PM 520 - Lecture 11

Examinable assignment 4 - part 1

- Use the Urn model starting with 2 red balls of weight 1, and one black (mutation) ball of weight w.
- Draw balls until you have 10 non-black balls.
- If all non-black balls are the same color at the end, what is the posterior distribution of the weight of the black ball?
- If we observe exactly 2 non-black colors at the end, what is the posterior distribution of the weight of the black ball?
- Use a Uniform[0,20] prior for the weight of the black ball

Bayesian Importance Sampling 2

- Suppose we have parameter(s), θ, prior π, and importance sampling distribution ξ, and observed data D. Iterate the following:
    1. Sample θ' from ξ, and simulate data D' using θ'.
    2. Accept θ' if D'=D. Otherwise reject θ' and return to 1.
    3. Add a point with **(Importance Sampling) weight** of $\pi(\theta)/\xi(\theta)$ to the posterior at θ'
    4. Rejection method: sample from prior $\pi(\theta)$ and construct posterior;

Result: empirical estimate of posterior distribution P(θ|D).
The sampler will be more efficient than naive rejection.
How do we choose ξ?

-

Examinable assignment - part 2

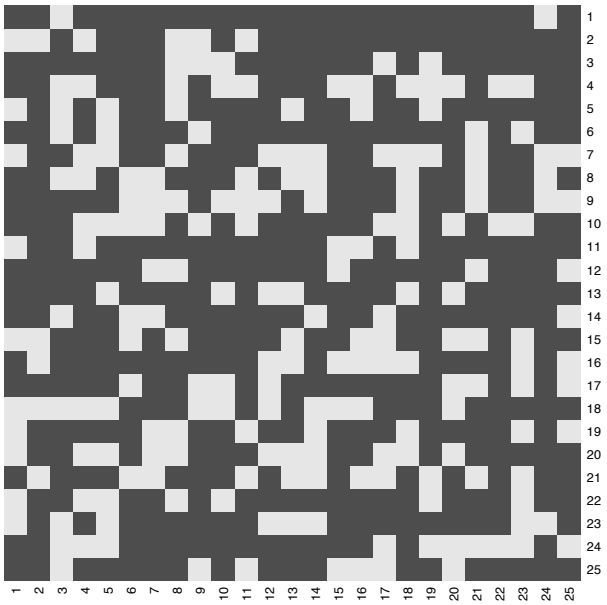
- Repeat the Urn model assignment to find the posterior **mass** of the black ball given that we observe just one (non-black0 color in an Urn containing 10 (non-black) balls.
- This time, sample θ' (black ball **mass**) from ξ, an expo(λ) distribution, rather than a Uniform distribution. (So this is Importance sampling.)
- If you accept θ' then add a point with (**Importance Sampling**) **weight** $\pi(\theta')/\xi(\theta')$ to the posterior distribution of θ.
- If we restrict θ' to be in the interval [0,20] (for simplicity's sake), then this means that we have:
  - $\pi(\theta') = 1/20$
  - $\xi(\theta') = \lambda e^{-\lambda\theta'} / (1-e^{-\lambda 20})$
- Compare efficiency to a rejection method that samples directly from a Uniform[0,20] distribution, by comparing the number of iterations you need to run in order to collect 10000 accepted θ's.

Remember: pseudocode on Blackboard "Pseudocode_Urn.R"

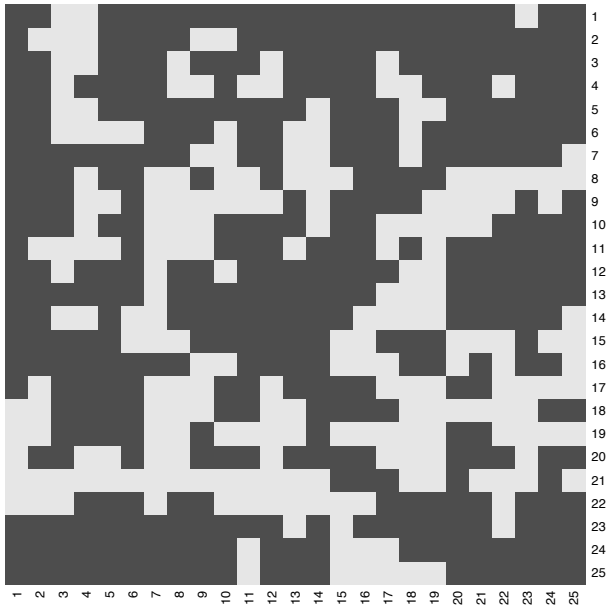
Examinable Assignment - Part 3

- Many tissues contain 2 (or more) cell-types.
- Investigator wants a way of testing whether each of 2 cell types in a given tissue is homogeneously distributed (in space). So:
  - H_0: cell types are homogeneously (randomly) arranged
  - H_1: cell types are not homogeneously arranged (i.e. they are clustered in some fashion).
- Your job, is to come up with such a test, using Monte Carlo methods.
  1. Formulate your test.
  2. Apply it to each of Grid1, Grid2, Grid3 and determine a p-value for rejecting the H_0 in favor of H_1.
  3. Write-up your test (i.e. Methods) and your Results.

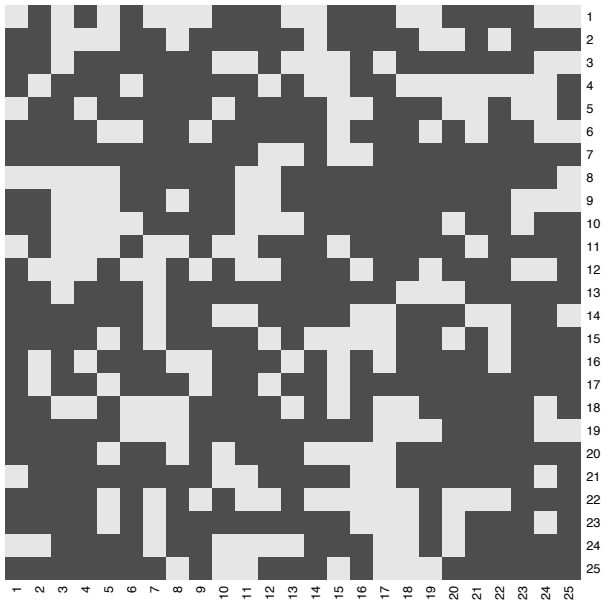
Examinable Assignment - Part 3



"Grid1.txt"



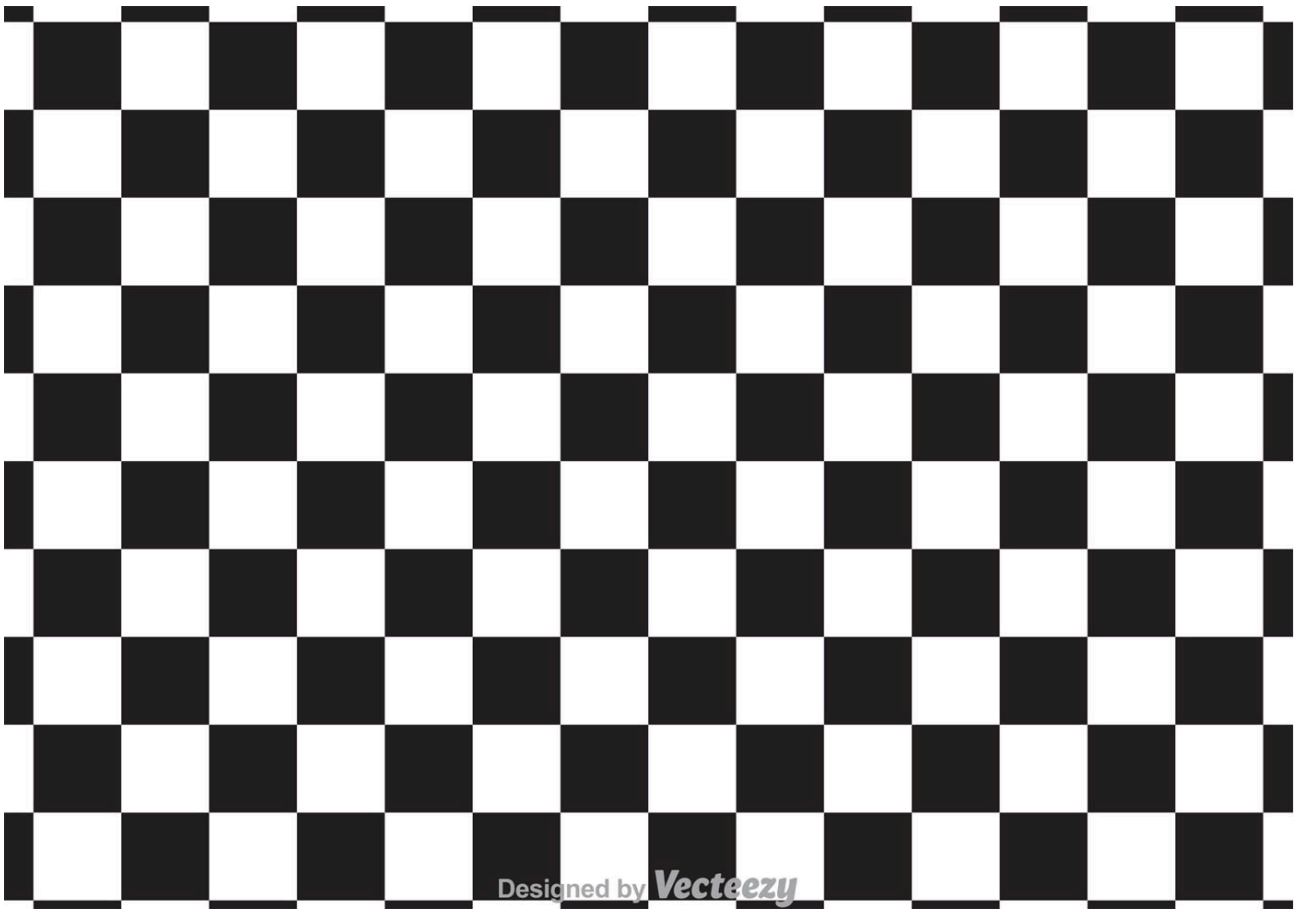
"Grid2.txt"



"Grid3.txt"

Something to think about

- What would your test conclude about this grid?…
-

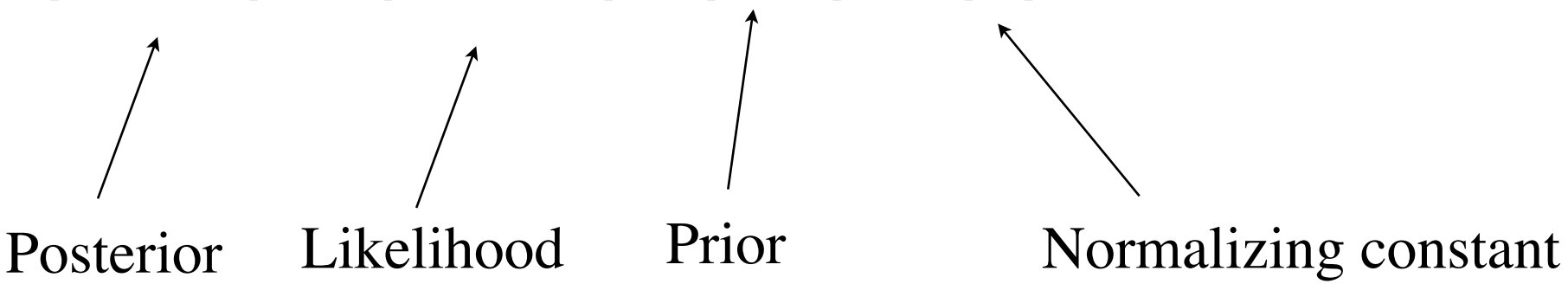


Bayesian analysis

- Given data D,
- Parameter(s) θ,
- Model M

- Wish to make inference re. f(θ|D).

- $f(\theta,M|D) = f(D|\theta,M)\,\pi(\theta,M) / f(D)$



Methods for obtaining f(θ|D)

- Likelihood estimation (Monte Carlo sim.)
- Rejection methods

proposed param. values not conditional on data

- Importance Sampling

- Markov chain Monte Carlo [MCMC] methods.

conditional on data

- All are “model-based”.

But what if we can't calculate P(D|θ'), or what if P(D'=D)~0 even when θ'=θ?….

Approximate Bayesian Computation

http://www.igfasouza.com/blog/what-is-big-data/

Intractability of likelihood

- Two possible responses:
  - Simplify the model so that it again becomes possible to calculate the likelihood;
  - Simplify the analysis by approximation.

"All models are wrong - some are useful." (Box)

"Far better an approximate answer to the right question, which is often vague, than an exact answer to the wrong question, which can always be made precise." (Tukey).

Rejection method III - Approximate Bayesian Computation

Suppose we have observed data D, a good summary statistic(s) of that data, S, and a model with parameter(s) θ that describes how the data got there. Do the following:

1. Sample parameter(s) θ' from prior π.
2. Simulate D' using θ'. Calculate S', the summary of D'.
3. Accept θ' if S'=S.
4. Return to 1.

Result: independent samples from f(θ|S)

Best case scenario: if S is *sufficient* for θ, then we have f(θ|S)= f(θ|D). [c.f. Last week's assignment.]

More generally - (Approximate Bayesian Computation)

Suppose:

A set of (normalized) summary statistics $S=\{S_1,\ldots,S_n\}$, that take values $S^o=\{S^o_1,\ldots,S^o_n\}$ on the observed data.

(A set of weights $w_1,\ldots,w_n$).

A *tolerance* ε.

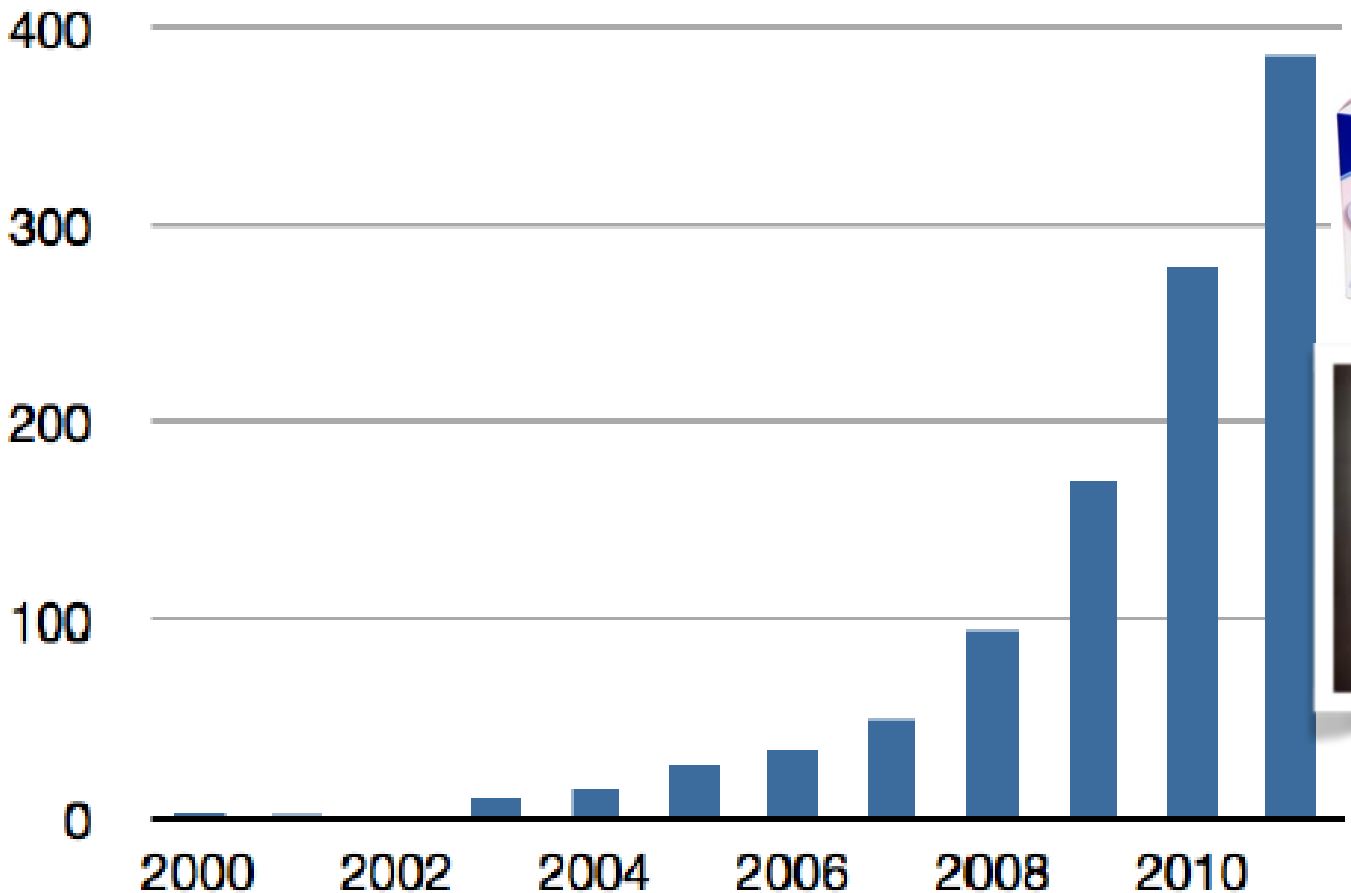
1. Sample *θ'* from prior π.
2. Simulate D′ using *θ'*. Calculate S^s (the value of S on D′)
3. Accept *θ'* if $\sum_i w_i(S_i^o - S_i^s)^2 < \varepsilon$
4. Return to 1.

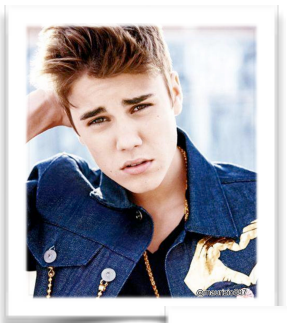
Results: independent samples from something we will call $\varphi(\theta|D)$.

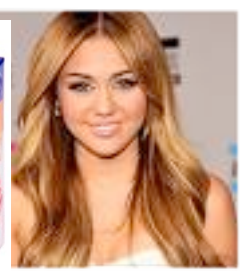
ABC: General 'philosophy'

- Follow Tukey's intuition and get samples from $\varphi(\theta|D)$ rather than $f(\theta|D)$.
- Generates an approximate answer to the right (original) question.
- Relies upon intensive computational simulation.
- Lack of theory for closeness of approximation of $\varphi(\theta|D)$ to $f(\theta|D)$. [So usually demonstrate this via simulation study.]

ABC pubs per year



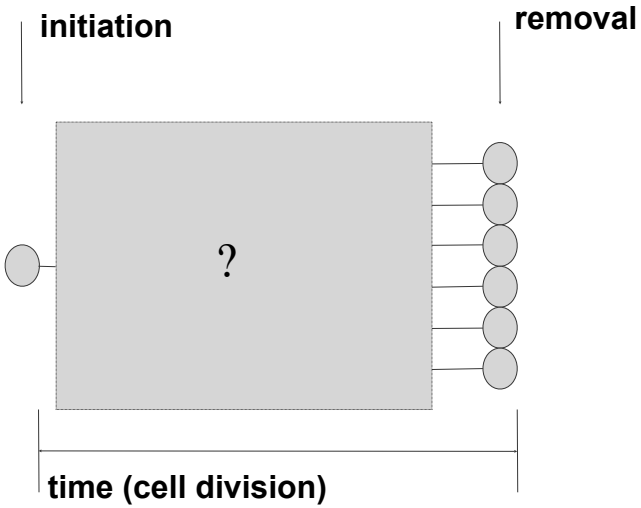


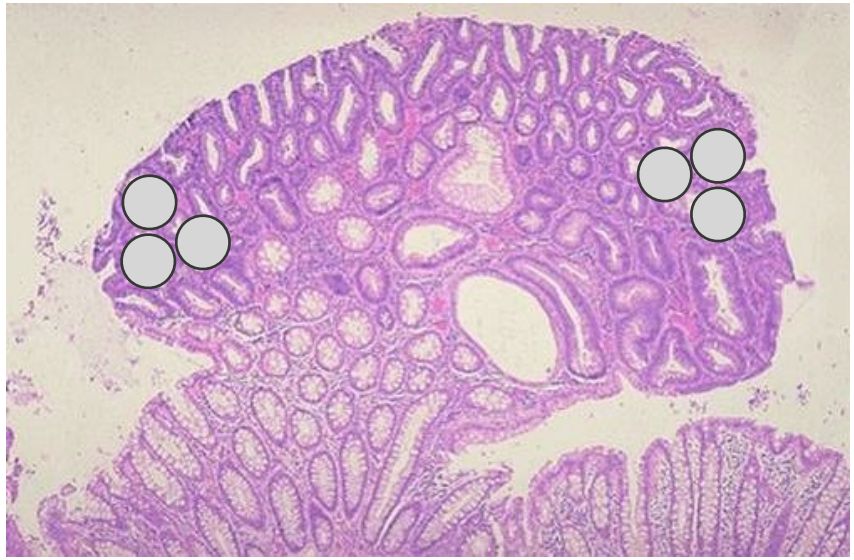


i'm lovin' it®

Example: Tumor History (Siegmund, Shibata, Curtis)

Cancer cell dynamics are not observed



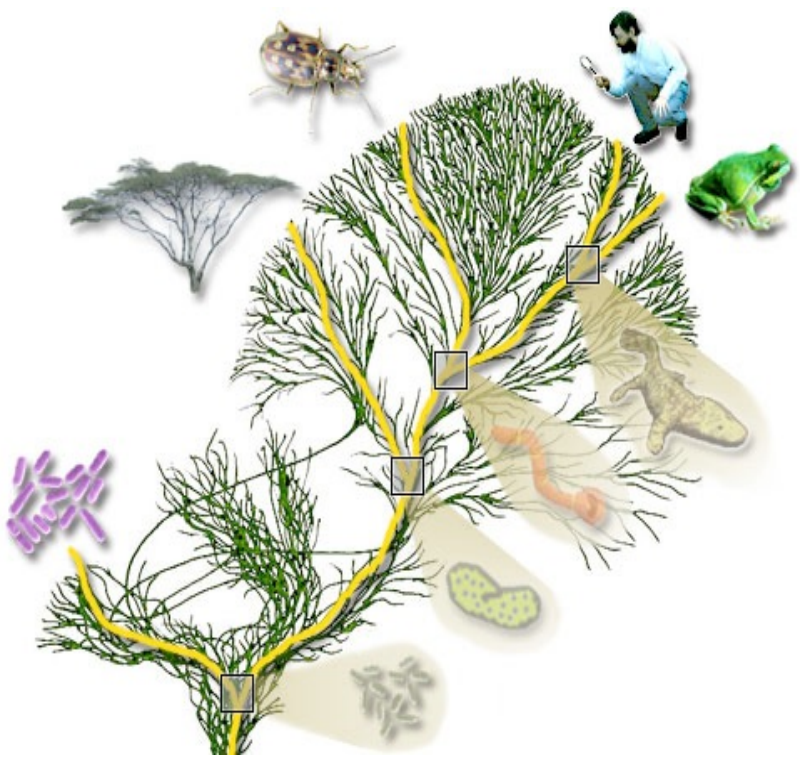


Basic questions:

1. How many times has the tumor undergone cell division?
2. How many long-lived tumor cells (cancer stem cells) are there?

Phylogenetic Trees

Living organisms sit at the tips of branches and are connected by a hierarchy of ancestors

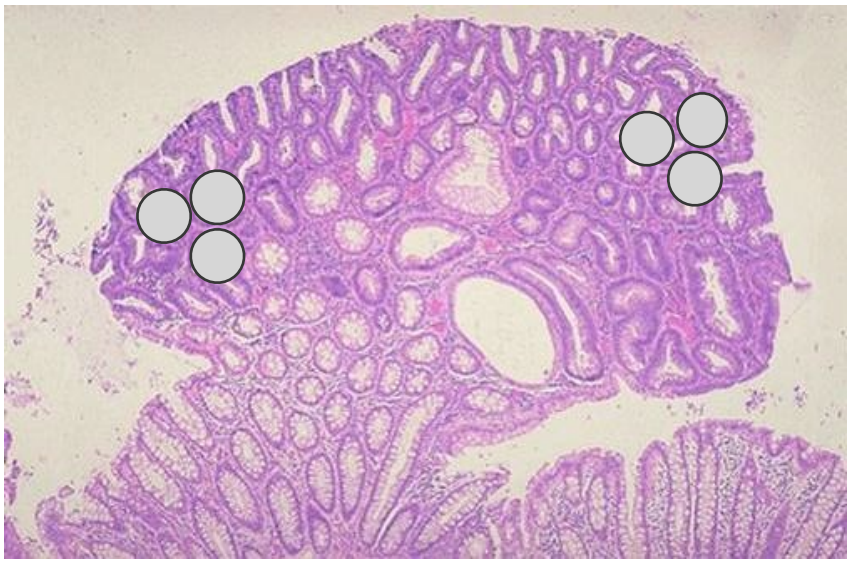


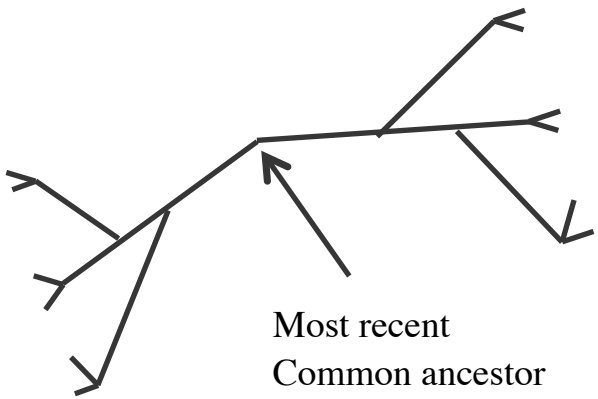
Genome comparisons can reconstruct genealogies of species

- Genomes are almost exact copies of prior copies
- Replication errors introduce small differences between generations

http://www.tolweb.org/tree/learn/concepts/whatisphylogeny.html

Somatic Cell Trees



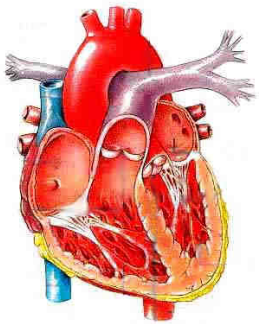


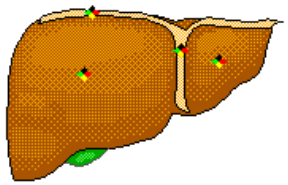
Epigenome comparisons can reconstruct genealogies from regions of a tumor

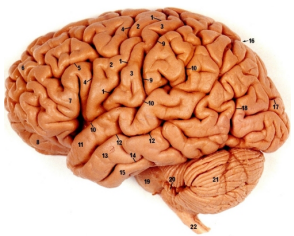
- DNA methylation patterns are inherited in cell division
- Replication errors introduce small differences between generations

Epigenetics

- Is derived from Greek for ‘upon’ genetics
- Each cell in our body has the same DNA
- Yet, different tissues perform different jobs

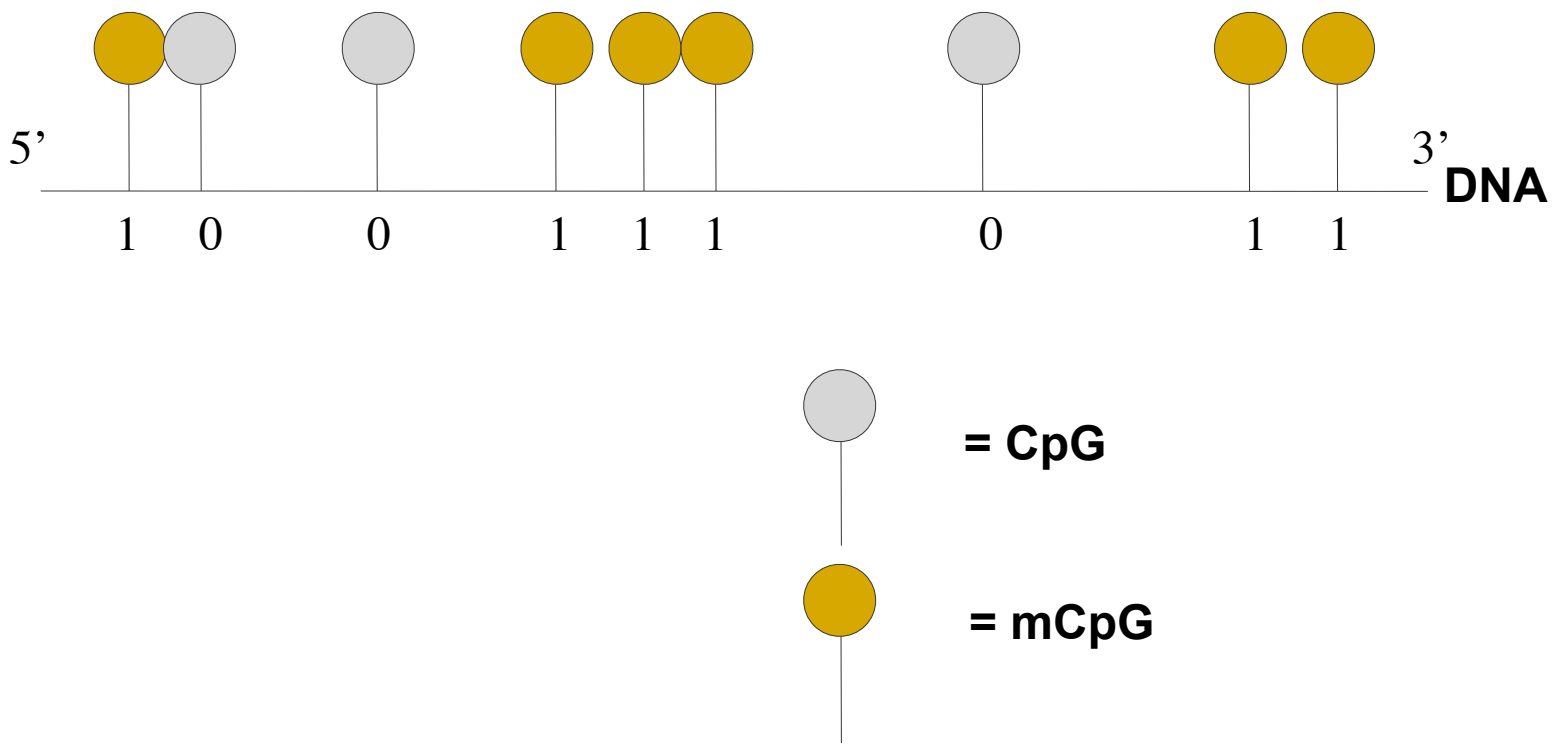






Measuring DNA methylation

Bisulphite Genomic Sequencing



DNA Methylation

A chemical modification of DNA that can silence gene expression

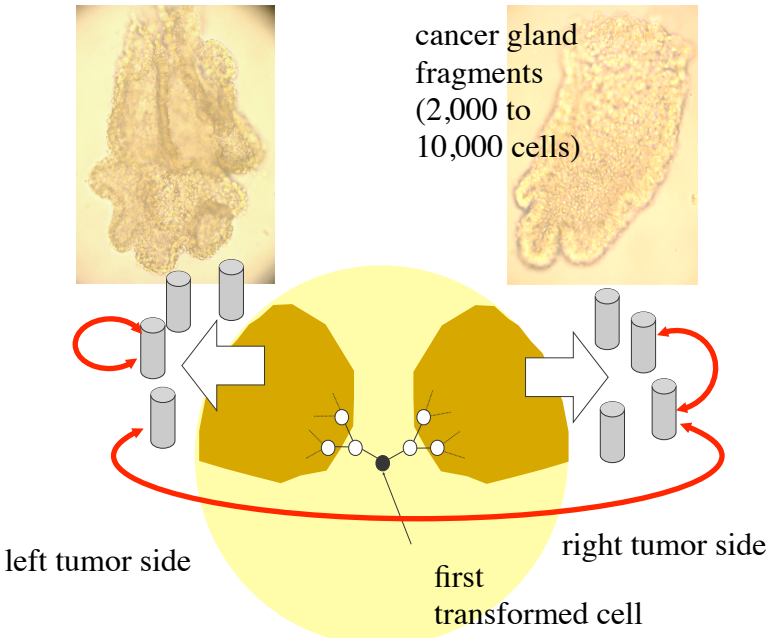
Normal function	Abnormal function
X-chromosome inactivation.	Silence tumor suppressor
Genetic imprinting.	genes in cancer.

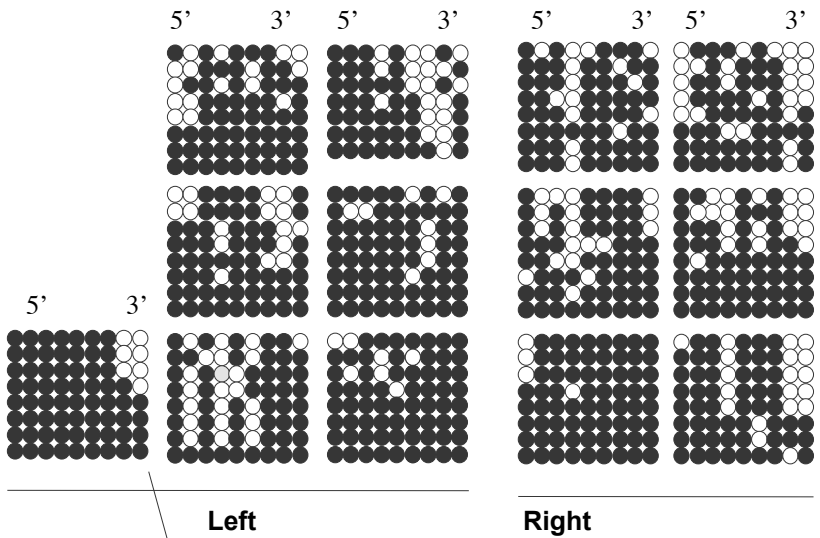
3 Key Features:

- inherited in cell division
- relatively stable
- 'mutates' at the right rate for our purpose

- The coloration of tortoise-shell (calico) cats is a visible manifestation of X-inactivation. The "black" and "orange" alleles of a fur coloration gene reside on the X chromosome. For any given patch of fur, the inactivation of an X chromosome that carries one gene results in the fur color of the other, active gene. That's why tortoise-shell cats are (almost always) female.

Summary statistics





Summary Measures

Percent methylated: 65/72=90%

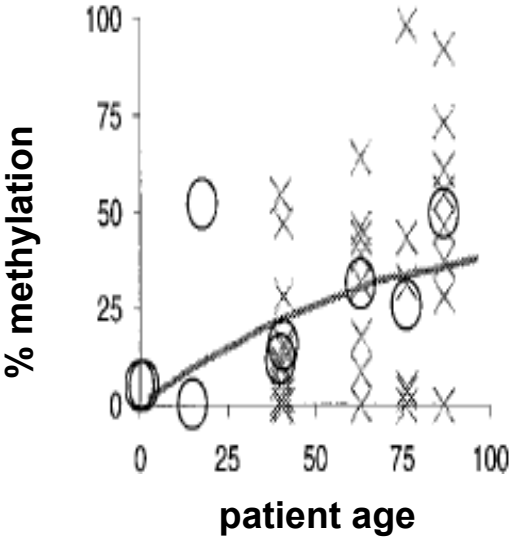
Number of unique patterns: =3

Avg. pairwise distance

intragland distance: 34/28=1.21

*inter*gland distance: 155/64=2.42

2 CpG-rich regions

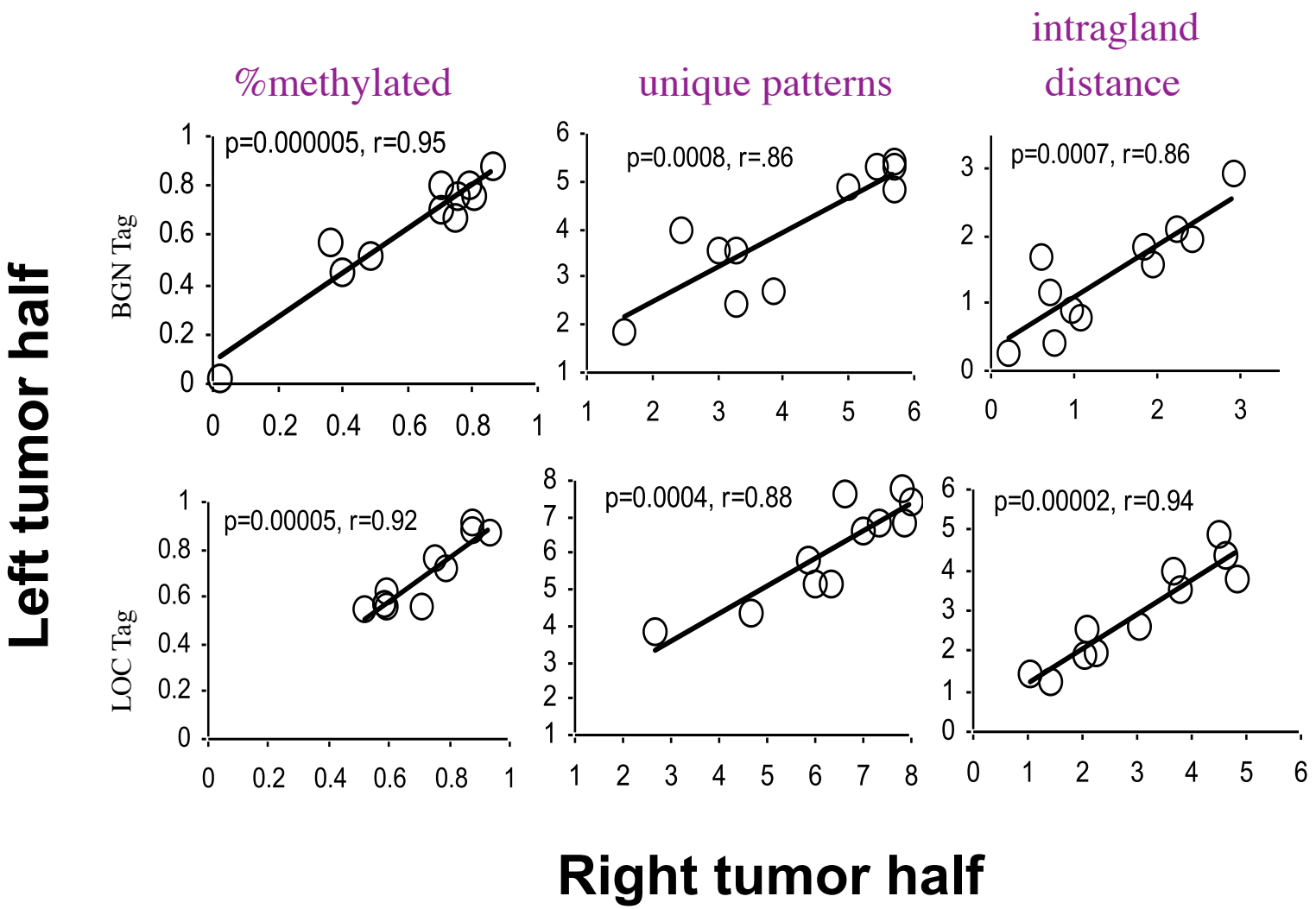


Yatabe *et al*. PNAS 2001

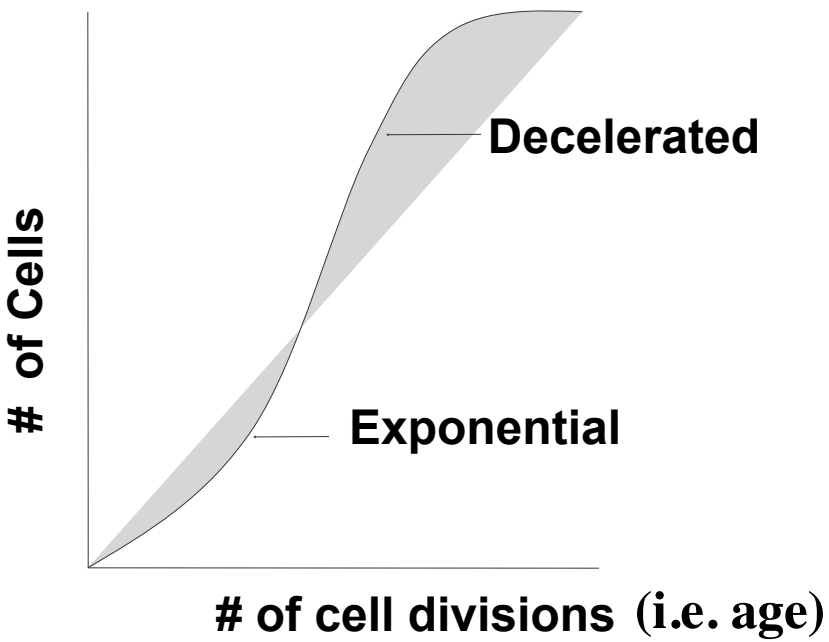
- Bisulfite genomic sequencing
- On X chromosome
- No selection in our regions
- Relatively high replication error rates
- Demonstrates age-related methylation in normal colon

('molecular clock' hypothesis)

DNA methylation by tumor half

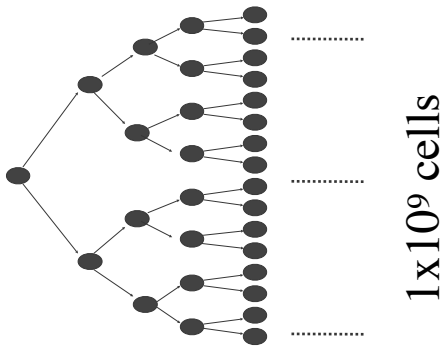


Tumor growth model



Gompertzian growth

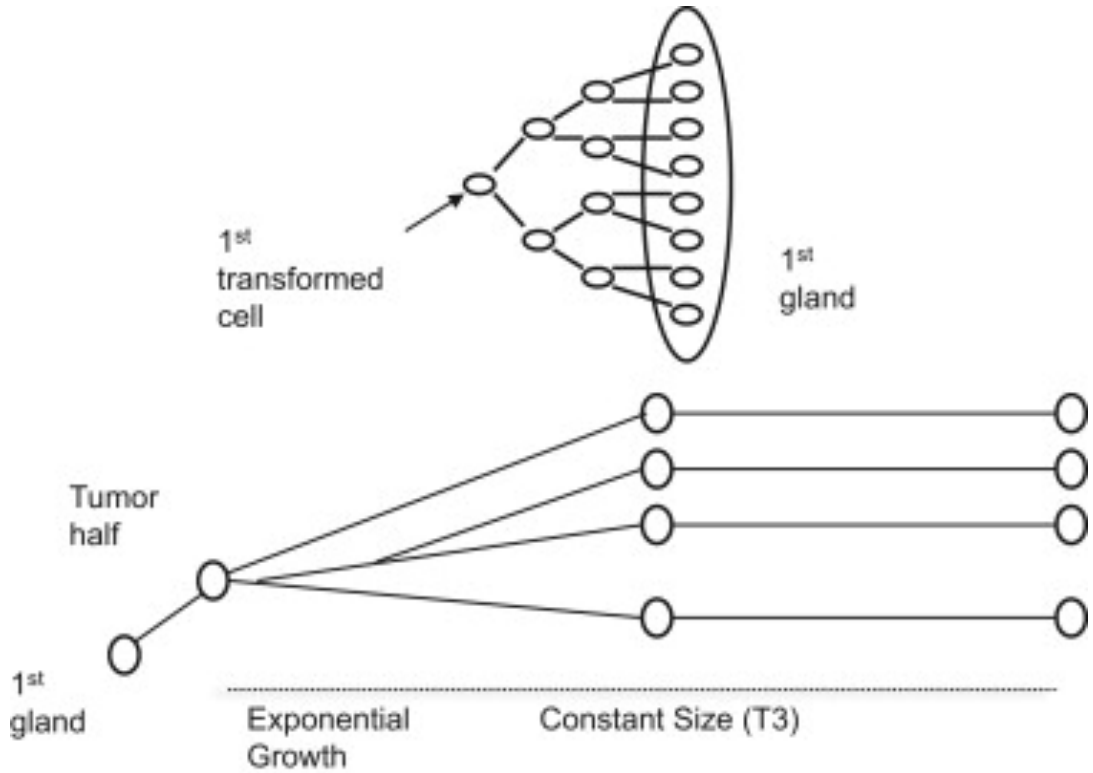
Exponential growth phase



Decelerated growth phase

Type of cell division

	Asymmetric	Symmetric		
Immortal:	100%	0%	0%	[R=1]
Random:	50%	25%	25%	[R=0.5]



Zhao, Junsong, et al. "Ancestral inference in tumors: How much can we know?." *Journal of Theoretical Biology* 359 (2014): 136-145.

Note: We don't simulate the growth of every cell in the tumor - just the cells that are ancestral to those that we sample.

Table 1
Parameters in our model.

Parameters	Possible value and prior distribution
DNA methylation error rate (MER)	U(0,0.1)
DNA demethylation error rate (DER)	U(0,0.1)
Number of cancer stem cells (NCSC)	2 to the power of 1,2,3,4,5,6,7,8,9
Probability of asymmetric division (PAD, R)	U(0.5,0.1)
Number of generations of constant size (T3)	10 to 365

Table 2
Summary statistics.

S_1	Percentage of methylation
S_2	Average number of unique tags
S_3	Average hamming distance within all glands
S_4	Average hamming distance among all glands
S_5	Average hamming distances between segments
S_6	Average number of transitions
S_7	Vector of sitewise percentage of methylation

See definition of each summary statistic in the Supplemental material.

Zhao, Junsong, et al. "Ancestral inference in tumors: How much can we know?." *Journal of Theoretical Biology* 359 (2014): 136-145.

Model calibration (does the ABC work?)

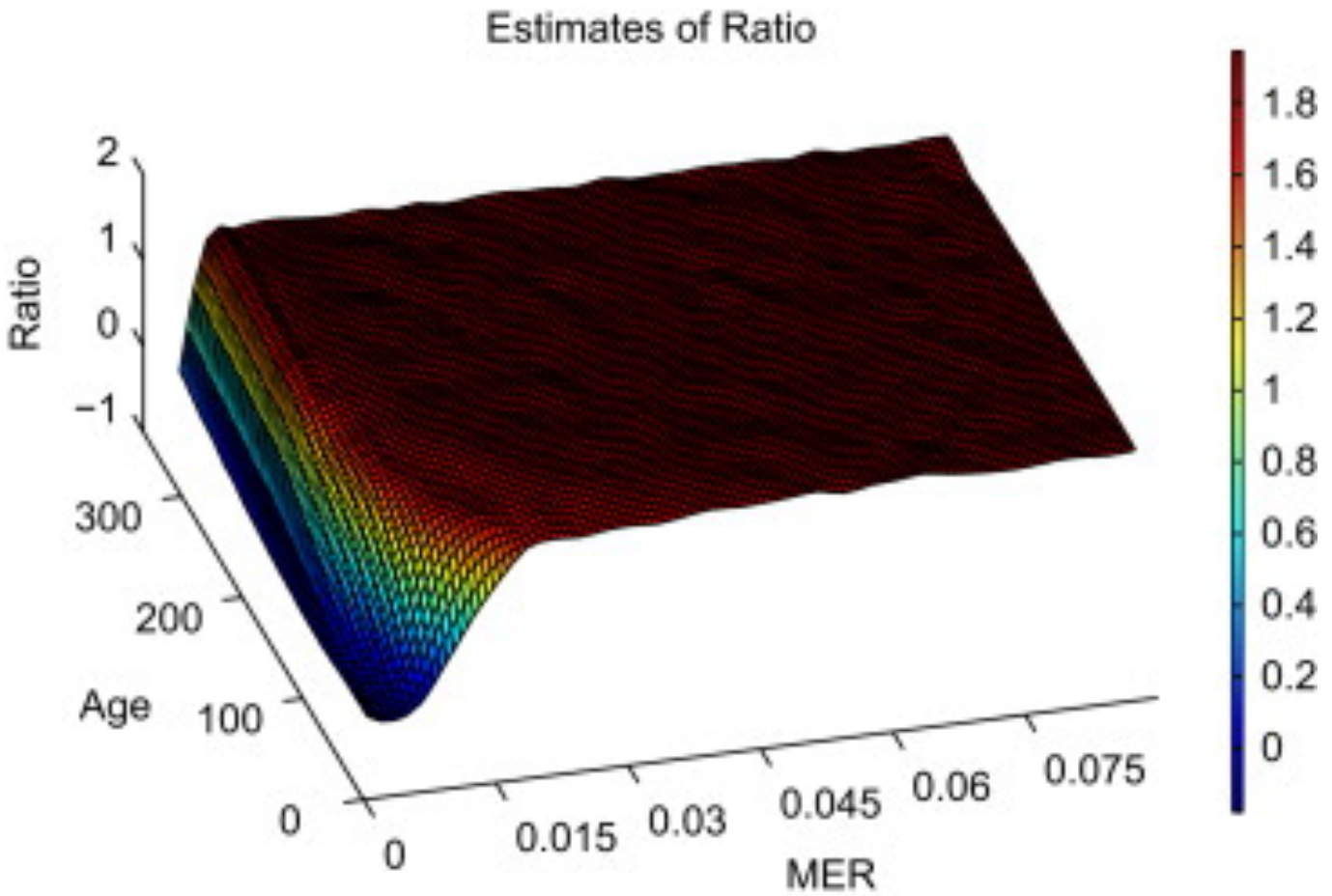


Fig. 5. The estimate of ratio of MER and DER. The tumors were generated using NCSC=256, R=0.8624, and an ancestor state of 11001001100100, with the ratio of MER and DER set to be 2.

http://www.sciencedirect.com/science/article/pii/S0022519314003002#gr5

Analysis of data

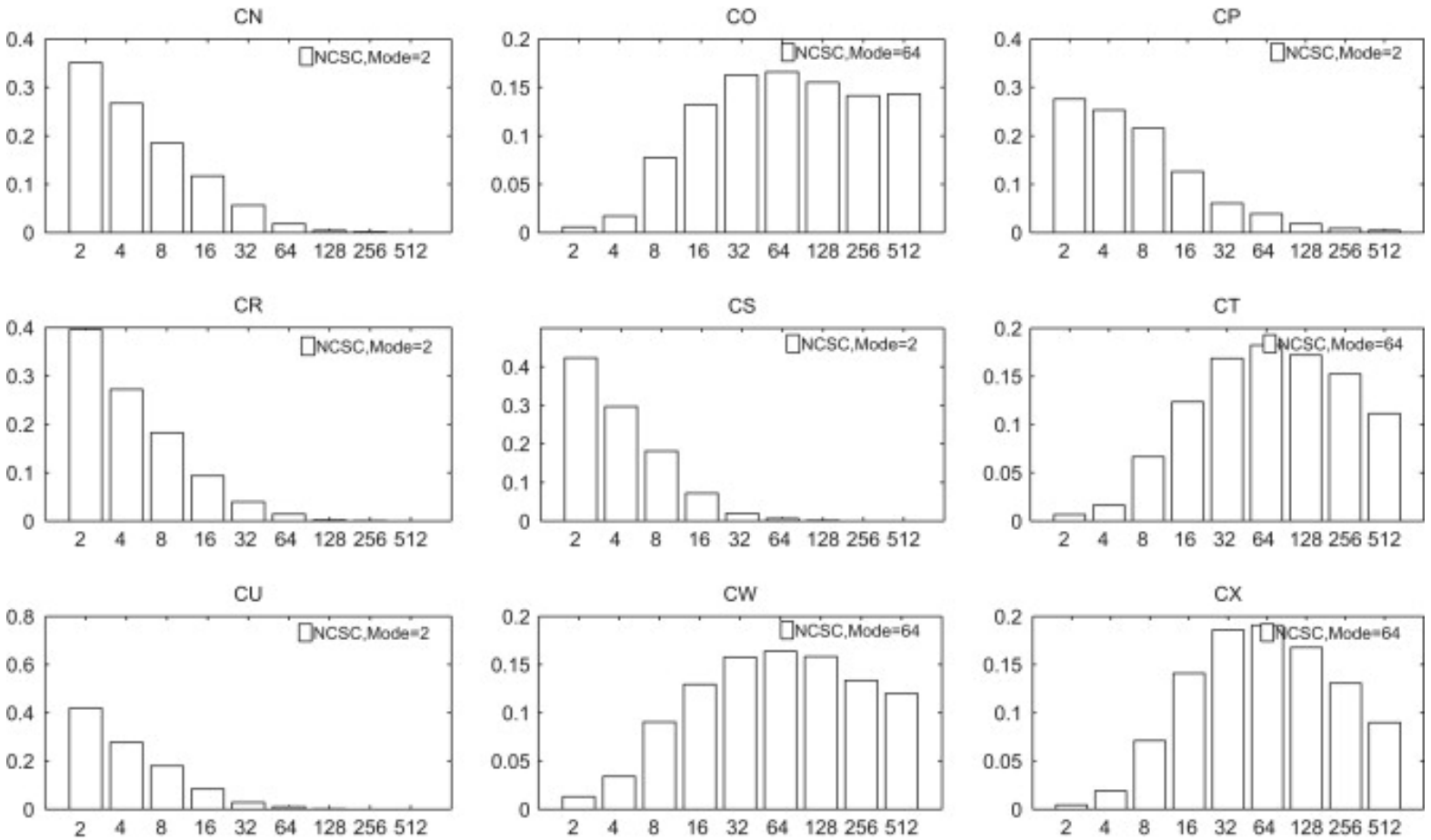


Fig. 11. The posterior distributions of the number of cancer stem cells in experimental data. The x-axis is the number of cancer stem cells.

"Born to be bad"

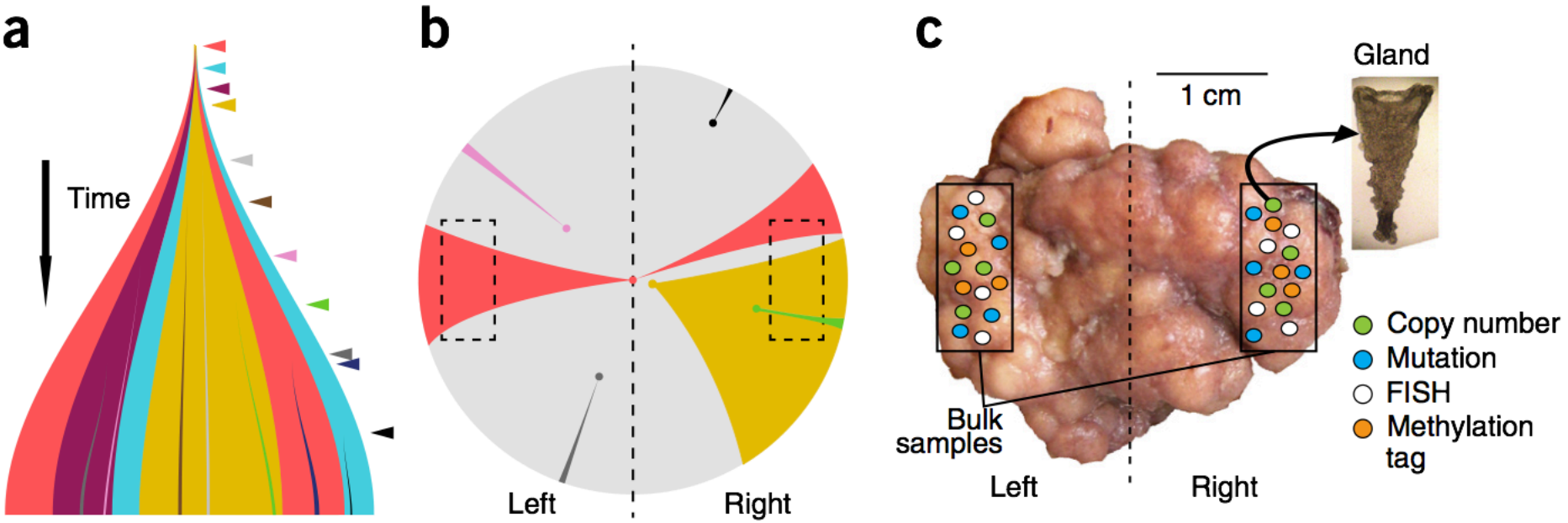


Fig 1. The Big Bang model of tumor growth.

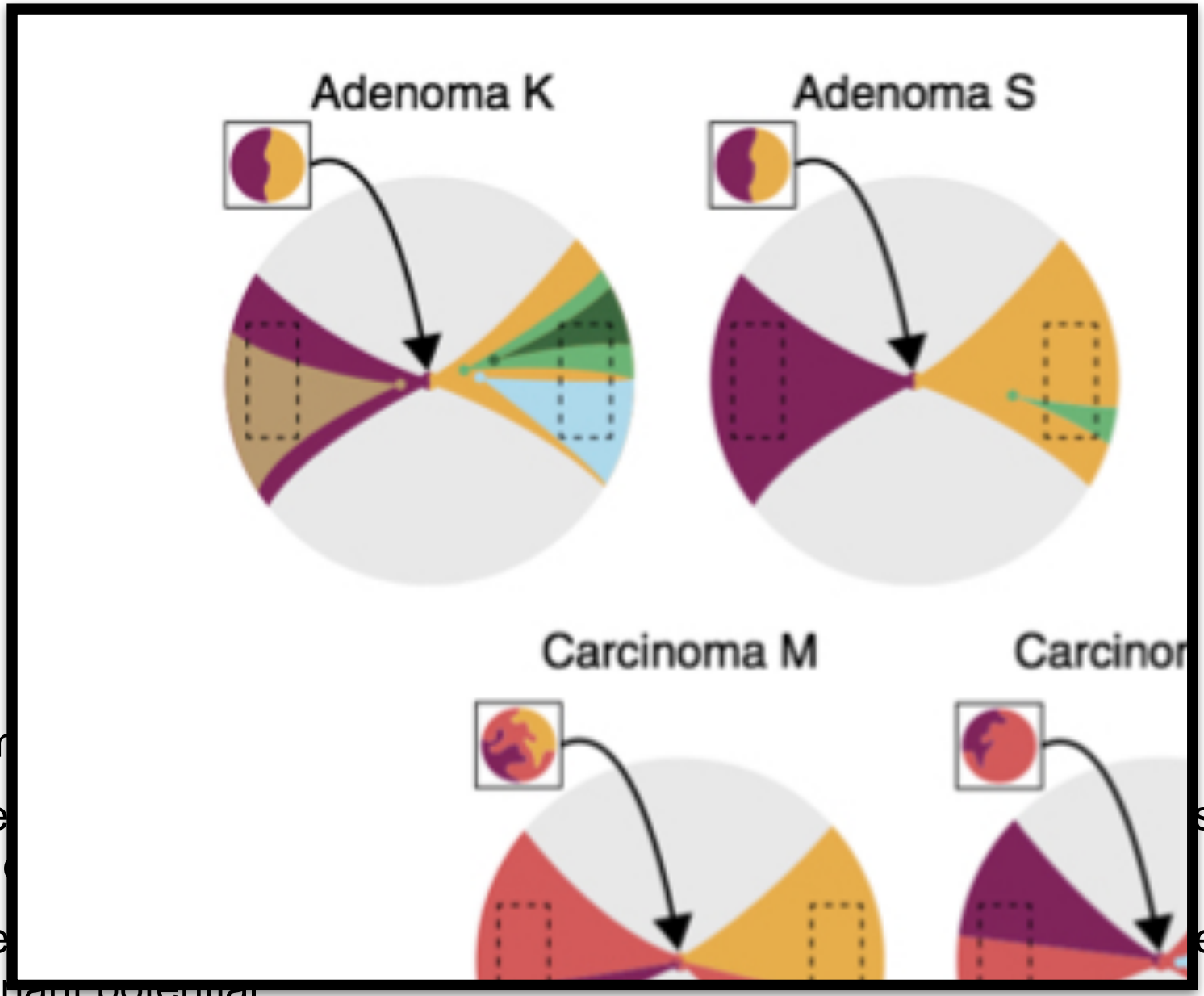
(a) After initiation, a tumor grows predominantly as a single expansion populated by numerous heterogeneous subclones.

(b) In the Big Bang model, the pervasiveness of private alterations depends on when the alteration occurs during growth, rather than on selection for that alteration.

(c) We sampled an average of 23 individual tumor glands (<10,000 cells) from distant regions (~0.5 cm^3 in size) and bulk (left and right) samples.

Sottoriva, A., et al. (2015). *Nature Genetics*, *47*(3), 209–216.

"Born to be bad"



- Our m
- We te subject to string
- Some e of early malig

Further reading

Siegmund, Kimberly D., et al. "Inferring clonal expansion and cancer stem cell dynamics from DNA methylation patterns in colorectal cancers." *Proceedings of the National Academy of Sciences* 106.12 (2009): 4828-4833.

Siegmund, Kimberly D., Paul Marjoram, and Darryl Shibata. "Modeling DNA methylation in a population of cancer cells." *Statistical applications in genetics and molecular biology* 7.1 (2008).

Siegmund, Kimberly D., et al. "High DNA methylation pattern intratumoral diversity implies weak selection in many human colorectal cancers." *PloS one* 6.6 (2011): e21657.

Siegmund, Kimberly D., et al. "Many colorectal cancers are “flat” clonal expansions." *Cell Cycle* 8.14 (2009): 2187-2193.

Zhao, Junsong, et al. "Ancestral inference in tumors: How much can we know?." *Journal of theoretical biology* 359 (2014): 136-145.

Sottoriva, Andrea, et al. "A Big Bang model of human colorectal tumor growth." *Nature genetics* 47.3 (2015): 209-216.

Zhao J, et al., “Early mutation bursts in colorectal tumors”. PLoS One 12.3 (2017): e0172516.

Non-examinable ABC task

- Use the Urn model starting with 2 red balls of mass 1, and one black (mutation) ball of mass w.
- Draw balls until you have 20 non-black balls.
- Find the posterior for the mass of the black ball using the following statistics (separately and together)
  - The number of different (non-black) colors in the urn was 5
  - There were 7 balls of the commonest color
  - The number of times the black ball was drawn was 7
- What do you conclude?

Use a Uniform[0,20] prior for the mass of the black ball

Next ABC Rejection example - Model fit

More generally - (Approximate Bayesian Computation)

Suppose:

A set of (normalized) summary statistics $S=\{S_1,\ldots,S_n\}$, that take values $S^o=\{S^o_1,\ldots,S^o_n\}$ on the observed data.

(A set of weights $w_1,\ldots,w_n$).

A *tolerance* ε.

1. Sample *θ'* from prior π.
2. Simulate D′ using *θ'*. Calculate S^s (the value of S on D′)
3. Accept *θ'* if $\sum_i w_i(S_i^o - S_i^s)^2 < \varepsilon$
4. Return to 1.

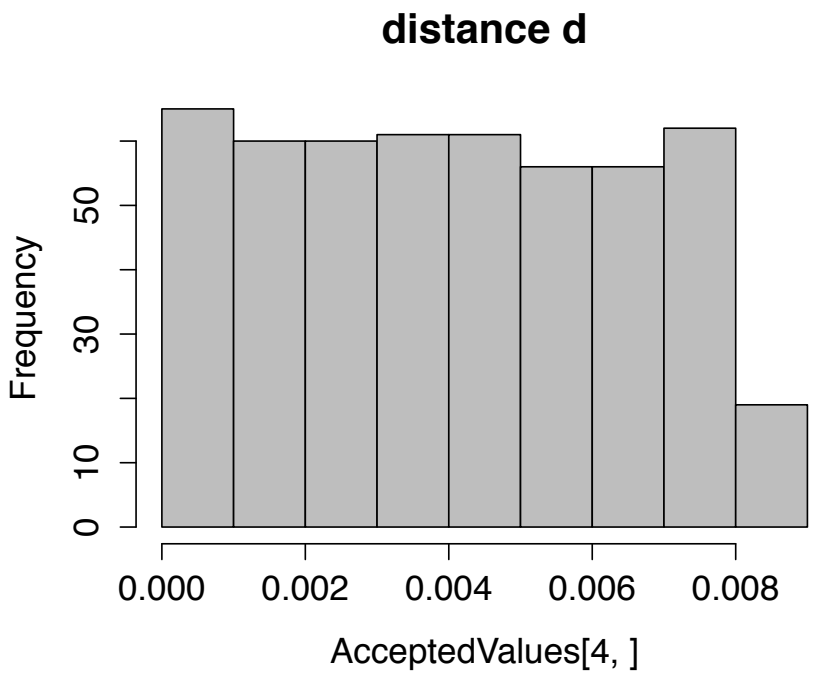
Results: independent samples from something we will call $\varphi(\theta|D)$.

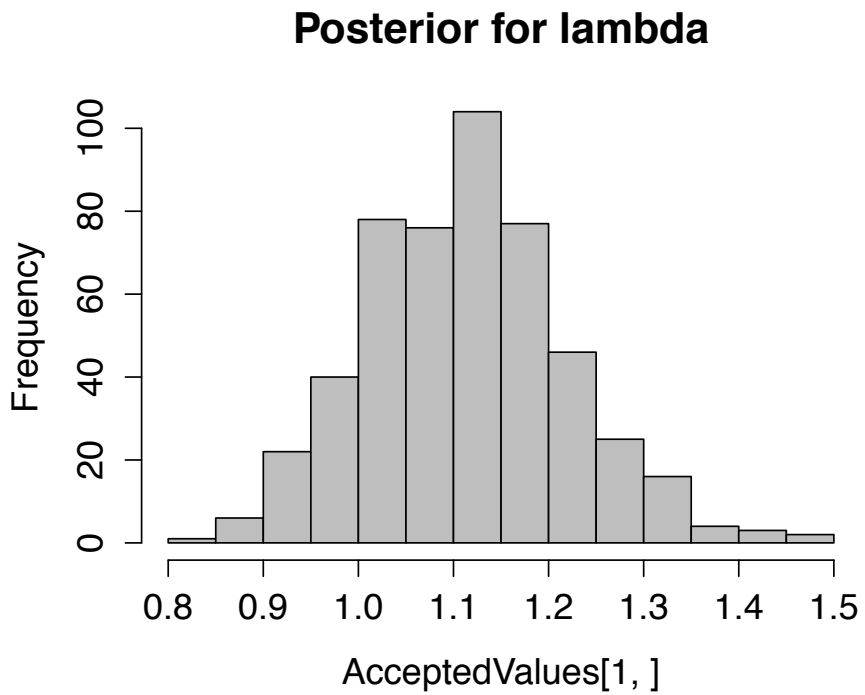
ABC Rejection example

Example ABC Rejection analysis:
Sample of 100 random variables with mean m=0.9 and std. dev. s=2.01.
Run an ABC Rejection analysis, generating 250K sets of 100 Exponential(λ) rvs, using a uniform[0,10] prior on λ.
Let m' the mean of a given simulated dataset.

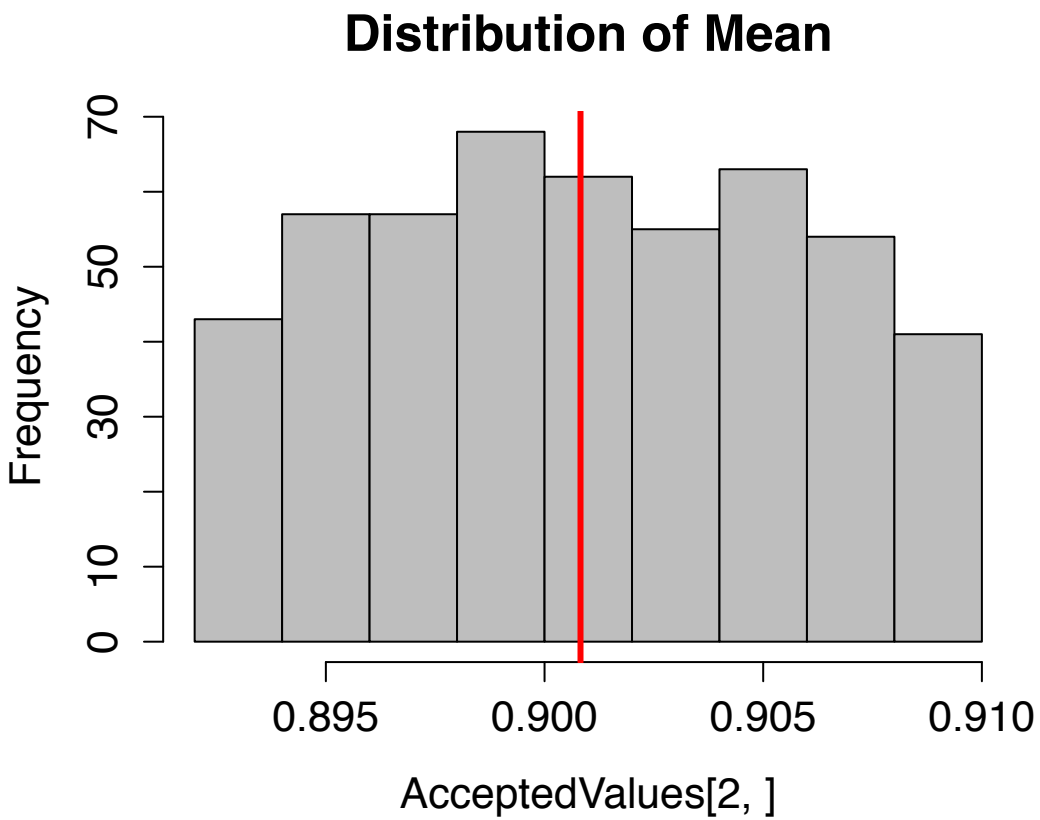
Define distance $d = \sqrt{(m-m')^2}$
Accept the 500 simulated datasets (and associated parameter values) with the smallest d (i.e. accept everything up to the 0.2 percentile).

Results

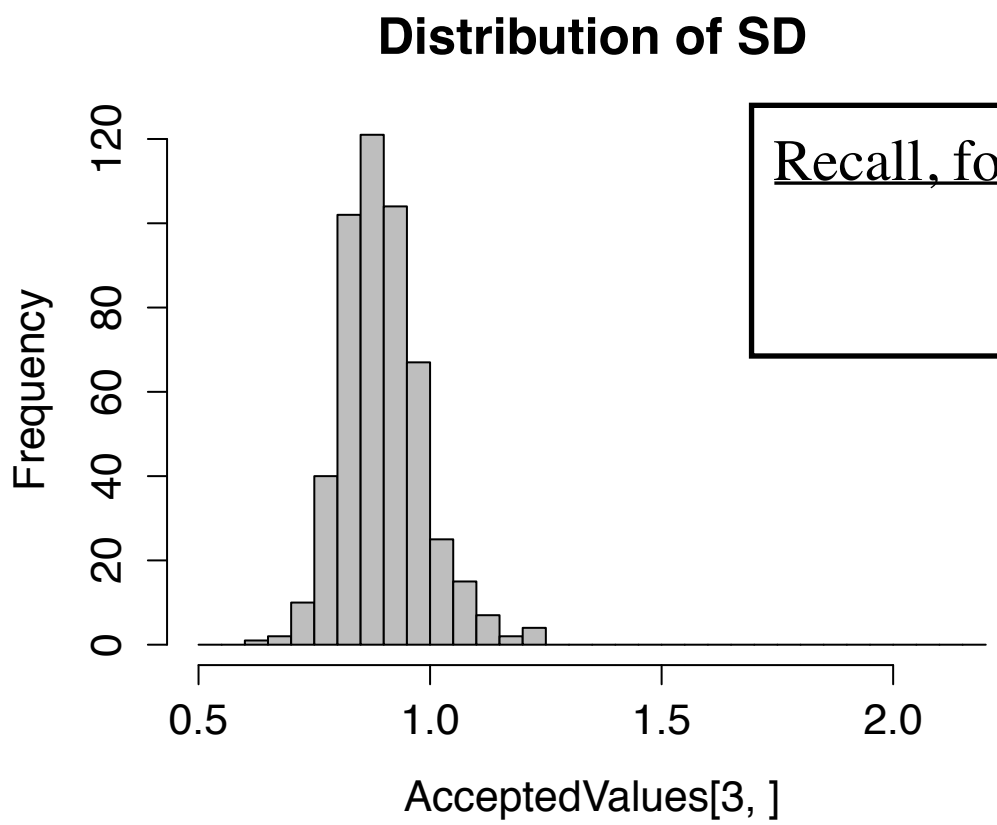

distance d
Frequency
0
10
30
50
0.000
0.002
0.004
0.006
0.008
AcceptedValues[4,]


Posterior for lambda
Frequency
0
20
40
60
80
100
0.8
0.9
1.0
1.1
1.2
1.3
1.4
1.5
AcceptedValues[1,]

Results


Distribution of Mean
Frequency
0
10
30
50
70
0.895
0.900
0.905
0.910
AcceptedValues[2,]

Results

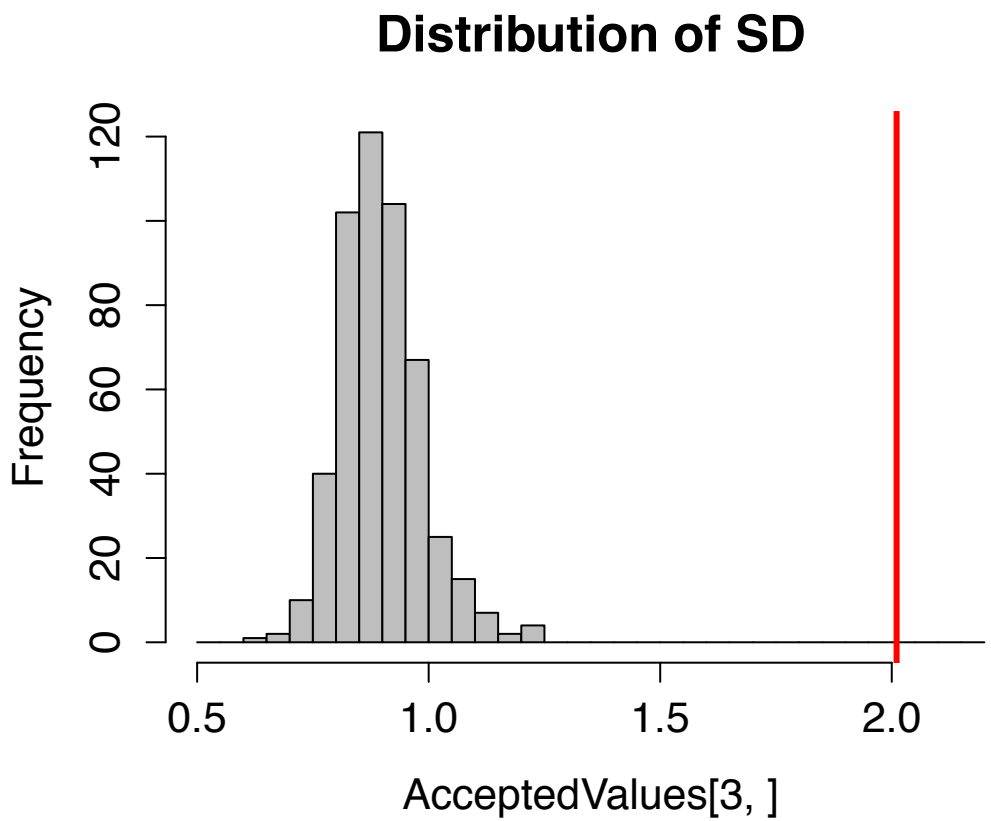


Recall, for exp(**λ**) distribution:

mean = 1/**λ**

sd = 1/**λ**

Results


Distribution of SD
Frequency
0
20
40
60
80
120
0.5
1.0
1.5
2.0
AcceptedValues[3,]

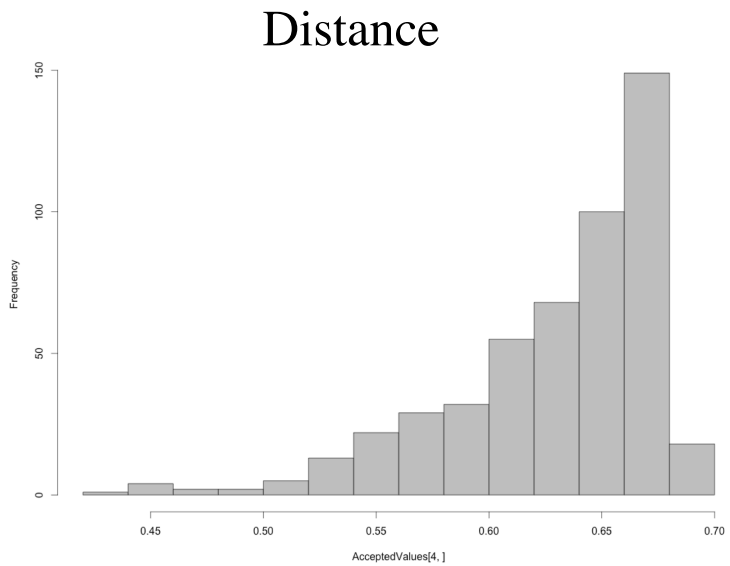
Model fit

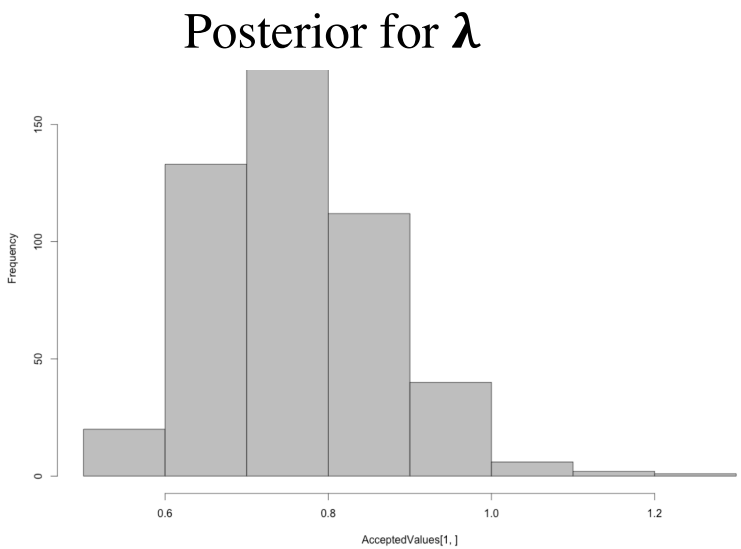
Example ABC Rejection analysis:
Sample of 100 random variables with mean m=0.9 and SD s=2.01.
Run an ABC Rejection analysis, generating 250K sets of 100 Exponential(λ) rvs, using a uniform[0,10] prior on λ.
Let m' and s' denote the mean and SD of a given simulated dataset.

Define distance $d = \sqrt{((m-m')^2+(s-s')^2)}$
Accept the 500 simulated datasets (and associated parameter values) with the smallest d (i.e. accept everything up to the 0.2 percentile).

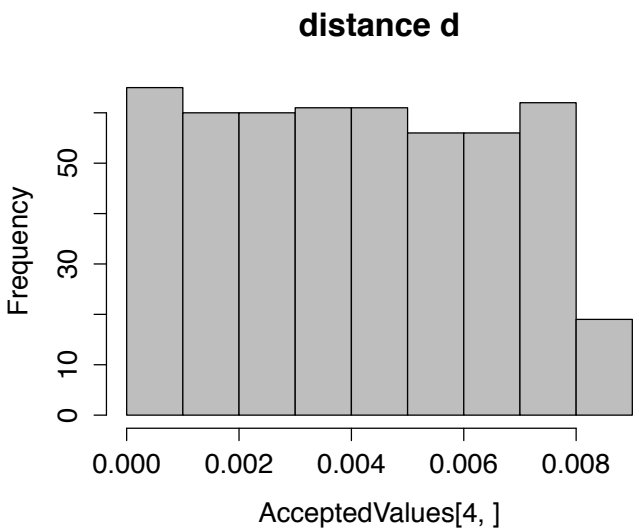
Results

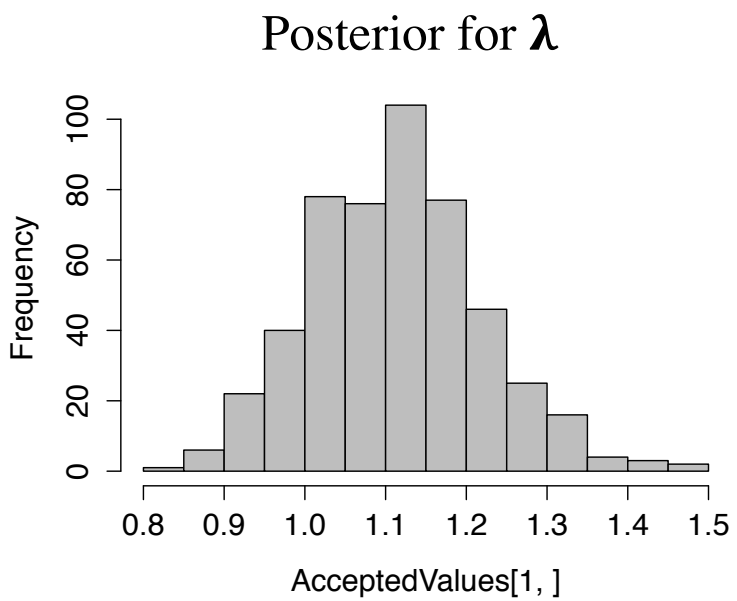
New fit



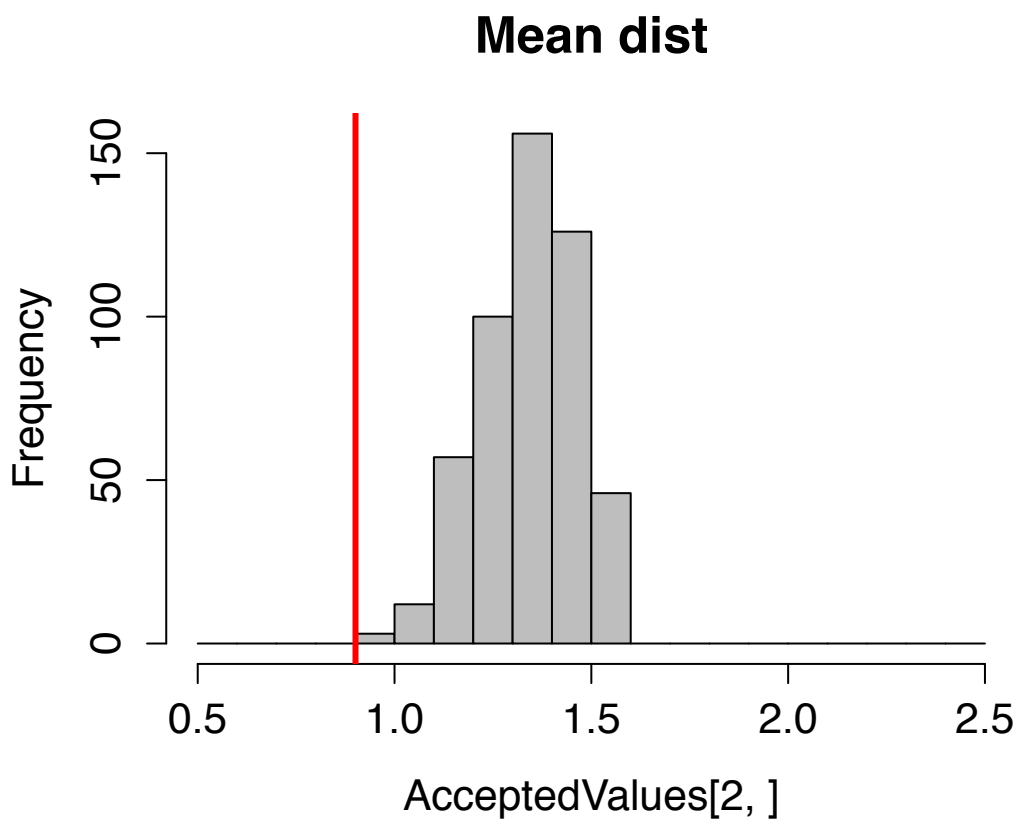


Old fit

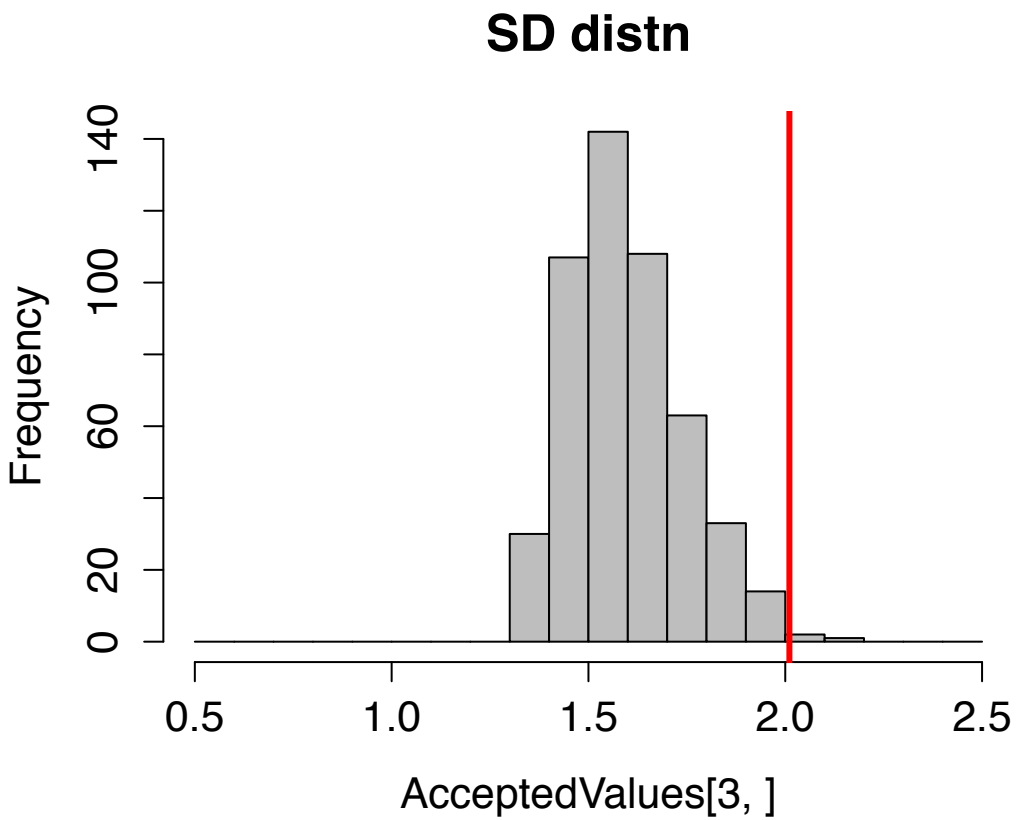




Results


Mean dist
Frequency
0
50
100
150
0.5
1.0
1.5
2.0
2.5
AcceptedValues[2,]

Results


SD distn
Frequency
0
20
60
100
140
0.5
1.0
1.5
2.0
2.5
AcceptedValues[3,]

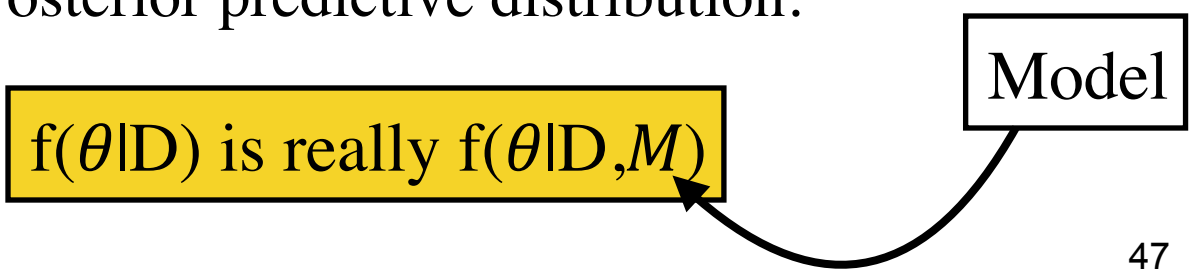
Model fit

A Bayesian model-based analysis will **always** return a posterior distribution for the parameters.

Important questions:

1. Is the approximated posterior close to the actual posterior? (Careful choice of summary statistics.)

2. Does the model actually fit the data? (model-fit)
   A. Prior predictive distribution;
   B. Posterior predictive distribution.



Model fit

Prior predictive distribution:

For summary statistic(s) S, look at

$$f(S) = \int f(S \mid \theta)\pi(\theta)d\theta$$

This is the distribution of the summary statistics under the prior for model parameter(s) θ.

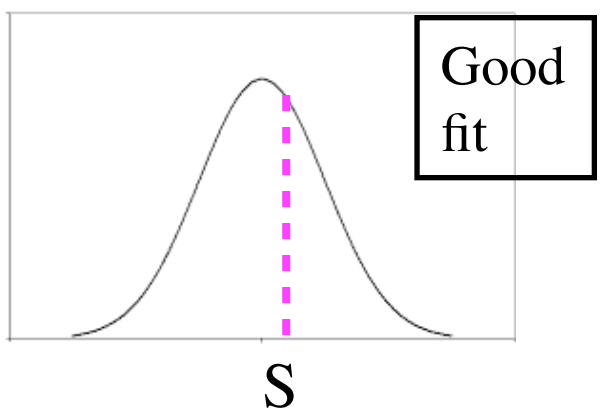
To do this, iterate the following:

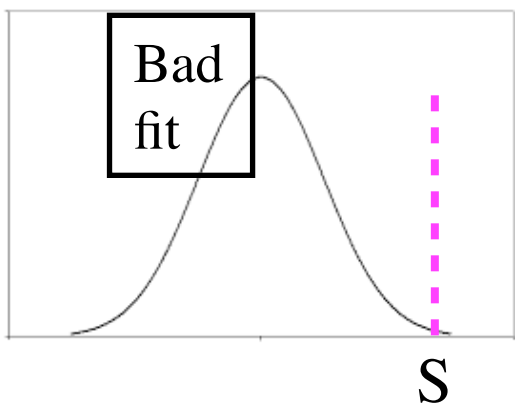
1. Sample θ' from the prior $\pi()$.
2. Simulate a dataset D' using θ'. Calculate S' (for D').
3. When finished, plot the distribution of S'.

> "Is it even possible for this model to produce data like that look like the observed data?"

Model fit

$$f(S) = \int f(S \mid \theta)\pi(\theta)d\theta$$





"Is it even possible for this model to produce data like that look like the observed data?"

Model fit

Posterior predictive distribution:

For summary statistic(s) S(D), look at

$$g(S) = \int f(S(D') \mid \theta) f(\theta \mid S(D)) d\theta$$

This is the distribution of the summary statistics under the posterior for model parameter(s) θ.

To do this, iterate the following:

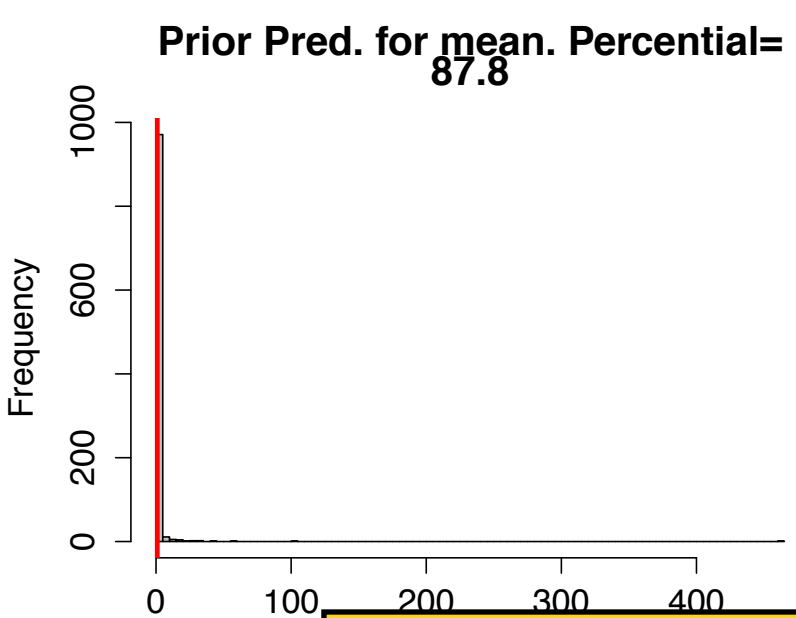
1. Sample θ' from the posterior for θ.
2. Simulate a dataset D' using θ'. Calculate S(D').
3. When finished, plot the distribution of S(D').

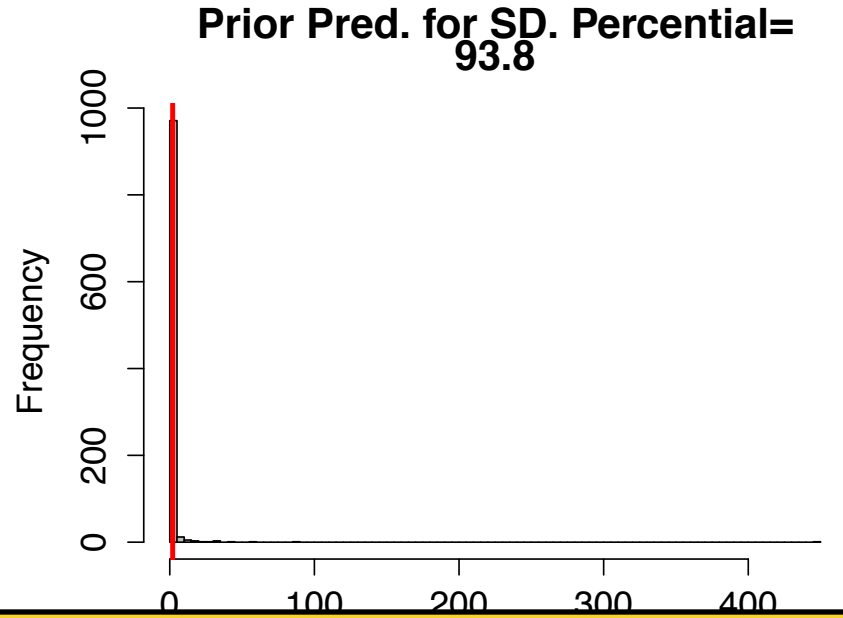
"With the 'right' parameter values, can the model produce data like that looks like the observed data *with reasonable probability*?"

Returning to the previous example

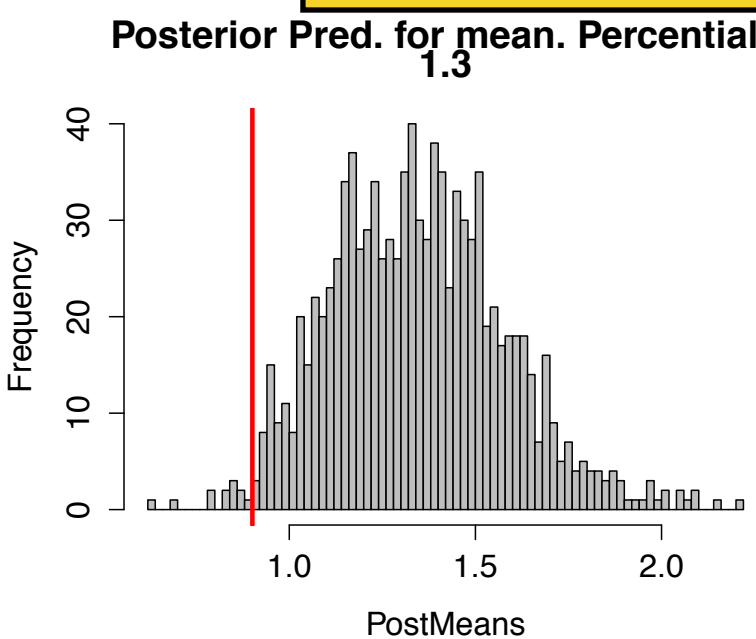
Example ABC Rejection analysis:
Sample of 100 random variables with mean m=0.9 and SD s=2.01.
Run an ABC Rejection analysis, generating 250K sets of 100 Exponential(λ) rvs, using a uniform[0,10] prior on λ.
Let m' and s' denote the mean and SD of a given simulated dataset.

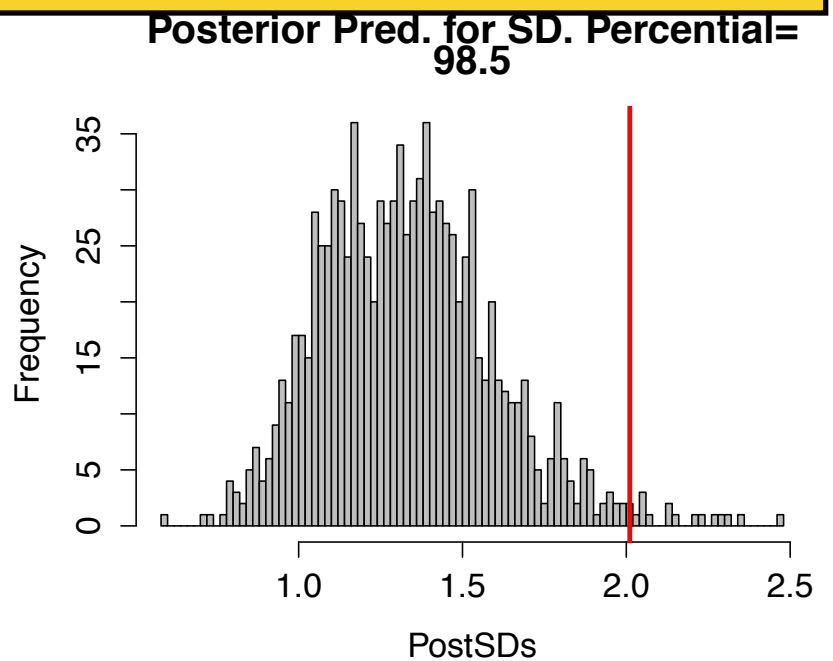
Define distance $d = \sqrt{((m-m')^2+(s-s')^2)}$
Accept the 500 simulated datasets (and associated parameter values) with the smallest d (i.e. accept everything up to the 0.2 percentile).





The data was not from an exponential distribution!





ABC Example 2 - Human Demography

PNAS PNAS

Statistical evaluation of alternative models of human evolution

Nelson J. R. Fagundes[†‡§], Nicolas Ray[§], Mark Beaumont[¶], Samuel Neuenschwander[§‖], Francisco M. Salzano[‡††], Sandro L. Bonatto[†,††], and Laurent Excoffier[§††]

[†]Laboratório de Biologia Genômica e Molecular, Faculdade de Biociências, Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), 90619-900 Porto Alegre, RS, Brazil; [‡]Departamento de Genética, Universidade Federal do Rio Grande do Sul, 91501-970 Porto Alegre, RS, Brazil; [§]Computational and Molecular Population Genetics (CMPG), Zoological Institute, University of Bern, CH-3012 Bern, Switzerland; [¶]School of Animal and Microbial Sciences, University of Reading, Reading RG6 6AJ, United Kingdom; and [‖]Department of Ecology and Evolution, University of Lausanne, Biophore, CH-1015 Lausanne, Switzerland

Contributed by Francisco M. Salzano, August 31, 2007 (sent for review August 1, 2007)

An appropriate model of recent human evolution is not only important to understand our own history, but it is necessary to disentangle the effects of demography and selection on genome diversity. Although most genetic data support the view that our species originated recently in Africa, it is still unclear if it completely replaced former members of the *Homo* genus, or if some interbreeding occurred during its range expansion. Several scenarios of modern human evolution have been proposed on the basis of molecular and paleontological data, but their likelihood has never been statistically assessed. Using DNA data from 50 nuclear loci

Africa and Asia raised claims for some degree of interbreeding between modern and archaic *Homo* forms (13, 14, 16, 17). Such interbreeding can occur under assimilation scenarios (Fig. 1*B*), where modern humans migrating out of Africa would have hybridized with local *Homo erectus* and incorporated old lineages (15, 18) or under multiregional scenarios (Fig. 1*C*), where migrants would have been continuously exchanged between Africa and Asia, leading to a synchronized emergence of modern anatomy. Note that these simple scenarios are somewhat arbitrary and that human evolution has certainly been more com-

Model selection

- Suppose we have data, D, and two (or more) possible models, $\mathcal{M}_1$ and $\mathcal{M}_2$, from a set $\mathcal{M}$. Suppose the models have parameters θ_1 and θ_2, with distributions π_1, π_2.
- Standard tool for model selection is the marginal likelihood of the data given the model:

$$w(D) = \int_{\Theta} \pi(\theta) f(D \mid \theta) d\theta$$

- In the Bayesian paradigm, for model selection (model 1 vs. model 2), we calculate the Bayes factor:

$$B_{12}(D) = \frac{w_1(D)}{w_2(D)} = \frac{\int_{\Theta_1} \pi_1(\theta_1) f_1(D \mid \theta_1) d\theta_1}{\int_{\Theta_2} \pi_2(\theta_2) f_2(D \mid \theta_2) d\theta_2}$$

f(D | Model 1)

f(D | Model 2)

- Bayes factor interpretation of evidence
  - 3: substantial
  - 10: decisive
  - 30: very strong (Jeffreys)

Model selection via ABC Rejection methods

Suppose:

A set of (normalized) summary statistics S={S_1,…,S_n}

A set of weights {w_1,…,w_n} and a *tolerance* ε.

A set of models, $\mathcal{M}$ [={$\mathcal{M}_1$,$\mathcal{M}_2$}, say].

1. Sample a model *M* from $\pi_{\mathcal{M}}$ (the prior distribution for models - typically uniform)
2. Sample *θ'* from prior $\pi_{\theta,M}$ (the prior for *θ* in model *M*)
3. Simulate D′ using *θ'*. Calculate S′.
4. Accept *θ'* and M if $\sum_i w_i(S_i^o - S_i^s)^2 < \varepsilon$
5. Return to 1.

Results: independent samples from something we will call φ(*θ,M|D)*.

P(M_1 | D) ~ proportion of accepted samples with M=$\mathcal{M}_1$. [Bayes factor will be the ratio of those two proportions for the two models.]

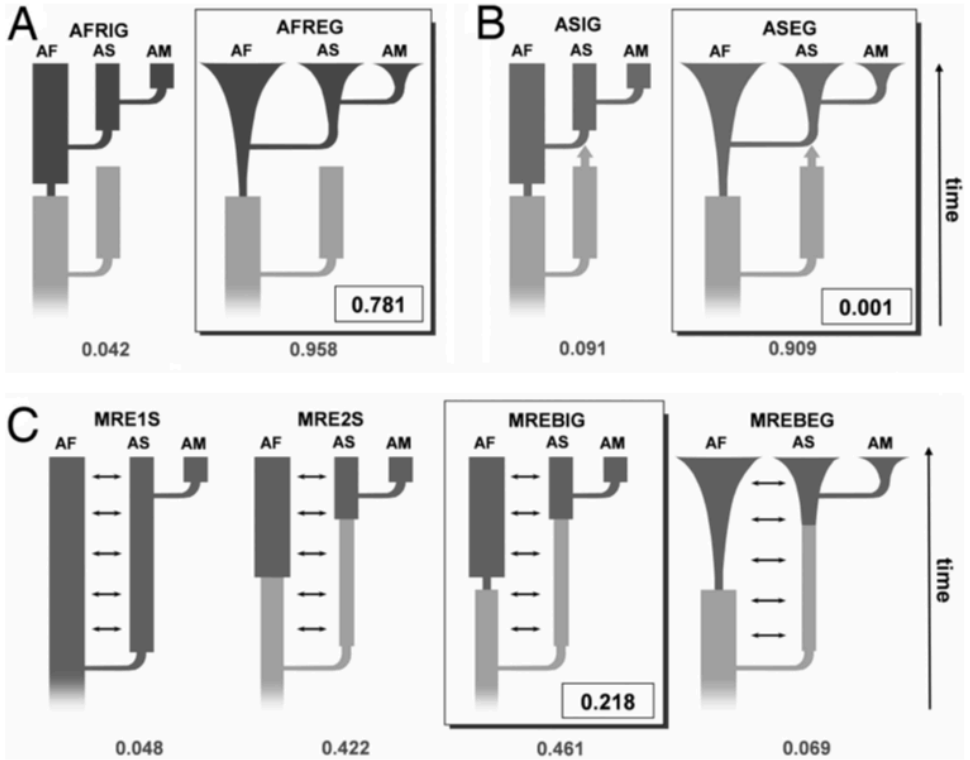


Fig. 1. Alternative scenarios of human evolution. (*A*) African replacement models: AFRIG, African replacement with instantaneous population growth; AFREG, African replacement with exponential population growth. (*B*) Assim-

For all models, the dark grays represent modern human populations, and lighter grays represent archaic populations.
AF, Africa;
AS, Asia;
AM, Americas.

The posterior probability of different models within each major scenario is given below each model. The posterior probabilities of the best model selected under each scenario are reported within boxes.

Fig 1: Alternative scenarios of human evolution.

(A) **African replacement models**: AFRIG, African replacement with instantaneous population growth; AFREG, African replacement with exponential population growth.

(B) **Assimilation models**: ASIG, assimilation with instantaneous population growth; ASEG, assimilation with exponential population growth.

(C) **Multiregional evolution (MRE) models**:

- MRE1S, MRE with constant population size in Africa/ Asia;
- MRE2S, MRE with two population sizes in Africa/Asia;
- MREBIG, MRE with bottleneck and instantaneous population growth;
- MREBEG, MRE with bottleneck and exponential population growth.

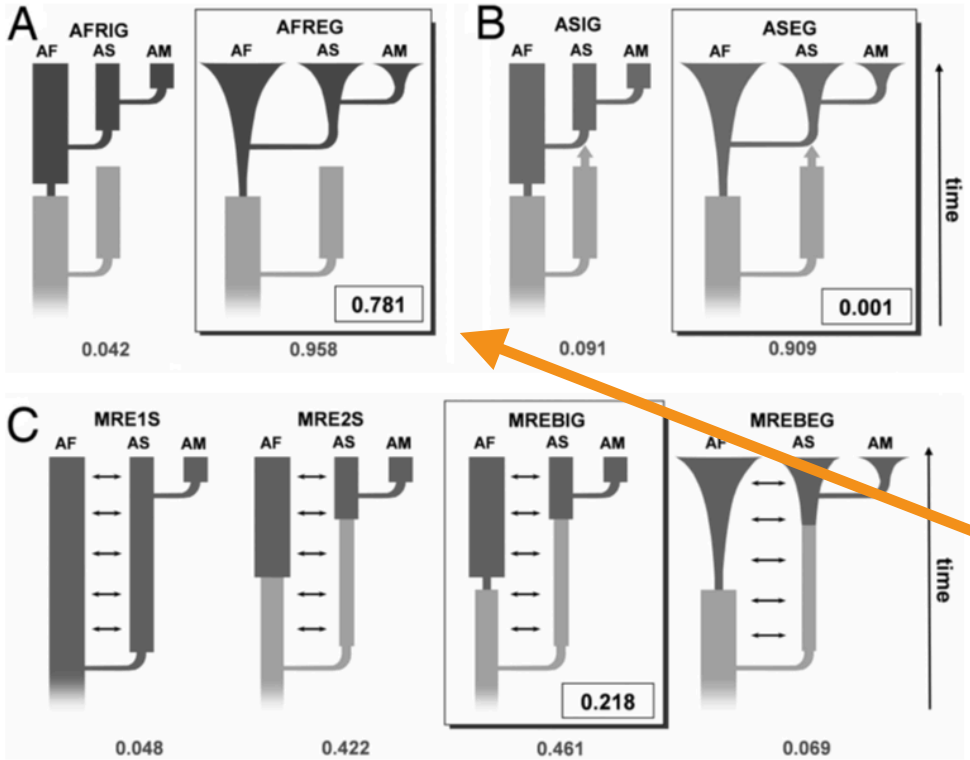


Fig. 1. Alternative scenarios of human evolution. (*A*) African replacement models: AFRIG, African replacement with instantaneous population growth; AFREG, African replacement with exponential population growth. (*B*) Assim-

For all models, the dark grays represent modern human populations, and lighter grays represent archaic populations.
AF, Africa;
AS, Asia;
AM, Americas.

The poste
within eac
model. Th
model sele
reported w

single and recent African origin for all modern humans (growth/no-growth)

Fig 1: Alternative scenarios of human evolution.

(A) **African replacement models**: AFRIG, African replacement with instantaneous population growth; AFREG, African replacement with exponential population growth.

(B) **Assimilation models**: ASIG, assimilation with instantaneous population growth; ASEG, assimilation with exponential population growth.

(C) **Multiregional evolution (MRE) models**:

- MRE1S, MRE with constant population size in Africa/ Asia;
- MRE2S, MRE with two population sizes in Africa/Asia;
- MREBIG, MRE with bottleneck and instantaneous population growth;
- MREBEG, MRE with bottleneck and exponential population growth.

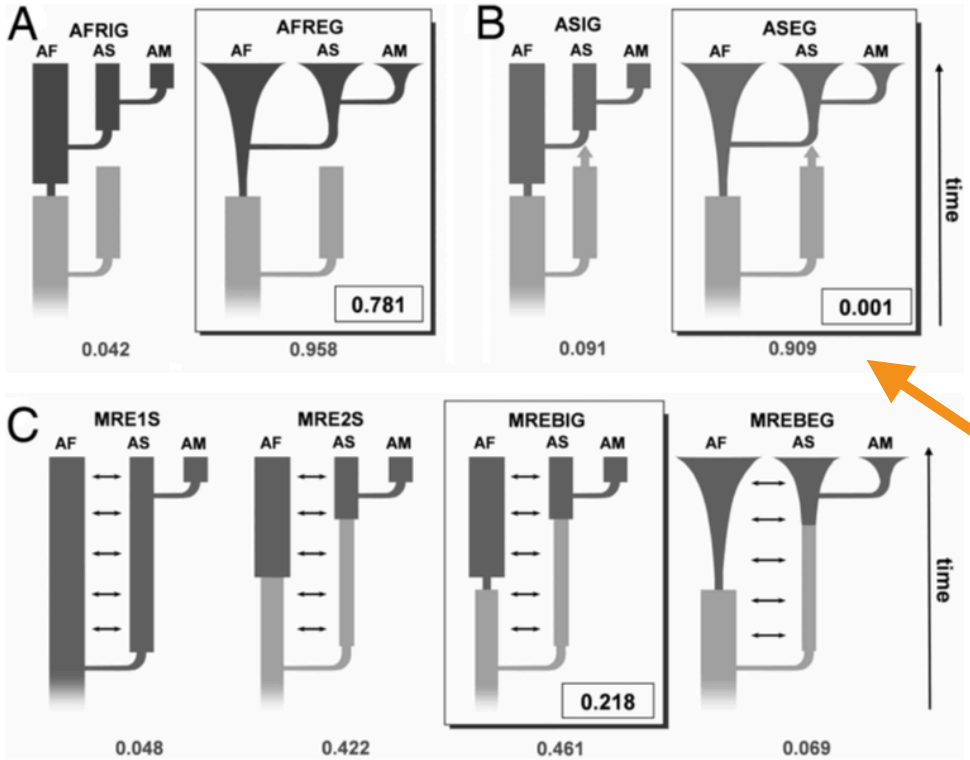


Fig. 1. Alternative scenarios of human evolution. (*A*) African replacement models: AFRIG, African replacement with instantaneous population growth; AFREG, African replacement with exponential population growth. (*B*) Assim-

For all models, the dark grays represent modern human populations, and lighter grays represent archaic populations.
AF, Africa;
AS, Asia;
AM, Americas.

The poste
within eac
model. Th
model sele
reported w

Assimilation scenarios. Modern humans migrate out of Africa & hybridize with local Homo erectus incorporating some old lineages

Fig 1: Alternative scenarios of human evolution.
(A) **African replacement models**: AFRIG, African replacement with instantan
population growth; AFREG, African replacement with exponential population growth.
(B) **Assimilation models**: ASIG, assimilation with instantaneous population growth; ASEG, assimilation with exponential population growth.
(C) **Multiregional evolution (MRE) models**:
- MRE1S, MRE with constant population size in Africa/ Asia;
- MRE2S, MRE with two population sizes in Africa/Asia;
- MREBIG, MRE with bottleneck and instantaneous population growth;
- MREBEG, MRE with bottleneck and exponential population growth.

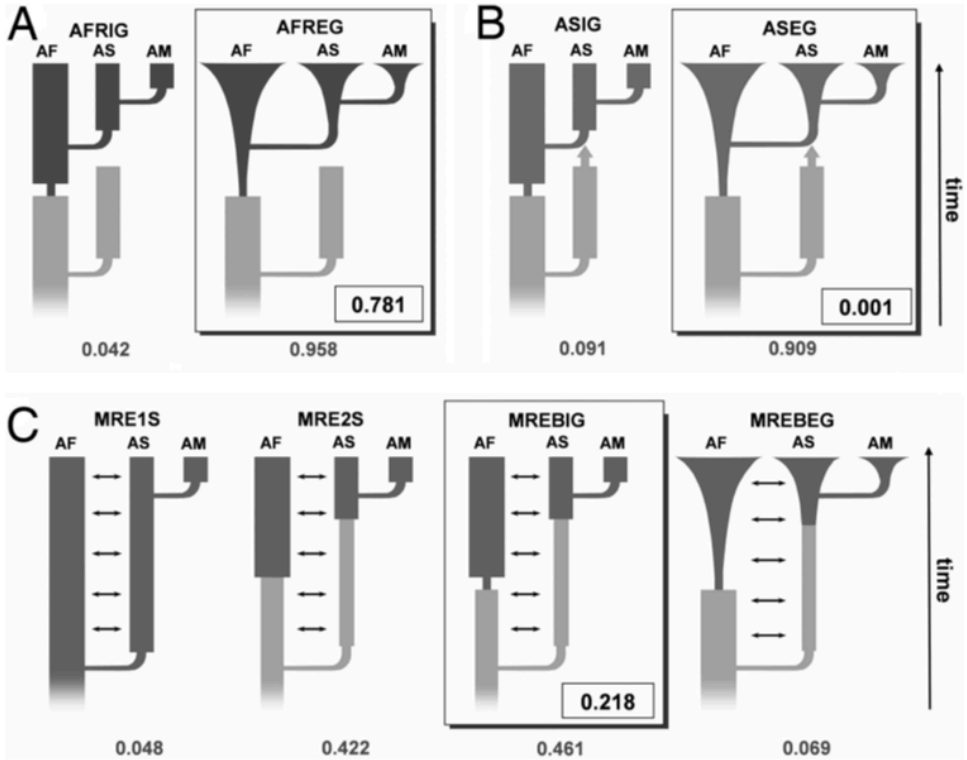


Fig. 1. Alternative scenarios of human evolution. (*A*) African replacement models: AFRIG, African replacement with instantaneous population growth; AFREG, African replacement with exponential population growth. (*B*) Assim-

For all models, the dark grays represent modern human populations, and lighter grays represent archaic populations.
AF, Africa;
AS, Asia;
AM, Americas.

The posterior probability of different models within each major scenario is given below each model. The posterior probabilities of the best model selected under each scenario are reported within boxes.

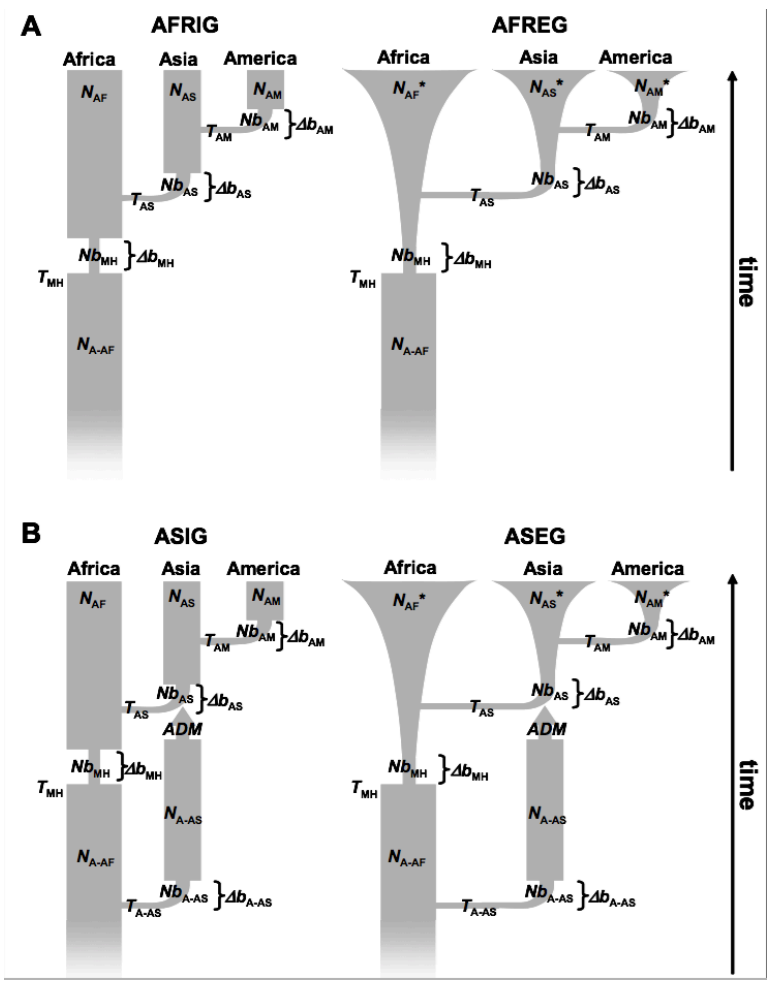
Fig 1: Alternative scenarios of human evolution.
(A) **African replacement models**: AFRIG, African replacement with instantaneo
population growth; AFREG, African replacement with exponential popula
(B) **Assimilation models**: ASIG, assimilation with instantaneous population gr
assimilation with exponential population growth.
(C) **Multiregional evolution (MRE) models**:
- MRE1S, MRE with constant population size in Africa/ Asia;
- MRE2S, MRE with two population sizes in Africa/Asia;
- MREBIG, MRE with bottleneck and instantaneous population growth;
- MREBEG, MRE with bottleneck and exponential population growth.

Multiregional scenarios. Migrants continuously exchanged between Africa and Asia -> synchronized emergence of modern anatomy.

Example 2 - Human Demography

Table 7. Prior distributions for the parameters of the tested evolutionary models

Parameter	Model								Values		
	AFRIG	AFREG	ASIG	ASEG	MRE1S	MRE2S	MREBIG	MREBEG	Min.	Max.	Distribution
Effective population sizes											
Current size in Africa N_{AF}	X		X			X	X		5,000	10^6	Log-uniform
Current size in Asia N_{AS}	X		X			X	X		1,000	10^5	Log-uniform
Current size in America N_{AM}	X		X			X	X		1,000	10^5	Log-uniform
Current size in Africa N_{AF}*		X		X				X	5,000	5×10^6	Log-uniform
Current size in Asia N_{AS}*		X		X				X	1,000	10^6	Log-uniform
Current size in America N_{AM}*		X		X				X	1,000	10^6	Log-uniform
Current size in Africa N_{AF}**					X				100	10^6	Log-uniform
Current size in Asia N_{AS}**					X				100	10^5	Log-uniform
Current size in America N_{AM}**					X				100	10^5	Log-uniform
Archaic size in Africa $N_{A\text{-}AF}$	X	X	X	X					1,000	10^5	Log-uniform



There are three pages like this (i.e. a lot of parameters)

Model selection via ABC Rejection methods

Suppose:

A set of (normalized) summary statistics $S=\{S_1,\ldots,S_n\}$

A set of weights $\{w_1,\ldots,w_n\}$ and a *tolerance* ε.

A set of models, $\{M_1,M_2\}$.

1. Sample a model M from π_M (the prior distribution for models - typically uniform)
2. Sample *θ'* from prior $\pi_{\theta,M}$ (the prior for θ in model M)
3. Simulate D′ using *θ'*. Calculate S′.
4. Accept *θ'* and M if $\sum_i w_i(S_i^o - S_i^s)^2 < \varepsilon$
5. Return to 1.

Results: independent samples from something we will call $\varphi(\theta,M|D)$.

"We retained the 5,000 simulations with smallest d-values computed on a total of 5 million simulations."

'Percentile threshold' [Here, accept nearest 0.1% of data.]

Analysis details

The data they collected: 50 loci, at each of which they have 500bp of sequence data.

"For each model, we first perform a large number of genetic simulations based on a demographic history that describes the model using the program SIMCOAL" [A coalescent simulator.]

The following summary statistics were computed for each locus in each sample:

S1: Number of segregating sites – number of sites that are polymorphic

S2: Nucleotide diversity π, – average proportion of sites that differ between pairs of individuals.

S3: Tajima's D – looks for signatures of selection [the difference between S2 and S1]

S4: F_{ST} – ratio of within and across sample genetic variation.

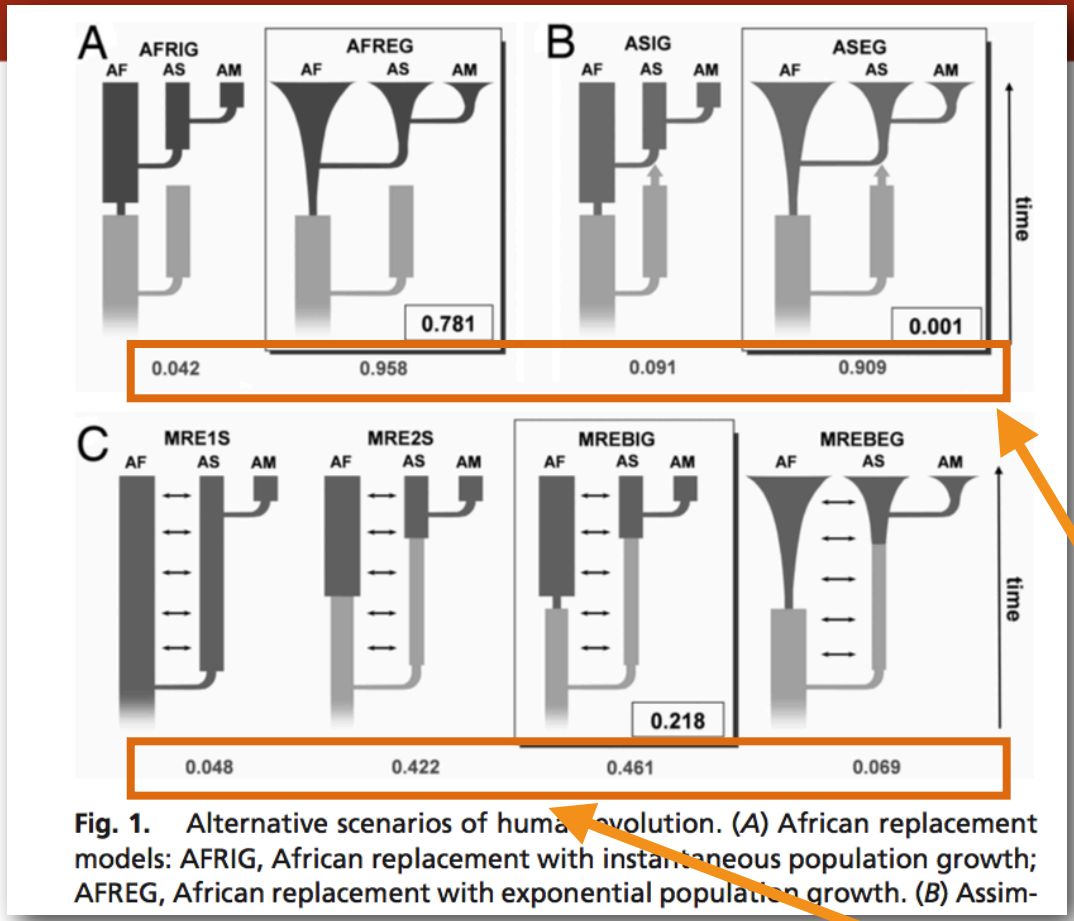


Fig. 1. Alternative scenarios of human evolution. (*A*) African replacement models: AFRIG, African replacement with instantaneous population growth; AFREG, African replacement with exponential population growth. (*B*) Assim-

For all models, the dark grays represent modern human populations, and lighter grays represent archaic populations.
AF, Africa;
AS, Asia;
AM, Americas.

The posterior probability of different models within each major scenario is given below each model. The posterior probabilities of the best model selected under each scenario are reported within boxes.

First look for the best model within each of the three scenarios…

Fig 1: Alternative scenarios of human evolution.
(A) **African replacement models**: AFRIG, African replacement with instantaneous population growth; AFREG, African replacement with exponential populatio
(B) **Assimilation models**: ASIG, assimilation with instantaneous population gr assimilation with exponential population growth.
(C) **Multiregional evolution (MRE) models**:
- MRE1S, MRE with constant population size in Africa/ Asia;
- MRE2S, MRE with two population sizes in Africa/Asia;
- MREBIG, MRE with bottleneck and instantaneous population growth;
- MREBEG, MRE with bot- tleneck and exponential population growth.

Simulation study to calibrate those three best models:

"To study the power of our model choice procedure in the context of human evolution, we simulated 1,000 random data sets under the best model for African replacement (AFREG), assimilation (ASEG), and multiregional (MREBIG) models and each time estimated the posterior probability of the three models. We find that the AFREG and MREBIG models are correctly recovered (have the highest posterior probability) in 79.3% and 80.1% of the cases, respectively, but the ASEG model is correctly identified in only 50.3% of the cases and, thus, seems to be the model most difficult to identify."

Example 2 - Human Demography

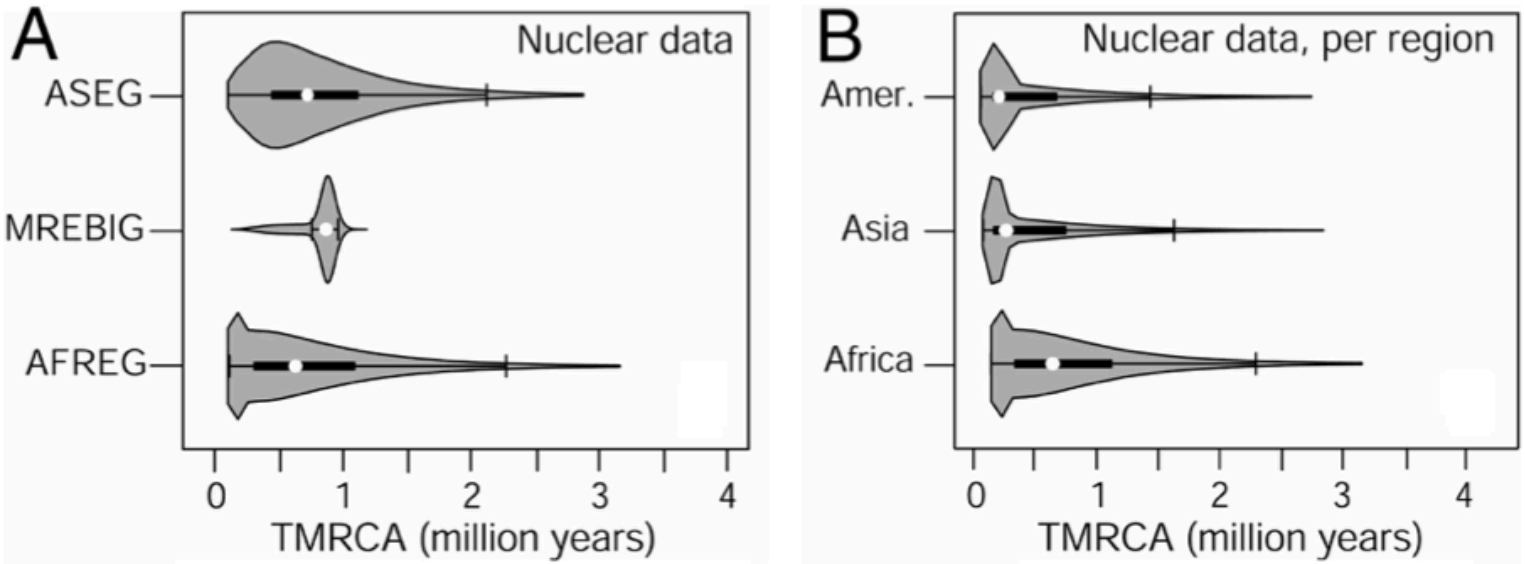


Fig. 2. Empirical TMRCA distribution obtained by simulation under different models. Parameter values were set to the median of the estimated marginal posterior distributions. Each distribution combines a mirrored estimated density surface in gray with a standard boxplot representation. Boxplots display the median of the distribution as a white dot, the interquartile range (IQR, 25–75%) as a thick line, and the region of 1.5 IQR as a thin line ending with vertical whiskers. To facilitate the comparison among models, all distributions (apart those from MREBIG model) were cut after the 99th percentile (full distributions are available in SI Tables 6, 7, and 8).

(A) Autosomal (non-sex chromosome) loci.

(B) Autosomal loci under the best model (AFREG), where only the samples of each of the three regions are considered.

Estimate parameters for best-fitting model

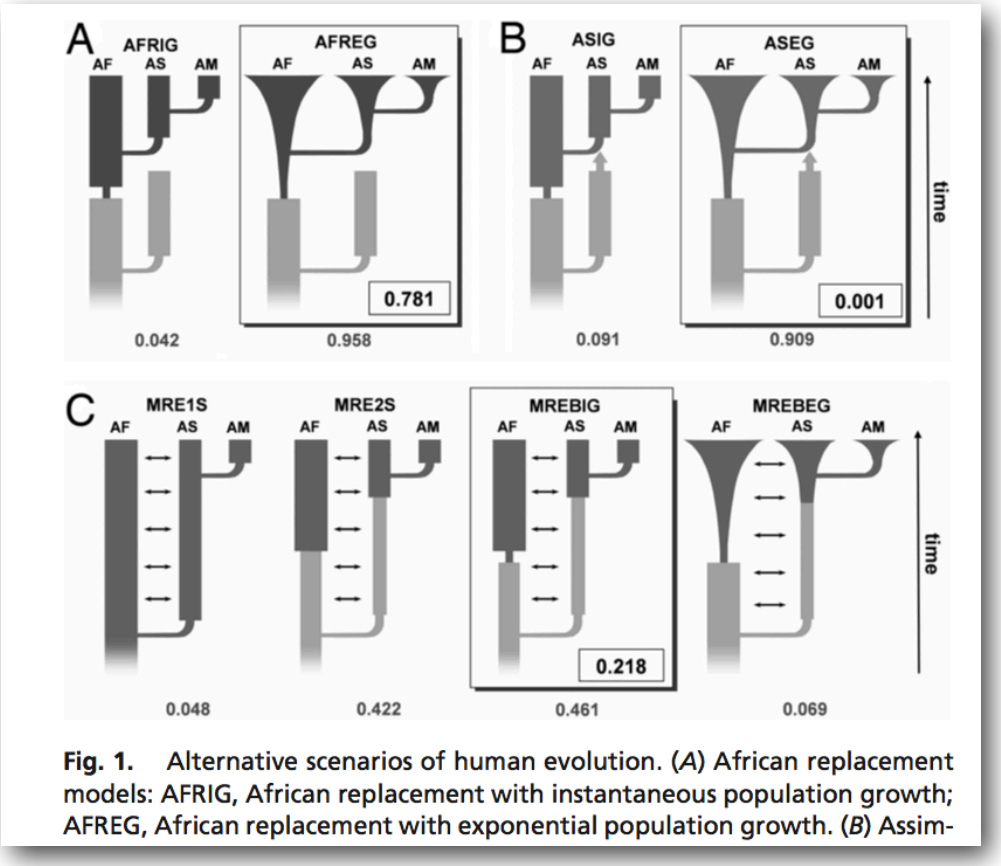


Fig. 1. Alternative scenarios of human evolution. (*A*) African replacement models: AFRIG, African replacement with instantaneous population growth; AFREG, African replacement with exponential population growth. (*B*) Assim-

"We then estimated the parameters of the overall best African replacement model (AFREG, Table 1 and SI Fig. 5) under an ABC framework based from **5 million simulations**.."

Example 2 - Human Demography

Table 1. Demographic and historical parameters estimated under the favored AFREG model

Parameters†	Median‡	95% HPD§
Speciation time for modern human, yr (T_{MH})	141,455	103,535–185,642
Exit out of Africa, yr (T_{AS})	51,102	40,135–70,937
Colonization of the Americas, yr (T_{AM})	10,280	7,647–15,945
Size of archaic African population ($N_{A\text{-}AF}$)	12,772	6,604–20,211
Bottleneck size during speciation (Nb_{MH})	600	76–1,620
Bottleneck size when leaving Africa (Nb_{AS})	462	64–1,224
Bottleneck size when leaving Asia (Nb_{AM})	452	71–1,280
Current African population size (N_{AF})	206,920	23,535–801,895
Current Asian population size (N_{AS})	20,262	1,938–62,726
Current American population size (N_{AM})	5,606	757–13,740

The estimates were calibrated by assuming a human-chimpanzee divergence of 6 million years and a generation time of 25 years.
†Population sizes are given in effective number of diploid individuals.

"Under this model, we find that an archaic African population of 12,800 effective individuals gave rise to modern humans 141 thousand years ago (Kya) after a bottleneck involving 600 effective individuals. The Out-of-Africa migration, initially involving only 450 effective individuals would have occurred some 51 Kya, and the Americas would have been colonized only 10.5 Kya by 450 individuals"

Biol Philos
DOI 10.1007/s10539-013-9391-1

The phylogeography debate and the epistemology of model-based evolutionary biology

Alfonso Arroyo-Santos · Mark E. Olson · Francisco Vergara-Silva

Received: 20 March 2012 / Accepted: 17 June 2013

Abstract Phylogeography, a relatively new subdicipline of evolutionary biology that attempts to unify the fields of phylogenetics and population biology in an explicit geographical context, has hosted in recent years a highly polarized debate related to the purported benefits and limitations that qualitative versus quantitative methods might contribute or impose on inferential processes in evolutionary biology. Here we present a friendly, non-technical introduction to the conflicting methods underlying the controversy, and exemplify it with a balanced selection of

Bayes factors

- In the Bayesian paradigm, for model selection (model 1 vs. model 2), we calculate the Bayes factor:

$$
\begin{aligned}
B_{12}(D) &= \frac{w_1(D)}{w_2(D)} \\
&= \frac{\int_{\Theta_1} \pi_1(\theta_1) f_1(D \mid \theta_1) d\theta_1}{\int_{\Theta_2} \pi_2(\theta_2) f_2(D \mid \theta_2) d\theta_2}
\end{aligned}
$$

- Bayes factor interpretation of evidence
  - 3: substantial
  - 10: decisive
  - 30: very strong (Jeffreys)

Bayes factors

- In the Bayesian paradigm, for model selection (model 1 vs. model 2), we calculate the Bayes factor:

$$B_{12}(D) = \frac{w_1(D)}{w_2(D)} = \frac{\int_{\Theta_1} \pi_1(\theta_1) f_1(D \mid \theta_1) d\theta_1}{\int_{\Theta_2} \pi_2(\theta_2) f_2(D \mid \theta_2) d\theta_2}$$

likelihood of data under model 1 - calculated by integrating over all param values with respect to the prior.

- Bayes factor interpretation of evidence
  - 3: substantial
  - 10: decisive
  - 30: very strong (Jeffreys)

ABC approximation to Bayes Factor

non-ABC:
$$B_{12}(D) = \frac{\int_{\Theta_1} \pi(\theta_1) f_1(D \mid \theta_1)}{\int_{\Theta_2} \pi(\theta_2) f_2(D \mid \theta_2)}$$

ABC:
$$\hat{B}_{12}(D) = \frac{\int_{\Theta_1} \pi(\theta_1) \hat{f}_1(S_1 \mid \theta_1)}{\int_{\Theta_2} \pi(\theta_2) \hat{f}_2(S_2 \mid \theta_2)}$$

S_1, S_2 Sufficient:
$$B_{12}(D) \sim \frac{f(D \mid S_1)}{f(D \mid S_2)} \hat{B}_{12}(D)$$

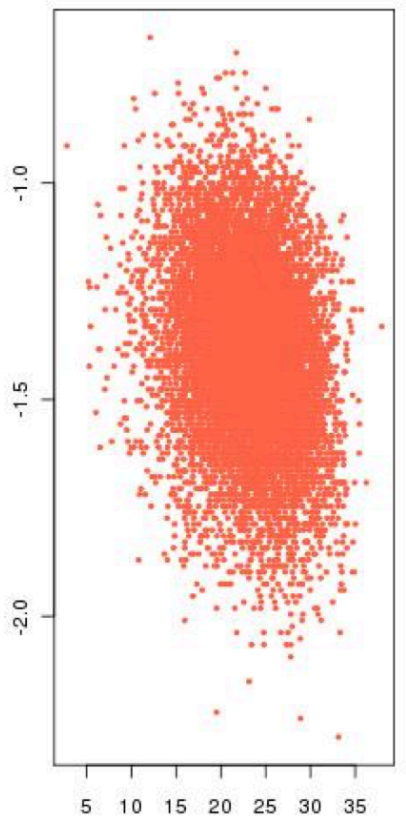
Sufficiency of statistics for models individually does not imply sufficiency for model selection!

ABC approximation to Bayes Factor

Example: Given count data that could come from a **Poisson or Geometric** distribution.

Use sufficient statistics for model parameters (individually): S = the sum of the counts.

Compare to importance sampling results (possible in this simple case)

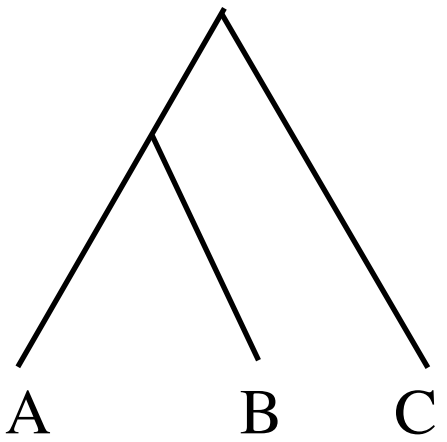


x-axis: true log Bayes factor
y-axis: estimated log Bayes factor

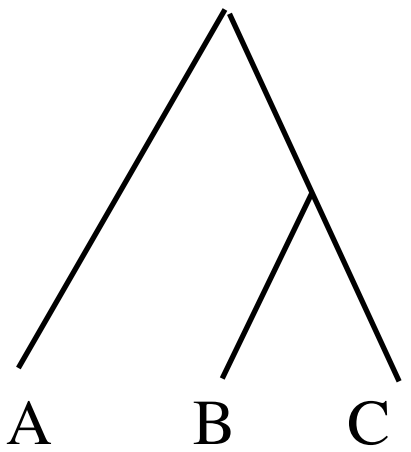
Robert, C. P., Cornuet, J.-M., Marin, J.-M., & Pillai, N. S. (2011). Lack of confidence in approximate Bayesian computation model choice. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(37), 15112–15117.

Pop. gen. example

- Three populations.
- 15 indivs per pop, genotyped at 15 microsat (stepwise mutation model).
- 24 summary statistics.
- Infer order of divergence between populations.
- First split: 60 gens ago
- Second split: 30 gens ago
- Compared Bayes Factors with those from an importance sampler.



Model 1



Model 2

Robert, C. P., Cornuet, J.-M., Marin, J.-M., & Pillai, N. S. (2011). Lack of confidence in approximate Bayesian computation model choice. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(37), 15112–15117.

Results

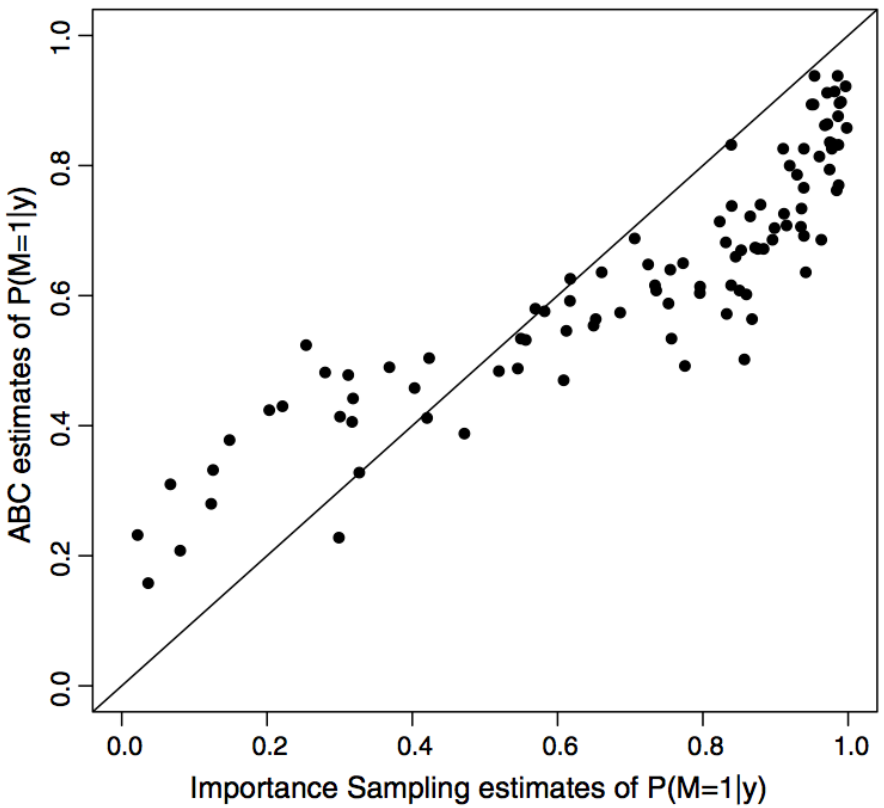


Fig. 1. Comparison of IS and ABC estimates of the posterior probability of scenario 1 in the first population genetic experiment, using 24 summary statistics.

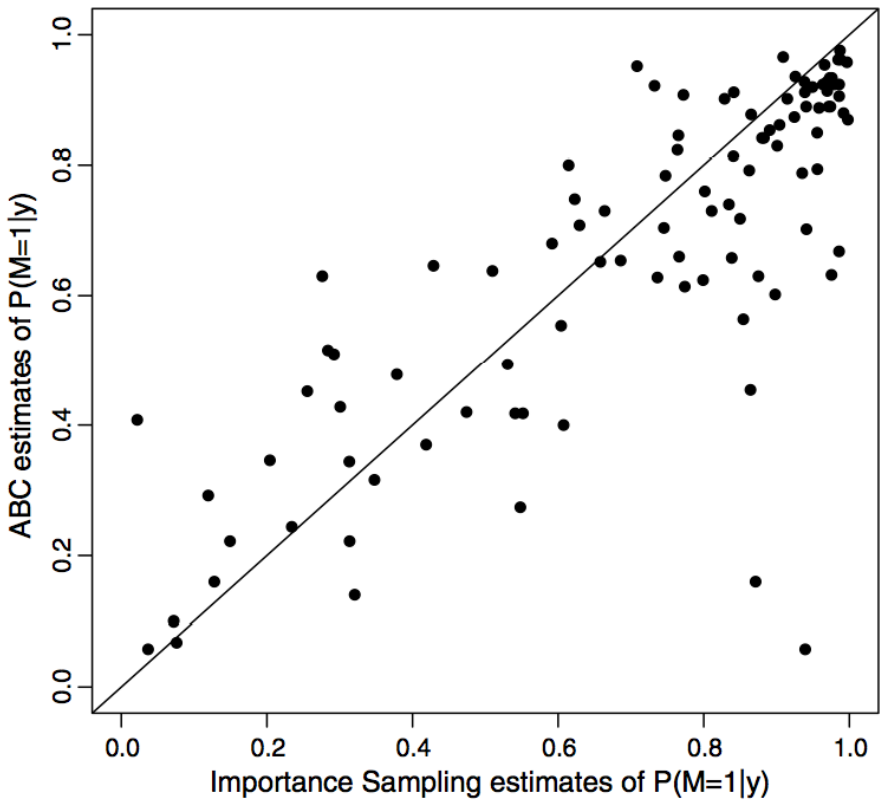


Fig. 2. Same caption as Fig. 1 when using 15 summary statistics.

So need to choose summary statistics carefully!

Robert, C. P., Cornuet, J.-M., Marin, J.-M., & Pillai, N. S. (2011). Lack of confidence in approximate Bayesian computation model choice. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(37), 15112–15117.

Stat Comput (2012) 22:1181–1197
DOI 10.1007/s11222-012-9335-7

Considerate approaches to constructing summary statistics for ABC model selection

Chris P. Barnes · Sarah Filippi · Michael P.H. Stumpf · Thomas Thorne

Paul Joyce
FormerDean of Science,
U. Idaho

Received: 30 June 2011 / Accepted: 10 May 2012 / Published online: 9 June 2012

Abstract For nearly any challenging scientific problem evaluation of the likelihood is problematic if not impossible. Approximate Bayesian computation (ABC) allows us to employ the whole Bayesian formalism to problems where we can use simulations from a model, but cannot evaluate the likelihood directly. When summary statistics of real and simulated data are compared—rather than the data

1 Introduction

Mathematical models are widely used to describe and analyze complex systems and processes across the natural, engineering and social sciences. Formulating a model to describe, e.g. a predator-prey system, geophysical process, communication system, or social network requires us to condense our assumptions and knowledge into a single coher

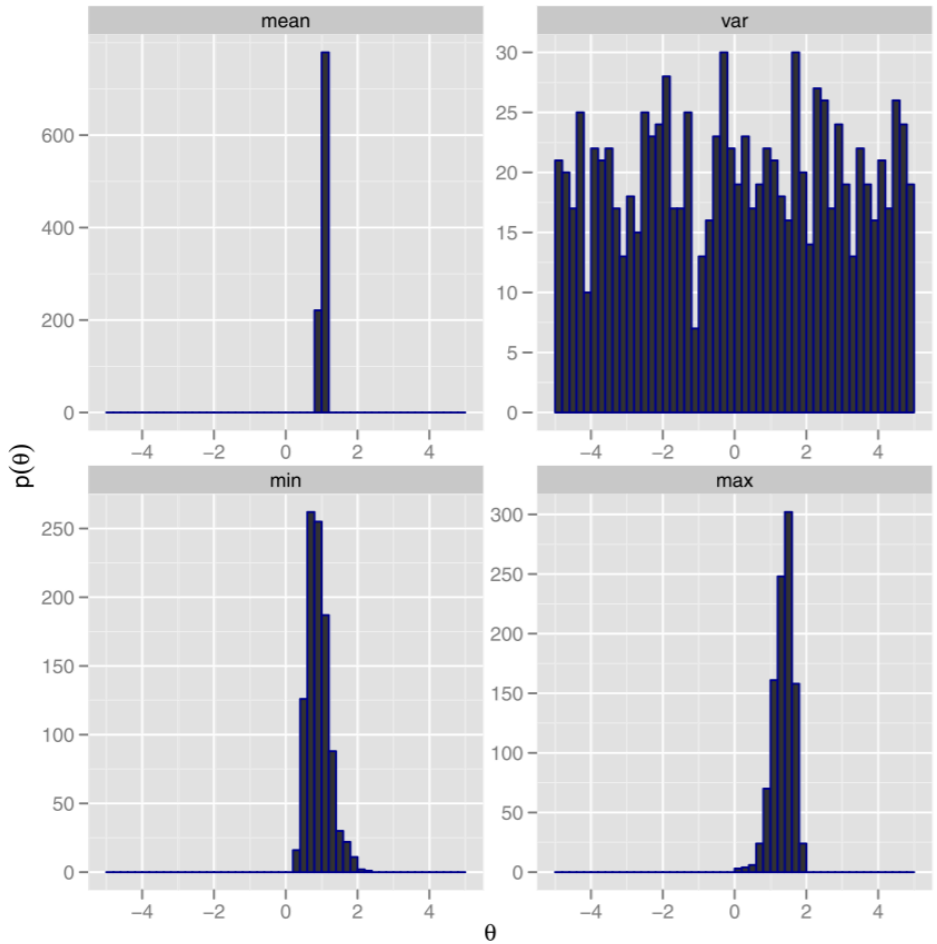


Fig. 1 Parameter inference for the mean of a normal model with known standard deviation, $\sigma^2 = 1$, using the mean, variance, maximum and minimum as a statistic. […] we generated 1000 acceptances of samples of size 10,000 with $\varepsilon = 0.001$.

The Normal distn from which the data were generated had a mean of 1

Which statistics to use?

Some statistics are more informative than others. If we can't tell which is which, we could argue as follows:

Including all statistics will make sure we are using the ones that matter.

If a statistic is non-informative, (independent of θ), it will not bias the estimation of posterior.

So….why not just include every statistic you can think of ?

Which statistics to use?

Some statistics are more informative than others. If we can't tell which is which, we could argue as follows:

Including all statistics will make sure we are using the ones that matter.

If a statistic is non-informative, (independent of θ), it will not bias the estimation of posterior.

So….why not just include every statistic you can think of ?

Adding unnecessary statistics will add empirical error, ultimately overwhelming the effect of the informative statistics:

Remember, we accept if $\sum_i w_i (S_i^o - S_i^s)^2 < \epsilon$

So we need something more principled. Their method:

Suppose we have a set of possible statistics S=$\{S_1,\ldots,S_w\}$.
Let $\mathbf{s}^* = \{\mathbf{s}_1^*, \ldots, \mathbf{s}_w^*\}$ be the values of the statistics on the dataset $\mathbf{x}^*$.

Goal: We aim to identify the subset $\mathcal{U}$ of S with minimum cardinality, such that $\mathcal{U}$ contains the same amount of information about θ as the whole set S *given the realization (data)* $\mathbf{x}^*$.

- In other words, conditional on $\mathcal{U}$, any non-included statistic S_k carries no further information about θ.
- Measure this by comparing the two posterior distributions $f(\theta|\mathcal{U})$ and $f(\theta|\mathcal{U} \cup S_k)$.

Pseudocode

Suppose we have a set of possible statistics $S=\{S_1,\ldots,S_w\}$. Define a tolerance δ.

1. Randomly choose a $v \in S$. Set $q= S \setminus v$.
2. While $q \neq \phi$ (the empty set)
   - I. Initialize d=0.
   - II. While ($d< \delta$)
     - A. Randomly choose $u \in q$. Set $q= q \setminus u$.
     - B. Perform ABC analysis to obtain $f(\theta \mid v \cup u)$.
     - C. Set $d = \| f(\theta \mid v \cup u), f(\theta \mid v) \|$ [They use Kullback - Leibler divergence for this metric.]
   - III. $v \leftarrow v \cup u$.
   - IV. $q \leftarrow s \setminus v$.
3. Return v. [This will be the set of statistics to use.]

Kullback - Leibler divergence:

$$KL(f,g) = \int_{-\infty}^{+\infty} f(y) \ln \frac{f(y)}{g(y)} dy$$

(A measure of how much information is lost when f is used to approximate g.)

Note: order in which statistics are tried may matter in previous algorithm. They offer a refinement to test for this as you go:

> After adding a statistic u to v, test all other statistics w $\in$ v\u to see whether you can now drop w without losing information.

END