Statistics of genetic variation in populations of variable size

B. Mehlig¹⁾

A. Eriksson^{1,2)}, M. Rafajlovic¹⁾, S. Sagitov³⁾, E. Schaper¹⁾ & A. Rimark¹⁾

¹⁾Department of Physics, University of Gothenburg, Sweden
 ²⁾Department of Zoology, University of Cambridge, UK
 ³⁾Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, Sweden

Alle Latarian

Financial support by Vetenskapsrådet, the "Centre for Theoretical Biology" at the University of Gothenburg, and the Bank of Sweden Tercenary Foundation is gratefully acknowledged.

Littorina saxatilis

Multi-locus genotypes of a mother and her progeny, *Littorina saxatilis* (microsatellite loci, step-wise mutation model).

How is demographic history and geographic structure reflected in the data?

Answer requires model for ancestral trees.

Note: high levels of multiple paternity.

Table 1 L scattills. Multi-locus genotypes (five loci) of a mother (denoted by 0) and n = 42 progeny from a clutch (Boström et al. 2009, unpublished). The original data contains more progeny than presented here

No.	L1	L2	1.3	L4	1.5
0	(151, 192)	(227, 242)	(225, 231)	(217, 223)	(199, 202)
1	(151, 204)	(236, 242)	(22.2, 231)	(217, 223)	(199, 202)
2	(151, 211)	(227, 242)	(210, 231)	(217, 223)	(199, 223)
3	(151, 195)	(227, 236)	(210, 225)	(217, 217)	(196, 202)
4	(151, 181)	(236, 242)	(213, 225)	(223, 223)	(199, 199)
5	(151, 195)	(227, 230)	(210, 225)	(217, 217)	(196, 202)
6	(151, 192)	(227, 236)	(22.2, 225)	(217, 217)	(184, 199)
7	(192, 208)	(227, 236)	(219, 231)	(217, 223)	(193, 199)
8	(151, 173)	(227, 227)	(213, 231)	(217, 232)	(193, 199)
9	(151, 208)	(236, 242)	(219, 231)	(217, 223)	(199, 199)
10	(151, 151)	(227, 230)	(216, 231)	(217, 223)	(184, 199)
11	(192, 208)	(227, 239)	(225, 225)	(184, 223)	(199, 223)
12	(173, 192)	(227, 236)	(213, 231)	(223, 226)	(193, 199)
13	(151, 201)	(236, 242)	(216, 231)	(217, 223)	(199, 202)
14	(151,201)	(236, 242)	(213, 231)	(184, 223)	(202, 223)
15	(169, 192)	(227, 227)	(219, 231)	(217, 220)	(202, 205)
16	(192,201)	(236, 242)	(219, 225)	(22.3, 22.3)	(199, 199)
17	(151, 192)	(236, 242)	(213, 225)	(223, 223)	(193, 199)
18	(192,201)	(227, 236)	(219, 231)	(217, 235)	(202, 202)
19	(151, 173)	(227, 242)	(225, 231)	(217, 226)	(193, 199)
20	(173, 192)	(236, 242)	(213, 231)	(184, 223)	(202, 202)
21	(182, 192)	(227, 236)	(213, 225)	(217, 223)	(190, 202)
22	(151,204)	(236, 242)	(22.2, 225)	(184, 217)	(202, 202)
23	(192, 204)	(227, 236)	(22.2, 225)	(217, 223)	(199, 199)
24	(151, 181)	(227, 236)	(219, 231)	(217, 217)	(199, 202)
25	(192,201)	(227, 236)	(213, 231)	(223, 223)	(199, 199)
26	(151, 208)	(236, 242)	(225, 231)	(217, 226)	(199, 202)
27	(151, 195)	(236, 242)	(210, 225)	(214, 217)	(184, 199)
28	(151, 208)	(239, 242)	(225, 231)	(184, 223)	(199, 199)
29	(151,201)	(227, 242)	(210, 231)	(217, 223)	(202, 223)
30	(192, 217)	(227, 236)	(216, 231)	(214, 223)	(199, 202)
31	(151, 195)	(236, 242)	(210, 225)	(217, 223)	(184, 202)
32	(192, 206)	(236, 242)	(219, 225)	(169, 217)	(196, 202)
33	(151, 208)	(236, 242)	(219, 231)	(217, 223)	(199, 205)
34	(151, 204)	(227, 239)	(210, 225)	(217, 217)	(196, 199)
35	(192, 194)	(230, 242)	(219, 225)	(217, 217)	(202, 202)
36	(192, 192)	(236, 242)	(22.2, 231)	(217, 223)	(199, 199)
37	(151, 194)	(227, 230)	(210, 225)	(202, 217)	(202, 202)
38	(151,206)	(227, 230)	(219, 225)	(217, 217)	(202, 205)
39	(151, 168)	(227, 236)	(22.2, 225)	(217, 220)	(199, 205)
40	(192, 195)	(227, 236)	(216, 225)	(217, 223)	(196, 202)
41	(192, 192)	(227, 236)	(213, 231)	(184, 217)	(199, 202)
42	(192, 217)	(236, 242)	(22.2, 225)	(220, 223)	(199, 199)

Eriksson, Mehlig, Panova, André & Johannesson (2009) Panova, Boström, Hofving, Areskoug, Eriksson, Mehlig, Mäkinen, André & Johannesson (2009)

Humans

Single-nucleotide polymorphisms (SNPs). Empirically observed patterns of genetic variation in selectively neutral DNA segments.

ACTTTCGGAA ACTTTCGCAA ACTGTCGGAA ACTGTCGCAA position along chromosome

Patterns of mutation (`genetic mosaic').

How do these patterns reflect the history of the population? Determine role of mutation, recombination, migration, population-size changes, selection,...

First: single-locus properties. Then: recombination.

Wright-Fisher model

Model for genealogy of selectively neutral loci.

- (a) discrete non-overlapping generations
- (b) constant population size N
- (c) freely mixing population
- (d) Mendelian inheritance (sampling of chromosomes with replacement)



Mutation model: assume that mutations occur randomly on ancestral lines with rate μ per base-pair per generation (•). Wright (1931), F

Wright (1931), Fisher (1930)

Consider sample genealogies (red).

Next: coalescent process. Model for sample genealogies of chromosomes in the limit of large $\,N$.

Coalescent process

Kingman (1982)

Wright-Fisher model. Probability P_2 that 2 individuals have distinct parents in previous generation: $P_2 = 1 - N^{-1}$. Similarly: $P_3 = (1 - N^{-1})(1 - 2N^{-1})$. In general:

14/14/14/14/14

$$P_n = \prod_{j=1}^{n-1} (1 - j/N)$$

Probability that a coalescent event (pairwise merger) occurs after t generations

$$(1-P_n)P_n^t \approx \frac{\binom{n}{2}}{N} e^{-\binom{n}{2}t/N}$$
 in the limit $N \to \infty$.
Used that $\log P_n \approx -\frac{1}{N}\binom{n}{2}$ provided $n \ll N$.

The time to the first coalescent event is exponentially distributed.

Sample genealogies of many microscopic population models are consistent with those of the coalescent for $n \ll N$ (in the limit of $N \to \infty$).

n = 20

Coalescent process

Random process for generating sample genealogies. Describes Wright-Fisher model in the limit of large N.



Kingman (1982)

Sample genealogies of many microscopic population models are consistent with those of the coalescent for $n \ll N$ (in the limit of $N \to \infty$).

Population-size fluctuations

Variable population size N(t). Expected time between coalescent events $\langle \tau_j \rangle = N/{j \choose 2}$. Coalescent proceeds faster when N(t) is small.



N

xN

N

Effective population size

Variable population size N(t). Genealogies nevertheless described by constant-N coalescent but with `effective population size'

$$N_{\text{eff}} = \left(\lim_{T \to \infty} \frac{1}{T} \int_0^T \frac{\mathrm{d}t}{N(t)}\right)^{-1}$$
?

Kaj & Krone (2003) Nordborg & Krone (2003) Jagers & Sagitov (2004)

Harmonic average: effect of catastrophic events (bottlenecks $x \ll 1$).

Whether or not the effective population-size approximation works depends on frequency ω of the size fluctuations.

Sjödin, Kaj, Krone, Lascoux & Nordborg (2005)

Other definitions of effective population size.

Ewens (1982) Sjödin, Kaj, Krone, Lascoux & Nordborg (2005) Wakeley & Sargsyan (2009)

Distribution of total branch length

Population subject to time-variable carrying capacity K(t)(time-changing evironment). Choose $K(t) = K_0 \sin(\omega t)$.



Result for moments of T_n

Result for moments of total branch length in sample of size n:

$$\langle T_n^k \rangle = k! \sum_{m_1=2}^n \sum_{m_2=2}^{m_1} \cdots \sum_{m_k=2}^{m_{k-1}} m_1 \cdots m_k$$
$$\int_0^\infty dt_1 f_{nm_1}(0, t_1) \cdots \int_{t_{k-1}}^\infty dt_k f_{m_{k-1}m_k}(t_{k-1}, t_k) \,.$$

Eriksson, Mehlig, Rafajlovic & Sagitov (2010)

where $f_{nm}(t_1, t_2) = g_{nm}(\Lambda(t_2) - \Lambda(t_1))$ is the probability that n ancestral lines at time t_1 coalesce to m lines at time t_2 (further in the past) in population of fluctuation size N(t).

Population-size intensity' function $\Lambda(t) = N_0 \int_0^t ds N^{-1}(s)$ and

 $g_{nm}(t_2 - t_1) = \operatorname{Prob}(\ell(t_2) = m | \ell(t_1) = n)$ in a population of constant size.

Tavaré (1984), Griffiths & Tavaré (1994), Slatkin (1996)

Example I

Population subject to time-changing carrying capacity $K(t) = K_0 \sin(\omega t)$.



Example 2

Population size randomly changing between two values, N and xN (for 0 < x < 1) with rates λ and λ_x . Sjödin, Kaj, Krone, Lascoux & Nordborg (2005)

Piecewise constant random process

x(t) = N(t)/N.



Compute moments $\overline{\langle T_n^k \rangle}$ by averaging $\langle T_n^k \rangle$ over x(t).

By way of example we have done this for k = 1, 2. For k = 1:

$$\overline{\langle T_n \rangle} = \sum_{j=1}^n d_{n;j} \frac{\lambda + \lambda_x + b_j/x}{b_j(\lambda/x + \lambda_x + b_j/x)}$$

for $b_j = j(j-1)/2$ and $d_{n;j} = (2j-1)(1+(-1)^j)\frac{\binom{2n-1}{n-j}}{\binom{2n-1}{n}}$

In the limit $\lambda = \lambda_x \to \infty$ recover effective population-size approximation.

Evolution punctuated by bottlenecks

Consider neutral evolution punctuated by severe bottlenecks



Example of Ξ - coalescent (simultaneous multiple mergers)

as opposed to Kingman coalescent:





Two-locus statistics (recombination)

Sample of n = 2 individuals (*i* and *j*), L = 2 loci (*x* and *y*). In a recombination, part of a gamete is inherited from one parent and the remaining part from the other parent. Scaled recombination rate R = 2Nr where *r* is recombination rate per generation per chromosome between loci in question.



Recombination causes decorrelation of times $\tau_{x(ij)}$ and $\tau_{y(ij)}$ to most recent common ancestor. Linkage equilibrium

Covariance of gene histories

Sample of n = 2 individuals (*i* and *j*), L = 2 loci (*x* and *y*). Recombination rate R = 2Nr.

In a population of constant size

 $\operatorname{cov}[\tau_{x(ij)}, \tau_{y(ij)}] = \frac{R+18}{R^2+13R+18}$ Griffiths (1981), Hudson (1983), Hudson, in Oxford Surveys of Evolutionary Biology (1990) Measure of linkage disequilibrium. Other measures \hat{r}^2 , σ_d^2 . McVean, Genetics 162 (2002) 987

For humans the covariance has been estimated from SNP data.

Jakobsson et al., Nature 451 (2008) 998

Influence of demographic history?



Jakobsson et al., Nature 451 (2008) 998

00

Fluctuating population size

Covariance of gene histories in population of fluctuating size?

Effective population-size approximation. Replace N by $N_{
m eff}$

$$N_{\text{eff}} = \left(\lim_{T \to \infty} \frac{1}{T} \int_0^T \frac{\mathrm{d}t}{N(t)}\right)^{-1}$$

and R by $R_{\text{eff}} = 2N_{\text{eff}}r$. Can this work? Time-scale separation?

Investigate covariance of gene-histories in bottleneck model.



Parameters:
$$R$$
, λ , λ_x , and x .

Time scales:
$$au = \lambda^{-1}$$
 and $au_x = \lambda_x^{-1}$

`Multiple mergers'

Weak versus strong bottlenecks



Results

Covariance of gene-history correlations



severe bottlenecks

Comparison to empirical data

Human demography: repeated founder events, correspond to repeated bottlenecks.

Expect that American LD curve (recent bottlenecks) should decay slower than African LD curve.



Jakobsson et al., Nature 451 (2008) 998

 \mathcal{X}

Multiple paternity in L. saxatilis

Multiple paternity in L. saxatilis.

Determine minimum number of sires of a batch of children from multiple matings of a single mother. migration m, colonisation c



mainland islands (carrying capacity <u>K</u>)

Female	Observed no of offspring	Analysed no of offspring	Most likely no of sires	Minimum no of sires
F2	87	77	23	21
F6	71	71	16	15
F8	69	53	15	12
F9	117	79	23	20

Four females and their offspring were genotyped at five microsatellite DNA loci. The most likely number was estimated using the likelihood-based software COLONY and the minimum number was calculated using MINSIRES.

doi:10.1371/journal.pone.0009640.t001

Table 1. The number of sires



2

Panova, Boström, Hofving, Areskoug, Eriksson, Mehlig, Mäkinen, André & Johannesson (2010)

Repeated founder events in L. saxatilis

Multiple paternity in L. saxatilis.

Study effect of multiple paternity on genetic variation. Compute population heterozygosity H_2 as a function of distance x from mainland.



Colonisation of empty islands

migration m, colonisation c



Steady state with migration

Rimark & Rafajlovic (2010)

Waves of genetic variation

Study time-dependence of population heterozygosity H_2 in steady state.



Conclusions

Single-locus sample genealogies in populations of fluctuating size

- how quickly (slowly) are the `effective population-size' limits are attained when $\omega \to 0$ and $\omega \to \infty$?
- allow to compute statistics of genetic variation (moments of S_n) in populations of changing size

Linkage disequilibrium in populations of fluctuating size

- Effective population-size approximation fails at large values of R
- Effect of repeated founder effects (sequence of bottlenecks) upon decay of linkage disequilibrium as a function of R.

Repeated founder effects in fragmented populations of L. saxatilis

- effect of multiple paternity?
- waves of population heterozygosity?
- ancestral recombination graph in fragmented populations?