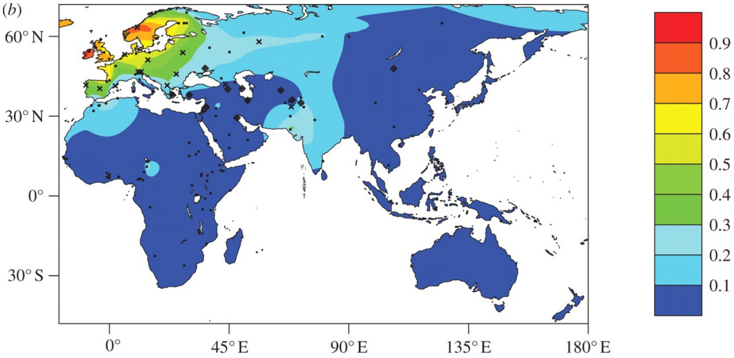
Inférence en génétique des populations IV.

François Rousset & Raphaël Leblois

M2 Biostatistiques 2015–2016

Modeling and analysis of covariances in allele frequencies

Spatial



Intra-family

Modeling and analysis of covariances in allele frequencies

Modeling: important parameters in selection processes

Modeling and analysis of covariances in allele frequencies

Modeling: important parameters in selection processes

- Artificial selection, or how to increase milk production by selecting bulls
- Wright's "shifting balance" theory
- General formal theory of selection with interactions between relatives

Modeling and analysis of covariances in allele frequencies

Modeling: important parameters in selection processes

- Artificial selection, or how to increase milk production by selecting bulls
- Wright's "shifting balance" theory
- General formal theory of selection with interactions between relatives

Statistical analysis: relatively simple (and routine) methods of inference

Modeling and analysis of covariances in allele frequencies

Modeling: important parameters in selection processes

- Artificial selection, or how to increase milk production by selecting bulls
- Wright's "shifting balance" theory
- General formal theory of selection with interactions between relatives

Statistical analysis: relatively simple (and routine) methods of inference

- Two main spatial models (Wright): island model, and isolation-by-distance
- Estimate Nm or the *neighborhood size* $4D\pi\sigma^2$
- A few statistically desirable properties
- Analytical understanding

Modeling and analysis of covariances in allele frequencies

Modeling: important parameters in selection processes

- Artificial selection, or how to increase milk production by selecting bulls
- Wright's "shifting balance" theory
- General formal theory of selection with interactions between relatives

Statistical analysis: relatively simple (and routine) methods of inference

- Two main spatial models (Wright): island model, and isolation-by-distance
- Estimate *Nm* or the *neighborhood size* $4D\pi\sigma^2$
- A few statistically desirable properties
- Analytical understanding
- Huge literature, often based on very naive ideas

Outline

- An example of the role of spatial structure in evolution
- Modeling of the role of spatial structure
- Formal modelling of covariances: diffusion vs. coalescence
- Genealogical interpretations of covariances and regressions
- An example of statistical inference based on covariances

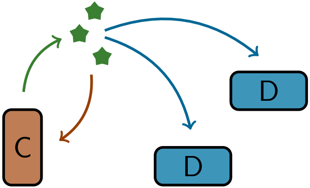
Outline

- An example of the role of spatial structure in evolution
- Modeling of the role of spatial structure
- Formal modelling of covariances: diffusion vs. coalescence
- Genealogical interpretations of covariances and regressions
- An example of statistical inference based on covariances

Example: evolution of social behaviour

Siderophore production as "public good"

Pseudomonas aeruginosa



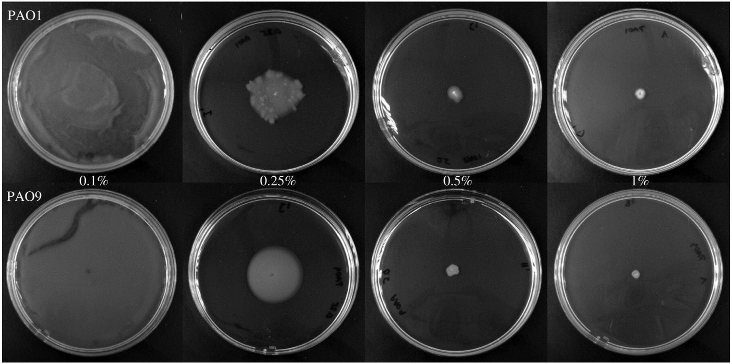
A Defector that does not produce the "public good" still benefits from those produced by the Cooperators.

⇒ Public good production should be counter-selected.

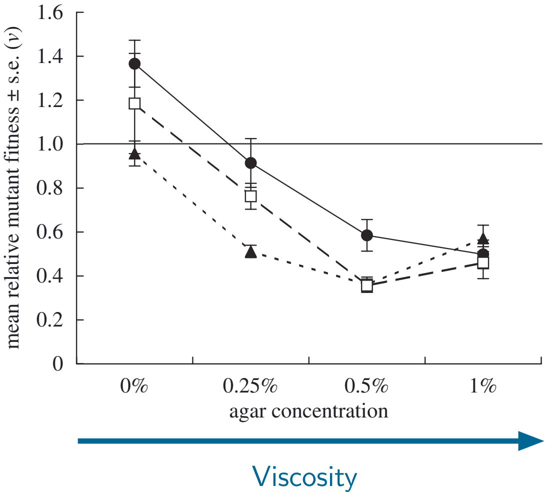
Biologists test theories about the importance of dispersal

Controlling bacterial dispersal

Different agar concentrations:



Public goods production

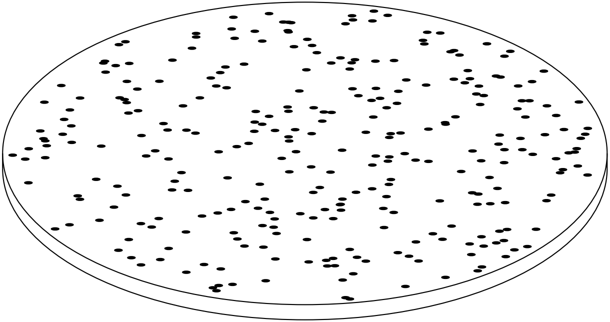


Siderophore-defective strain wins

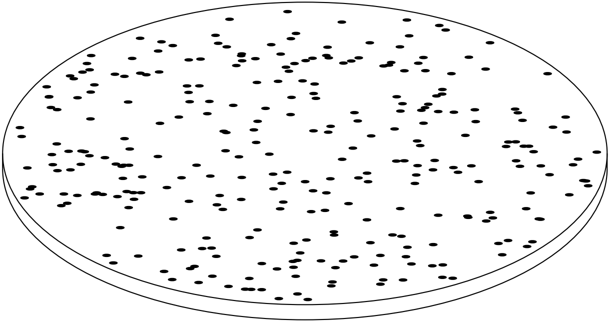
Siderophore-producing strain wins

Kümmerli et al, PRSB, 2009

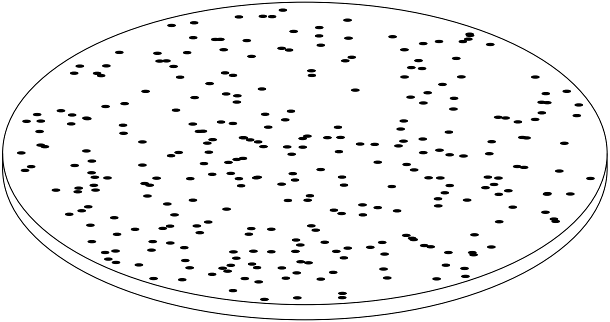
The effect of dispersal *almost* explained



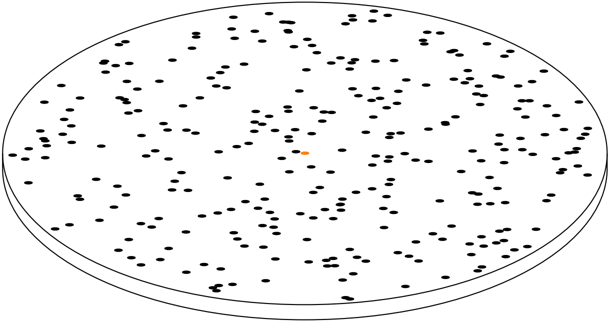
The effect of dispersal *almost* explained



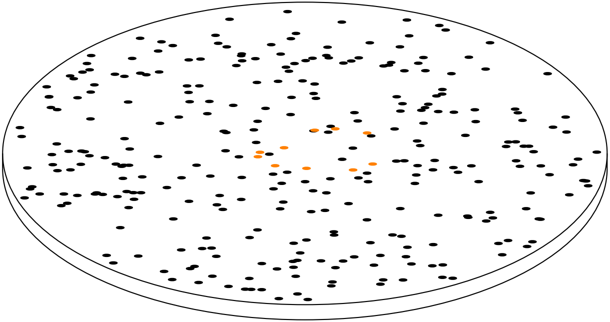
The effect of dispersal *almost* explained



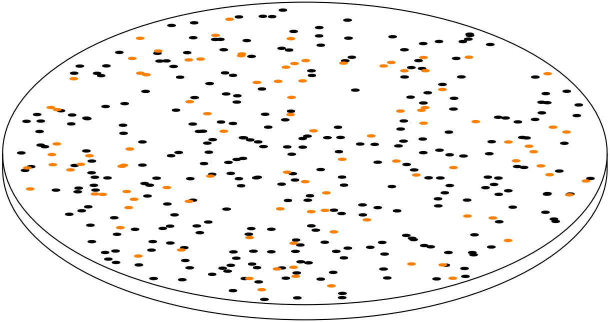
The effect of dispersal *almost* explained



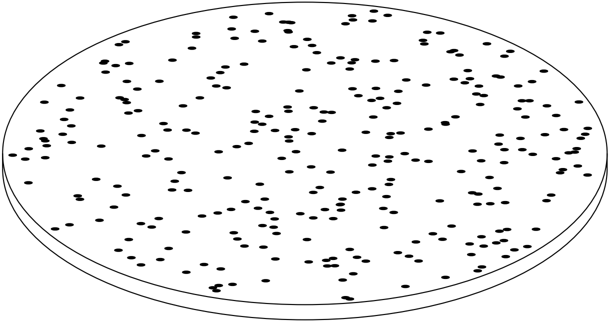
The effect of dispersal *almost* explained



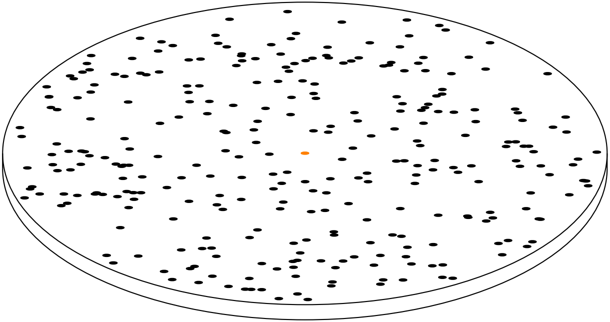
The effect of dispersal *almost* explained



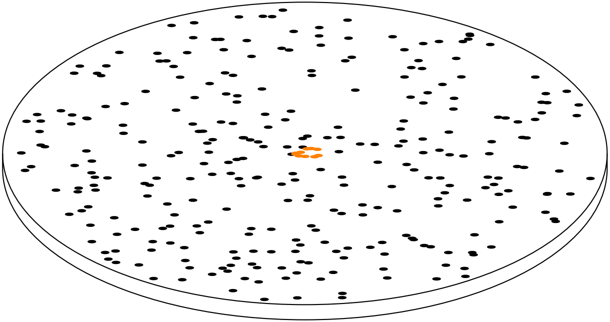
The effect of dispersal *almost* explained



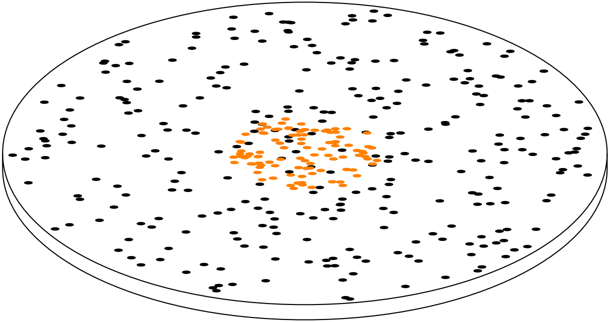
The effect of dispersal *almost* explained



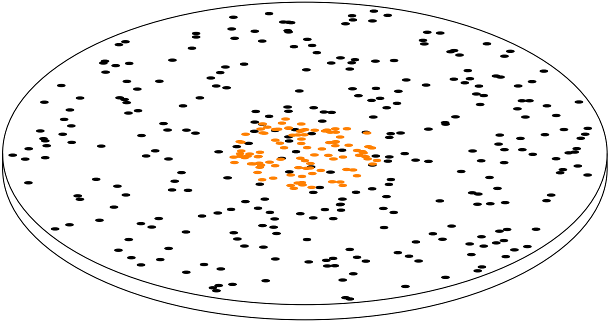
The effect of dispersal *almost* explained



The effect of dispersal *almost* explained



The effect of dispersal *almost* explained



- Same logic important for virtually any trait evolving in a population spatially subdivided in (small) subpopulations.

Outline

- An example of the role of spatial structure in evolution
- Modeling of the role of spatial structure
- Formal modelling of covariances: diffusion vs. coalescence
- Genealogical interpretations of covariances and regressions
- An example of statistical inference based on covariances

Formulation in terms of partial regression

We have seen that (for deterministic models) the change in allele frequency can be written in terms of allelic fitnesses $w_{\textcolor{red}{\bullet}}$ and $w_{\textcolor{blue}{\bullet}}$, as

$$\begin{aligned} p'_{\textcolor{red}{\bullet}} - p_{\textcolor{red}{\bullet}} =& (w_{\textcolor{red}{\bullet}} - w_{\textcolor{blue}{\bullet}}) p_{\textcolor{red}{\bullet}} (1 - p_{\textcolor{red}{\bullet}}) \\ =& \beta_{w,\mathbf{1}_{\textcolor{red}{\bullet}}} \operatorname{Var}(\mathbf{1}_{\textcolor{red}{\bullet}}) = \operatorname{Cov}[w_i, \mathbf{1}_{\textcolor{red}{\bullet}}(i))] \end{aligned}$$

where $\beta_{w,\mathbf{1}_{\textcolor{red}{\bullet}}}$ is the regression coefficient of an individual's allelic fitness w_i on her genotype $\mathbf{1}_{\textcolor{red}{\bullet}}(i)$

Formulation in terms of partial regression

We have seen that (for deterministic models) the change in allele frequency can be written in terms of allelic fitnesses $w_{\color{red}\bullet}$ and $w_{\color{blue}\bullet}$, as

$$\begin{aligned} p'_{\color{red}\bullet} - p_{\color{red}\bullet} =& (w_{\color{red}\bullet} - w_{\color{blue}\bullet}) p_{\color{red}\bullet} (1 - p_{\color{red}\bullet}) \\ =& \beta_{w,\mathbf{1}_{\color{red}\bullet}} \operatorname{Var}(\mathbf{1}_{\color{red}\bullet}) = \operatorname{Cov}[w_i, \mathbf{1}_{\color{red}\bullet}(i))] \end{aligned}$$

where $\beta_{w,\mathbf{1}_{\color{red}\bullet}}$ is the regression coefficient of an individual's allelic fitness w_i on her genotype $\mathbf{1}_{\color{red}\bullet}(i)$

We can apply the standard recursive relationship for regression coefficients between a response variable W and two predictors F and N:

$$\beta_{W,F} = \beta_{W,F.N} + \beta_{N,F} \beta_{W,N.F}.$$

here for $F := \mathbf{1}_{\color{red}\bullet}$, the indicator variable for the (focal) individual's genotype, and $N :=$ the mean value of the indicator variable for a group of social partners

Formulation in terms of partial regression

We have seen that (for deterministic models) the change in allele frequency can be written in terms of allelic fitnesses $w_{\color{red}\bullet}$ and $w_{\color{blue}\bullet}$, as

$$\begin{aligned} p'_{\color{red}\bullet} - p_{\color{red}\bullet} =& (w_{\color{red}\bullet} - w_{\color{blue}\bullet}) p_{\color{red}\bullet}(1 - p_{\color{red}\bullet}) \\ =& \beta_{w,\mathbf{1}_{\color{red}\bullet}} \operatorname{Var}(\mathbf{1}_{\color{red}\bullet}) = \operatorname{Cov}[w_i, \mathbf{1}_{\color{red}\bullet}(i))] \end{aligned}$$

where $\beta_{w,\mathbf{1}_{\color{red}\bullet}}$ is the regression coefficient of an individual's allelic fitness w_i on her genotype $\mathbf{1}_{\color{red}\bullet}(i)$
This entails the two-predictor regression expression for the change in allele frequency:

$$p'_{\color{red}\bullet} - p_{\color{red}\bullet} = (\beta_{W,F.N} + \beta_{N,F}\beta_{W,N.F}) \operatorname{Var}(\mathbf{1}_{\color{red}\bullet})$$

where $\beta_{N,F}$ is the regression coefficient of the $\color{red}\bullet$ allele frequency among social partners to $\color{red}\bullet$ allele presence in the focal individual.

$\beta_{N,F}$ is (one definition of) relatedness in the genetic literature.

Outline

- An example of the role of spatial structure in evolution
- Modeling of the role of spatial structure
- Formal modelling of covariances: diffusion vs. coalescence
- Genealogical interpretations of covariances and regressions
- An example of statistical inference based on covariances

Probability of identity

- Probability that two gene copies are of same allelic state
  = Identity in state

Probability of identity

- Probability that two gene copies are of same allelic state
  = Identity in state
- Cf Wright-Fisher model with two alleles

$$\mathrm{a} \underset{v}{\overset{u}{\rightleftarrows}} \mathrm{A}$$

Probability of identity

- Probability that two gene copies are of same allelic state
  = Identity in state
- Cf Wright-Fisher model with two alleles

$$\mathrm{a} \overset{u}{\underset{v}{\rightleftarrows}} \mathrm{A}$$

- pdf of allele frequency $f(p) =$Beta$(\alpha = 2Nu, \beta = 2Nv)$

Probability of identity

- Probability that two gene copies are of same allelic state $=$ Identity in state
- Cf Wright-Fisher model with two alleles

$$\mathrm{a} \overset{u}{\underset{v}{\rightleftarrows}} \mathrm{A}$$

- pdf of allele frequency $f(p) =$Beta$(\alpha = 2Nu, \beta = 2Nv)$
-

$$\mathbb{P}(\text{identity}) = \int_0^1 \qquad\qquad f(x)\,\mathrm{d}x$$

Probability of identity

- Probability that two gene copies are of same allelic state = Identity in state
- Cf Wright-Fisher model with two alleles

$$\mathrm{a} \overset{u}{\underset{v}{\rightleftarrows}} \mathrm{A}$$

- pdf of allele frequency $f(p) =$Beta$(\alpha = 2Nu, \beta = 2Nv)$
- $$\mathbb{P}(\text{identity}) = \int_0^1 [x^2 + (1-x)^2] f(x)\,\mathrm{d}x = \frac{u + 2Nu^2 + v + 2Nv^2}{(u+v)[1+2N(u+v)]}$$
  $$\approx \frac{1}{1+4Nu} \text{ if } u = v$$

Probability of identity

- Probability that two gene copies are of same allelic state
  = Identity in state
- Cf Wright-Fisher model with two alleles

$$\mathrm{a} \overset{u}{\underset{v}{\rightleftarrows}} \mathrm{A}$$

- pdf of allele frequency $f(p) =$Beta$(\alpha = 2Nu, \beta = 2Nv)$
- $$\mathbb{P}(\text{identity}) = \int_0^1 [x^2 + (1-x)^2] f(x)\,\mathrm{d}x = \frac{u + 2Nu^2 + v + 2Nv^2}{(u+v)[1 + 2N(u+v)]}$$
  $$\approx \frac{1}{1+4Nu} \text{ if } u = v$$
- Mathematical overkill!

Probability of identity: coalescent argument

- Infinite allele model $\Rightarrow$ identity by descent

Probability of identity: coalescent argument

- Infinite allele model ⇒ identity by descent
- For c_t the probability that coalescence of a a given pair of genes occurred t generations ago,

$$\mathbb{P}(\text{identity}) = Q = \sum_{t=0}^{\infty} c_t \times \ldots ?$$

Probability of identity: coalescent argument

- Infinite allele model ⇒ identity by descent
- For c_t the probability that coalescence of a a given pair of genes occurred t generations ago,

$$\mathbb{P}(\text{identity}) = Q = \sum_{t=0}^{\infty} c_t (1-\mu)^{2t} = \sum_{t=0}^{\infty} c_t \gamma^t \text{ for } \gamma \equiv (1-\mu)^2$$

Probability of identity: coalescent argument

- Infinite allele model ⇒ identity by descent
- For c_t the probability that coalescence of a a given pair of genes occurred t generations ago,

$$\mathbb{P}(\text{identity}) = Q = \sum_{t=0}^{\infty} c_t (1-\mu)^{2t} = \sum_{t=0}^{\infty} c_t \gamma^t \text{ for } \gamma \equiv (1-\mu)^2$$

- Generating function

Probability of identity: coalescent argument

- Infinite allele model ⇒ identity by descent
- For c_t the probability that coalescence of a a given pair of genes occurred t generations ago,

$$\mathbb{P}(\text{identity}) = Q = \sum_{t=0}^{\infty} c_t (1-\mu)^{2t} = \sum_{t=0}^{\infty} c_t \gamma^t \text{ for } \gamma \equiv (1-\mu)^2$$

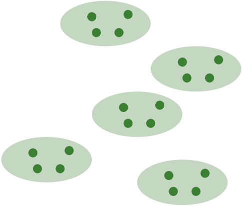
- Generating function
- For Wright-Fisher model:

$$Q = \sum_{1}^{\infty} \frac{1}{N} \left(1 - \frac{1}{N}\right)^{t-1} \gamma^t = \frac{\gamma}{\gamma + N(1-\gamma)}$$

$$\rightarrow \frac{1}{1+2N\mu} \text{ when } N \rightarrow \infty \text{ and } \mu \rightarrow 0$$

Infinite island model; Wright's F_{ST}

A very simple, easily solved model



Immigration m

Frequency of some focal allele ● locally, $p = p_i$ for given "island" i, and globally, $\bar{p}$.

Infinite island model; Wright's F_{ST}

Diffusion approximation for frequency in each "island" *i*:

$$P_i \sim \mathrm{Const} \quad p_i^{2Nm\bar{p}-1}(1-p_i)^{2Nm(1-\bar{p})-1} = \mathrm{Beta}(2Nm\bar{p}, 2Nm(1-\bar{p}))$$

Infinite island model; Wright's F_{ST}

Diffusion approximation for frequency in each "island" i:

$P_i \sim \mathrm{Const} \quad p_i^{2Nm\bar{p}-1}(1-p_i)^{2Nm(1-\bar{p})-1} = \mathrm{Beta}(2Nm\bar{p}, 2Nm(1-\bar{p}))$

We have seen that the stationarity distribution satisfies:

$$0 = \frac{\partial f(p,t)}{\partial t} = -\frac{\partial a(p) f(p,t)}{\partial p} + \frac{\partial^2 b(p) f(p,t)}{2\partial p^2}$$

$$f(p) \propto \frac{exp\left(2\int^p a(x)/b(x)\,\mathrm{d}x\right)}{b(p)}$$

Infinite island model; Wright's F_{ST}

Diffusion approximation for frequency in each "island" i:

$$P_i \sim \mathrm{Const} \quad p_i^{2Nm\bar{p}-1}(1-p_i)^{2Nm(1-\bar{p})-1} = \mathrm{Beta}(2Nm\bar{p}, 2Nm(1-\bar{p}))$$

Wright-Fisher model:
$b(p) = p(1-p)$ and
$a(p) = N(-vp + u(1-p)) = -N[(u+v)(p - u/(u+v))]$

Frequency $P \sim \mathrm{Const} \quad p^{2Nu-1}(1-p)^{2Nv-1} = \mathrm{Beta}(\alpha = 2Nu, \beta = 2Nv)$

Infinite island model; Wright's F_{ST}

Diffusion approximation for frequency in each "island" i:

$$P_i \sim \mathrm{Const} \quad p_i^{2Nm\bar{p}-1}(1-p_i)^{2Nm(1-\bar{p})-1} = \mathrm{Beta}(2Nm\bar{p}, 2Nm(1-\bar{p}))$$

Wright-Fisher model:
$b(p) = p(1-p)$ and
$a(p) = N(-vp + u(1-p)) = -N[(u+v)(p - u/(u+v))]$

Frequency $P \sim \mathrm{Const} \quad p^{2Nu-1}(1-p)^{2Nv-1} = \mathrm{Beta}(\alpha = 2Nu, \beta = 2Nv)$

Island model
$b(p_i) = p_i(1-p_i)$ and $a(p_i) = -N(m(p_i - \bar{p}))$

Frequency $P \sim \mathrm{Beta}[\alpha = 2Nm\bar{p}, \beta = 2Nm(1-\bar{p})]$

Infinite island model; Wright's F_{ST}

Diffusion approximation for frequency in each "island" i:

$$P_i \sim \mathrm{Const} \quad p_i^{2Nm\bar{p}-1}(1-p_i)^{2Nm(1-\bar{p})-1} = \mathrm{Beta}(2Nm\bar{p}, 2Nm(1-\bar{p}))$$

$$F_{\mathrm{ST}} \equiv \frac{\mathrm{Var}\, p_i}{\bar{p}(1-\bar{p})} = \frac{\mathrm{E}(p_i^2) - \bar{p}^2}{\bar{p}(1-\bar{p})} = \frac{1}{1+2Nm}$$

More generally variances and covariances of allele frequencies can be expressed in terms of probabilities of identity of pair of genes. Here locally,

$$\mathbb{P}(\bullet\bullet | p_i; \bar{p}) = F_{\mathrm{ST}} p_i + (1 - F_{\mathrm{ST}})\bar{p}^2$$

(with mutation at a low enough rate: proven in next slides); hence globally

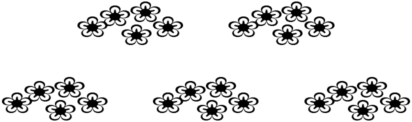
$$\mathbb{P}(\bullet\bullet; \bar{p}) = F_{\mathrm{ST}} \bar{p} + (1 - F_{\mathrm{ST}})\bar{p}^2$$

.

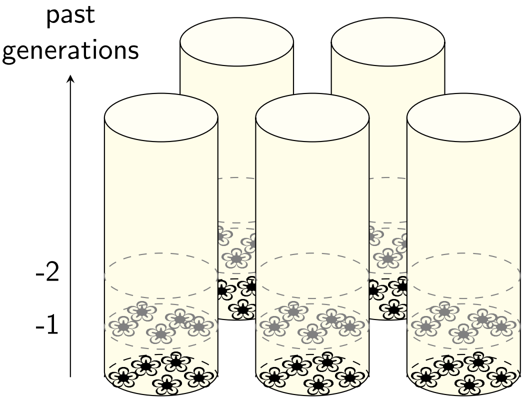
Outline

- An example of the role of spatial structure in evolution
- Modeling of the role of spatial structure
- Formal modelling of covariances: diffusion vs. coalescence
- Genealogical interpretations of covariances and regressions
  Infinite island model – F_{ST}, separation of time scales
  Finite island model – coalescence times, separation of time scales
  Isolation by distance – separation of time scales
  $\Rightarrow$ Robustness of regression coefficients
- An example of statistical inference based on covariances

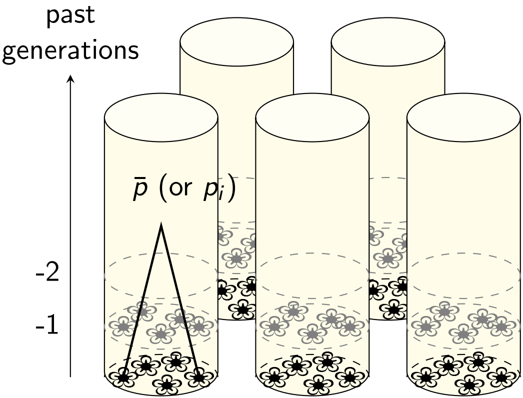
Genealogical interpretation of F_{ST}



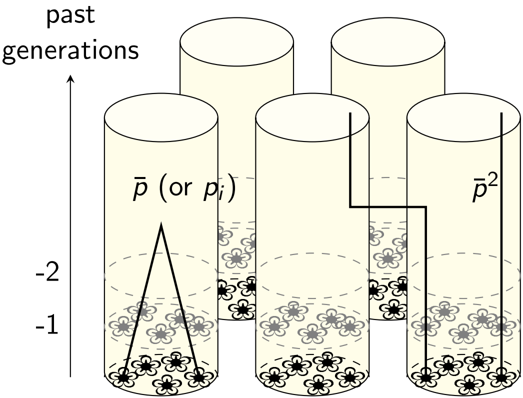
Genealogical interpretation of F_{ST}



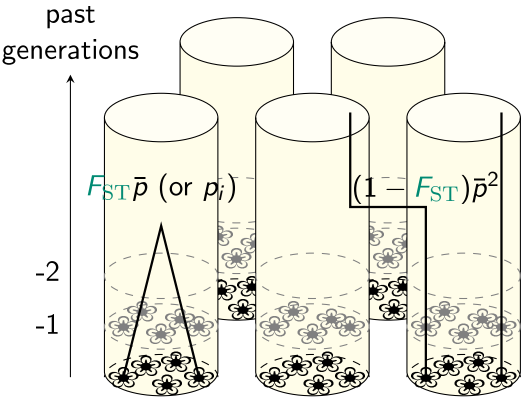
Genealogical interpretation of F_{ST}



Genealogical interpretation of F_{ST}



Genealogical interpretation of F_{ST}



Wright's F_{ST} redefined

(For any genetic marker)

One can describe the correlation of allele types within some given sub-population relative to the allele types of random genes in the total population as

$$F_{\mathrm{ST}} = \frac{Q_{\mathrm{w}} - Q_{\mathrm{b}}}{1 - Q_{\mathrm{b}}}$$

in terms of probabilities of identity Q_{w} and Q_{b} within and between subpopulations.

This leads to relatively simple analysis.

General facts about Wright's *F*-statistics

$$F_{\mathrm{ST}} = \frac{Q_{\mathrm{w}} - Q_{\mathrm{b}}}{1 - Q_{\mathrm{b}}}$$

$$F_{\mathrm{IS}} = \frac{Q_0 - Q_{\mathrm{w}}}{1 - Q_{\mathrm{w}}}$$

General facts about Wright's F-statistics

$$\hat{F}_{\mathrm{ST}} = \frac{\hat{Q}_{\mathrm{w}} - \hat{Q}_{\mathrm{b}}}{1 - \hat{Q}_{\mathrm{b}}}$$

$$F_{\mathrm{IS}} = \frac{Q_0 - Q_{\mathrm{w}}}{1 - Q_{\mathrm{w}}}$$

Naive but straightforward estimators

General facts about Wright's *F*-statistics

$$F_{\mathrm{ST}} = \frac{Q_{\mathrm{w}} - Q_{\mathrm{b}}}{1 - Q_{\mathrm{b}}}$$

$$F_{\mathrm{IS}} = \frac{Q_0 - Q_{\mathrm{w}}}{1 - Q_{\mathrm{w}}}$$

Relationship with average coalescence times

$$Q = \sum_{t=0}^{\infty} c_t \gamma^t \text{ for } \gamma \equiv (1-\mu)^2 = 1 - 2\mu\, \mathsf{E}(T) + O(\mu^2)$$

$$F_{\mathrm{ST}} \to \frac{\mathsf{E}(T_{\mathrm{b}}) - \mathsf{E}(T_{\mathrm{w}})}{\mathsf{E}(T_{\mathrm{b}})} \text{ when } \mu \to 0.$$

Modelling in terms of probabilities of identity Q

Example: finite island model
Relationship of Q_{w} with Q_{w}, Q_{b} one generation earlier?

where

and

Modelling in terms of probabilities of identity Q

Example: finite island model
Relationship of Q_{w} with Q_{w}, Q_{b} one generation earlier?

$$\mathbf{Q}(t+1) = \gamma\mathbf{A}[\mathbf{Q}(t) + \mathbf{c}(t)]$$

where

and

Modelling in terms of probabilities of identity Q

Example: finite island model
Relationship of Q_{w} with Q_{w}, Q_{b} one generation earlier?

$$\mathbf{Q}(t+1) = \gamma \mathbf{A}[\mathbf{Q}(t) + \mathbf{c}(t)]$$

where

$$\mathbf{Q}(t) = \begin{pmatrix} Q_{\mathrm{w}}(t) \\ Q_{\mathrm{b}}(t) \end{pmatrix} \text{ and } \mathbf{c}(t) = \begin{pmatrix} \frac{1-Q_{\mathrm{w}}(t)}{N} \\ 0 \end{pmatrix}$$

and

Modelling in terms of probabilities of identity Q

Example: finite island model
Relationship of Q_{w} with Q_{w}, Q_{b} one generation earlier?

$$\mathbf{Q}(t+1) = \gamma \mathbf{A}[\mathbf{Q}(t) + \mathbf{c}(t)]$$

where

$$\mathbf{Q}(t) = \begin{pmatrix} Q_{\mathrm{w}}(t) \\ Q_{\mathrm{b}}(t) \end{pmatrix} \text{ and } \mathbf{c}(t) = \begin{pmatrix} \frac{1-Q_{\mathrm{w}}(t)}{N} \\ 0 \end{pmatrix}$$

and

$$\mathbf{A} = \begin{pmatrix} a_{11} = (1-m)^2 + \frac{m^2}{n_{\mathrm{d}}-1} & 1 - a_{11} \\ 1 - \frac{a_{11}}{n_{\mathrm{d}}-1} & \frac{a_{11}}{n_{\mathrm{d}}-1} \end{pmatrix}$$

is the transition matrix for the "random walk" of two gene lineages between the state "within same deme" and the complementary state.

The stationary solution in an informative form

$$\mathbf{Q} = \gamma(\mathbf{I} - \gamma\mathbf{A})^{-1}\mathbf{A}\mathbf{c}$$

Eigensystem of **A**

$\lambda_1 = 1$, $\mathbf{e}_1 = (1, 1)^\top$

$\lambda_2 = \left(1 - m\frac{n_\mathrm{d}}{n_\mathrm{d}-1}\right)^2$, $\mathbf{e}_2 = (n_\mathrm{d} - 1, -1)^\top$

$$\mathbf{c} = \begin{pmatrix} \frac{1-Q_\mathrm{w}}{N} \\ 0 \end{pmatrix} = \frac{1 - Q_\mathrm{w}}{N}\frac{1}{n_\mathrm{d}}(\mathbf{e}_1 + \mathbf{e}_2)$$

The stationary solution in an informative form

$$\mathbf{Q} = \gamma(\mathbf{I} - \gamma\mathbf{A})^{-1}\mathbf{A}\mathbf{c}$$

Eigensystem of **A**

$\lambda_1 = 1$, $\mathbf{e}_1 = (1, 1)^\top$

$\lambda_2 = \left(1 - m\frac{n_\mathrm{d}}{n_\mathrm{d}-1}\right)^2$, $\mathbf{e}_2 = (n_\mathrm{d} - 1, -1)^\top$

$$\mathbf{c} = \begin{pmatrix} \frac{1-Q_\mathrm{w}}{N} \\ 0 \end{pmatrix} = \frac{1 - Q_\mathrm{w}}{N}\frac{1}{n_\mathrm{d}}(\mathbf{e}_1 + \mathbf{e}_2)$$

$$\mathbf{Q} = \frac{1 - Q_\mathrm{w}}{N}\frac{1}{n_\mathrm{d}}\left(\frac{\gamma}{1-\gamma}\mathbf{e}_1 + \frac{\gamma\lambda_2}{1-\gamma\lambda_2}\mathbf{e}_2\right)$$

The stationary solution in an informative form

$$\mathbf{Q} = \gamma(\mathbf{I} - \gamma\mathbf{A})^{-1}\mathbf{A}\mathbf{c}$$

Eigensystem of **A**

$\lambda_1 = 1$, $\mathbf{e}_1 = (1, 1)^\top$

$\lambda_2 = \left(1 - m\frac{n_\mathrm{d}}{n_\mathrm{d}-1}\right)^2$, $\mathbf{e}_2 = (n_\mathrm{d} - 1, -1)^\top$

$$\mathbf{c} = \begin{pmatrix} \frac{1-Q_\mathrm{w}}{N} \\ 0 \end{pmatrix} = \frac{1 - Q_\mathrm{w}}{N}\frac{1}{n_\mathrm{d}}(\mathbf{e}_1 + \mathbf{e}_2)$$

$$\mathbf{Q} = \frac{1 - Q_\mathrm{w}}{N}\frac{1}{n_\mathrm{d}}\left(\frac{\gamma}{1-\gamma}\mathbf{e}_1 + \frac{\gamma\lambda_2}{1-\gamma\lambda_2}\mathbf{e}_2\right)$$

$$\frac{Q_\mathrm{w} - Q_\mathrm{b}}{1 - Q_\mathrm{w}} = \frac{1}{N}\frac{\gamma\lambda_2}{1-\gamma\lambda_2}$$

The stationary solution in an informative form

$$\mathbf{Q} = \gamma(\mathbf{I} - \gamma\mathbf{A})^{-1}\mathbf{A}\mathbf{c}$$

Eigensystem of **A**

$\lambda_1 = 1$, $\mathbf{e}_1 = (1,1)^\top$

$\lambda_2 = \left(1 - m\frac{n_\mathrm{d}}{n_\mathrm{d}-1}\right)^2$, $\mathbf{e}_2 = (n_\mathrm{d} - 1, -1)^\top$

$$\mathbf{c} = \begin{pmatrix} \frac{1-Q_\mathrm{w}}{N} \\ 0 \end{pmatrix} = \frac{1 - Q_\mathrm{w}}{N}\frac{1}{n_\mathrm{d}}(\mathbf{e}_1 + \mathbf{e}_2)$$

$$\mathbf{Q} = \frac{1 - Q_\mathrm{w}}{N}\frac{1}{n_\mathrm{d}}\left(\frac{\gamma}{1-\gamma}\mathbf{e}_1 + \frac{\gamma\lambda_2}{1-\gamma\lambda_2}\mathbf{e}_2\right)$$

$$F_\mathrm{ST} \equiv \frac{Q_\mathrm{w} - Q_\mathrm{b}}{1 - Q_\mathrm{b}} = \frac{\gamma\lambda_2}{\gamma\lambda_2 + N(1-\gamma\lambda_2)} \approx \frac{1}{1+2Nm}$$

The stationary solution in another informative form

in terms of eigenvalues of another matrix **G**:

$$\mathbf{Q} = \gamma \mathbf{A}(\mathbf{Q} + \mathbf{c}) = \gamma[\mathbf{G}\mathbf{Q} + (\mathbf{A} - \mathbf{G})\mathbf{1}] = \gamma[\mathbf{G}\mathbf{Q} + (\mathbf{I} - \mathbf{G})\mathbf{1}]$$

where **G** is the matrix whose ijth element is the coefficient of the jth element of **Q** in the ith element of $\mathbf{A}[\mathbf{Q} + \mathbf{c}]$.

Denote ℓ_j and $\mathbf{u}_j$ the eigenvalues and eigenvectors of **G**; and write $(\mathbf{I} - \mathbf{G})\mathbf{1}$ as $\sum_j x_j \mathbf{u}_j$. Then

The stationary solution in another informative form

in terms of eigenvalues of another matrix **G**:

$$\mathbf{Q} = \gamma\mathbf{A}(\mathbf{Q}+\mathbf{c}) = \gamma[\mathbf{G}\mathbf{Q} + (\mathbf{A}-\mathbf{G})\mathbf{1}] = \gamma[\mathbf{G}\mathbf{Q} + (\mathbf{I}-\mathbf{G})\mathbf{1}]$$

where **G** is the matrix whose *ij*th element is the coefficient of the *j*th element of **Q** in the *i*th element of $\mathbf{A}[\mathbf{Q}+\mathbf{c}]$.

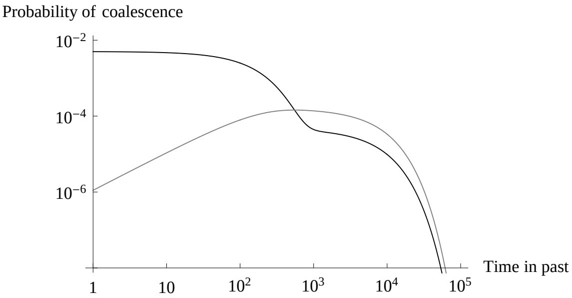
Denote ℓ_j and $\mathbf{u}_j$ the eigenvalues and eigenvectors of **G**; and write $(\mathbf{I}-\mathbf{G})\mathbf{1}$ as $\sum_j x_j\mathbf{u}_j$. Then

$$\mathbf{Q} = \gamma(\mathbf{I}-\gamma\mathbf{G})^{-1}\sum_j x_j\mathbf{u}_j = \sum_j \frac{\gamma x_j\mathbf{u}_j}{1-\gamma\ell_j} = \sum_t \gamma^t \sum_j \ell_j^{t-1}x_j\mathbf{u}_j$$

i.e. for *i*th element of **Q**, $c_{i,t} = \sum_j \ell_j^{t-1}x_j u_{ji}$.

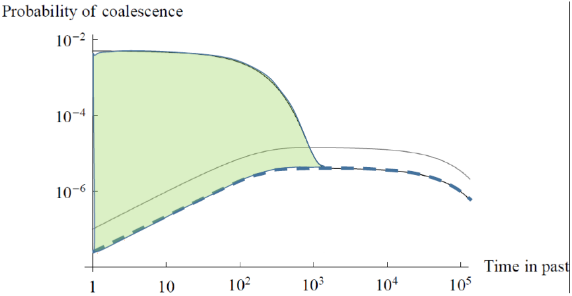
Genealogical interpretation of F-statistics

$(N = 100,\ m = 10^{-3}, \qquad n_{\mathrm{d}} = 10)$



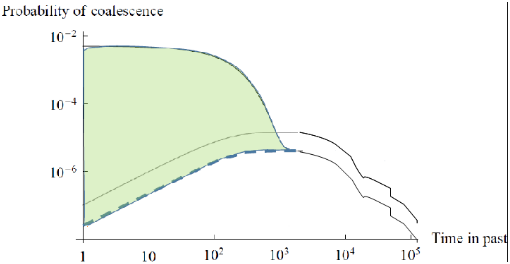
Genealogical interpretation of F-statistics

⇒ Interpretation of F_{ST} from comparison of two distributions of coalescence times



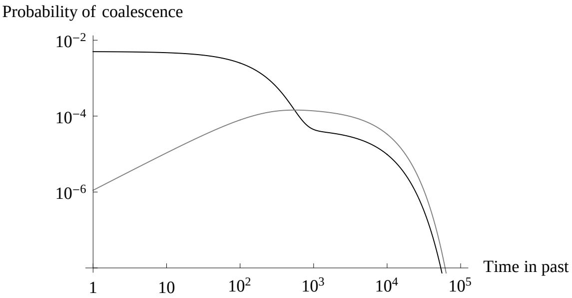
Genealogical interpretation of F-statistics

$\Rightarrow$ robustness: ancient events have no effect on F_{ST}



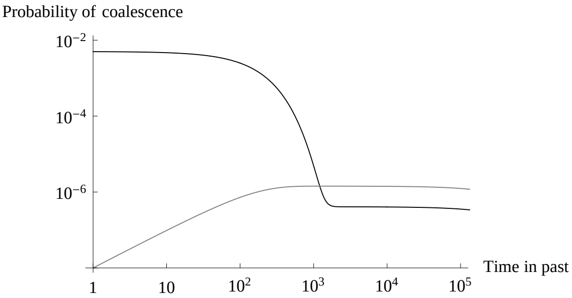
Genealogical interpretation of F-statistics

$(N = 100,\ m = 10^{-3}, \qquad n_{\mathrm{d}} = 10)$



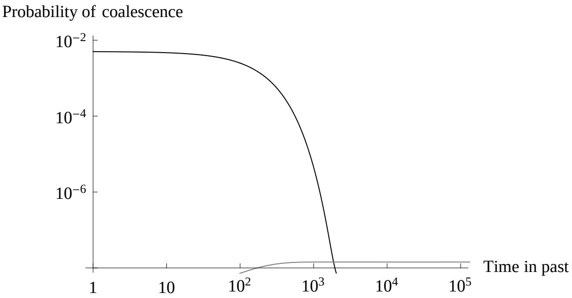
Genealogical interpretation of F-statistics

$(N = 100,\ m = 10^{-3}, \qquad n_{\mathrm{d}} = 1000)$

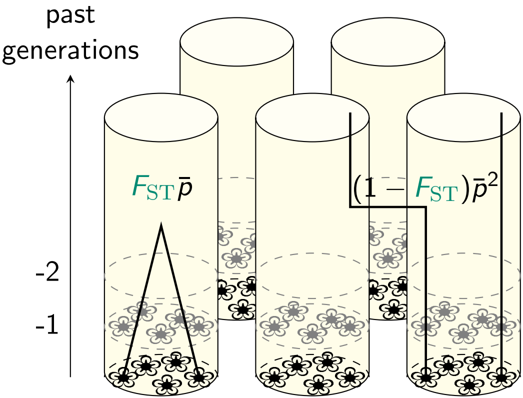


Genealogical interpretation of F-statistics

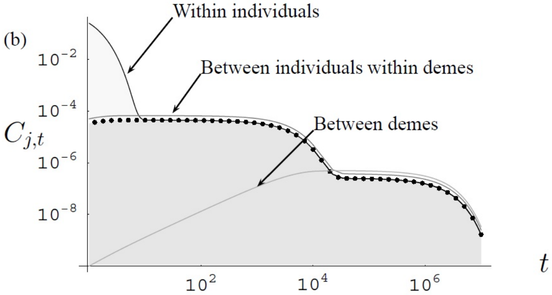
$(N = 100,\ m = 10^{-3}, \qquad n_{\mathrm{d}} = 100000)$



Genealogical interpretation of F_{ST}



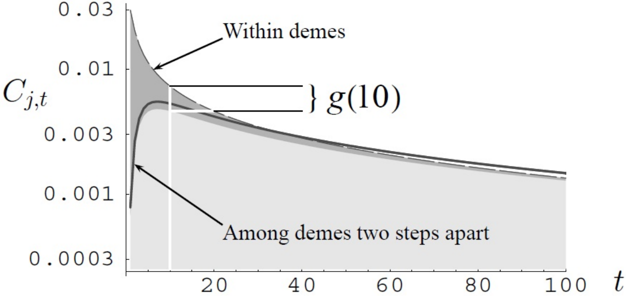
Island model with self-fertilization ("selfing")



Three eigenvalues of $\mathbf{G}$ = different time scales dependent on magnitude of $n_\mathrm{d}N$ vs. N vs selfing rate.

Isolation by distance

Pattern of isolation by distance



Isolation by distance: a quantitative approximation

- Generating function of one-lineage dispersal $\psi(z) := \sum_k m_k z^k$;

Isolation by distance: a quantitative approximation

- Generating function of one-lineage dispersal $\psi(z) := \sum_k m_k z^k$;
- Generating function of two-lineage increments in distance dispersal $\psi^2(z)$;

Isolation by distance: a quantitative approximation

- Generating function of one-lineage dispersal $\psi(z) := \sum_k m_k z^k$;
- Generating function of two-lineage increments in distance dispersal $\psi^2(z)$;
- In terms of the eigensystem of **A** for "any" distribution of dispersal distance:

$$\frac{\mathbf{Q}}{1-Q_{\mathrm{w}}} = \frac{1}{Nn_{\mathrm{d}}} \sum_j \frac{\gamma\lambda_j}{1-\gamma\lambda_j}\mathbf{e}_j$$

$\lambda_k = \psi^2\left(e^{\imath 2\pi k/n_x}\right)$ and $\mathbf{e}_k = (\cos(2\pi k i/n_x))$ for $(i = 0, \dots, n_x - 1)$;

Isolation by distance: a quantitative approximation

- Generating function of one-lineage dispersal $\psi(z) := \sum_k m_k z^k$;
- Generating function of two-lineage increments in distance dispersal $\psi^2(z)$;
- In terms of the eigensystem of **A** for "any" distribution of dispersal distance:

$$\frac{\mathbf{Q}}{1-Q_{\mathrm{w}}} = \frac{1}{Nn_{\mathrm{d}}} \sum_j \frac{\gamma\lambda_j}{1-\gamma\lambda_j} \mathbf{e}_j$$

  $\lambda_k = \psi^2\left(e^{\imath 2\pi k/n_x}\right)$ and $\mathbf{e}_k = (\cos(2\pi k i/n_x))$ for $(i = 0, \ldots, n_x - 1)$;
- Two dimensional $\lambda_{kl} = \psi^2\left(e^{\imath 2\pi k/n_x}\right)\psi^2\left(e^{\imath 2\pi l/n_y}\right)$;

Isolation by distance: a quantitative approximation

$$
\begin{aligned}
\frac{Q_{\mathrm{w}} - Q_{ij}}{1 - Q_{\mathrm{w}}} &= \frac{1}{N n_{\mathrm{d}}} \sum_k \sum_l \frac{\gamma \lambda_{kl}}{1 - \gamma \lambda_{kl}} (\mathbf{e}_{kl,0} - \mathbf{e}_{kl,ij}) \\
&\approx \frac{1}{N\pi^2} \int_0^\pi \int_0^\pi \frac{\gamma \psi^2(e^{\imath x}) \psi^2(e^{\imath y})}{1 - \gamma \psi^2(e^{\imath x}) \psi^2(e^{\imath y})} [1 - \cos(kx)\cos(ly)] \,\mathrm{d}x \,\mathrm{d}y \\
\frac{Q_{\mathrm{w}} - Q_d}{1 - Q_{\mathrm{w}}} &\approx \frac{\log(d)}{2N\pi\sigma^2} + \mathrm{constant}
\end{aligned}
$$

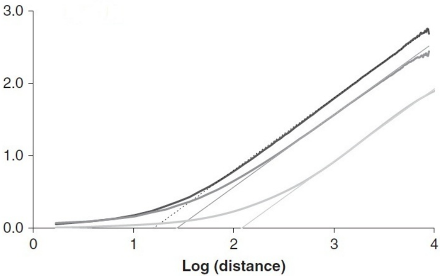
Isolation by distance: a quantitative approximation

$$\frac{Q_\mathrm{w}-Q_{ij}}{1-Q_\mathrm{w}} = \frac{1}{N n_\mathrm{d}} \sum_k \sum_l \frac{\gamma\lambda_{kl}}{1-\gamma\lambda_{kl}}(\mathbf{e}_{kl,0}-\mathbf{e}_{kl,ij})$$

$$\approx \frac{1}{N\pi^2}\int_0^\pi\int_0^\pi \frac{\gamma\psi^2\left(e^{\imath x}\right)\psi^2\left(e^{\imath y}\right)}{1-\gamma\psi^2\left(e^{\imath x}\right)\psi^2\left(e^{\imath y}\right)}\left[1-\cos(kx)\cos(ly)\right]\mathrm{d}x\,\mathrm{d}y$$

$$\frac{Q_\mathrm{w}-Q_d}{1-Q_\mathrm{w}} \approx \frac{\log(d)}{2N\pi\sigma^2} + \text{constant}$$

$$2N\pi\sigma^2\frac{Q_\mathrm{w}-Q_d}{1-Q_\mathrm{w}}$$

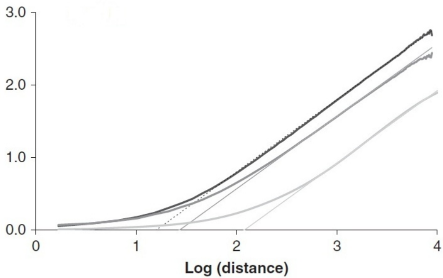


Outline

- An example of the role of spatial structure in evolution
- Modeling of the role of spatial structure
- Formal modelling of covariances: diffusion vs. coalescence
- Genealogical interpretations of covariances and regressions
- An example of statistical inference based on covariances

A simple inference under isolation by distance

$$2N\pi\sigma^2 \frac{Q_{\mathrm{w}} - Q_d}{1 - Q_{\mathrm{w}}}$$



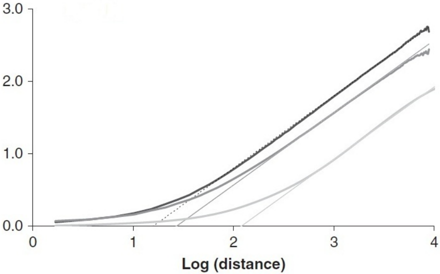
$\Rightarrow$ regress

$$\frac{Q_{\mathrm{w}} - Q_d}{1 - Q_{\mathrm{w}}}$$

against log(distance) to obtain an estimate of $1/2N\pi\sigma^2$.

A simple inference under isolation by distance

$$2N\pi\sigma^2 \frac{Q_{\mathrm{w}} - Q_d}{1 - Q_{\mathrm{w}}}$$



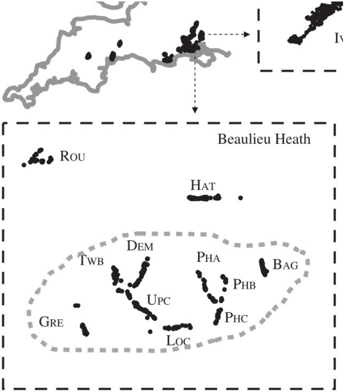
$\Rightarrow$ regress

$$\frac{\hat{Q}_{\mathrm{w}} - \hat{Q}_d}{1 - \hat{Q}_{\mathrm{w}}}$$

against log(distance) to obtain an estimate of $1/2N\pi\sigma^2$.

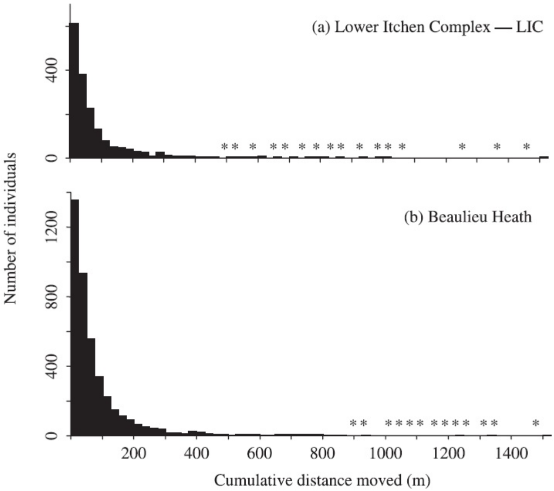
Statistical inference: an illustration

Statistical inference: an illustration

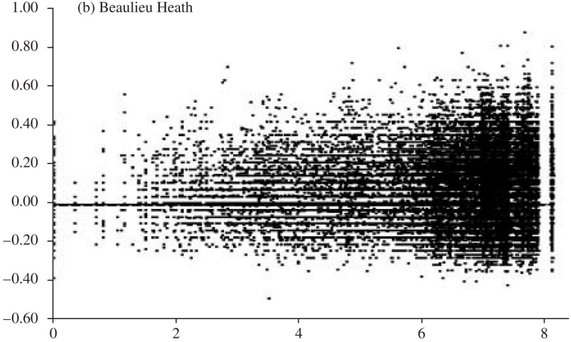


Statistical inference: an illustration

744 P. C. WATTS *ET AL.*

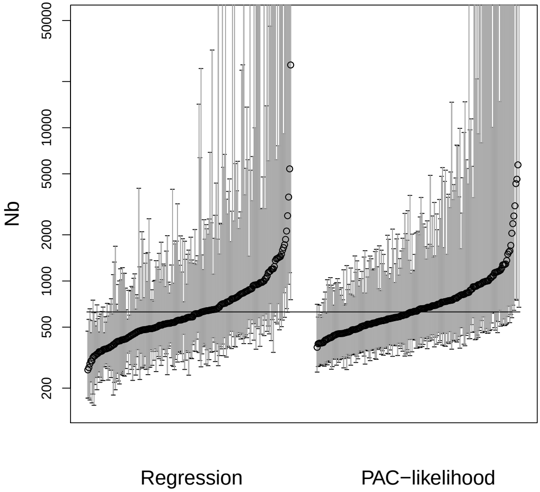


Statistical inference: an illustration



1/regression slope = estimate of “neighborhood size” Nb$\equiv 4D\pi\sigma^2$
Here one finds Nb$\hat{=}$753 with CI 319–3162 obtained by bootstrapping over loci.
(mark-recapture experiments yielded Nb$\hat{=}$555)

Moment vs. likelihood method



Summary

- Various evolutionary processes can be understood in terms of covariances in allele frequencies;
- Formal analysis in terms of probabilities of identity and distributions of coalescence times;
- Wright's F_{ST} and its variants can characterize recent events;
- They can provide some reasonable data analyses, although not the most precise ones.