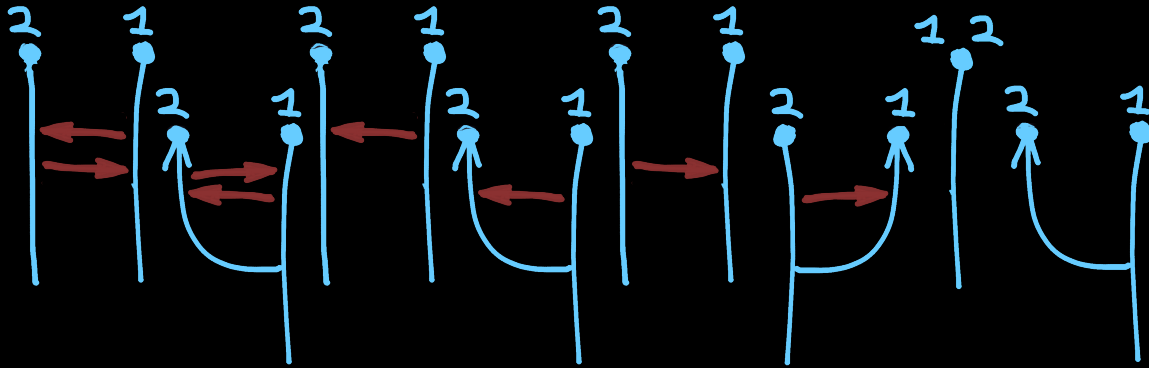
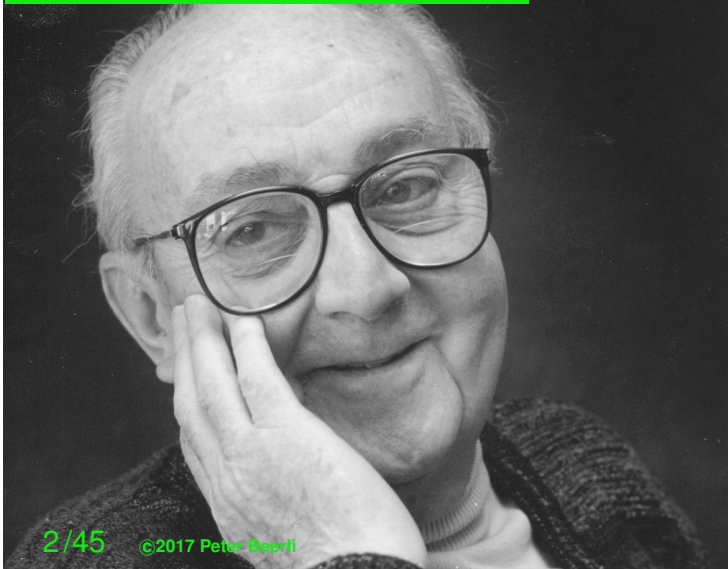


All models are good, but only some are useful



Peter Beerli Scientific Computing, Florida State University Twitter:@peterbeerli

On models

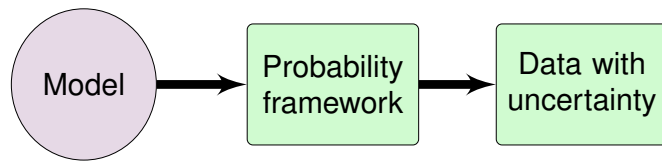
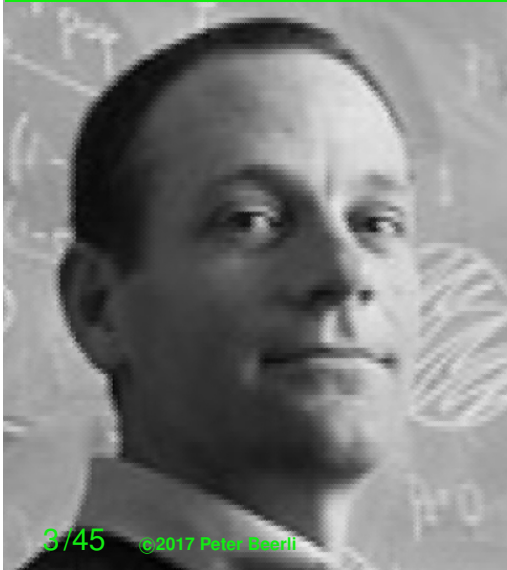


2/45 ©2017 Peter Beerli

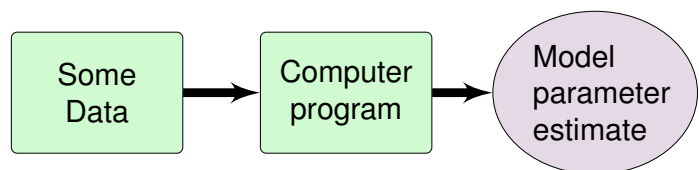
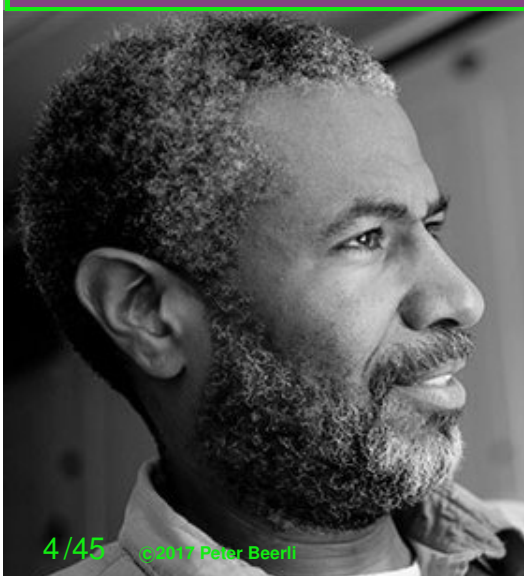
Essentially, all models are wrong, but some are useful.

Box, George E. P.; Norman R. Draper (1987). Empirical Model-Building and Response Surfaces, p. 424, Wiley.

On models and data



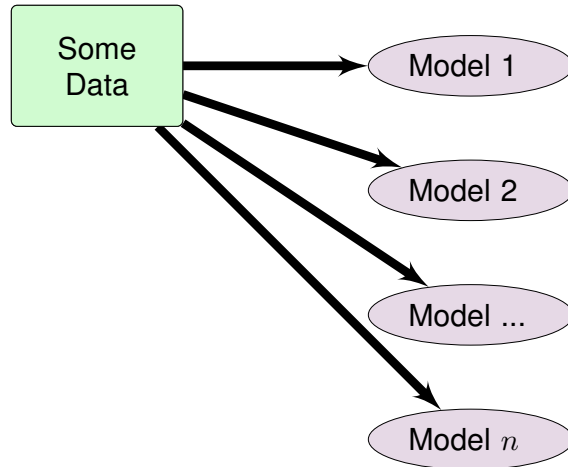
On data and models



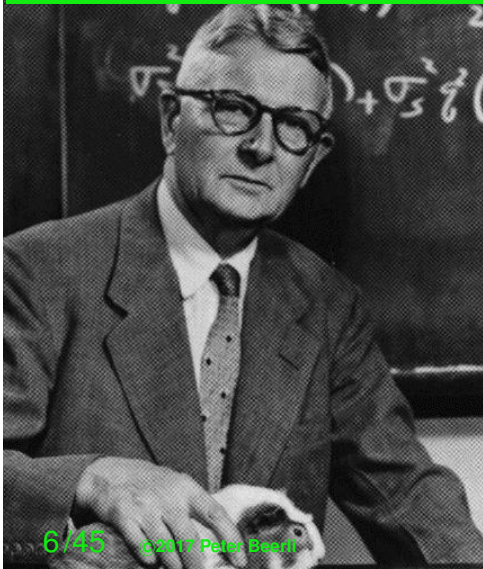
On models and data



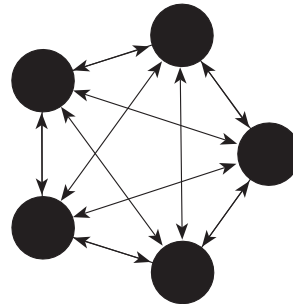
5/45 ©2017 Peter Beerli



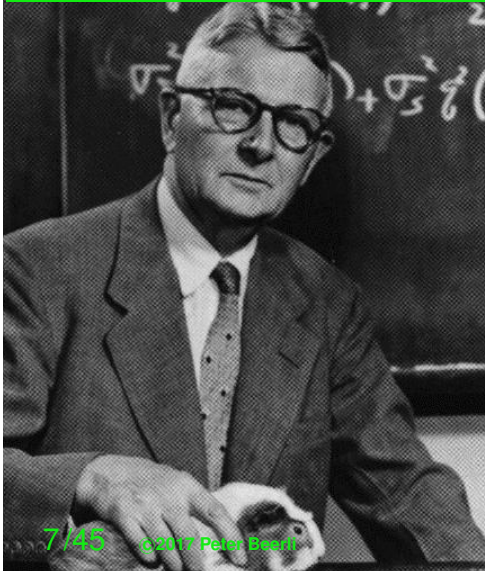
Population genetics models



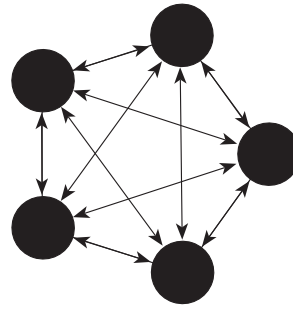
6/45 ©2017 Peter Beerli



Population genetics models



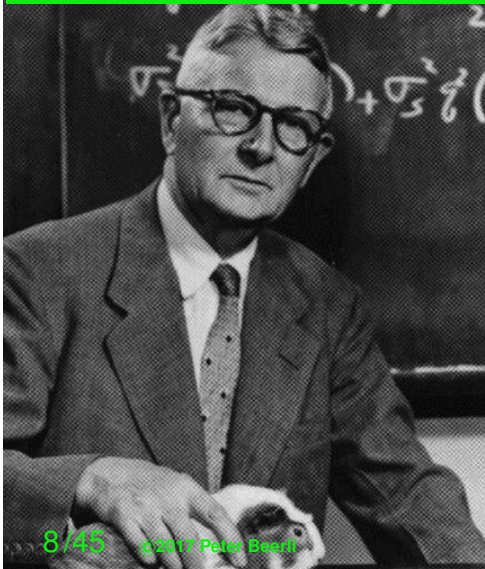
7/45 ©2017 Peter Beerli



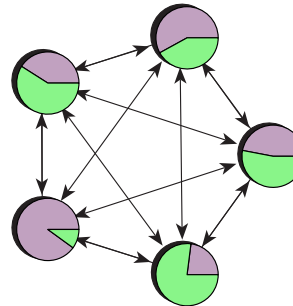
$$F_{ST} = \frac{\sigma^2(p)}{p(1-p)} \approx \frac{H_T - \bar{H}_S}{H_T}$$

$$F_{ST} \approx \frac{1}{4Nm+1}$$

Population genetics models



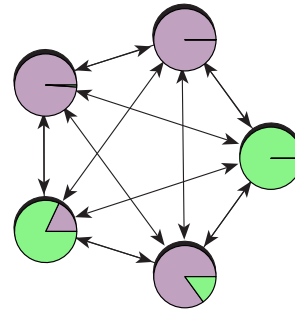
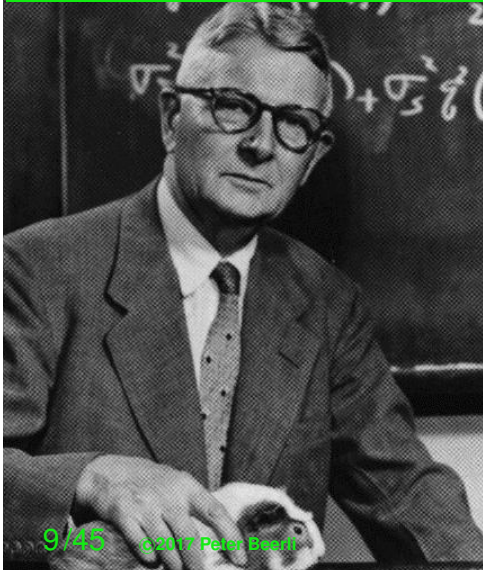
8/45 ©2017 Peter Beerli



$$F_{ST} = \frac{\sigma^2(p)}{p(1-p)} = \mathbf{0.25}$$

$$Nm \approx \frac{1}{4} \left(\frac{1}{F_{ST}} - 1 \right) \approx \mathbf{0.76}$$

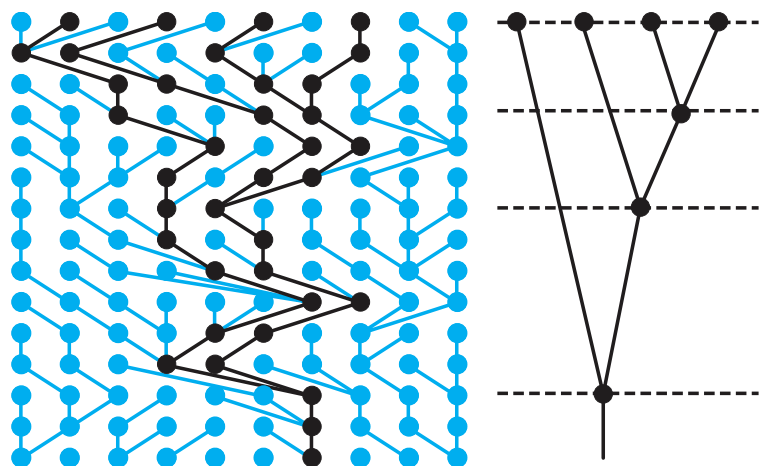
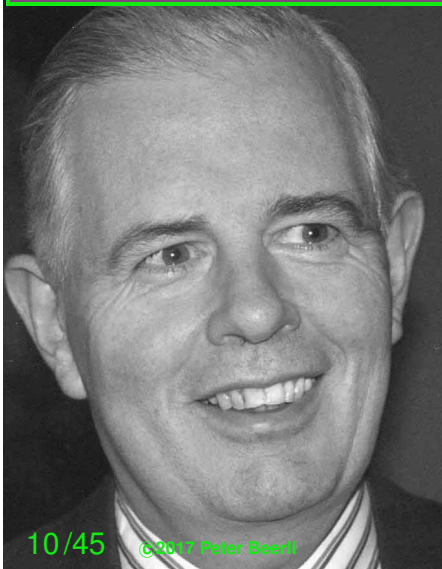
Population genetics models



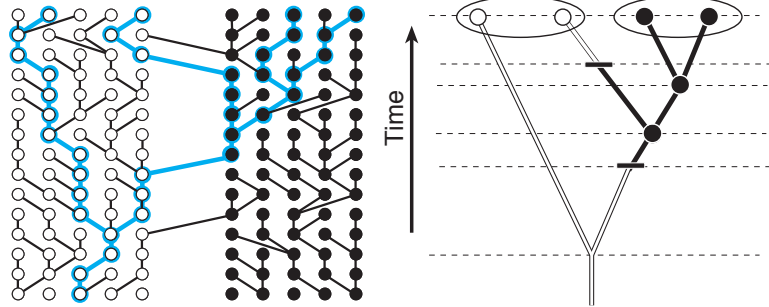
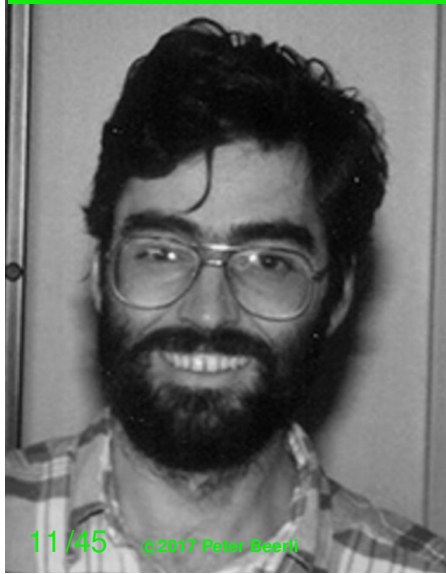
$$F_{ST} = \frac{\sigma^2(p)}{p(1-p)} = 0.95$$

$$Nm \approx \frac{1}{4} \left(\frac{1}{F_{ST}} - 1 \right) \approx 0.01$$

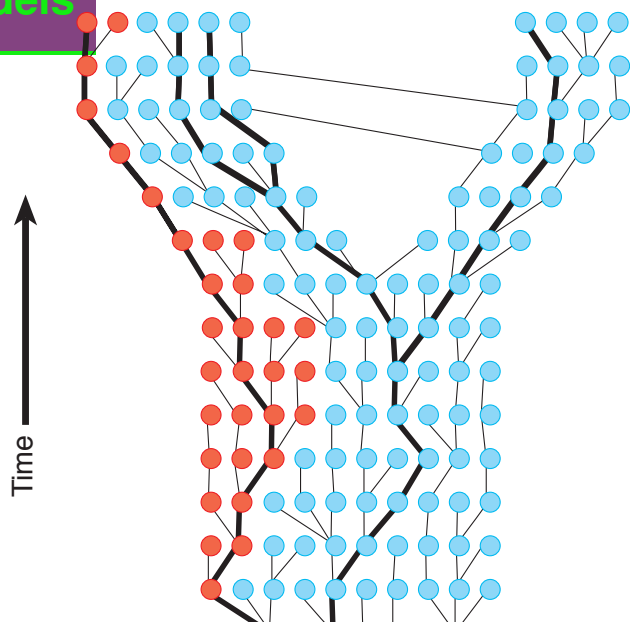
Population genetics models



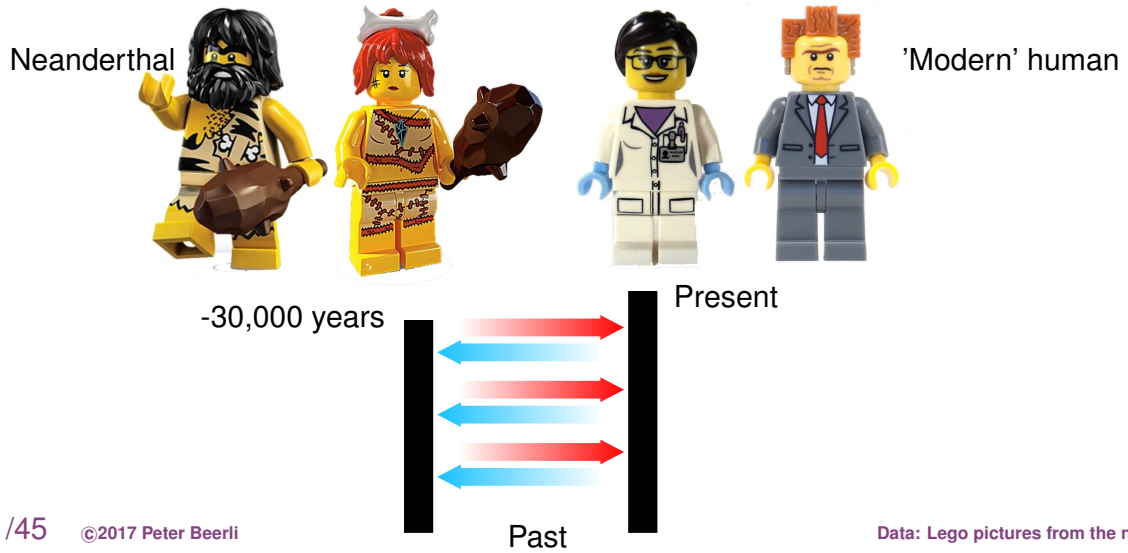
Population genetics models



Population genetics models



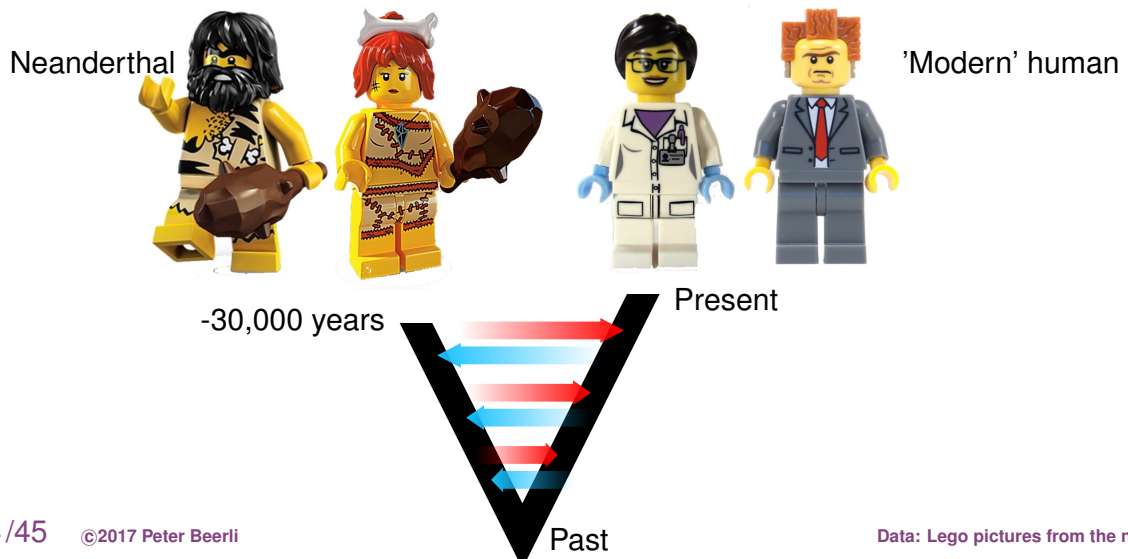
Gene flow



13/45 ©2017 Peter Beerli

Data: Lego pictures from the net

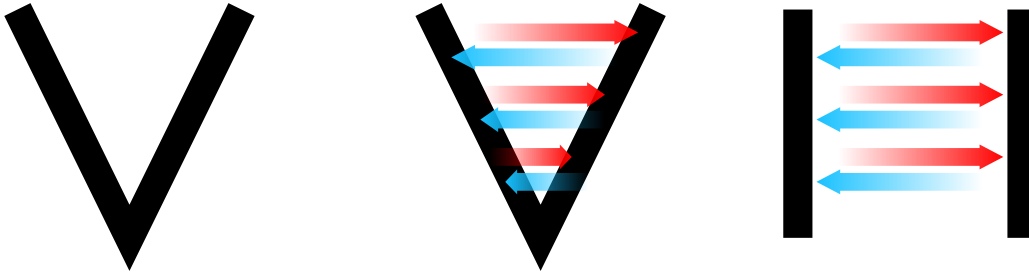
Divergence and Gene flow



14/45 ©2017 Peter Beerli

Data: Lego pictures from the net

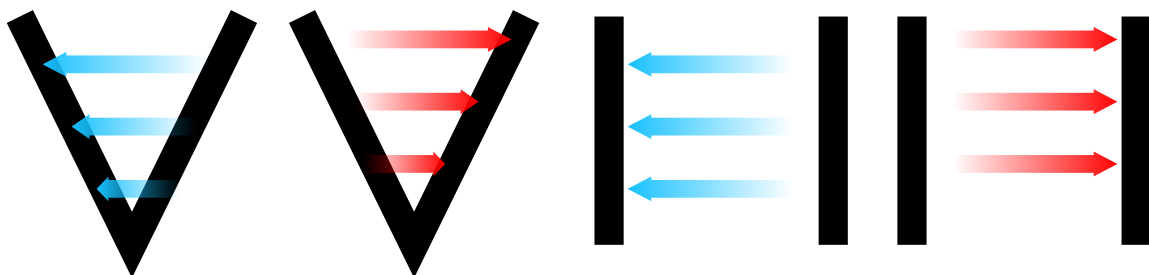
Different structural models



15/45 ©2017 Peter Beerli

Summary

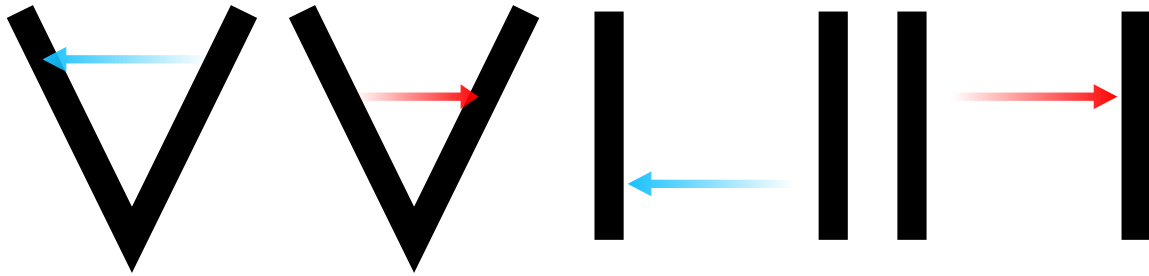
More different structural models



16/45 ©2017 Peter Beerli

Summary

Even more different structural models



17/45 ©2017 Peter Beerli

Summary

So many models – so little time



18/45 ©2017 Peter Beerli

Photo CC Wikimedia Wolfgang Sauber

Model comparison

- ◆ Several tests that establish whether two locations belong to the same population exist. The test by Hudson and Kaplan (1995) seemed particularly powerful even with a single locus.
- ◆ These days researchers mostly use the program STRUCTURE to establish the number of populations.
- ◆ A procedure that not only can handle panmixia versus all other gene flow models would help.

19/45 ©2017 Peter Beerli

Model comparison

With a criterium such as likelihood we can compare nested models. Commonly we use a likelihood ratio test (LRT) or Akaike's information criterion (AIC) to establish whether phylogenetic trees are statistically different or mutation models have an effect on the outcome, etc.

Kass and Raftery (1995) popularized the [Bayes Factor](#) as a Bayesian alternative to the LRT.

20/45 ©2017 Peter Beerli

Bayesian inference

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

21/45 ©2017 Peter Beerli

Bayes factor

Theoretically, we can calculate the posterior probability density of the model

$$p(M_1|X) = \frac{p(M_1)p(X|M_1)}{p(X)}$$

22/45 ©2017 Peter Beerli

Bayes factor

Theoretically, we can calculate the posterior probability density of the model 1 and model 2

$$p(M_1|X) = \frac{p(M_1)p(X|M_1)}{p(X)}$$

$$p(M_2|X) = \frac{p(M_2)p(X|M_2)}{p(X)}$$

23/45 ©2017 Peter Beerli

Bayes factor

Theoretically, we can calculate the posterior probability density of the model 1 and model 2

$$\frac{p(M_1|X)}{p(M_2|X)} = \frac{\frac{p(M_1)p(X|M_1)}{p(X)}}{\frac{p(M_2)p(X|M_2)}{p(X)}}$$

24/45 ©2017 Peter Beerli

Bayes factor

We could look at the **posterior odds ratio** or equivalently the **Bayes factors**.

$$\frac{p(M_1|X)}{p(M_2|X)} = \frac{p(M_1)}{p(M_2)} \times \frac{p(X|M_1)}{p(X|M_2)}$$

$$BF = \frac{p(X|M_1)}{p(X|M_2)} \quad LBF = 2 \ln BF = 2 \ln \left(\frac{p(X|M_1)}{p(X|M_2)} \right)$$

25/45 ©2017 Peter Beerli

Bayes factor

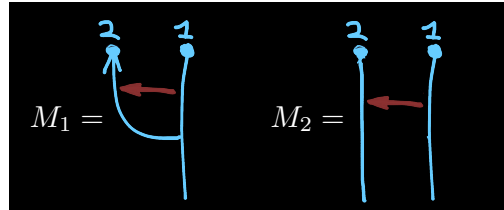
$$BF = \frac{p(X|M_1)}{p(X|M_2)} \quad LBF = 2 \ln BF = 2 \ln \left(\frac{p(X|M_1)}{p(X|M_2)} \right)$$

The magnitude of BF gives us evidence against or for hypothesis M_2

$$LBF = 2 \ln BF = z \quad \begin{cases} 0 < |z| < 2 & \text{No real difference} \\ 2 < |z| < 6 & \text{Positive} \\ 6 < |z| < 10 & \text{Strong} \\ |z| > 10 & \text{Very strong} \end{cases}$$

26/45 ©2017 Peter Beerli

Bayes factor example



$$\text{LBF} = 2 \ln \text{BF} = 2 \ln \left(\frac{p(X|M_1)}{p(X|M_2)} \right) = 2(-9638.69) - (-9641.01) = 4.64$$

The magnitude of BF gives us evidence against or for hypothesis M_2

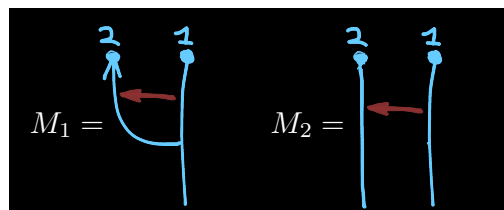
$$\text{LBF} = 2 \ln \text{BF} = z \quad \begin{cases} 0 < |z| < 2 & \text{No real difference} \\ 2 < |z| < 6 & \text{Positive} \\ 6 < |z| < 10 & \text{Strong} \\ |z| > 10 & \text{Very strong} \end{cases}$$

27/45 ©2017 Peter Beerli

Posterior model probability

(example continued)

Instead of calculating the Bayes factor we could use the probability of all tested models M_i and use them as weights (cf. Burnham and Anderson, 1998)



$$p_i^* = \frac{p(X|M_i)}{\sum_j p(X|M_j)}, \quad \sum_i p_i^* = 1, \quad \ell_1 = -9638.61, \quad \ell_2 = -9641.01$$

$$p_1^* = \frac{\exp(\ell_1)}{\exp(\ell_1) + \exp(\ell_2)} = 0.911$$

$$p_2^* = \frac{\exp(\ell_2)}{\exp(\ell_1) + \exp(\ell_2)} = 0.089$$

28/45 ©2017 Peter Beerli

Marginal likelihood

So why are we not all running BF analyses instead of the other model selection measures, such as

- ◆ LRT: Likelihood ratio test
- ◆ AIC: Akaike's information criterion
- ◆ BIC: Bayesian information criterion
- ◆ DIC: Deviance information criterion
- ◆ others

29/45 ©2017 Peter Beerli

Marginal likelihood

Typically, it is rather difficult to calculate the marginal likelihoods with good accuracy, because most often we only approximate the posterior distribution using Markov chain Monte Carlo (MCMC).

In MCMC we need to know only differences and therefore we typically do not need to calculate the denominator to calculate the Posterior distribution $p(\Theta|X)$:

$$p(\Theta|X, M) = \frac{p(\Theta)p(X|\Theta)}{p(X|M)} = \frac{p(\Theta)p(X|\Theta)}{\int_{\Theta} p(\Theta)p(X|\Theta)d\Theta}$$

where $p(X|M)$ is the marginal likelihood, **which we need for our model selection!**

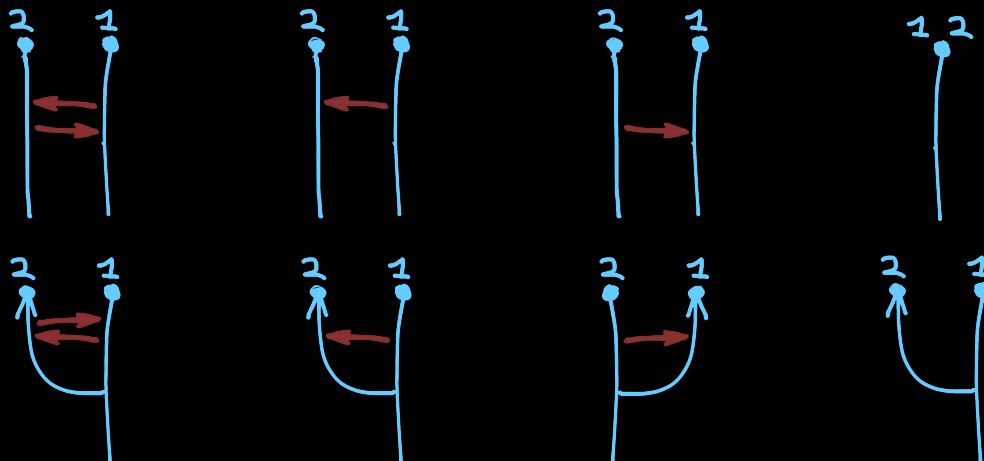
30/45 ©2017 Peter Beerli

Estimation of the marginal likelihood

- ◆ Harmonic mean estimator [Kass and Raftery 1995]: method is easy and used in many programs, results are biased and overestimate the marginal likelihood, variance of estimates can be very large.
- ◆ Thermodynamic integration (Path sampling) [Gelman and Meng 1997, Lartillot et al. 2006]: method is tedious to compute because several MCMC chains are needed. Results are accurate and reproducible with small variance when MCMC runs were run long enough.
- ◆ Stepping stone approach (Xie et al. 2011)

31/45 ©2017 Peter Beerli

Population models

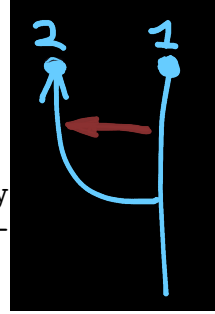


32/45 ©2017 Peter Beerli

Simulated data

Two loci simulated from model x0Dx:

Model	Log(mL)	LBF*	Model-probability
1: xxxxx:	-9662.42	-23.73	0.0000
2: xDxx:	-9661.98	-23.29	0.0000
3: xxDx:	-9661.52	-22.83	0.0000
4: xd0x:	-9656.51	-17.82	0.0000
5: xD0x:	-9649.33	-10.64	0.0000
6: xx0x:	-9648.93	-10.24	0.0000
7: x0dx:	-9641.77	-3.08	0.0402
8: x0xx:	-9641.01	-2.32	0.0859
9: x0Dx:	-9638.69	0.00	0.8739

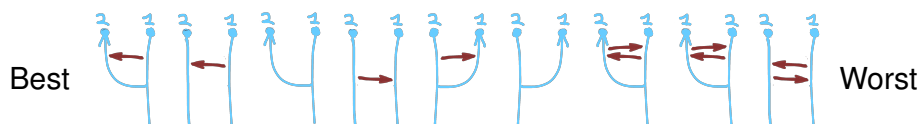
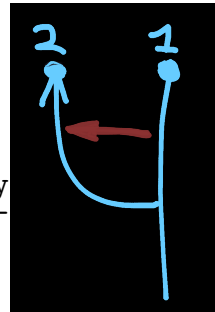


33/45 ©2017 Peter Beerli

Simulated data

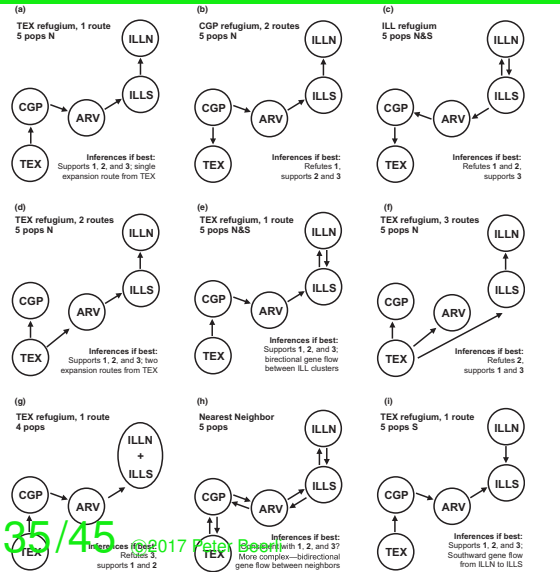
Two loci simulated from model x0Dx:

Model	Log(mL)	LBF*	Model-probability
1: xxxxx:	-9662.42	-23.73	0.0000
2: xDxx:	-9661.98	-23.29	0.0000
3: xxDx:	-9661.52	-22.83	0.0000
4: xd0x:	-9656.51	-17.82	0.0000
5: xD0x:	-9649.33	-10.64	0.0000
6: xx0x:	-9648.93	-10.24	0.0000
7: x0dx:	-9641.77	-3.08	0.0402
8: x0xx:	-9641.01	-2.32	0.0859
9: x0Dx:	-9638.69	0.00	0.8739



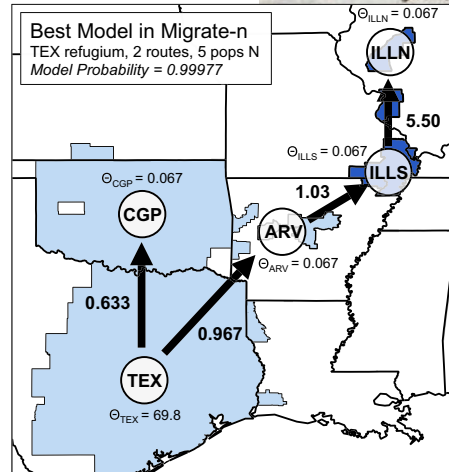
34/45 ©2017 Peter Beerli

A real example



Frog picture: <http://mdc.mo.gov/discover-nature/field-guide>

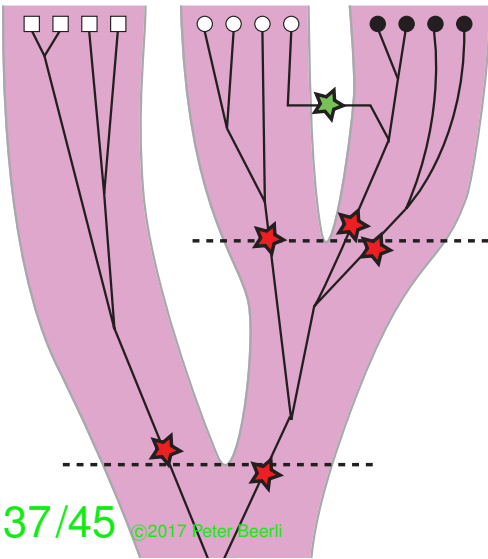
Lisa N. Barrow, A. T. Bigelow, C. A. Phillips, and E. Moriarty Lemmon (2015) Phylogeographic inference using Bayesian model comparison across a fragmented chorus frog species complex. Molecular Ecology






Model components

- ◆ Effective population size
- ◆ Sample size
- ◆ Processes that add variants:
 - ◆ Mutation rate
 - ◆ Migration or Admixture
- ◆ Processes that remove variants:
 - ◆ Genetic drift
 - ◆ Population splitting
 - ◆ (Selection) [I reserve this for another day – we work on that]

Population splitting model



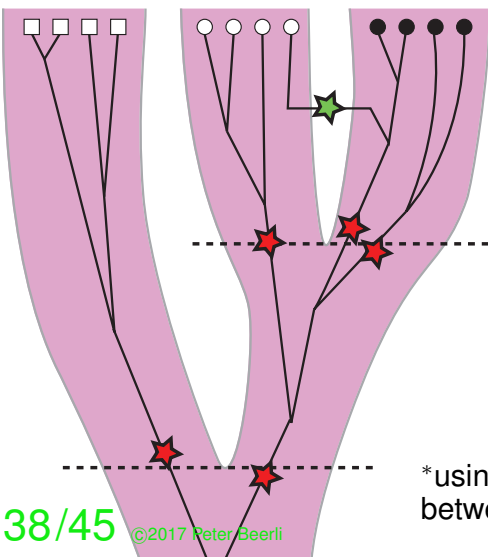
The relationship among individuals can be expressed, looking backward in time, by a waiting process where random lineages

-  coalesce
-  migrate between populations
-  split off an ancestral population

Each of these processes can be expressed as a waiting time process with rate λ for N populations and k_j lineages in population j .

37/45 ©2017 Peter Beerli

Population genetics



Each of these processes can be expressed as a waiting time process with rate λ for N populations and k_j lineages in population j :

$$\lambda_{\text{two lineages coalesce}} = \sum_{j=1}^N \frac{k_j(k_j - 1)}{4N}$$

$$\lambda_{\text{lineages migrate}} = \sum_{j=1}^N \sum_{i=1, i \neq j}^N k_j m_{ij}$$

$$\lambda_{\text{a lineage splits off*}} = \frac{\sqrt{\frac{2}{\pi}} e^{-\frac{(t-\mu)^2}{2b^2}}}{b \left(1 - \text{erf} \left(\frac{t-\mu}{\sqrt{2}b}\right)\right)}$$

*using a Normal distribution to model the splitting time between two populations.

38/45 ©2017 Peter Beerli

Combining the parts

$$P(\Theta | \mathbf{D}_1, \mathbf{D}_2, \dots, \mu) = \frac{P(\Theta)P(\mathbf{D}_1, \mathbf{D}_2, \dots | \Theta)}{P(\mathbf{D}_1, \mathbf{D}_2, \dots)} = \frac{P(\Theta) \int_G P(G | \Theta) \prod_i^{n_{\text{Loci}}} P(\mathbf{D}_i | \Theta, \mu) dG}{\int_{\Theta} P(\Theta) \int_G P(G | \Theta) \prod_i^{n_{\text{Loci}}} P(\mathbf{D}_i | \Theta, \mu) dG d\Theta}$$

$$P(G | \Theta) = \prod_{i=1}^K \lambda_x \exp(-t_i [\lambda_{\text{coalescence}} + \lambda_{\text{migration}} + \lambda_{\text{splitting}}])$$

- Θ vector of parameters for population size, migration and splitting parameters.
- $\mathbf{D}_1, \mathbf{D}_2, \dots$ independent genetic sequence data,
- μ mutation model,
- G nuisance genealogies that we integrate out (we are interested in the parameters not the trees).
- x the particular event on the genealogy
- K number of total events on the genealogy

39/45 ©2017 Peter Beerli

Finally....

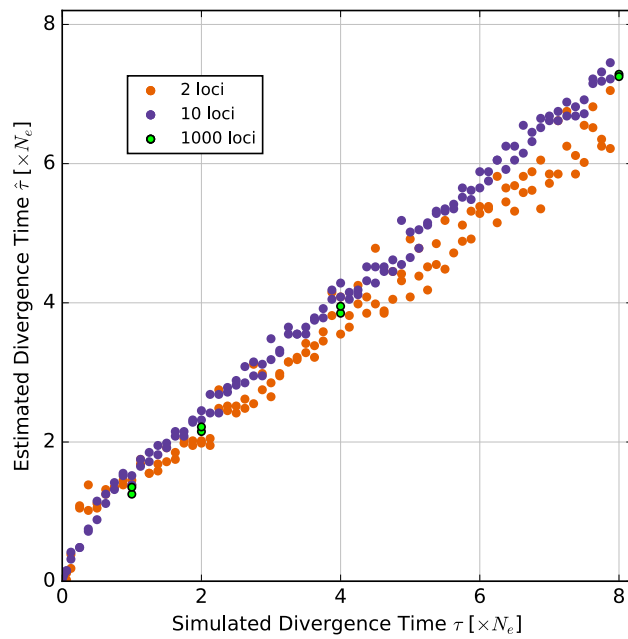
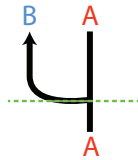
$$p(\mathbf{D} | \Theta) = \int_G p(G | \Theta) p(\mathbf{D} | G) dG$$

The number of possible genealogies is very large and for realistic data sets, programs need to use Markov chain Monte Carlo methods.

40/45 ©2017 Peter Beerli

Population splitting

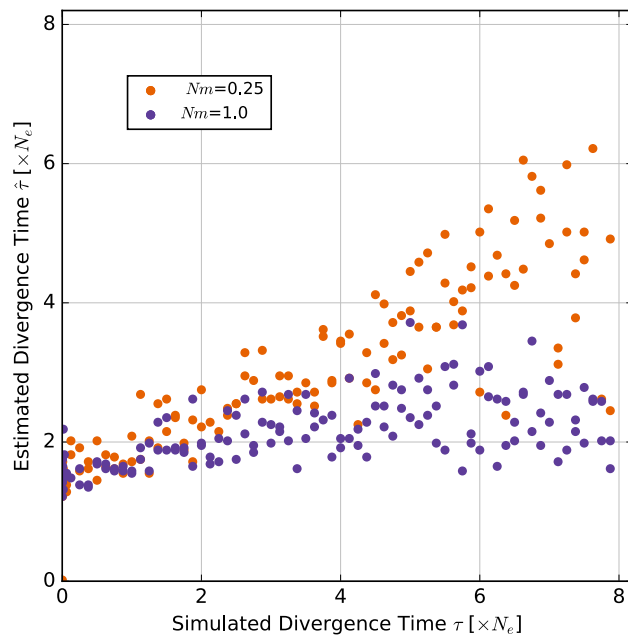
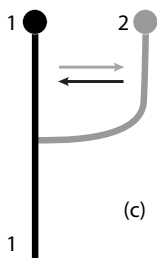
Comparison of estimated versus simulated divergence times for different number of loci



41/45 ©2017 Peter Beerli

Population splitting

Sampled and Analyzed



42/45 ©2017 Peter Beerli

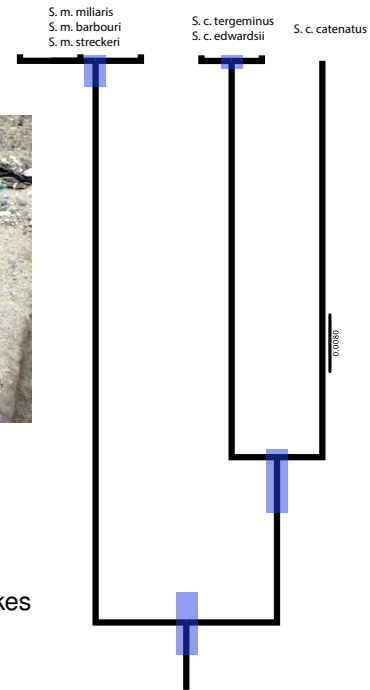
Population splitting



Model	Log(mL)	LBF	Model-probability
1: 3 species:	-15887.49	0.00	1.0000
2: 6 species:	-15961.95	-74.46	0.0000

Estimation of splitting dates of 6 subspecies of pygmy rattlesnakes using MIGRATE (data from Kubatko et al. 2011)

43/45 ©2017 Peter Beerli



Summary



44/45 ©2017 Peter Beerli

- ◆ You may be surprised that your favored model does not win in a model comparison competition, but figuring out the model order leads oftentimes to new insights about the problem.
- ◆ Models by themselves are not true or wrong. BUT they may not fit your data well, OR they describe your data even when you “know” that the model is insufficient.

Thank you



Lucrezia Bieler



National Science
Foundation



Michal Palzcewski, <http://popgen.sc.fsu.edu>
Haleh Ashki,
Justin Bricker,
Somayeh Mashayekhi,
Kyle Shaw

