Stochastic Demography, Coalescents, and Effective Population Size

Steve Krone

University of Idaho
Department of Mathematics & IBEST

Demography

- Demographic effects (bottlenecks, expansion, fluctuating population size, population structure) affect polymorphism data
- Ex) Detecting selective sweeps confounded by demography and structure
- Effective population size: When is it meaningful? What is effect of demography?
- Appropriate scaling comes from what is observable in the coalescent (i.e., what has an explicit effect on data).

Wright–Fisher model

- discrete time (generations)
- constant population size N
- panmictic
- no selection, no recombination
- ancestry: each individual chooses (haploid) parent at random (prob $1/N$ each) from previous generation

Effective population size

Other population models (reproduction, variable pop size, structure, . . .) sometimes behave in certain respects like a W-F model with an "effective population size" N_e.

- inbreeding effective size (probability of identity by descent)
- variance effective size (variance in reproductive success)
- eigenvalue effective size (leading non-unit eigenvalue for allele frequency transition matrix)
- "coalescent effective size" (if it exists) supersedes all of these. Exists when scaled ancestral process converges to linear time change of Kingman's coalescent; demographic fluctuations "average" out.

The coalescent

- P(2 indiv choose same parent) $= 1/N$
- Takes $O(N)$ generations to find common ancestor (per pair)
- Measure time in units of N generations . . . $[Nt]$
- $A_N(\tau) = \#$ ancestors τ generations in past
- $A_N([Nt]) \Rightarrow A(t)$ . . . Kingman coalescent

All genetic information about a sample (polymorphism data) is embedded in the coalescent.

Fu and Li's F statistic

$$F = F(\pi, \eta_s, S) = \frac{\pi - (\frac{n-1}{n})\eta_s}{\sqrt{c_1 S + c_2 S^2}}$$

where $n =$ sample size

$\pi =$ ave. $\#$ pairwise differences (influenced by deep branches)

$\eta_s = \#$ singletons (influenced by external branches)

$S = \#$ segregating sites

Tajima's D statistic

$$D = D(\pi, \eta_s, S) = \frac{\pi - \frac{S}{a_n}}{\sqrt{c_1' S + c_2' S^2}}$$

where

$a_n = \sum_{i=1}^{n-1} \frac{1}{i}$

Both statistics have mean ≈ 0, variance ≈ 1.

Deviations from assumptions (neutrality, constant pop size, panmixia,...) produce changes in F and D.

Relative time scales

Coalescence events have prob $\sim O(1/N)$.

- Events that are "faster" have prob $\sim O(1/N^\alpha)$, where $0 \leq \alpha < 1$. Effects appear in coalescent only in average sense. (All demographic processes "fast" $\Rightarrow$ coalescent effective size exists.)
- Events with prob $\sim O(1/N)$ are incorporated in the coalescent and affect pattern of variation in nonhomogeneous way. (No coalescent effective size)

Fluctuating population size

(backward) size process $M_N(1), M_N(2), M_N(3), \ldots$

Markov chain with state space $\{N_1, N_2, \ldots\}$

$N_i = N x_i$

How does this affect the coalescent?

Depends on time it takes for "large" size changes (i.e., $O(N)$) to occur.

Fast size fluctuations–linear time change

Large pop size changes occur quickly (e.g., every generation);

size process stationary distribution $(\gamma_1, \gamma_2, \ldots)$;

W-F reproduction:

$$\begin{aligned} & P_2(\text{no coalescence in } [Nt] \text{ generations}) \\ = \; & E\Big[\prod_{\tau=1}^{[Nt]} \big(1 - \frac{1}{M_N(\tau)}\big)\Big] \\ \sim \; & \Big(1 - \sum_i \gamma_i \cdot \frac{1}{N x_i}\Big)^{[Nt]} \to \exp\{-t \sum \gamma_i / x_i\} \end{aligned}$$

Limiting coalescent . . . linear time change of standard coalescent:

$$A_N([Nt]) \Rightarrow A(ct)$$

where $c = \sum \frac{\gamma_i}{x_i}$ . . . pairwise coalescence rate

$\Rightarrow$ pairwise coalescence prob $\approx \frac{1}{N} \sum \frac{\gamma_i}{x_i} \equiv \frac{1}{N_e}$

$\Rightarrow N_e = \frac{N}{c} = \big(\sum \frac{\gamma_i}{N_i}\big)^{-1}$ ... harmonic mean of sizes

This is the "coalescent effective size": $N_e = N/c$

Intermediate fluctuations–stochastic time change

What if macroscopic changes in pop. size (i.e., $O(N)$) occur on coalescent time scale (i.e., $O(N)$ generations)?

Pop. size τ generations in past (Markov chain):

$$M_N(\tau) = N X_N(\tau),$$

where relative size proc. $X_N([Nt]) = \frac{M_N([Nt])}{N} \Rightarrow X(t)$

... cont-time Markov (e.g., diffusion proc. or cont-time jump chain)

Reproduction

Let

$$c_N\big(M_N(\tau-1), M_N(\tau)\big)$$

denote prob. that two lineages coalesce when going from gen. $\tau-1$ to gen. τ (in past). Assume

$$c_N(k,m) = \frac{1}{N} H_N(\frac{k}{N}, \frac{m}{N}),$$

where $H_N(\frac{k}{N}, \frac{m}{N}) \to H(x,y)$ as $k/N \to x$ and $m/N \to y$.

Time change becomes

$$\int_0^t H(X_s, X_s) ds.$$

Ex: Wright–Fisher model

$$c_N\big(M_N(\tau-1), M_N(\tau)\big) = \frac{1}{M_N(\tau)} = \frac{1}{NX_N(\tau)}$$

So $c_N(k,m) = \frac{1}{m} = \frac{1}{N} H_N(\frac{k}{N}, \frac{m}{N})$,

where $H_N(x,y) = \frac{1}{y} = H(x,y)$

Ex: Cannings model

$$c_N\big(M_N(\tau-1), M_N(\tau)\big) = \frac{1}{(M_N(\tau-1))_2} \sum_{i=1}^{M_N(\tau)} E\big[(\nu_i^{(\tau)})_2\big]$$

$\nu_i^{(\tau)}$ . . . number of offspring produced by ith indiv in gen τ. With exchangeable reproduction, get

$$H_N(\frac{k}{N}, \frac{m}{N}) = \Big(\frac{k}{N}\big(\frac{k}{N} - \frac{1}{N}\big)\Big)^{-1} \frac{md}{N} \to \frac{yd}{x^2} \equiv H(x,y)$$

"Large" size changes occur on same time scale as coalescence events; do not "average out." Limiting coalescent is of form

$$A_N([Nt]) \Rightarrow A(Y(t)),$$

where the time change

$$Y(t) \equiv \int_0^t H(X(s), X(s)) ds$$

is nonlinear and stochastic (coalescence intensity). [WF case: $Y(t) = \int_0^t \frac{1}{X(s)} ds$] (Kaj and Krone 2003; Donnelly and Kurtz 1999; Griffiths and Tavaré 1994.)

No (coalescent) effective size! Behavior different from any standard W-F model. Effects should show up in polymorphism data.

Idea for W-F model

$$\begin{aligned} & P_2(\text{no coalescence in } [Nt] \text{ generations} | \{M_N(\cdot)\}) \\ = & \prod_{\tau=1}^{[Nt]} \big(1 - \frac{1}{M_N(\tau)}\big) \\ = & \prod_{\tau=1}^{[Nt]} \big(1 - \frac{1}{NX_N(\tau)}\big) \\ \sim & \exp\big(-\frac{1}{N} \sum_{\tau=1}^{[Nt]} \frac{1}{X_N(\tau)}\big) \Rightarrow \exp\big(-\int_0^t \frac{1}{X(s)} ds\big) \end{aligned}$$

Full convergence theorem

$$(X_N([Nt]), A_N([Nt])) \Rightarrow (X(t), A(Y_t))$$

in $D_{S\times\{1,\ldots,n\}}[0,\infty)$, whenever $X_N(0) \Rightarrow X(0)$ in S.

Transition semigroup $\mathcal{T}_t f(x,i) = E^{(x,i)}\big[f(X(t), A(Y_t))\big]$ can be decomposed as

$$\mathcal{T}_t f(x,i) = \sum_{j=1}^{i} \sum_{\ell=j}^{i} C_\ell(i,j) E^{(x,i)}\big[f(X(t), j)\, e^{-\binom{\ell}{2} Y_t}\big]$$

$$C_\ell(i,j) \equiv \prod_{j+1\le s\le i}\binom{s}{2}\prod_{j\le r\le i, r\neq \ell}\frac{1}{\binom{r}{2}-\binom{\ell}{2}}$$
$$= \frac{(2\ell-1)(-1)^{\ell-j}j_{(\ell-1)}(i)_\ell}{j!(\ell-j)!i_{(\ell)}}, \quad j\le\ell\le i.$$

$$\mathcal{L}f(x,i) = \frac{d}{dt}\mathcal{T}_t f(x,i)\big|_{t=0}$$
$$= Lf(x,i)+\binom{i}{2}H(x,x)\big(f(x,i-1)-f(x,i)\big)$$

Idea of Proof

For $1\le i\le n$, $x\in\mathbb{Z}_N$ and $r\ge 0$, transition operator for (X_N, A_N):

$$\mathcal{T}_r^N f(x,i) = E^{(x,i)}\big[f(X_N(r), A_N(r))\big], \qquad (1)$$

Show uniform convergence of semigroups:

$$\sup_{x,i}|\mathcal{T}^N_{[Nt]}f(x,i)-\mathcal{T}_t f(x,i)|\to 0, \quad N\to\infty$$

For any $t\ge 0$ fixed and $1\le j\le i\le n$, as $N\to\infty$,

$$P^{(k,i)}\big((M_N([Nt]),A_N([Nt]))=(m,j)\big)$$
$$=\sum_{\ell=j}^{i}C_\ell(i,j)\left(P-\binom{\ell}{2}\hat{P}\right)^{[Nt]}(k,m)+\mathcal{O}\left(\frac{1}{N}\right),$$

Combinatorial term is same in discrete semigroup, so decomposition implies enough to show uniform convergence of discrete Feynman–Kac semigroups to

$$E^{(x,i)}\big[f(X(t),j)\,e^{-\binom{\ell}{2}Y_t}\big]$$

Local time interpretation of time change

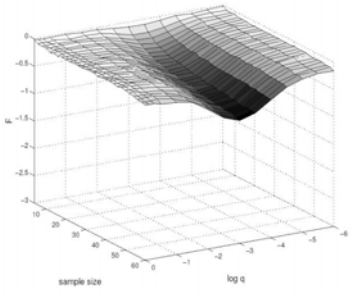
$$\int_0^t\frac{1}{X_s}ds=\int_E\frac{1}{x}\cdot L_t^x\,m(dx)$$

L_t^x . . . diffusion local time

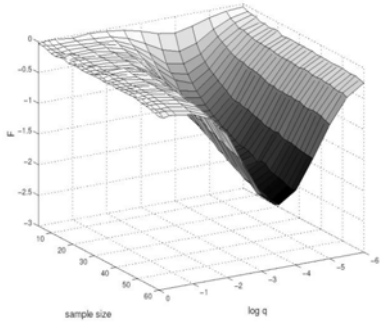
$m(dx)$ . . . speed measure

Simulations for fluctuating size

2 sizes N_1, N_2; equal prob of size change $q_1=q_2\equiv q$; mutation prob $u=.001$; 10,000 runs per data pt.; stationary starting size. Plot of Fu and Li's F



$N_1=10^3, N_2=10^4$

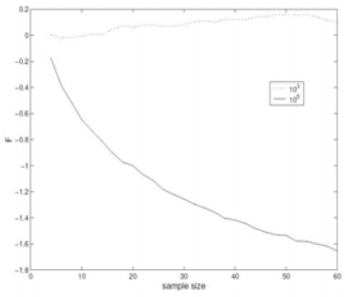


$N_1=10^3, N_2=10^5$

Rule of thumb: $q\in(\frac{10^{-1}}{N_2},\frac{10^1}{N_1})$ $\Rightarrow$ no averaging; too close to coalescent scale.

Dependence on initial size

"Why do polymorphism data always seem to suggest population expansion, and not population contraction?"



$N_1 = 10^3, N_2 = 10^5;\ q_1 = q_2 = 10^{-4}$.

Top curve: initial size 10^3
Bottom curve: initial size 10^5

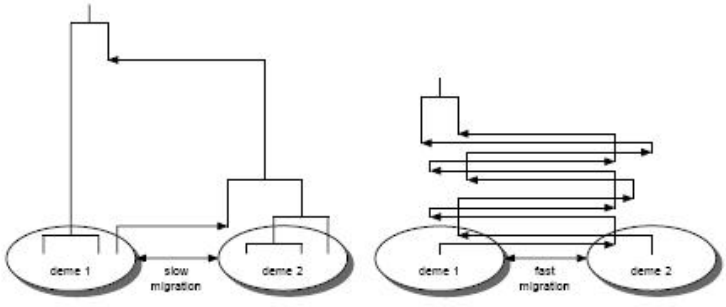
General Population Structure

Many deviations from basic Wright–Fisher model can be thought of as examples of "population structure."

- geographic structure
- age classes
- diploidy (Nordborg and Donnelly 1997)
- males and females

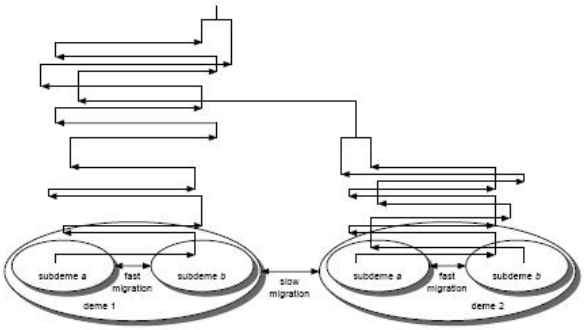
Population divided into "groups" that are connected by "migration."

- Fast migration ... effects only present in averaged sense.
- Slow migration ... effects explicitly appear in coalescent.
- Differences in scaling ... hierarchical structuring of population



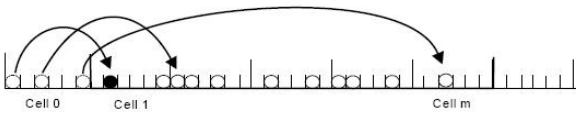
(a) No collapse . . . structured coalescent.

(b) Full collapse to Kingman coalescent.



Partial collapse to structured coalescent.

Rates depend on configuration of ancestral lineages



Ex. "Ancestral urn" for age structure.

While there are r ancestors ($r = 1, \ldots, n$), the configuration process moves among the configurations in *level* r:

$$S_r \equiv \{(x_1, \ldots, x_L) : x_1 + \cdots + x_L = r\}.$$

Starting with sample of size n, the state space for the configuration process is $S = S_1 \cup \cdots \cup S_n$. For any configuration $(x_1, \ldots, x_L) \in S$, specify probabilities of jumping to other configurations due to migration and/or coalescence of ancestors.

"Proof of convergence"

Stationary distribution of backward migration process: $\gamma = (\gamma_1, \ldots, \gamma_L)$.

Stationary distribution for level-r configuration process:

$$\pi_r(x) = \frac{r!}{x_1! \cdots x_L!} \gamma_1^{x_1} \cdots \gamma_L^{x_L}.$$

Transition matrix for whole configuration process on $S = S_1 \cup \cdots \cup S_n$:

$$\Pi_N = I + \frac{1}{N^\alpha} B + \frac{1}{N} C + o\left(\frac{1}{N}\right),$$

where I is the identity matrix, B is a block diagonal matrix

$$B = \begin{bmatrix} B_{11} & 0 & 0 & \cdots & 0 & 0 \\ 0 & B_{22} & 0 & \cdots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdots & \cdot & \cdot \\ & & & \ddots & & \\ \cdot & \cdot & \cdot & & \cdot & \cdot \\ 0 & 0 & 0 & \cdots & B_{n-1,n-1} & 0 \\ 0 & 0 & 0 & \cdots & 0 & B_{n,n} \end{bmatrix}$$

... from backward migration jumps,

C is a block matrix of the form

$$C = \begin{bmatrix} -C_{11} & 0 & 0 & \cdots & 0 & 0 & 0 \\ C_{21} & -C_{22} & 0 & \cdots & 0 & 0 & 0 \\ 0 & \cdot & \cdot & \cdots & \cdot & \cdot & \cdot \\ & & & \ddots & & & \\ \cdot & \cdot & \cdot & & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \cdots & 0 & C_{n,n-1} & -C_{n,n} \end{bmatrix}$$

... due to coalescence jumps.

Möhle's Lemma (1998)

Case $\alpha = 0$:

$$\Pi_N^{[Nt]} = \left(A + \frac{1}{N} C + o\left(\frac{1}{N}\right)\right)^{[Nt]} \to P - I + e^{tG},$$

where $P = \lim_{k \to \infty} A^k$ and $G = PCP$.

Structured Populations

Population of total size N, subdivided into L demes, connected by migration. Pop. size in deme k is $N_k = Na_k$ $(a_1 + \cdots + a_L = 1)$.

- Migration on same time scale as coalescence events (i.e., migration prob. for lineage $b_{ij} = \beta_{ij}/N$)

$\Rightarrow$ limiting coalescent is "structured." (no averaging, no coalescent effective size)

- Fast migration (i.e., $b_{ij} = \beta_{ij}/N^{\alpha}, 0 \leq \alpha < 1$), and stationary distribution for locations $(\gamma_1, \gamma_2, \ldots, \gamma_L)$

$\Rightarrow$ averaging occurs w/ coalescent time change

$$c = \sum_{k=1}^{L} \frac{\gamma_k^2}{a_k}.$$

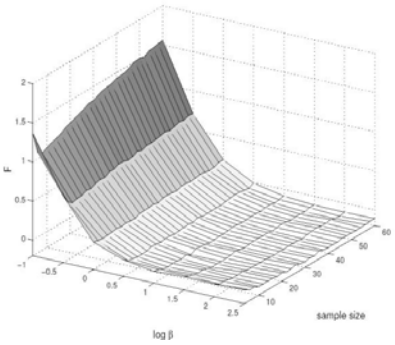
$\Rightarrow$ coalescent effective size is

$$N_e = \frac{N}{c} = \Big(\sum \frac{\gamma_k^2}{N_k}\Big)^{-1} \qquad \text{"harmonic mean"}$$

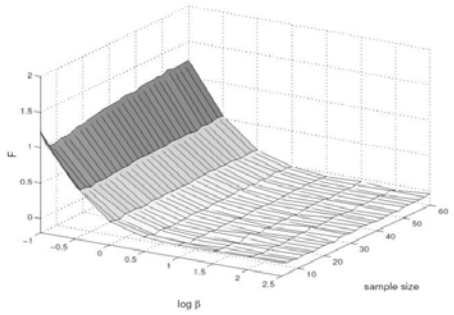
In case of fast migration, structured model can be thought of as panmictic W-F model with pop. size N_e.

Simulations for population subdivision

2 demes, equal size, equal migration rate $\beta = 2Nb$



$N = 10^3$



$N = 10^4$

I. Kaj, Uppsala Univ. (Mathematics)

M. Nordborg, USC (Molecular and Computational Biology)

M. Lascoux, Uppsala Univ. (Cons. Biology & Genetics)

P. Sjödin, Uppsala Univ. (Cons. Biology & Genetics)

NSF DMS-00-72198

NIH P20 RR16448

- P. Sjödin, I. Kaj, S. Krone, M. Lascoux, M. Nordborg (2005) On the meaning and existence of an effective population size. *Genetics* **169**: 1061-1070.
- I. Kaj and S. Krone (2003) The coalescent process in a population with stochastically varying size. *J. Appl. Probab.* **40**: 33-48.
- M. Nordborg and S. Krone (2002) Separation of time scales and convergence to the coalescent in structured populations. In *Modern Developments in Theoretical Population Genetics*. M. Slatkin and M. Veuille (eds.).