal model

$$B = \frac{p(X \mid \theta \neq 0)}{p(X \mid \theta = 0)}$$



Compound Bayes

Tentative formalization of asymptotic calibration

- an infinite, non-random sequence $(\theta_i)_{i \in \mathbb{N}}$
- a random observable sequence $X_i \sim p(X_i \mid \theta_i)$
- for any interval A, $N \in \mathbb{N}$ and $\alpha \in (0, 1)$:
- define $q_A^N(\alpha)$, $r_A^N(\alpha)$ as previously, based on first N observations
- define calibration error:

$$\epsilon_A^N(\alpha) \quad = \quad q_A^N(\alpha) - r_A^N(\alpha)$$

- behavior of $\epsilon_A^N(\alpha)$ for large N?
- conditions on $(\theta_i)_{i \in \mathbb{N}}$ for which $\epsilon \to 0$ in some useful sense?