

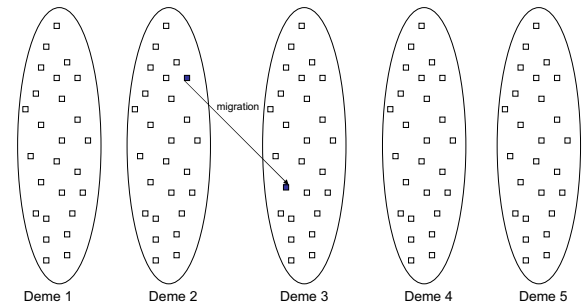
# A Coalescent Analysis of the Population Structure Statistic $F_{st}$

Sivan Leviyang  
Georgetown University

Cornell Probability Summer School 2008

## Population Structure

Population structure - the population can be partitioned into components across which migration is restricted.



## The Biological Motivation

### Basic Questions

- How can we experimentally determine if a population has structure?
- If a population has structure, how can we experimentally determine the migration rate(s)?

## A Common Statistical Approach

One popular statistic related to population structure is  $F_{st}$ .

### Population Structure Hypothesis Test

- $H_0$ : no structure.
- if  $H_0$  is true then  $F_{st} \approx 0$ .
- accept hypothesis if  $F_{st} < \epsilon$ .
- usually  $\epsilon$  is determined by randomization resampling techniques.

### Migration Estimation

- $F_{st} = f \Rightarrow m = \phi(f)$ .

## Mathematical Question

**Question:**What is the probability distribution of  $F_{st}$  for a certain model of a structured population?

**Motivation:**When do  $F_{st}$  based statistical tests perform well?

## Outline of Talk

- 1 Define  $F_{st}$
- 2 Introduce evolutionary model. (infinite allele, island model)
- 3 Result for  $F_{st}$  in high mutation case
- 4 Result and analysis for  $F_{st}$  in low mutation case
- 5 Stepping stone model

## Population and Sampling Variables

We consider a specific multiallelic gene in a haploid population.

$D$  = number of demes.

$N$  = number of individuals in each deme.

$d$  = number of demes sampled.

$n$  = number of individuals sampled from each deme.

$$x_{k,j} = \text{allele type of } j\text{th sampled individual in } k\text{th sampled deme.} \quad (1)$$

## Homozygosity: A Diversity Measure

Within deme homozygosity

$$H_k = \frac{1}{n^2} \sum_{j,j'=1}^n \chi(x_{k,j} = x_{k,j'}) \quad (2)$$

Average within deme homozygosity

$$H_W = \frac{1}{d} \sum_{k=1}^d H_k. \quad (3)$$

Total population homozygosity

$$H_T = \frac{1}{d^2} \sum_{k,k'=1}^d \frac{1}{n^2} \sum_{j,j'=1}^n \chi(x_{k,j} = x_{k',j'}). \quad (4)$$

## $F_{st}$

For  $H_T \neq 1$

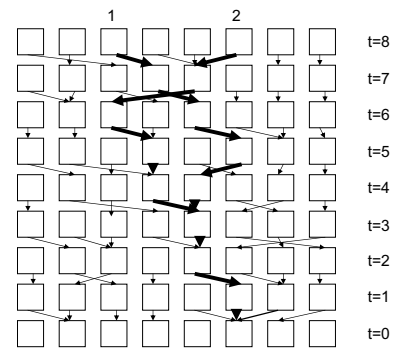
$$F_{st} = \frac{H_W - H_T}{1 - H_T}. \quad (5)$$

**Fact :**  $H_W > H_T$

$$F_{st} \in [0, 1]$$

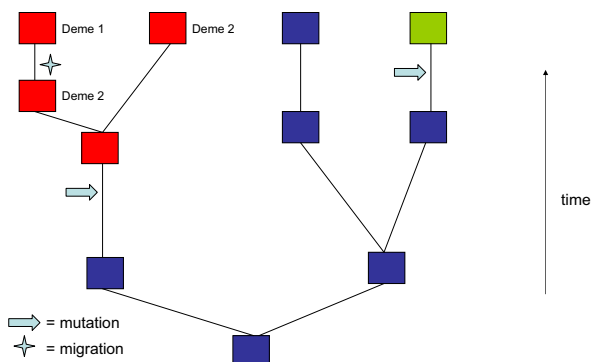
## The Coalescent

In the case of data sampling we only need consider the genealogy of our samples...



## Coalescent Model of Evolution

On the sample genealogy we have coalescent events, migration events, mutation events.



## The Coalescent

The coalescent is a Markov chain  $S(t)$

$$\mathbb{X} = \cup_k \cup_j \{x_{k,j}\} = \text{sampled individuals.} \quad (6)$$

$$G = \{1, 2, \dots, D\} = \text{demes.} \quad (7)$$

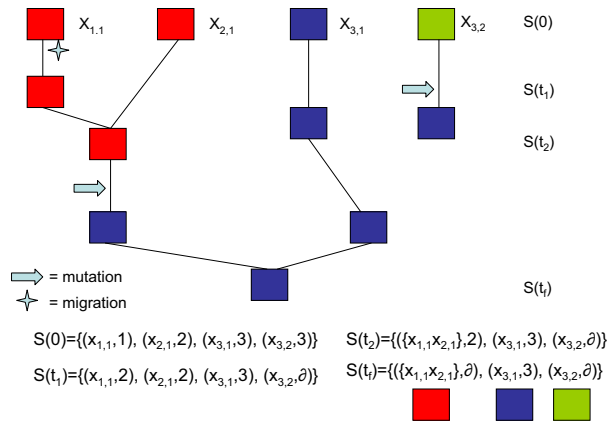
$\pi \in S$ :

$$\pi = \{(A_1, g_1), (A_2, g_2), \dots, (A_k, g_k) \mid \cup A_i = \mathbb{X}, g_i \in G\} \quad (8)$$

Jump rates

- 1 if  $g_1 = g_2$  then  $(A_1, g_1), (A_2, g_2) \rightarrow (A_1 \cup A_2, g_1)$  at rate  $\frac{2}{N}$ . (coalescent event)
- 2 take  $g \neq g_1$  then  $(A_1, g_1) \rightarrow (A_1, g)$  at rate  $\frac{\mu}{D}$ . (migration event under island model)
- 3 With rate  $\mu$  a block in  $S(t)$  is marked. A block with a mark does not coalesce with other blocks. (mutation event).

## Example Coalescent



## The Limiting Distribution

What is the distribution of  $F_{st}$ ?

$L(t)$  = number of unmarked blocks in  $S(t)$ .

$$T_M = \inf\{t : L(t) = 1\},$$

Large population, large sample, many demes limit [Wakeley 98]

$$\lim_{N, D, n, d \rightarrow \infty} F_{st}(S(T_M))$$

## Strong Mutation

$$\begin{aligned} \Gamma &= Nm \\ \theta_T &= \mu ND \\ \theta_W &= \mu N \end{aligned}$$

### Proposition (Strong Mutation)

Fix  $\Gamma$ . Suppose  $\theta_T \log(d) \rightarrow \infty$  and  $\theta_W \rightarrow 0$ . Then

$$\lim_{N, D, n, d \rightarrow \infty} F_{st} \rightarrow \frac{1}{1 + \Gamma} \quad (9)$$

## Proof of Strong Mutation Result

Recall that in mean field theory case we have

$$F_{st} = \frac{H_W - H_T}{1 - H_T} \approx H_W \rightarrow E[H_W] \quad (10)$$

When is  $H_T$  small?  $E[H_T] \approx P(x_{1,1} = x_{2,1})$ .

- P(enter same deme)  $\approx 2m \frac{1}{D} = \frac{2\Gamma}{ND}$
- P(mutation) =  $2\mu$

$$E[H_T] \leq \frac{2m \frac{1}{D}}{2m \frac{1}{D} + 2\mu} = \frac{1}{1 + \frac{\theta_T}{\Gamma}} \quad (11)$$

Mean field theory holds if  $\theta_T \rightarrow \infty$ .

## Proof of Strong Mutation Result

If  $\theta_T \log(d) \rightarrow \infty$  with  $\theta \rightarrow 0$  then we are not in mean field theory case. But we still have a collapse in correlation:

$$E[H_T^k] = C \left(\frac{1}{k}\right)^\theta \quad (12)$$

And a Taylor series argument works to show  $V[F_{st}] \rightarrow 0$ .

$$V[F_{st}] = V[(H_W - H_T)(1 + H_T + H_T^2 + \dots + H_T^n)] + o(1). \quad (13)$$

## Weak Mutation

$F_{st}$  is undefined when no mutations occur

$$F_{st} = \frac{H_W - H_T}{1 - H_T} = \frac{1 - 1}{1 - 1} \quad (14)$$

$$\lim_{\mu \rightarrow 0} F_{st} \Big|_{H_T \neq 1} = F_{st} \Big|_{\text{exactly one mutation in genealogy}} \quad (15)$$

Level of coalescent:  $L(t)$  = number of non-mutated blocks in  $S(t)$ .

## Weak Mutation

### Proposition (Weak Mutation)

Fix  $\Gamma$ . Let  $L = \lambda d$  be the level of the mutation. Then taking  $N, D, n, d \rightarrow \infty$  with  $\frac{n}{N} \rightarrow 0, \frac{d^2 n^2}{D} \rightarrow 0, \frac{\log^3(n)}{d} \rightarrow 0$  we have the following limit:

$$F_{st} \rightarrow F_{st}^{\text{lim}}(\lambda) \quad (16)$$

where

$$F_{st}^{\text{lim}}(\lambda) = \frac{\sum_{i=1}^{\infty} (X_i^{(1)})^2 + (X_i^{(2)})^2 + \dots + (X_i^{(W_i)})^2}{\sum_{i=1}^{\infty} X_i^{(1)} + X_i^{(2)} + \dots + X_i^{(W_i)}} \quad (17)$$

and

$W_i = \text{Poisson}(\frac{V}{\lambda}),$   
 $V = \text{Exponential}(1),$   
 $X_i = \beta_i \prod_{j=1}^{i-1} (1 - \beta_j),$   
 $\beta_j = \text{Beta}(1, \Gamma).$

July 2, 2008 19 / 31

## Weak Mutation

### Corollary (Low Level Mutations)

$$\lim_{\lambda \rightarrow 0} F_{st}^{\text{lim}}(\lambda) = \frac{1}{1 + \Gamma} \quad (18)$$

### Corollary (High Level Mutations)

$$\lim_{\lambda \rightarrow \infty} F_{st}^{\text{lim}}(\lambda) = 0 \quad (19)$$

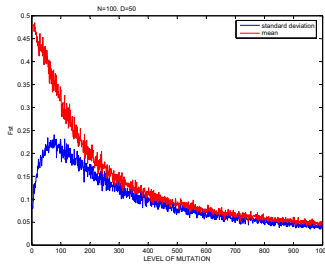
### Proposition

If  $L = d^{1-\epsilon}$  for  $\epsilon > 0$  then  $F_{st} \rightarrow \frac{1}{1+\Gamma}$ .

July 2, 2008 20 / 31

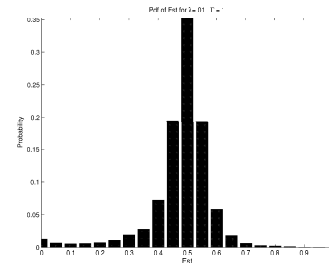
## $F_{st}$ and the Level of the Mutation

The mean and standard deviation of  $F_{st}$  for  $\Gamma = 1, d = 50$ .



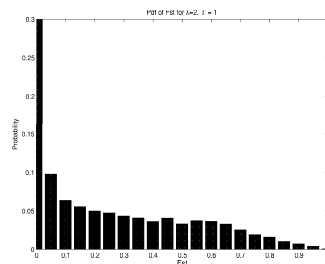
July 2, 2008 21 / 31

## $\lambda = .01$



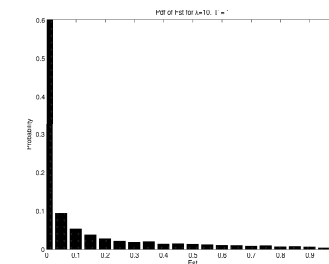
July 2, 2008 22 / 31

## $\lambda = 2$



July 2, 2008 23 / 31

## $\lambda = 10$



July 2, 2008 24 / 31

## Sketch of Proof for Weak Mutation Case

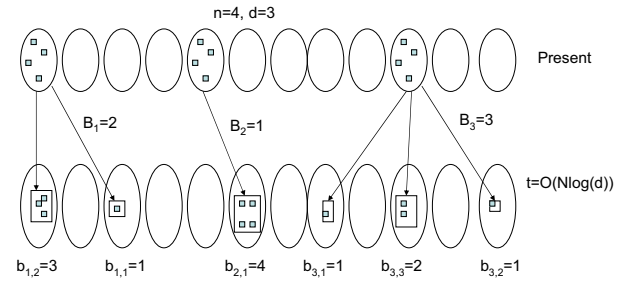
Uses an idea of Wakeley.

$n_i$  = the number of sampled individuals that are mutants in sample deme  $i$

$$F_{st} \rightarrow \frac{\sum_{i=1}^d \left(\frac{n_i}{n}\right)^2}{\sum_{i=1}^d \frac{n_i}{n}} \quad (20)$$

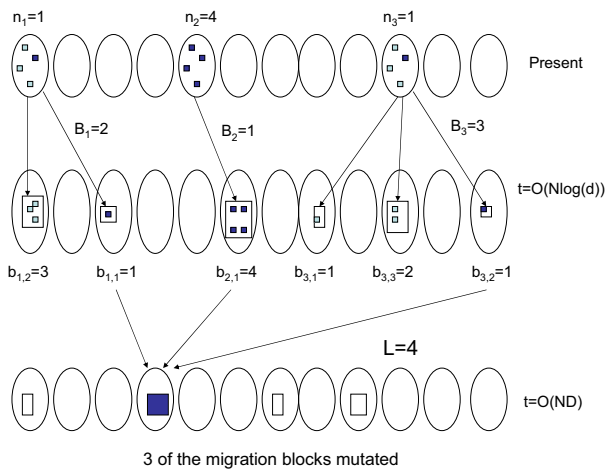
We need the distribution of the number of mutants from each sampled deme.

## Dispersal Phase/Ewens Sampling Formula Phase



$B_1 + B_2 + B_3 = 6$  MIGRATION BLOCKS

## Collecting Phase/Kingman Coalescent Phase



The  $B_i, b_{i,j}$  are i.i.d. over  $i$  and obey Ewens Sampling Formula.

**Lemma (a result of Tavaré and Donnelly)**

For fixed  $j$ ,

$$\left(\frac{b_{i,1}}{n}, \frac{b_{i,2}}{n}, \dots, \frac{b_{i,j}}{n}\right) \rightarrow (X_1, X_2, \dots, X_j) \quad (21)$$

where  $X_j = \text{Beta}(1, \Gamma) \prod_{j=1}^{j-1} (1 - \text{Beta}(1, \Gamma))$  (as defined in proposition).

**Lemma**

$$B_i \approx \log(n) \quad (22)$$

## Number of Migration Blocks that Mutate

**Lemma**

Let  $h$  be the number of migration blocks that mutate.  $B = B_1 + B_2 + \dots + B_d$ . As  $d \rightarrow \infty$ ,  $\frac{hL}{B} \rightarrow \text{Exp}[1]$ . ( $h \approx \text{Exp}[\frac{B}{L}]$ ).

idea of proof:

let  $k_1, k_2, \dots, k_L$  be the number of descendants for blocks 1, 2, ...,  $L$  at level  $L$ . Then,

$$P(k_1, k_2, \dots, k_L) = \left(\frac{B-1}{L-1}\right)^{-1}. \quad (23)$$

So simple combinatorics gives  $P(k_1) = P(h)$ . Then just take limit of  $\frac{k_1 L}{B}$ .

## Putting It All Together

How many  $j$ th migration blocks will be mutants?

- There are  $d$ ,  $j$ th migration blocks.
- There are  $B$  migration blocks total.
- $h$  out of the  $B$  migration blocks are mutants.
- $h \approx V \frac{B}{L}$
- all migration blocks are equally likely to be mutants.
- $E[\text{number of mutant } j\text{th migration blocks}] \approx \left(\frac{d}{B}\right) V \frac{B}{L} = \frac{V}{\lambda}$ .

**Consequence:** Number of mutant  $j$ th migration blocks is  $\text{Poisson}\left[\frac{V}{\lambda}\right]$ .

**Final Fact:** each migration block that mutates comes from a different sampled deme. This makes the mutant migration blocks independent.

## Stepping Stone Models

Stepping stone model - demes are placed on a square subset of  $\mathbb{Z}^2$ . [Zahle, Cox, Durrett 05]

- still have a dispersal and collecting phase
- collecting phase satisfies Kingman coalescent description
- dispersal phase will not satisfy ESF.

Under the stepping stone model, the distribution of  $F_{st}$  in the weak mutation case will be different than the island model case.