« Environmental Genetics » doctoral course ABIES-GAIA

Models in Population Genetics: a Reminder (or an Overview?)

Renaud Vitalis
Centre de Biologie pour la Gestion des Populations
INRA ; Montpellier
E-mail : vitalis@supagro.inra.fr

What is population genetics?

Population genetics

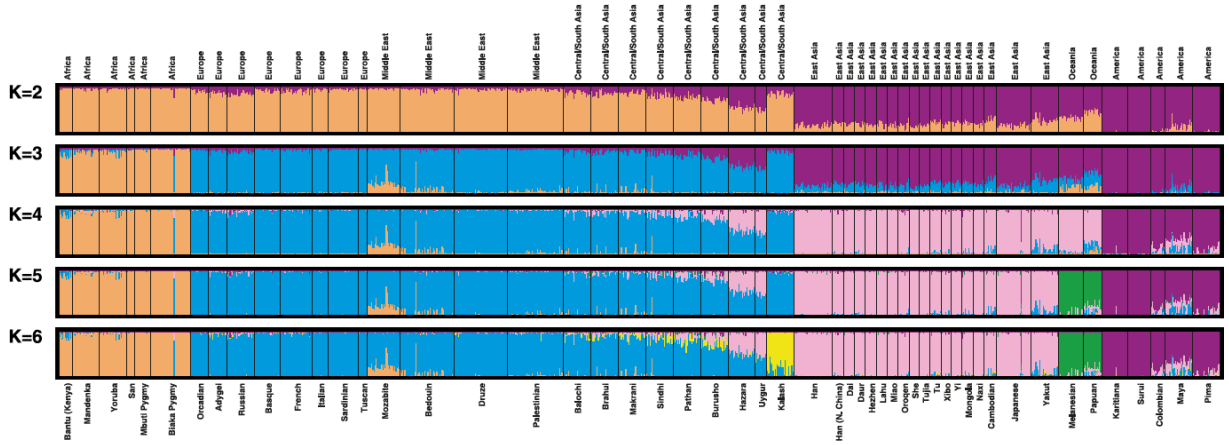
- Studying genetic variation in populations. Two aspects have been considered in the models:

- **Predictive**: predicting the future composition of a population from its current composition

- **Retrospective**: understanding what determined the current composition of a population

Population genetics

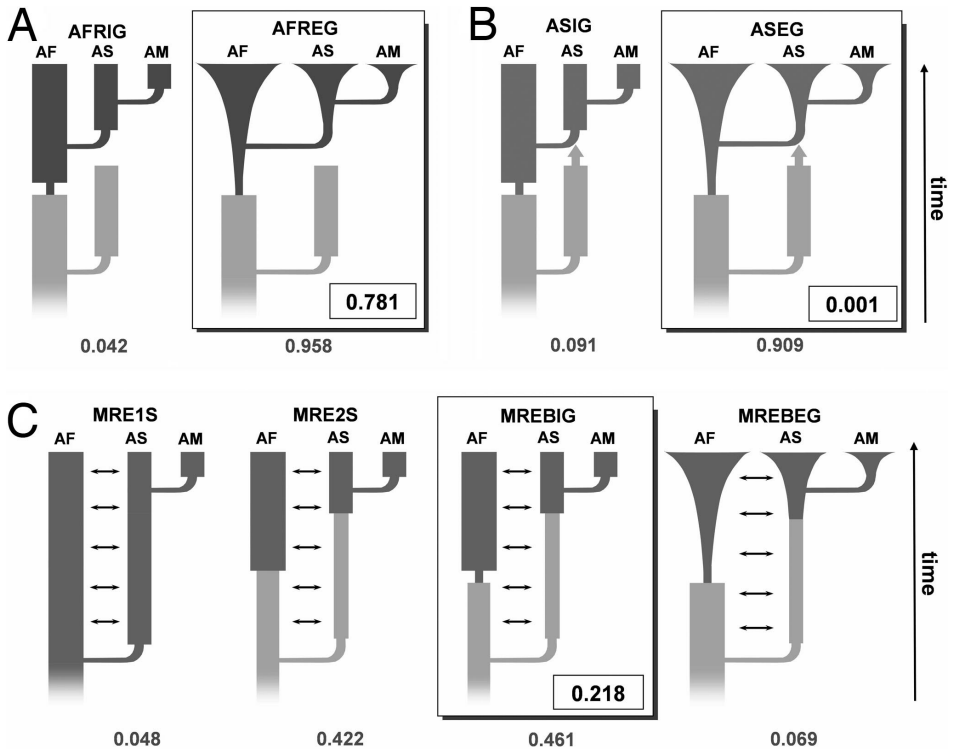
- **Data analysis relies on:**
- **Descriptive statistics** (characterizes the structure of the data)
- Define groups of individuals, quantify the distances among them



Population genetics

- **Data analysis relies on:**

- **Inference methods** (requires evolutionary models)



Credits : Fagundes et al. (2007) *PNAS* **104**: 17614-17619

Population genetics

- Define groups of individuals, quantify the distances among them

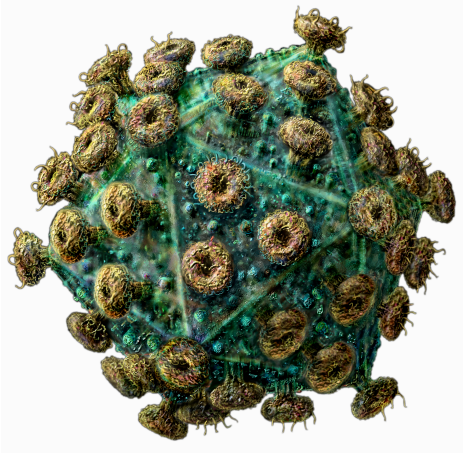
- Infer population history: since how long populations have diverged? Do they exchange migrants? Is there evidence of admixture between some populations? Etc.

What is a genetic marker?

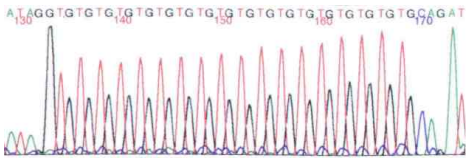
Types of genetic data

- A mutation (***single nucleotide polymorphism,* or SNP**) in the *MC1R* gene (melanocortin-1 receptor): TT homozygotes at position 478 tend to have freckles and red hairs

- A 32bp **deletion** in the CCR5 gene (CCR5-Δ32) confers resistance to HIV-1



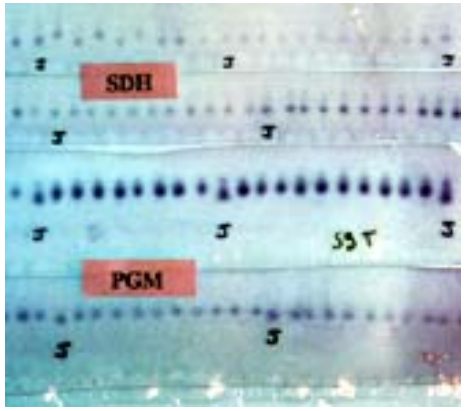
- **Microsatellites** markers (short tandem repeats, e.g.: AGAGAGAGAGAG…), dispersed in the genomes



Detecting differences in genotypes

- Until the 1960's, only the phenotypic differences can be observed: this is the golden age of **ecological genetics**

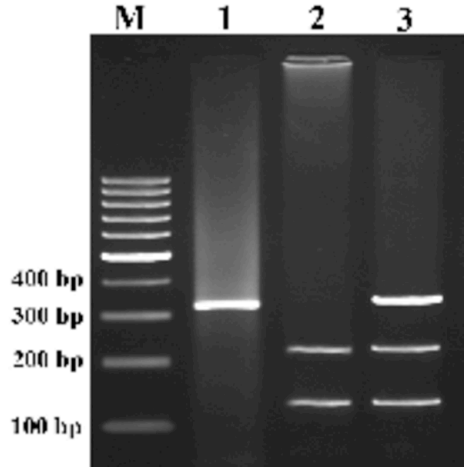
- In the 1960's, protein electrophoresis is developed in a number of (non-model) species



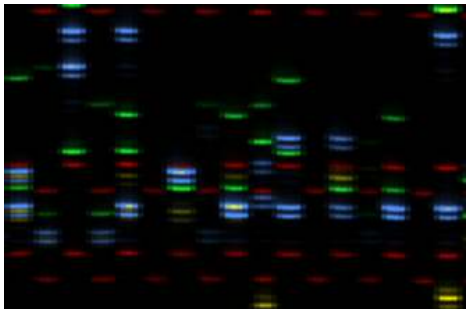
- In the 1980's, PCR and Sanger sequencing are used to sequence DNA (both mitochondrial and nuclear)

Detecting differences in genotypes

- Restriction enzymes (discovered in the 1970's) are used to develop RFLP markers



- In the 1990's: genotyping of microsatellite markers



- More recently: ***next generation sequencing*** (NGS)

The NGS revolution

- Publication of the first two human genome assemblies in 2001

- 2001 (human genome): 7 years for 3 10^9 $
- 2007 (horse genome): 18 months for 3 10^6 $
- In 2013, a resequenced human genome costs 1,000 $

- 1,000 genome projects in humans (McVean et al. 2011), rice (McNally et al. 2014), cattle (Hayes et al. 2014)

- **Marker availability is no more limiting…**

Genetic markers

- We use genetic markers to analyse the distribution of genetic polymorphism within individuals, within populations and among populations…

- An “ideal” genetic markers should:
- be **polymorphic!**
- have a **simple and known heredity!**
- be **co-dominant** (yet few methods exist for dominant)
- be **neutral** (only to infer demography)

What is the distribution of genotypes in populations?

Allele and genotype frequencies

- With panmixia (random union of gametes), Hardy-Weinberg equilibrium is reached in one generation:

Male gametes \ Female gametes	A $p[t]$	a $q[t]$
A $p[t]$	AA $p[t]^2$	Aa $p[t]q[t]$
a $q[t]$	aA $q[t]p[t]$	aa $q[t]^2$

AA	$D[t+1] = p[t]^2$
Aa	$H[t+1] = 2\,p[t]\,q[t]$
aa	$R[t+1] = q[t]^2$

- You may check that allele frequencies are constant

HWE: test of conformity

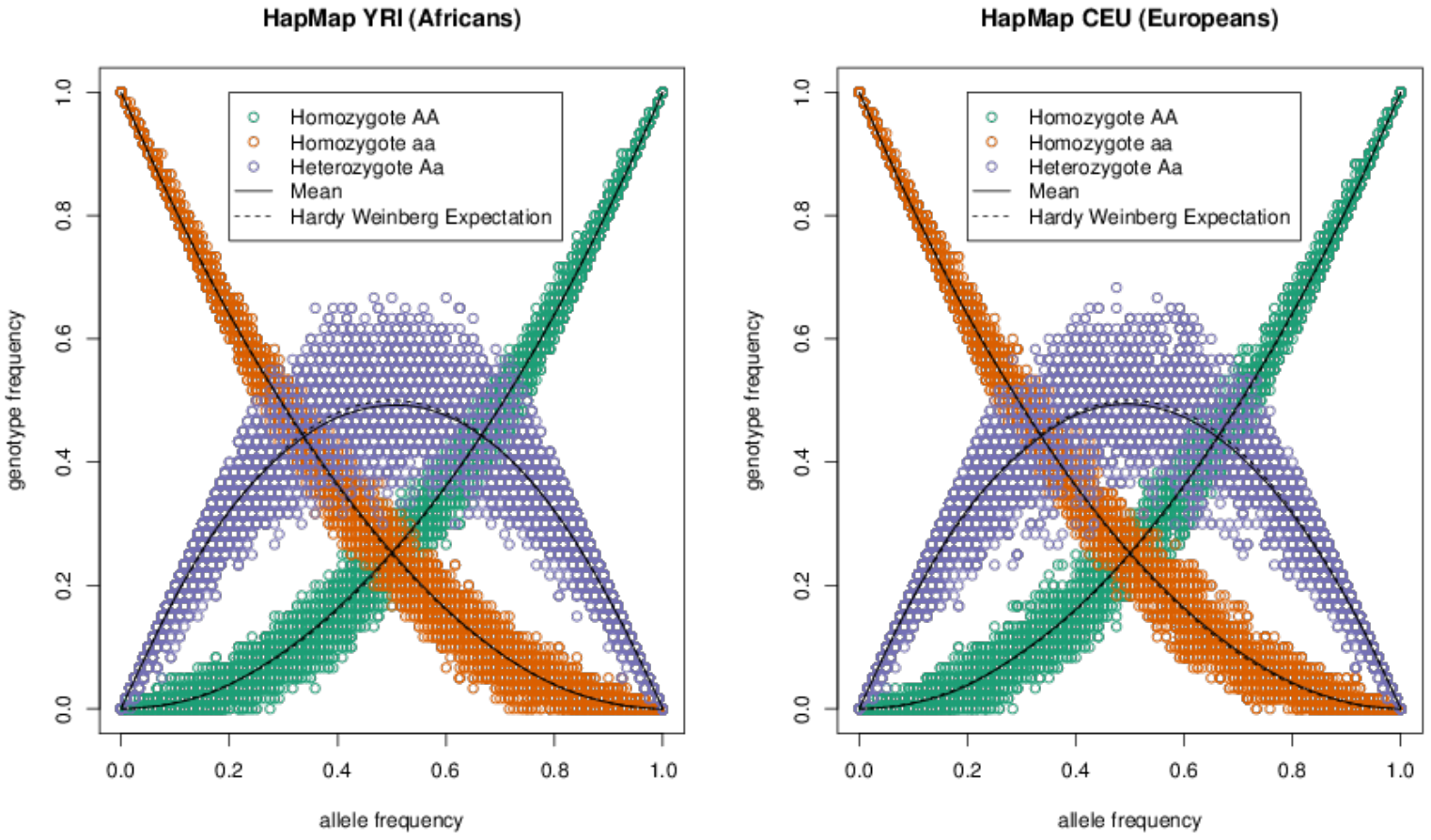
- For a bi-allelic locus, in a sample of size *N*:

	Expected	*Observed*
AA	Np^2	N_1
Aa	$2Npq$	N_2
aa	Nq^2	N_3

- The chi-square test statistic *X* measures the difference between the observed and the expected numbers. This statistic is distributed as a χ^2 with 1 degree of freedom (number of genotypes – number of constraints):

$$\chi^2 = \sum_{\text{genotypes}} \frac{(\text{expected - observed})^2}{\text{expected}}$$

HWE in humans



- 10,000 SNPs from the HapMap CEU European and YRI African populations fit pretty well to expectations

Credits : Graham Coop (https://gcbias.org/2011/10/13/population-genetics-course-resources-hardy-weinberg-eq/)

Hardy-Weinberg equilibrium

- **panmixia**: gametes encounter each other randomly
- All individuals reproduce simultaneously and then die (no overlapping generations)
- **Isolated populations**: no migration
- **Infinite** population size
- No **mutation**
- No **selection**

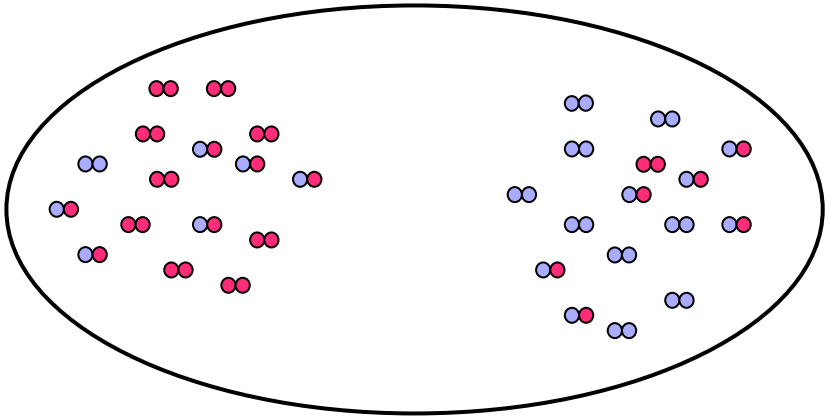
Deviations from HWE

- Assortative mating
- Mating systems (e.g., selfing, clonality, etc.)
- Population structure
- Selection

The Wahlund effect (1923)

- Over all populations, the **observed** frequencies are:

AA	$\sum_{i=1}^{n} p_i^2 / n$
Aa	$\sum_{i=1}^{n} 2 p_i q_i / n$
aa	$\sum_{i=1}^{n} q_i^2 / n$



- If the population was panmictic, the **expected** frequencies would be (with $\overline{p} = \sum_{i=1}^{n} p_i / n$):

AA	$\overline{p}^2$
Aa	$2\overline{p}\overline{q}$
aa	$\overline{q}^2$

The Wahlund effect (1923)

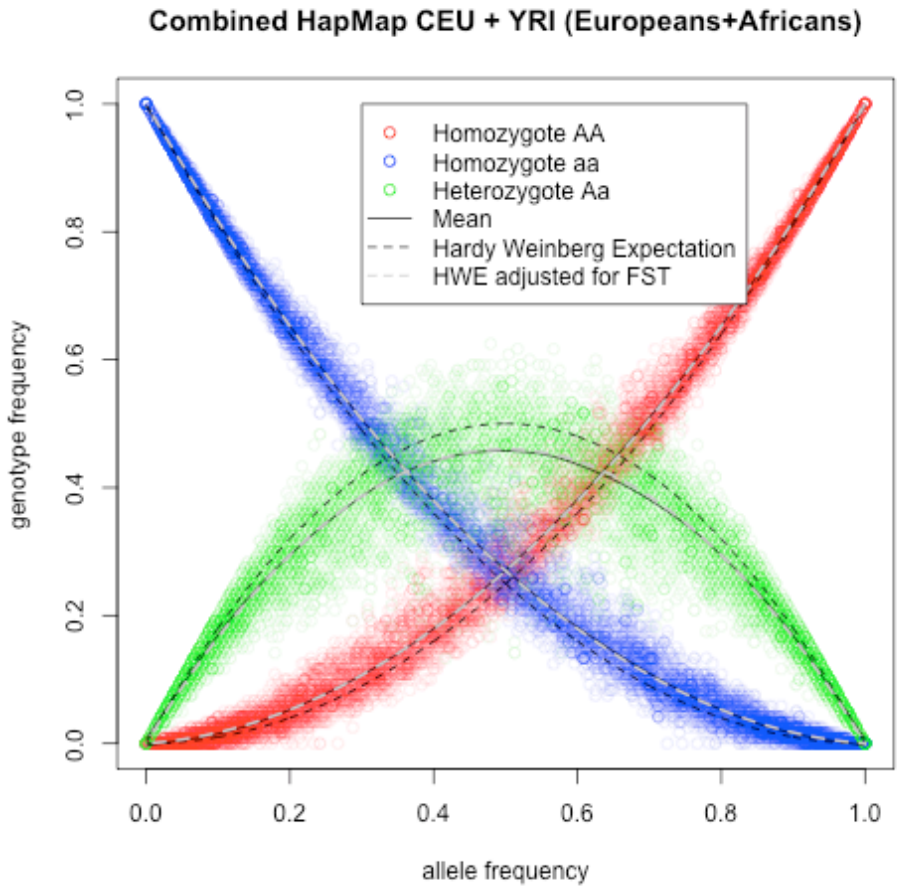
- The overall proportion of heterozygotes is:

$$
\begin{aligned}
H_{\mathrm{O}} &= \sum_{i=1}^{n} 2p_i q_i / n = 2\sum_{i=1}^{n} (p_i - p_i^2) / n \\
&= 2\sum_{i=1}^{n} p_i / n - 2\sum_{i=1}^{n} p_i^2 / n \\
&= 2\overline{p} - 2(\sigma_p^2 + \overline{p}^2) \\
&= 2\overline{p}\overline{q} - 2\sigma_p^2 \\
&= 2\overline{p}\overline{q}(1 - \sigma_p^2 / \overline{p}\overline{q})
\end{aligned}
$$

- Which is less than the expected proportion in a panmictic unit. We note $F_{\mathrm{ST}} = \sigma_p^2 / \overline{p}\overline{q}$, where σ_p^2 is the variance of *p* between populations, and hence:

$$H_{\mathrm{O}} = 2\overline{p}\overline{q}(1 - F_{\mathrm{ST}})$$

HWE in humans



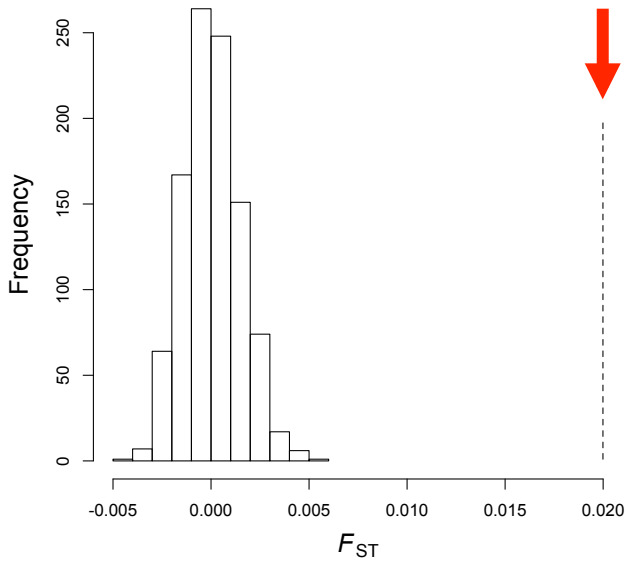
- 10,000 SNPs from the HapMap CEU European and YRI African populations fit pretty well to expectations

Credits : Graham Coop (https://gcbias.org/2011/11/05/population-genetics-course-resources-f-statistics/)

Hypothesis testing

- Testing the null hypothesis that "genes (or genotypes) are drawn from the same distribution in all populations" using exact tests (see, e.g., Genepop)

- Testing the departure from a null distribution of F_{ST} generated by random permutations of multilocus genotypes across populations (see, e.g., Genetix or Fstat)



Assignment methods

- With genotype data from *L* biallelic loci for *K* populations, the likelihood of an individual's genotype g_l in population *k* is (assuming HWE):

$$L(g_l \mid \text{pop } k) = p_{k,l}^2 \qquad \text{if } g_l = \text{AA}$$
$$L(g_l \mid \text{pop } k) = 2p_{k,l}(1 - p_{k,l}) \qquad \text{if } g_l = \text{Aa}$$
$$L(g_l \mid \text{pop } k) = (1 - p_{k,l})^2 \qquad \text{if } g_l = \text{aa}$$

(where $p_{k,l}$ is the frequency of allele A in population *k*)

- Assuming that the *L* loci are independent, the likelihood that the individual belongs to pop *k* reads:

$$L(\text{ind.} \mid \text{pop } k) = \prod_{l=1}^{L} L(g_l \mid \text{pop } k)$$

Assignment methods

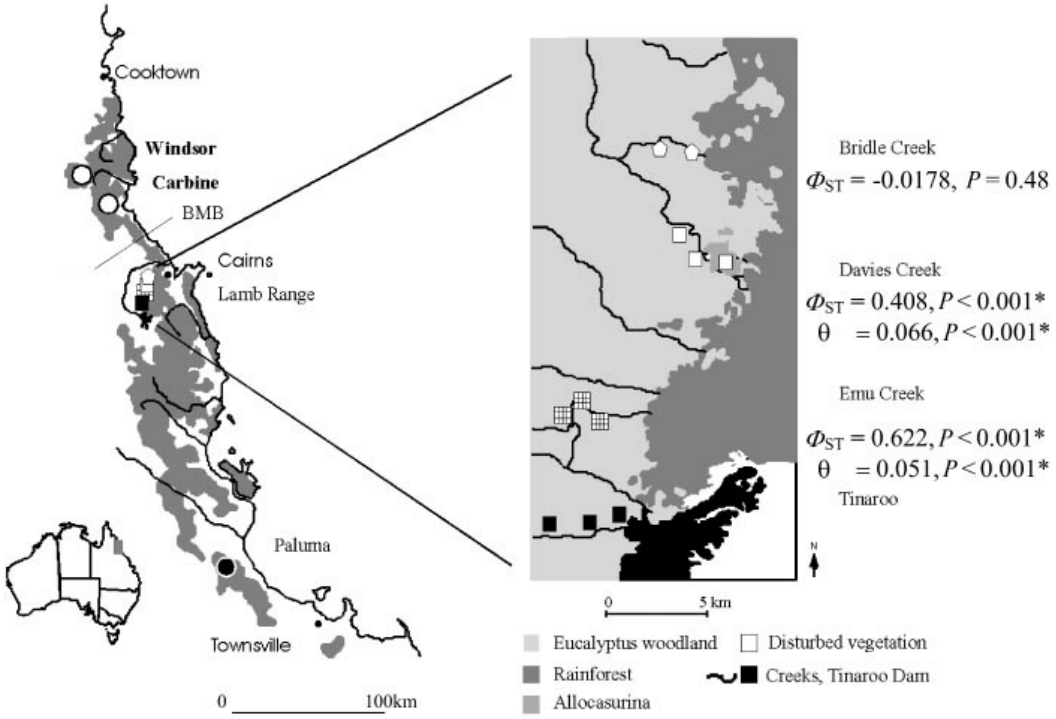
- Using Bayes' rule, one can also compute the posterior probability that the individual comes from population *k*:

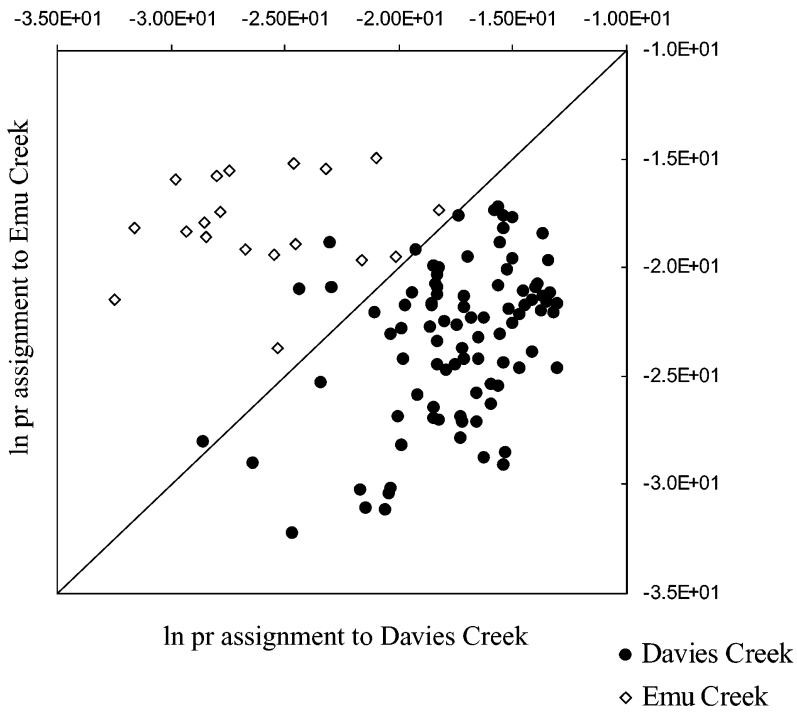
$$\mathrm{P}(\text{pop } k \mid \text{ind.}) = \frac{\mathrm{L}(\text{ind.} \mid \text{pop } k)\mathrm{P}(\text{pop } k)}{\sum_{k=1}^{K} \mathrm{L}(\text{ind.} \mid \text{pop } k)\mathrm{P}(\text{pop } k)}$$

- P(pop *k*) is the prior probability. With no prior knowledge, assume P(pop *k*) = 1 / *K*

Assignment methods

- Phylogeography and population structure of an ecotonal marsupial, *Bettongia tropical*





- Evidence for significant structure: most individuals are assigned to their sampling location

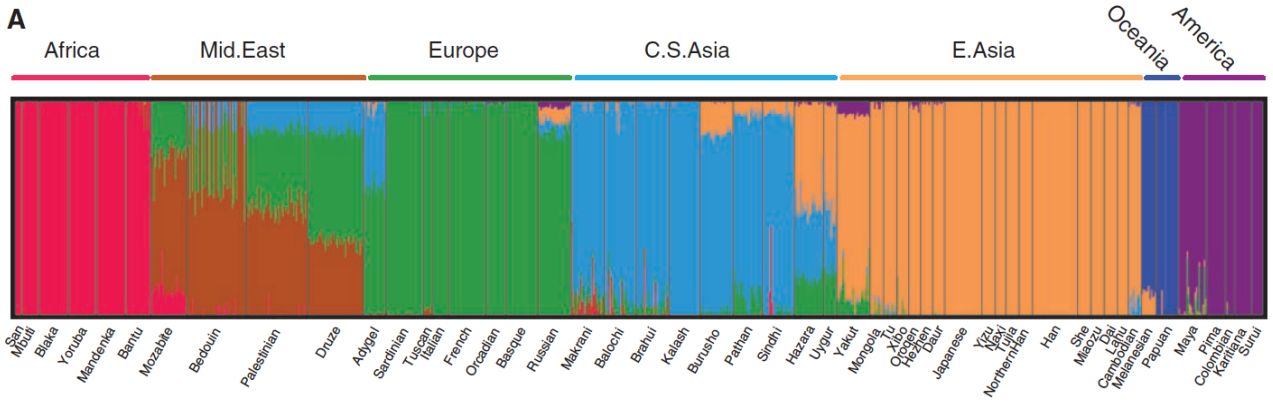
Credits: Pope et al. (2000) *Molecular Ecology* **9**: 2041-2053

Clustering

- Start with a random assignment of individuals to groups (clusters). Given assignments probabilities, the allele frequencies at all loci are computed for each population

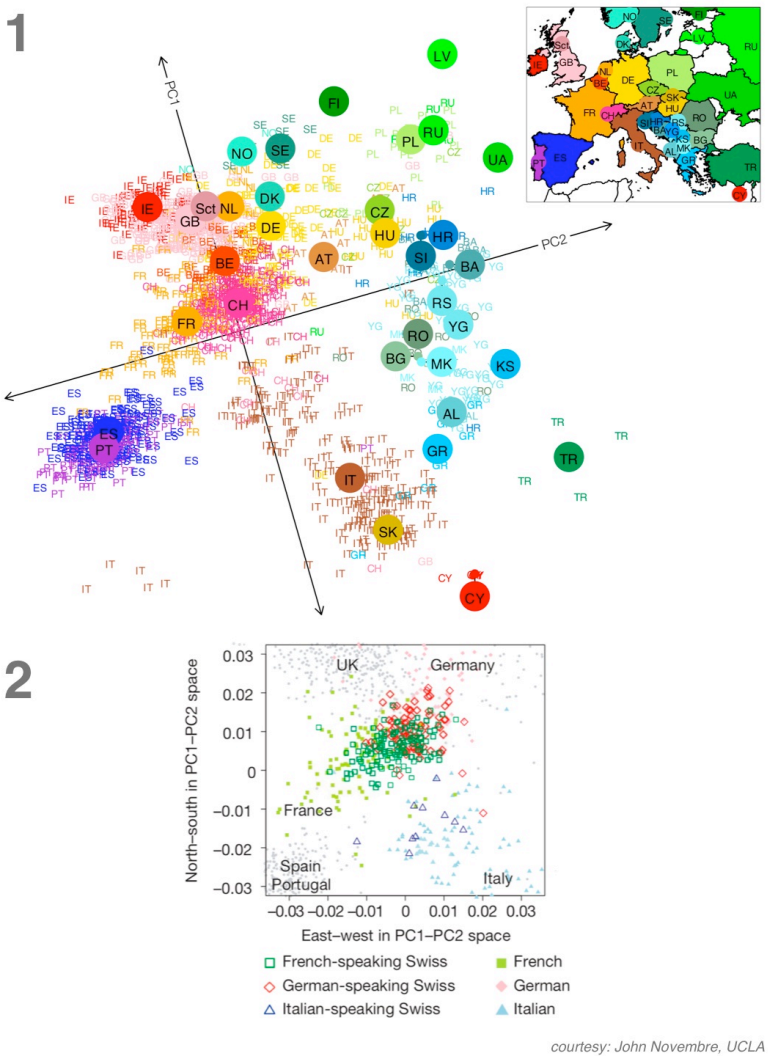
- Given these allele frequencies, each individual is reassigned to population *k* with probability

- These steps are iterated many times. In a Bayesian framework, prior distributions are defined for allele frequencies (e.g., a beta distribution)



Credits : Li et al. (2008) *Science* **319**: 1100-1104

Principal component analysis



courtesy: John Novembre, UCLA

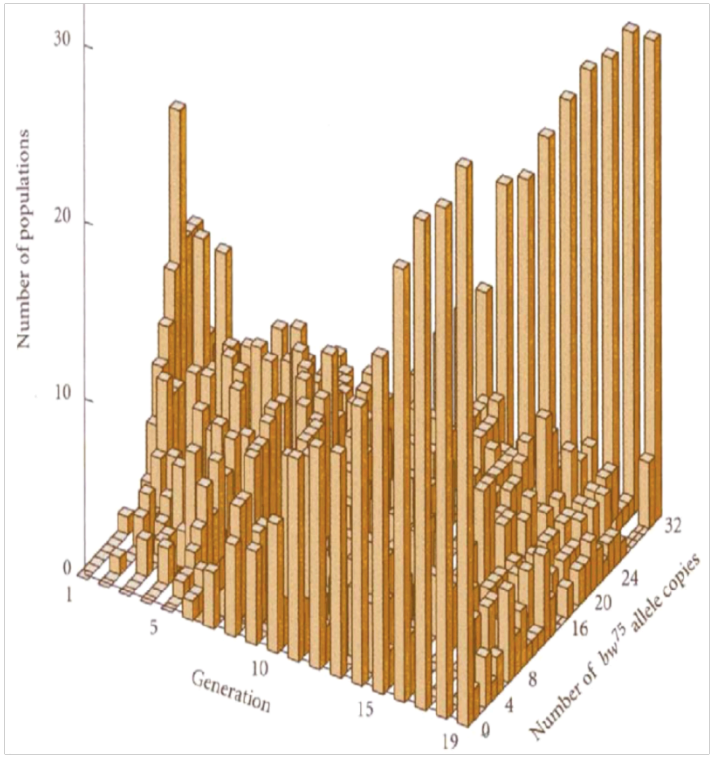
- The data consist of *N* individuals genotyped at *L* bi-allelic SNPs. One individual's genotype data at a locus takes value 0, 1 or 2 (corresponding, e.g., to the number of copies of the reference allele).
- Principal component analysis (PCA) of this data *N* x *L* matrix covers the major axes of genotype variance in the sample
- PCA reduces the dimension of the dataset. Descriptive approach (but see McVean 2009)

Credits : Novembre *et al*. (2008) *Nature* **456**: 98-101

How do evolutionary forces affect the distribution of polymorphisms in populations?

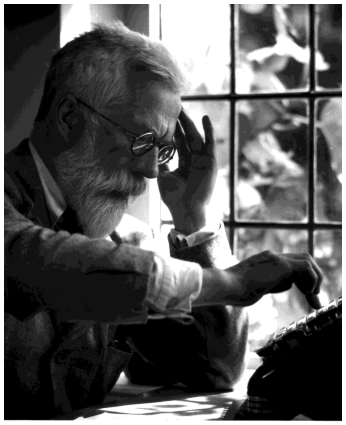
Evolution in finite populations

- Buri's experiment (1956): 107 Drosophila populations, each of which was funded with 16 individuals heterozygote for the 'brown eye' mutation (*bw*^{75})

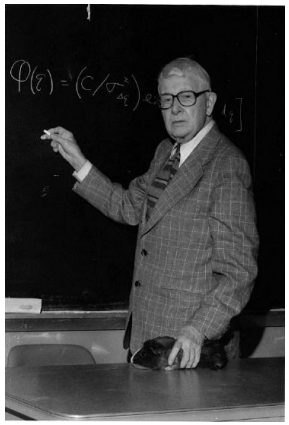


- As time goes on, the **variation** within populations **decreases**

The Wright-Fisher model



Ronald A Fisher

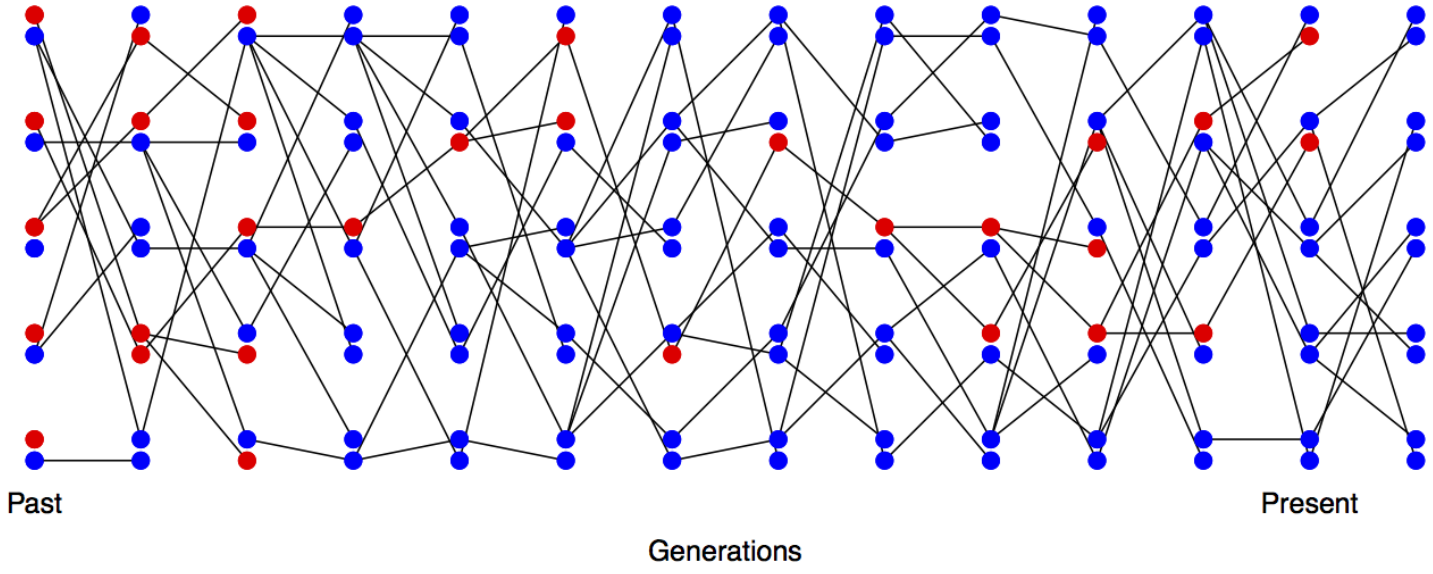


Sewall Wright

- Consider a **haploid**, **isolated** population of size *N*
- Consider a **biallelic locus** (alleles *A* and *a*). Let's note $p[t]$ the frequence of *A* and $q[t] = (1 - p[t])$ the frequency of *a* at time *t*
- No **mutation**
- Each generation, each individual produces a **large number of gametes**, (same expectation = neutrality)
- Draw *N* gametes to make the next generation (random draw in a **gametic urn** of infinite size)

The Wright-Fisher model

- In a finite and constant-size population, each gene does not provide **exactly** one copy of itself in the next generation, but rather a **variable** number of copies



Credits : Graham Coop (https://gcbias.org/2015/02/11/figures-of-genetic-drift/)

The Wright-Fisher model

- At time (t+1), draw N genes in a **infinite gametic urn** made of alleles A at frequency $p[t]$ and alleles a at frequency $q[t]$

- The **random variable** $X[t+1]$ that gives the number of A copies follows a **binomial distribution**, with parameters N **and** $p[t] = X[t] / N$:

$$\Pr\left(X\left[t+1\right]=k\right)=\binom{N}{k}p[t]^{k}\left(1-p[t]\right)^{N-k}$$

The Wright-Fisher model

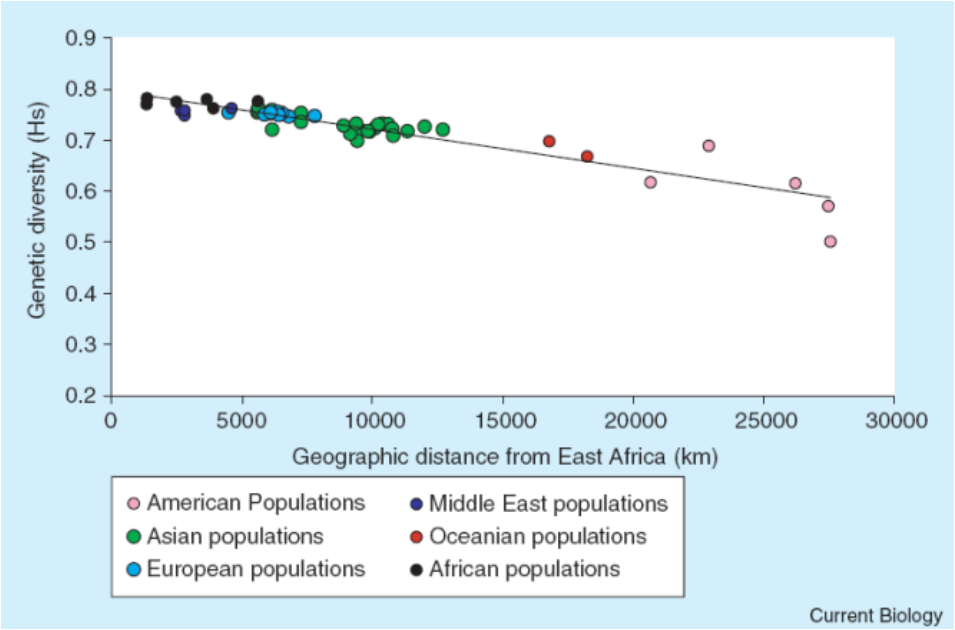
- Let $Y[t+1] = X[t+1] / N = p[t+1]$, be the **frequency of *A*** at generation $(t+1)$:

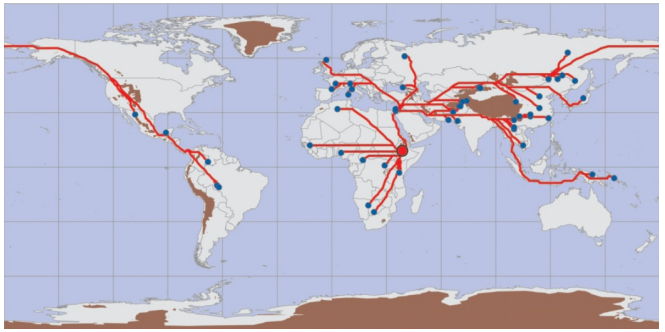
$$E\left(Y[t+1]\right) = E\left(p[t+1]\right) = E\left(\frac{X[t+1]}{N}\right) = \frac{E\left(X[t+1]\right)}{N} = p[t]$$

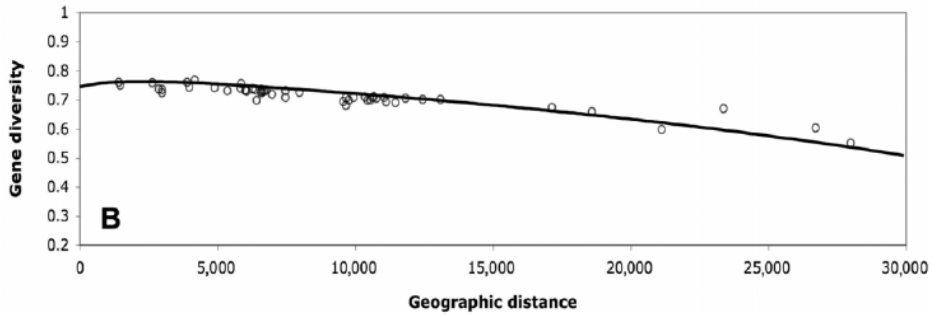
- In **expectation**, the frequency **is constant** from one generation to the next, but the **variance** increases as ***N*** **decreases:**

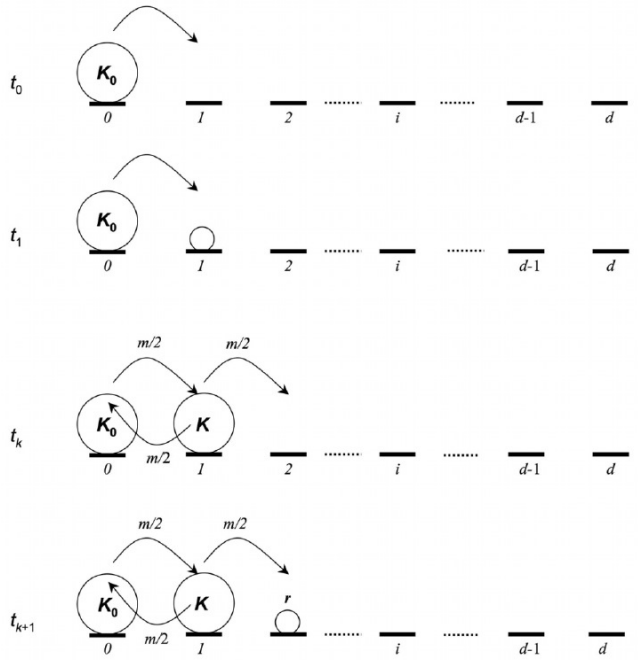
$$V\left(Y[t+1]\right) = V\left(p[t+1]\right) = V\left(\frac{X[t+1]}{N}\right) = \frac{V\left(X[t+1]\right)}{N^2} = \frac{p[t]q[t]}{N}$$

Founder effects in humans



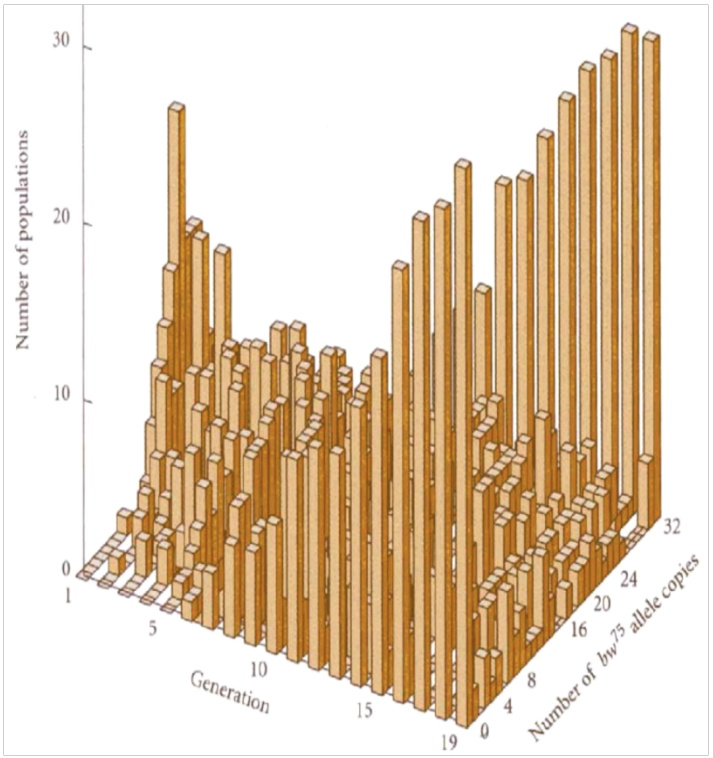


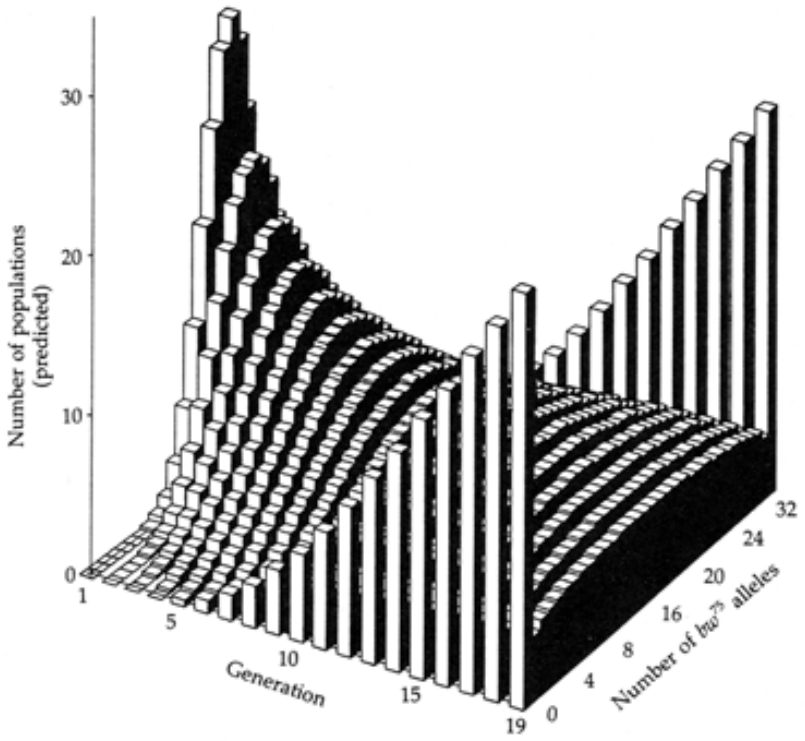




- Heterozygosity decreases as the distance from Africa increases: Prugnolle *et al*. (2005) Curr. Biol. 15: R159-R160; Liu *et al*. 2006 *Am. J. Hum. Genet.* 79: 230-237

The Wright-Fisher model





- The model with $2N = 32$ predicts **less populations that are fixed**, as compared to the observations: the **variance of reproductive success** is about 70% larger than what is supposed in the Wright-Fisher's model

Effective population size

- Effective population size (denoted N_e) is defined as the **size of an ideal Wright-Fisher's population** (*) where genetic drift would have the **same intensity** (**) as compared to the population under scrutiny

- (*) *constant-size, randomly mating population, hermaphrodite individuals, no fitness differences between allalic types, etc.*

- (**) *same rate of drift, same increase in inbreeding, same increase in the variance of allele frequencies, etc.*

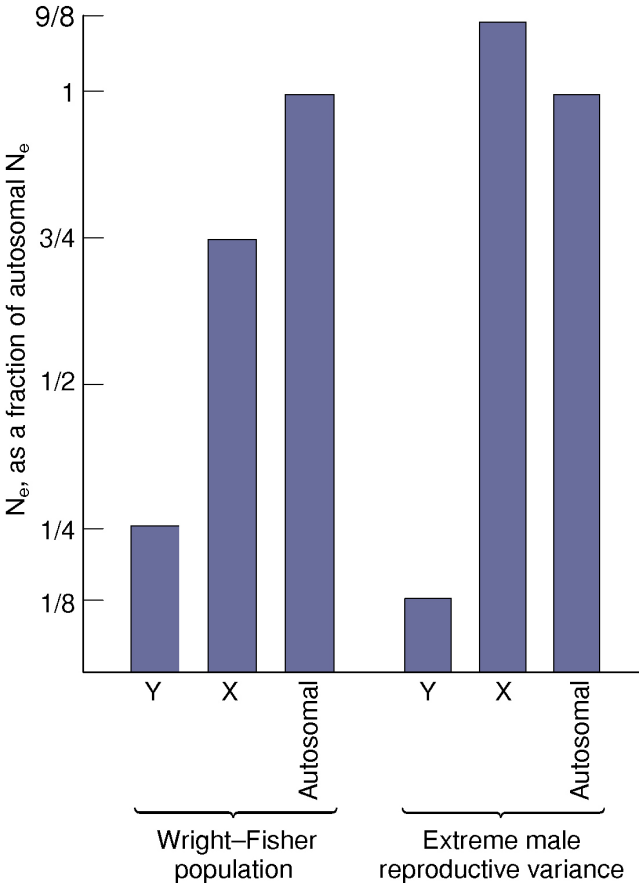
Effective population size

- Many definitions, and therefore many estimators of effective population size:
- Inbreeding effective size (related to the rate of increase in inbreeding)
- Variance effective size (related to the rate of allele frequency change)
- Coalescent effective size (related to the asymptotic rate of coalescence of pairs of genes)

Effective population size

- **Many factors influence effective population size**:

- reproductive system: **selfing** reduces effective size
- class-structure: e.g., **biased sex-ratio** reduces effective size
- **age-structure**: e.g., diapause or dormancy tend to increase effective size
- **variance of reproductive success** reduces effective size



Interaction of evolutionary forces: drift and mutation

- The **loss** of diversity due to drift might **compensated** by new **mutations**

- A useful way to characterize the amount of polymorphism in populations is to use probabilities of genetic identity

- 2 genes drawn at random are **identical if** one of them (or both) have **mutated**:

$$Q\left[t+1\right]=\left[\frac{1}{2N}+\left(1-\frac{1}{2N}\right)Q[t]\right]\left(1-\mu\right)^2$$

- A equilibrium:

$$Q\approx\frac{1}{1+4N\mu}$$

$$H=1-Q\approx\frac{4N\mu}{1+4N\mu}=\frac{\theta}{1+\theta}$$

Interaction of evolutionary forces: drift and migration

- If we consider a **structured** population (geography, age-classes, sex, etc.), we can always define probabilities of gene identity within a class (Q_w) and between classes (Q_b)

- We can then use a generic definition of *F*-statistics, which depends on both identities, to measure the differentiation between classes:

$$F \equiv \frac{Q_w - Q_b}{1 - Q_b}$$

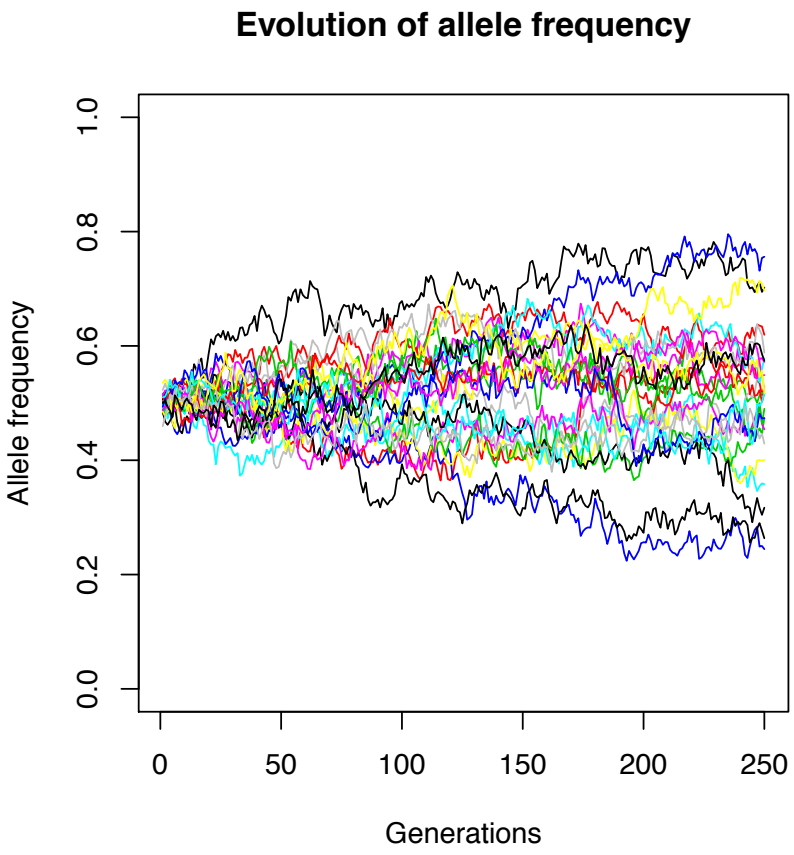
- For a spatially structured population (infinite island model), we get at equilibrium:

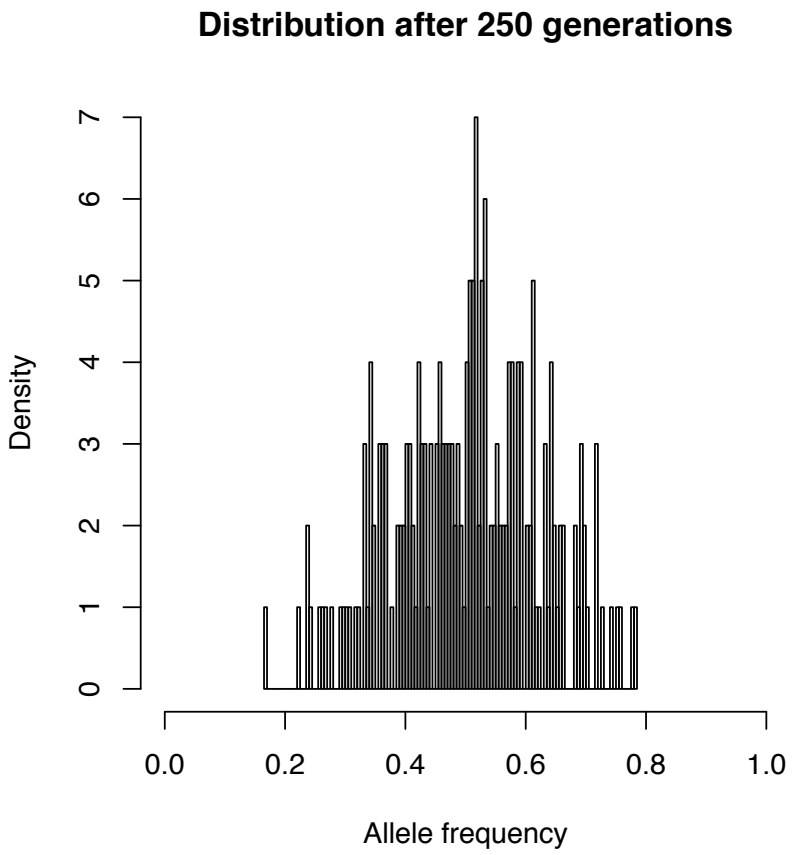
$$F_{ST} = \frac{\gamma(1-m)^2}{\gamma(1-m)^2 + 2N\left[1-\gamma(1-m)^2\right]}$$

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

How to characterize the distribution of polymorphisms?

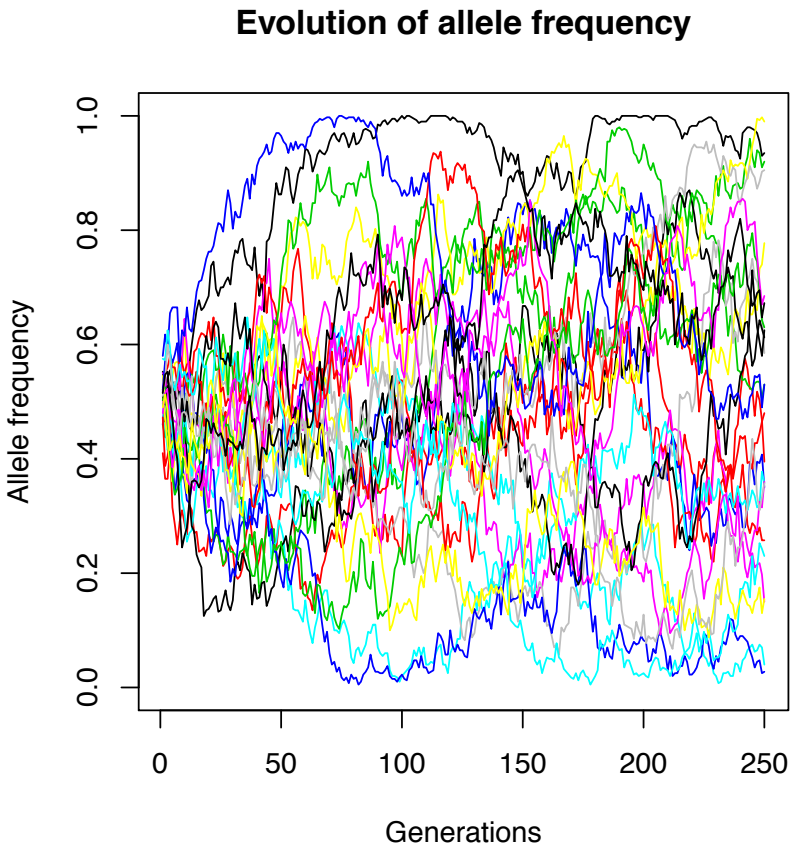
Diffusion approximation

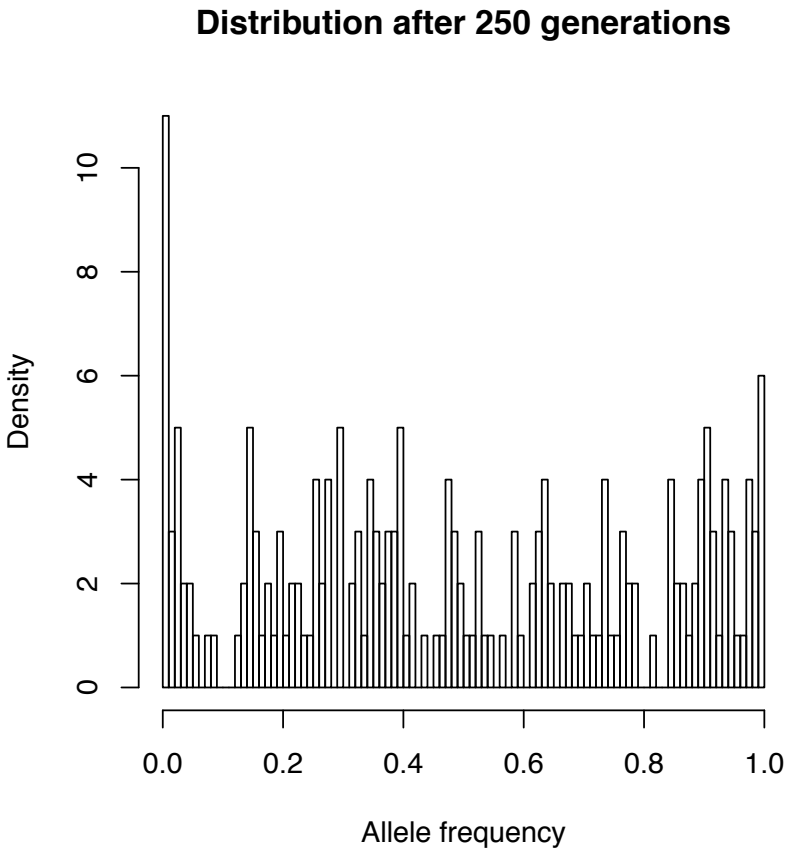




- Evolution of allele frequency in an island model (6 demes, N = 1000, F_{ST} = 0.001): 50 genes sampled in one deme

Diffusion approximation





- Evolution of allele frequency in an island model (6 demes, N = 200, F_{ST} = 0.25): 50 genes sampled in one deme

Diffusion approximation

- Deriving the allele frequency distribution $f(p,t)$ from the Wright-Fisher model is a complex problem…

- Solution: approximating the Wright-Fisher (discrete) process by a continuous (diffusion) approximation (assuming N tends to infinity), which satisfies the forward Kolmogorov equation:

$$\frac{\partial f(p,t)}{\partial t} = -\frac{\partial M(p) f(p,t)}{\partial p} + \frac{1}{2}\frac{\partial^2 V(p) f(p,t)}{\partial p^2}$$

Where $M(p)$ and $V(p)$ are the 1st and the 2nd moments of change in p per unit of time (i.e., the *drift* and the *diffusion* coefficients)

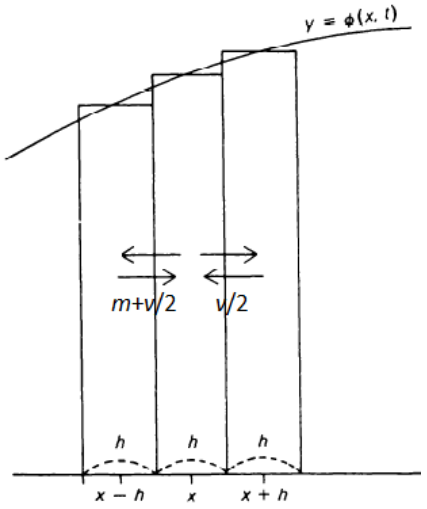


Figure 8.3.1. Diagram to show the meaning of terms in the Kolmogorov forward (Fokker–Planck) equation as applied to population genetics. (From Kimura, 1955).

Diffusion approximation

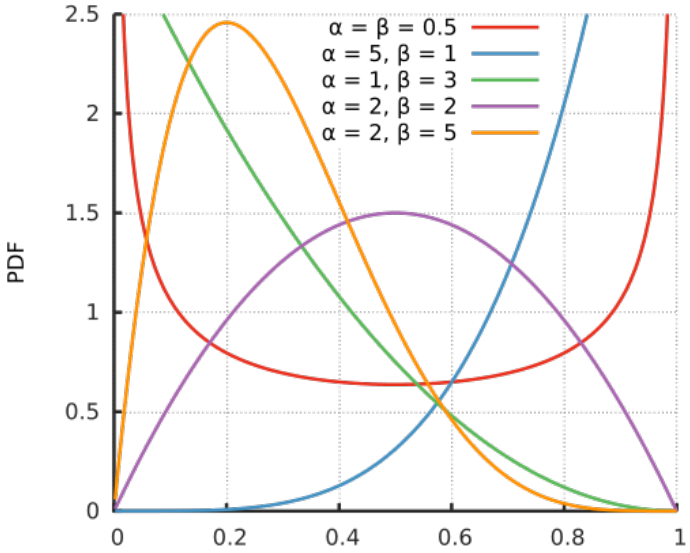
- In the Wright-Fisher model with mutation: $M(p) = -\nu p + (1 - p)\mu$ and $V(p) = p(1 - p) / N$ [μ is the mutation rate from a to A; ν is the mutation rate from A to a]

- Stationary distribution:

$$\frac{\partial f(p,t)}{\partial t} = -\frac{\partial M(p) f(p,t)}{\partial p} + \frac{1}{2}\frac{\partial^2 V(p) f(p,t)}{\partial p^2} = 0$$

$$f(p,t) \sim \mathrm{C}\, p^{2N\nu-1}(1-p)^{2N\mu-1}$$

- i.e., a beta distributions with parameters $2N\nu$ and $2N\mu$

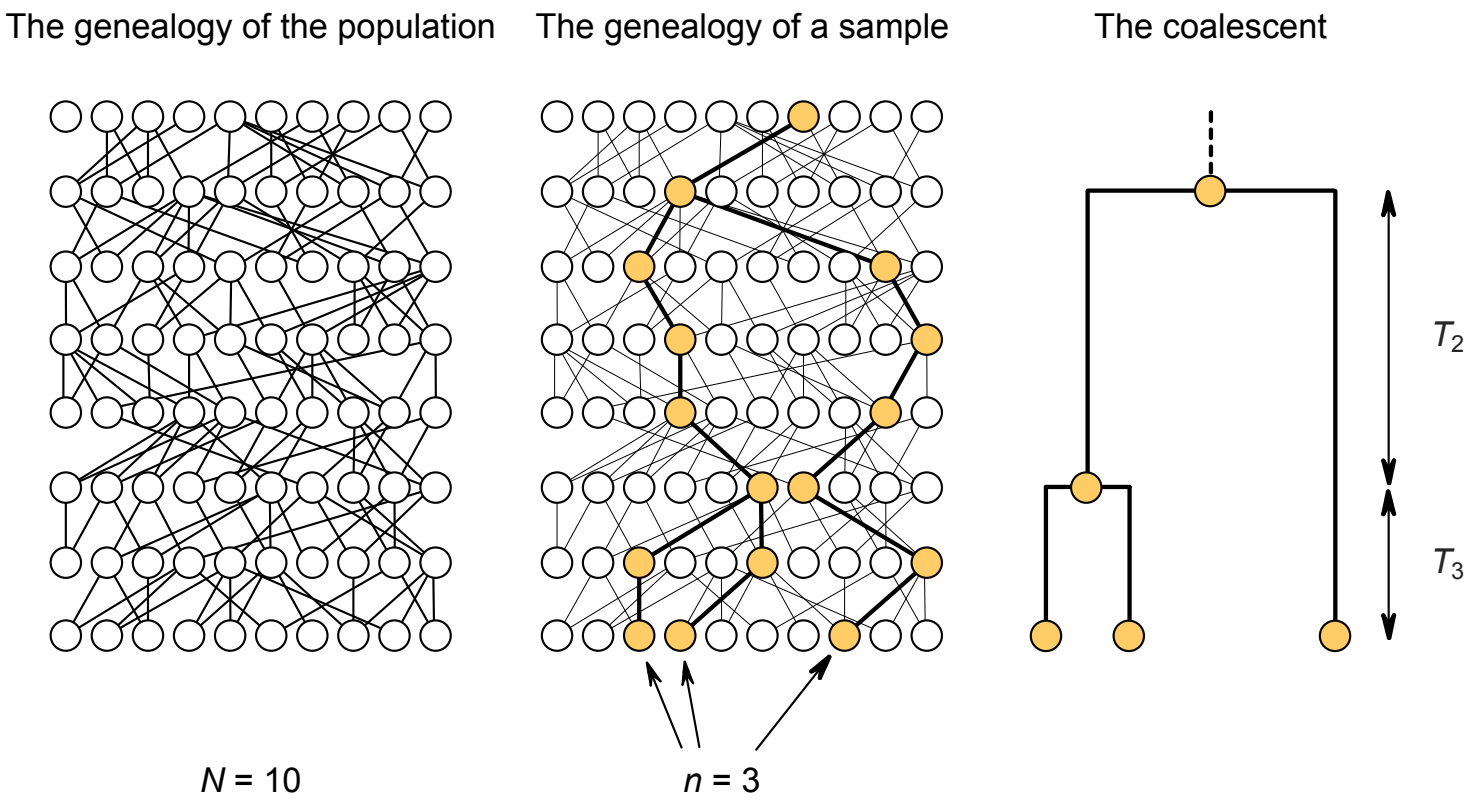


Diffusion approximation

- Diffusion theory provides distributions of allele frequencies in simple models. It is usually restricted to stationary solutions

- An alternative way to characterize the distribution of variation in populations is given by coalescent theory

The coalescent

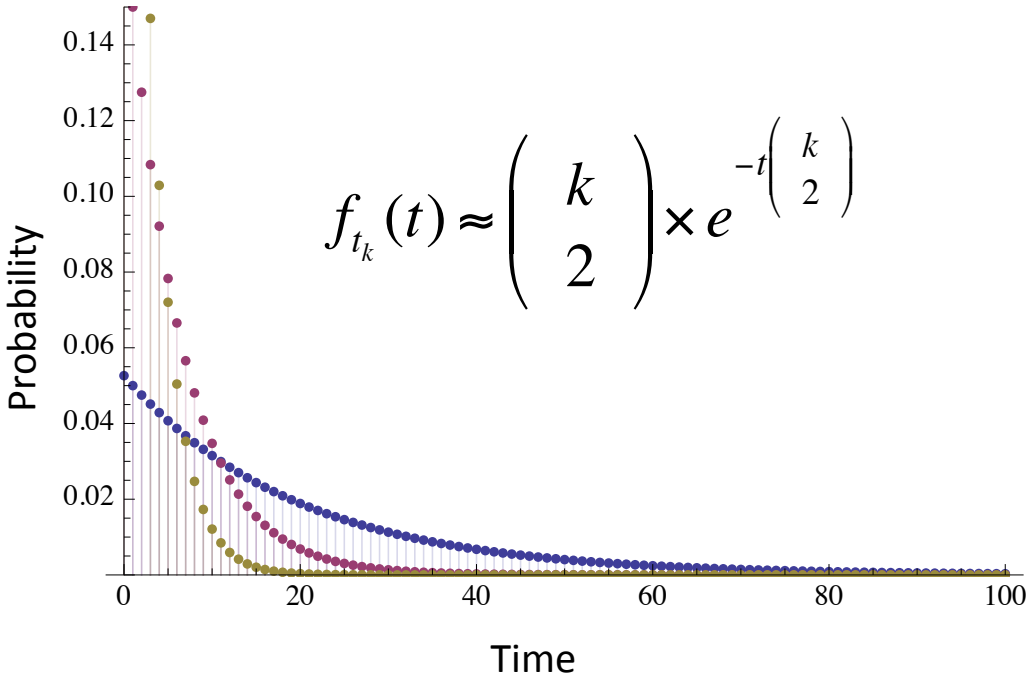


- In neutral models, mutations have no impact on genealogies of genes; therefore the *mutation* process can be *decoupled* from the *genealogical* process

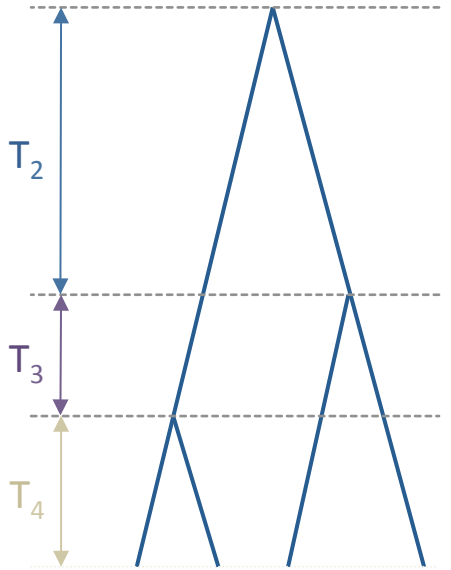
The coalescent

- The branch lengths (coalescence times) are exponentially distributed (*k* is the number of lineages, time is scaled with the population size *N*)

$$E(t_{MRCA}) = 2 \times \left(1 - \frac{1}{k}\right)$$

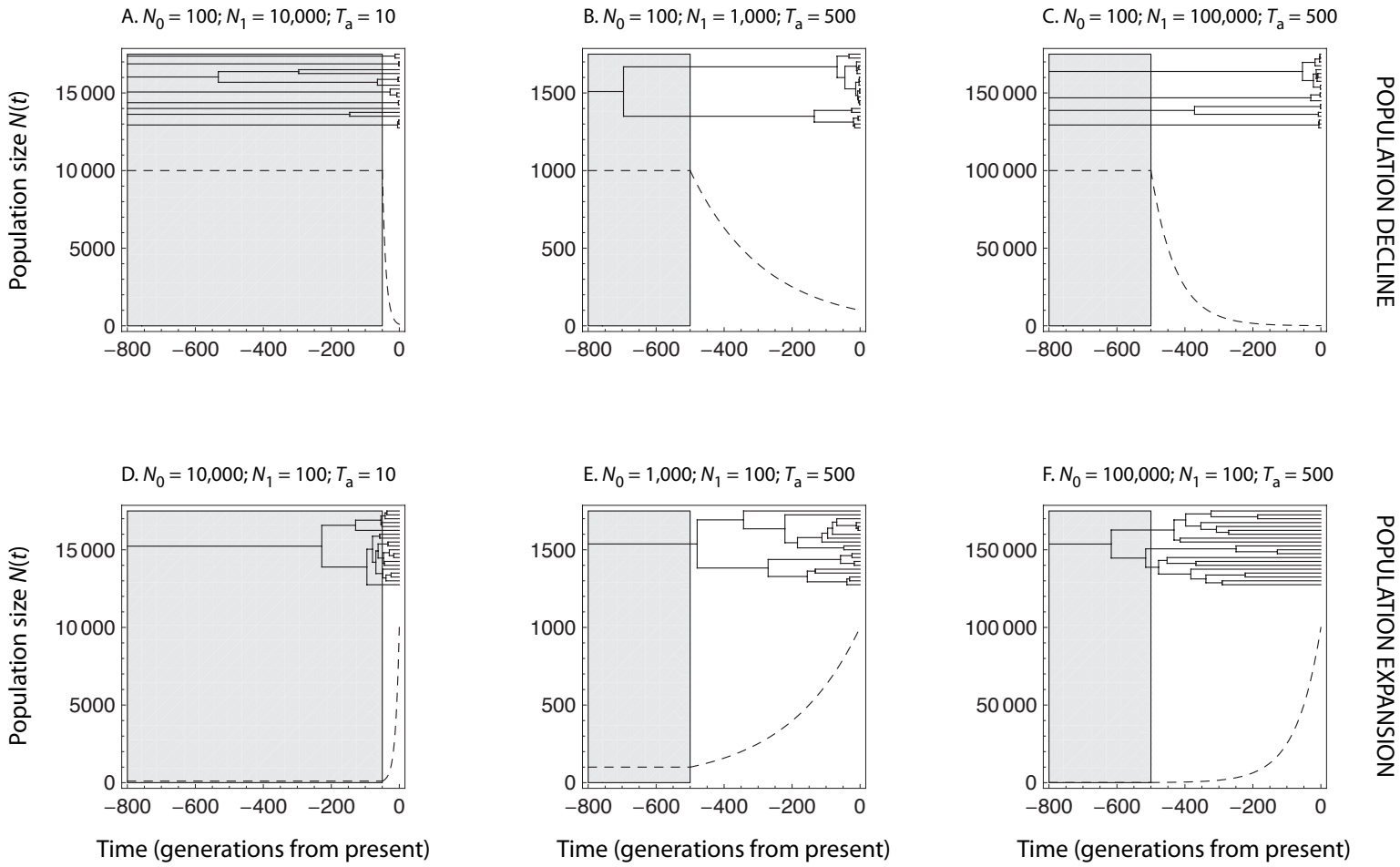


$$f_{t_k}(t) \approx \binom{k}{2} \times e^{-t\binom{k}{2}}$$



$$E(t_k) = \frac{2}{k(k-1)}$$

The coalescent



- Population size changes affect the shape of coalescent trees: “star-shaped” genealogies for expanding populations, and “shallow” genealogies for declining ones

The coalescent

- Coalescent theory provides a **probabilistic model** for gene genealogies
- It may simplify the analysis of population genetics models and/or their interpretation
- It is largely used to **simulate efficiently** the genetic variation (simulations of gene samples rather than full populations)
- It paves the way for new techniques to infer population parameters

How to infer parameters of interest from polymorphism data?

Maximum likelihood

- Maximum likelihood approaches are based on a **stochastic model** for the evolution of gene frequencies in populations, specified by some **parameters**

- The aim is to estimate these parameters from the data *D* (the allele counts at different molecular markers)

- To that end, one computes the likelihood of the parameters, given the observed data *D* (i.e, the probability of the data given those parameter values)

Likelihood in the island model: Wright's formula (1940)

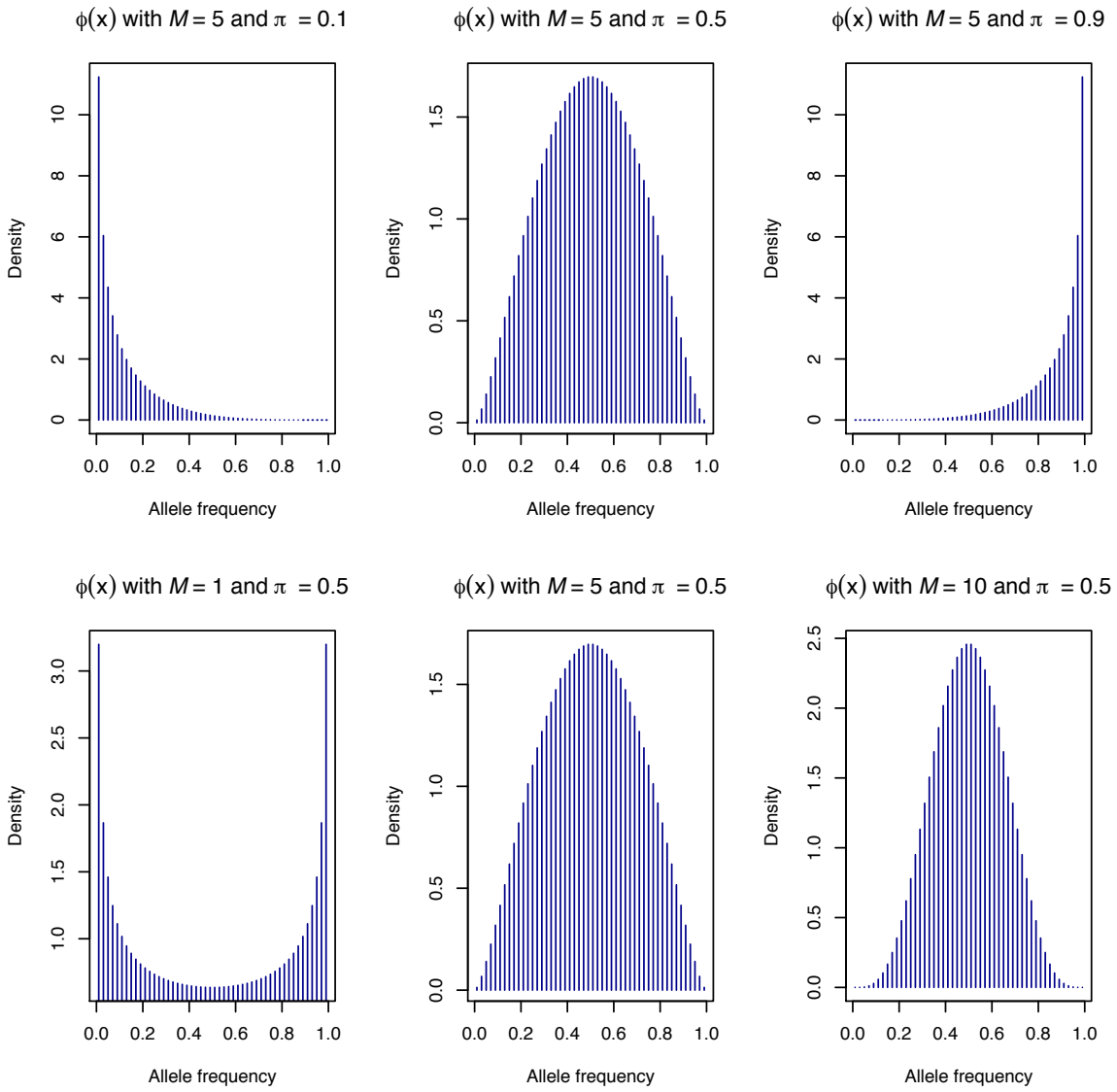
- In an island model with 2 alleles (*A* and *a*), the distribution of the frequency of allele *A* in a deme is given by (Wright, 1940) :

$$\phi(x) = \frac{\Gamma(M)}{\Gamma(M\pi)\Gamma(M(1-\pi))} x^{M\pi-1}(1-x)^{M(1-\pi)-1}$$

- This is the probability that the frequency *x* of allele *A* in a deme that receives $M = 4Nm$ migrants per generation

- The above formula assumes large *N*, and small *m* (diffusion approximation)

Likelihood in the island model: Wright's formula (1940)



Maximum likelihood

- The probability to observe *k* alleles *A* in a *n*-sized sample of a population where the frequency of *A* is *x*, is given by (binomial distribution):

$$\Pr(k \text{ alleles} \mid x) = \binom{n}{k} x^k (1-x)^{n-k}$$

- Integrating over the distribution of allele frequencies:

$$\Pr(k \text{ alleles}) = \int \Pr(k \text{ alleles} \mid x)\phi(x)\mathrm{d}x$$

Maximum likelihood

- The likelihood of a *n*-sized sample with *k* alleles A is therefore given by the following distribution (beta-binomial):

$$\Pr(k \text{ alleles}) = \frac{\Gamma(M)}{\Gamma(M+n)}\binom{n}{k}\frac{\Gamma(M\pi+k)}{\Gamma(M\pi)}\frac{\Gamma(M(1-\pi)+(n-k))}{\Gamma(M(1-\pi))}$$

- This is for one deme and one locus: with multiple demes and loci, multiply the likelihoods (conditional independence of demes and loci)

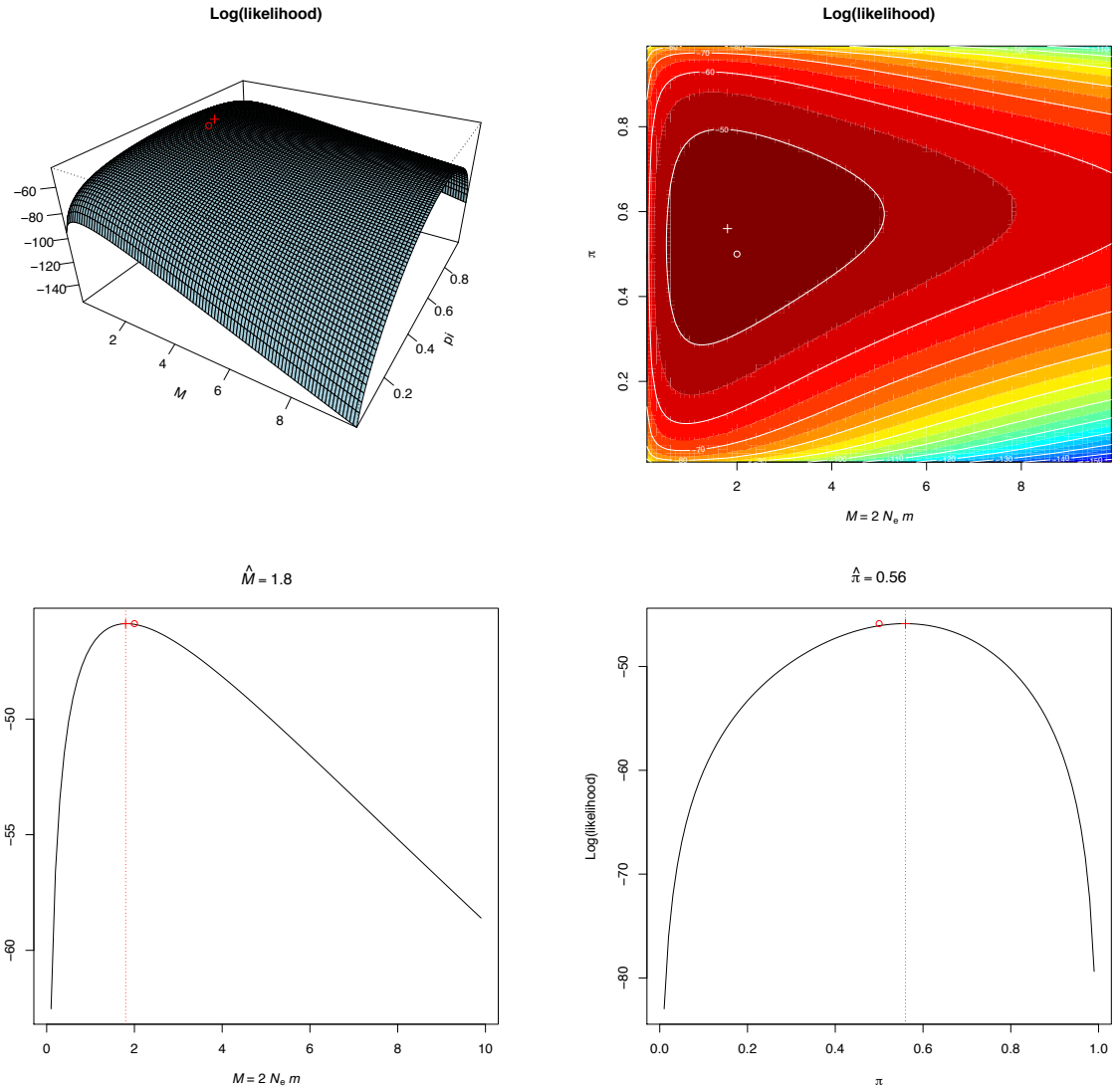
- Characterize the values of M and π that **maximize this likelihood function** (maximum likelihood estimates)

Maximum likelihood

- Let's simulate some data (at a single locus), from 10 sampled demes with:
- $M = 4Nm = 2$ and $\pi = 0.5$
- The data (counts of alleles *A* and *a* among 100 sampled genes) are:

A **:**	**92**	**88**	**71**	**76**	**60**	**12**	**21**	**94**	**70**	**74**
a :	8	12	29	24	40	88	79	6	30	26

Maximum likelihood



- Likelihood profile for one parameter, considering the maximum likelihood for the other parameter…

Likelihood ratios

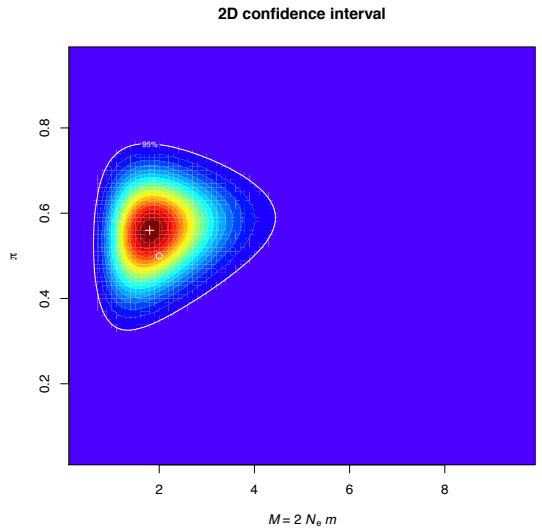
- We may not only calculate point estimates (maximum likelihood) but also compute confidence intervals (based on likelihood ratios):

$$LR = -2\log\left(\frac{L}{L_{\max}}\right)$$

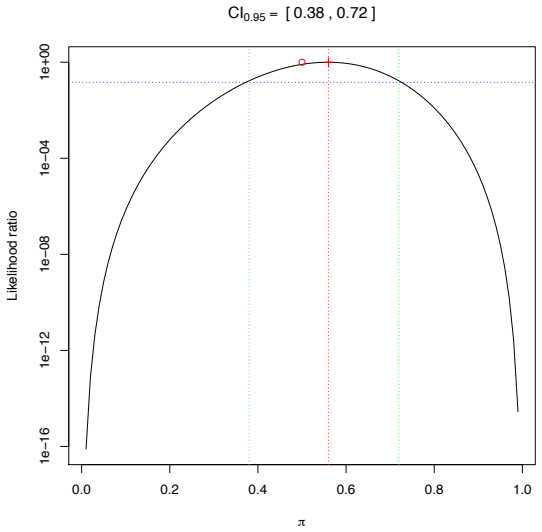
- The *likelihood ratio (LR)* is chi-squared distributed with *k* degrees of freedom (*k* being the number of parameters in the model)

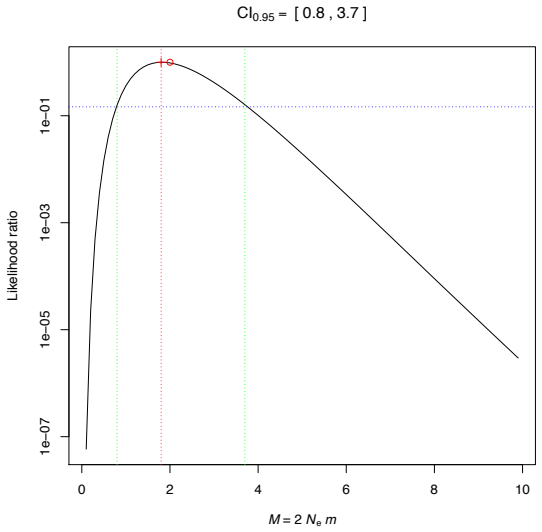
- So a parameter value is included in the confidence interval if *LR* is above a given bound, which is given by the chi-square distribution with *k* degrees of freedom.

Confidence intervals



- Were this procedure to be repeated on multiple samples, the 95%CI would contain the **true value** of the parameter 95% of the time (frequentist interpretation)

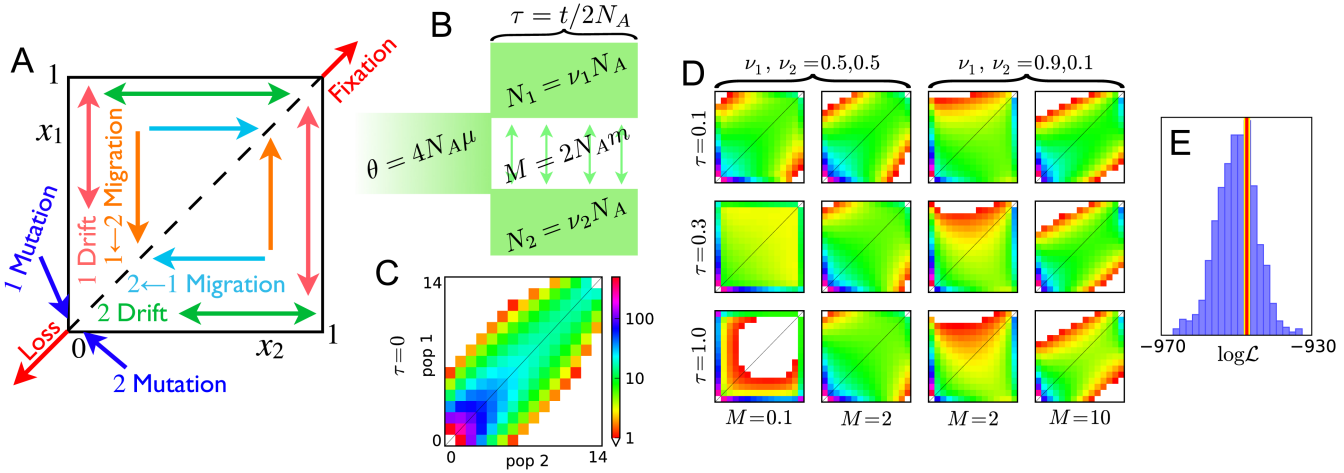




Maximum likelihood

- Limits: it is sometime very difficult (not to say impossible) to maximize the likelihood that way (too many parameters, complicated mathematical expression, etc.)

- Yet, some recent attempts to achieve maximum-likelihood inference in relatively complex models (the likelihood of the allele frequency spectrum is computed numerically using diffusion approximation)



Credits: Gutenkunst et al. (2009) PLoS Genet 5(10): e1000695

Bayesian methods

- In Bayesian statistics, it is assumed that **the parameters have a probability distribution**

- From Bayes' inversion formula:

$$\mathrm{P}(\Theta \mid D) = \frac{\mathrm{P}(D \mid \Theta)\mathrm{P}(\Theta)}{\int \mathrm{P}(D \mid \Theta)\mathrm{P}(\Theta)\, d\Theta} = \frac{\mathrm{P}(D \mid \Theta)\mathrm{P}(\Theta)}{\mathrm{P}(D)}$$

Constant, which only depends upon the data

$$\mathrm{P}(\Theta \mid D) \propto \mathrm{L}(\Theta; D)\mathrm{P}(\Theta)$$

Likelihood

Prior distribution

- To sample from the *posterior* distribution of the parameters, a Markov chain is constructed, with stationary distribution $\mathrm{P}(\Theta \mid D)$

Metropolis-Hasting's algorithm

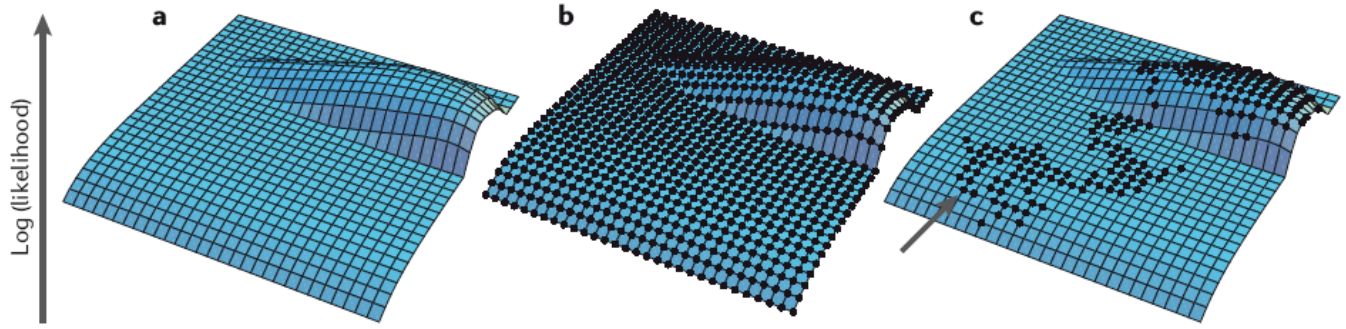


Image courtesy of Peter Beerli, Florida State University, USA.

① In the parameter space, start at Θ

② Propose a new value Θ' following $q\left(\Theta \rightarrow \Theta'\right)$

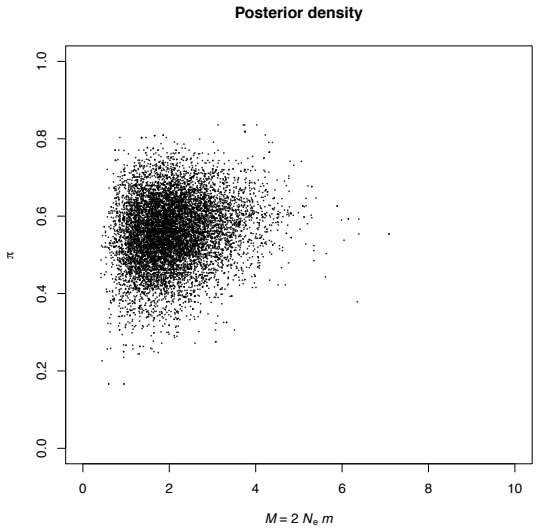
③ Accept this new value with probability:

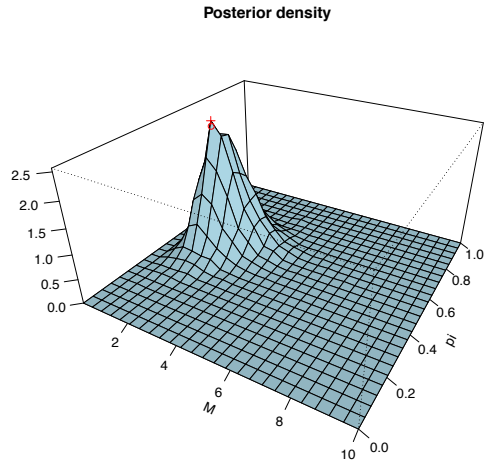
$$h = \min\left(1, \frac{L\left(\Theta';D\right)}{L\left(\Theta;D\right)} \frac{\mathrm{P}\left(\Theta'\right)}{\mathrm{P}\left(\Theta\right)} \frac{q\left(\Theta' \rightarrow \Theta\right)}{q\left(\Theta \rightarrow \Theta'\right)}\right)$$

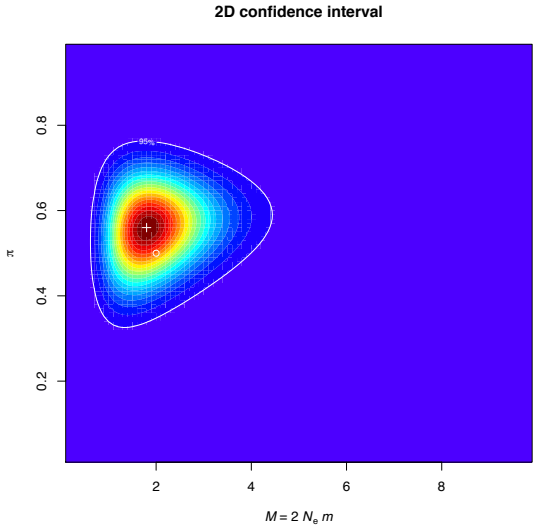
④ Go to (1)

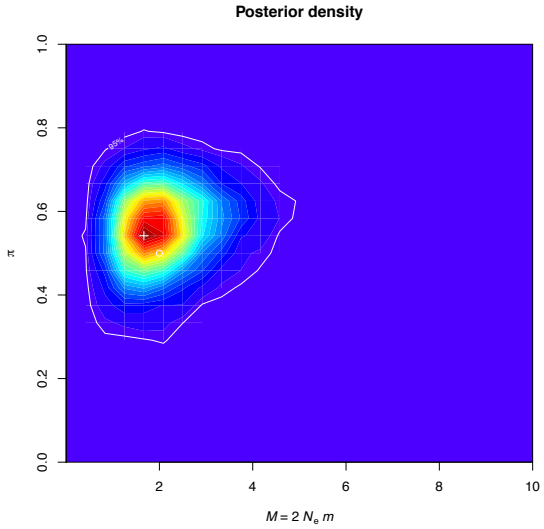
Credits : Excoffier et Heckel (2006) *Nature Reviews Genetics* **7** : 745-758

Markov chain Monte Carlo





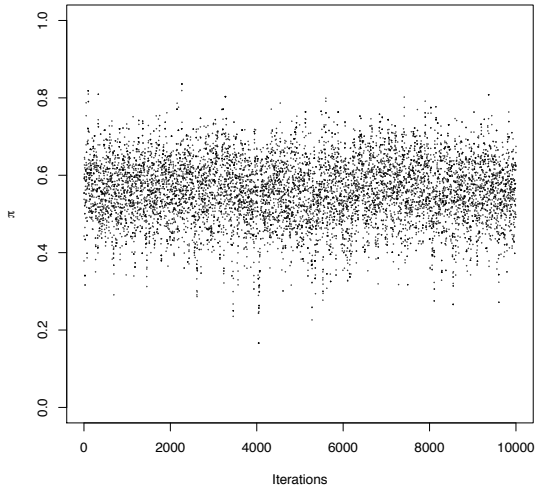


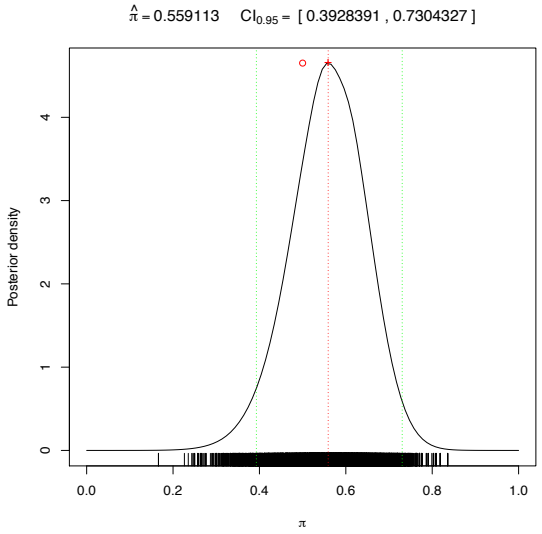


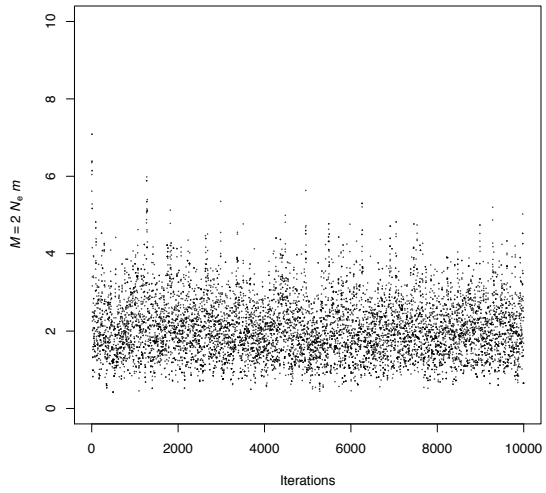
- Joint posterior distribution of the parameters

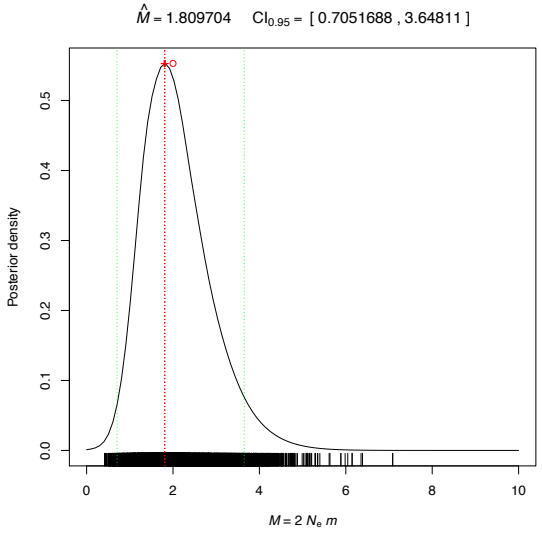
- Comparison with the likelihood: stochastic process, influenced by convergence and mixing properties of the Markov chains…

Markov chain Monte Carlo









- Marginal posterior distributions of the parameter
- Point estimates from the mean, or the mode, or the median
- A 95% credible interval defines an interval where the probability that the parameters lies equals 95%

Felsenstein's equation

- For most models, there is **no mathematical expression** for the likelihood of the parameters:

$$\mathrm{L}(\Theta;D) = \mathrm{P}(D \mid \Theta)$$

- Yet, it is still possible to compute the probability of observing the data, conditionally on the parameters **and the genealogy** (using coalescent theory):

$$\mathrm{P}(D \mid \Theta, G)$$

- Therefore, the likelihood can expressed as a sum over all possible genealogies (**Felsenstein's equation**):

$$\mathrm{L}(\Theta;D) = \mathrm{P}(D \mid \Theta) = \int_G \mathrm{P}(D \mid \Theta, G)\,\mathrm{P}(G \mid \Theta)\,dG$$

Mutation

Coalescent theory

Coalescent-based MCMC

- An important point to consider: genealogies are considered as **nuisance parameters**: these are important quantities in the computation, that we do not try/need to estimate

- Although we are dealing with trees, this approach is **very different from phylogenetic approaches** (where trees are the objects we want to estimate)

Markov chain Monte Carlo

- MSVAR: a demographic model with population size change: Beaumont (1999) *Genetics* 153: 2013-2029

- An application example with orang-utans: Goossens *et al*. (2006) *PLoS Biology* 4(2): e25

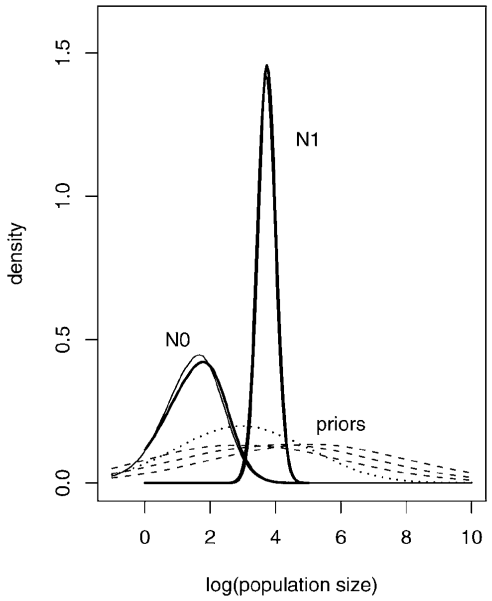


Figure 2. Ancestral and Present Population Sizes

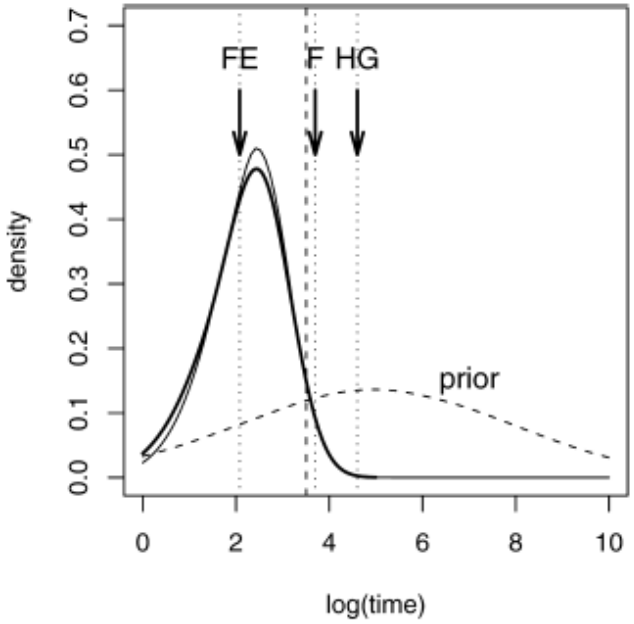


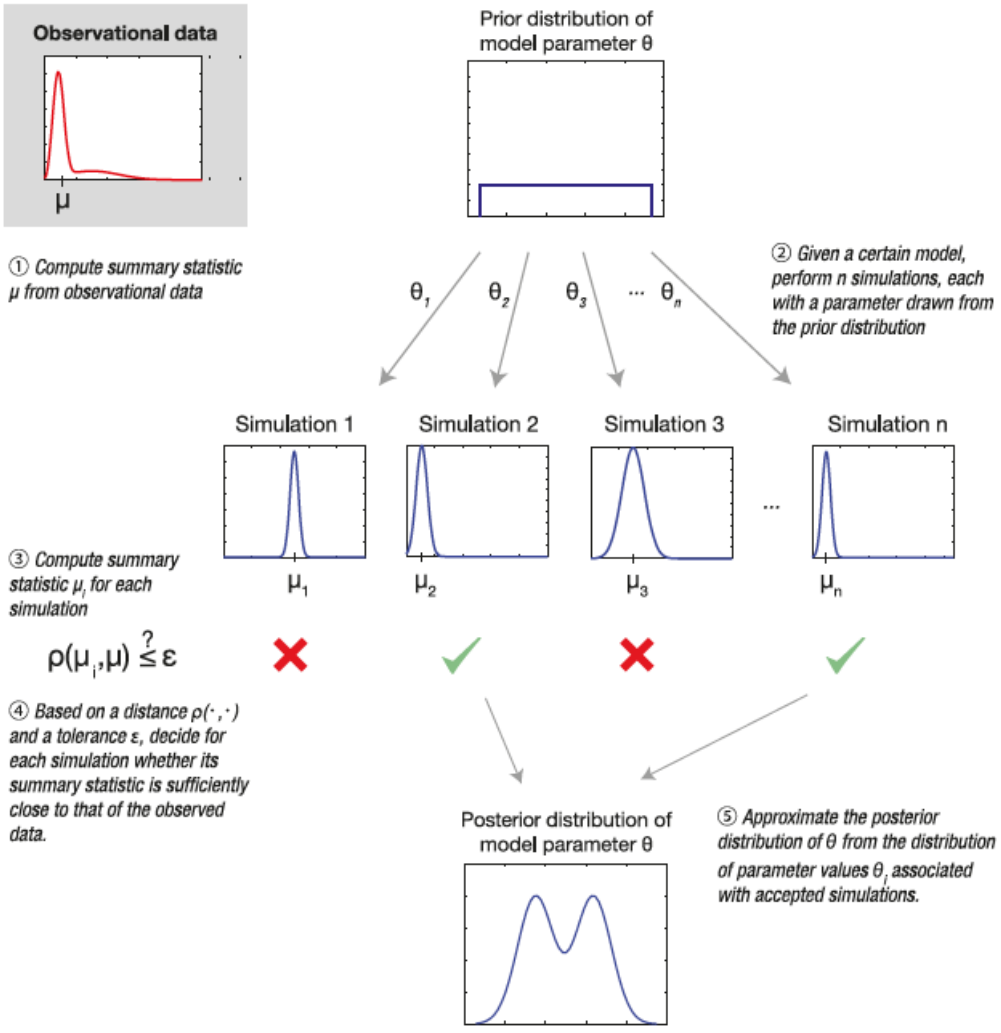
Figure 3. Time since the Population Collapse

FE : *Forest exploitation*
F : *Farmers*
HG : *Hunter-gatherers*

Approximate Bayesian Computation (ABC)

- An alternative with complex models (when the likelihood is impossible to compute):

- Approximate Bayesian Computation: Beaumont *et al*. (2002) *Genetics* 162: 2025-2035

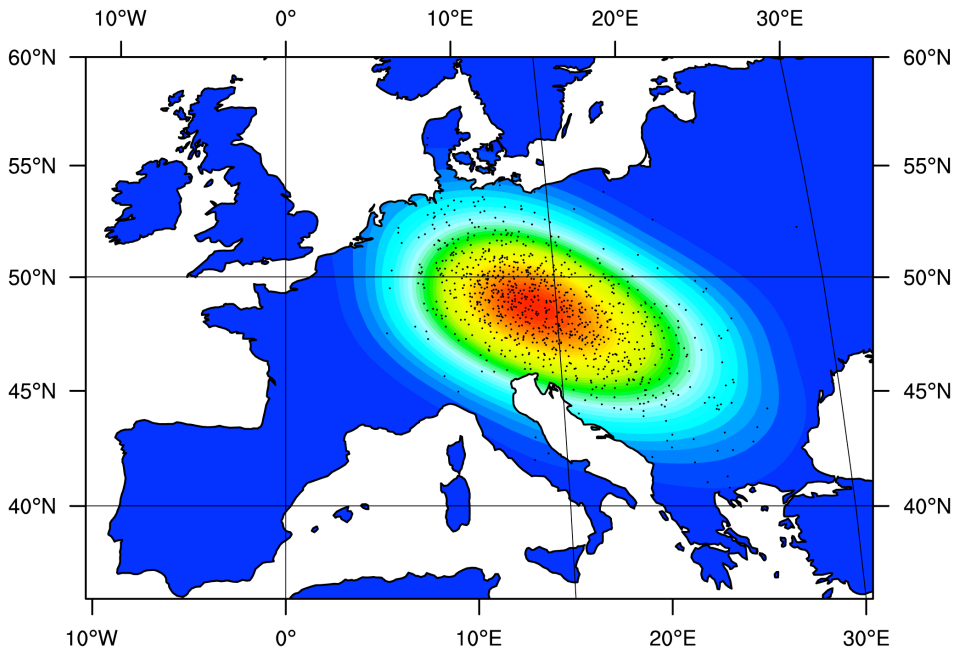


Sunnaker *et al*. (2013) *PLoS Computational Biology* **e1002803**

Approximate Bayesian Computation (ABC)

- A spatially explicit model to characterize the origins of the “lactase persistent” phenotype in Europe, using both genetic and archeological data:

- **The origin of the co-evolution between lactase persistence and dairy culture traces back to 7,500 yrs ago somewhere between Central Europe and the Balkans**



Itan *et al*. (2009) *PLoS Computational Biology* **e1000491**

Conclusions

- Likelihood-based approaches make full use of the data (not limited to some summary statistics)
- They provide point estimates and confidence intervals, but also the likelihood (frequentist approaches) or the full posterior distribution (Bayesian approaches)
- These approaches may be much more difficult to implement (depending on whether the likelihood can or cannot be derived)
- *Approximate Bayesian Computation* (ABC) may be an alternative