# Statistics of genetic variation in populations of variable size <br> B. Mehlig ${ }^{1)}$ <br> A. Eriksson ${ }^{\text {1,2) }}$, M. Rafajlovic ${ }^{1}$, S. Sagitov ${ }^{3)}$, E. Schaper ${ }^{1)}$ \& A. Rimark ${ }^{\text {1) }}$ <br> ${ }^{1)}$ Department of Physics, University of Gothenburg, Sweden <br> ${ }^{2}$ ) Department of Zoology, University of Cambridge, UK <br> ${ }^{3}$ )Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, Sweden 

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.

Eriksson, Mehlig, Panova, André \& Johannesson (2009)
 (berobed by 09 and $n=42$ progeny from a clanch (Bostrìn it a
 then prowerted here

| No. | 1.1 | 1.2 | 13 | 1.4 | 15 |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 0 | (151, 192) | $(227,242)$ | (225, 231) | (217, 223) | (199, 202) |
| 1 | (151, 204) | (236, 242) | (222, 231) | (217, 223) | (199, 202) |
| 2 | (151, 211) | $(227,242)$ | (210, 231) | (217, 2231 | (199, 2231 |
| 3 | (151, 196) | $(227,236)$ | (210. 235) | (217, 217) | (196. 202) |
| 4 | (151, 181) | (236,242) | (213.25) | (223, 2231 | (199, 199) |
| 5 | (151, 196) | $(227,230)$ | (210. 25) | (217, 217) | (196, 202) |
| 6 | (151, 192) | $(227,236)$ | (22, 235) | (217, 217) | (184. 199) |
| 7 | (192, 208) | $(227,236)$ | (219, 231) | (217, 2231 | (198, 199) |
| 8 | $(151,173)$ | $(227,227)$ | (213, 231) | (217, 232) | (19, 199) |
| 9 | ( 151,208 ) | (236,242) | (219, 231) | (217, 223) | (199, 199) |
| 10 | (151, 151) | $(227,230)$ | (216, 231) | (217, 2231 | (184. 199) |
| 11 | (192, 208) | $(227,239)$ | (225.235) | (184, 2231 | (199, 2231 |
| 12 | $(173,192)$ | $(227,236)$ | (213.231) | (233, 226) | (19, 199) |
| 13 | (151, 201) | $(236,242)$ | (216, 231) | (217, 2231 | (199, 202) |
| 14 | (151, 201) | $(236,242)$ | (213, 231) | (184, 2231 | (202, 223) |
| 15 | (109, 192) | $(227,227)$ | (219, 231) | (217, 2201) | (202, 205) |
| 16 | (192, 201) | $(236,242)$ | (219, 235) | (223, 2231 | (199, 199) |
| 17 | (151, 192) | (236, 242) | (213, 255) | (223, 2231 | (199, 199) |
| 18 | (192, 201) | $(227,236)$ | (219, 231) | (217, 235) | (202, 202) |
| 19 | $(151,173)$ | $(227,242)$ | (225. 231) | (217, 226) | (19, 199) |
| 2) | (173,192) | $(236,24)$ | (213.231) | (184, 2231 | (202, 202) |
| 21 | (182, 192) | $(227,236)$ | (213,25) | (217, 2331 | (190, 202) |
| 22 | (151, 204) | $(236,242)$ | (222, 235) | (184, 217) | (202, 202) |
| 23 | $(192,204)$ | $(227,236)$ | (222, 25) | (217, 2231 | (199, 199) |
| 24 | (151, 181) | $(227,236)$ | (219, 231) | (217, 217) | (199, 202) |
| 25 | (192, 201) | $(227,236)$ | (213, 231) | (223, 2231 | (199, 199) |
| 26 | (151, 208) | (236,242) | (235. 231) | (217, 226 ) | (199, 202) |
| 2 | (151, 16) | (236,24) | (210. 225) | (214, 217) | (184. 199) |
| 28 | (151, 208) | (239,24) | (225. 231) | (184, 2231 | (199, 199) |
| 29 | (151, 201) | $(227,242)$ | (210, 231) | (217, 2231 | (202, 2231 |
| 30 | (192, 217) | $(227,236)$ | (216, 231) | (214, 2231 | (199, 202) |
| 31 | (151, 196) | $(236,242)$ | (210, 235) | (217, 223) | (184. 202) |
| 32 | (192, 206) | (236,24) | (219, 25) | (169, 217) | (196, 202) |
| 33 | (151, 208) | (236,24) | (219, 231) | (217, 2231 | (199, 205) |
| 34 | (151, 204) | $(227,239)$ | (210. 235) | (217, 217) | (196. 199) |
| 3 | (192, 194) | (230.24) | (219, 235) | (217, 217) | (202, 202) |
| 36 | $(192,192)$ | $(236,242)$ | (22, 231) | (217, 2231 | (199, 199) |
| 37 | (151, 194) | $(227,230)$ | (210. 235) | (202, 217) | (202, 202) |
| 38 | (151, 206) | $(227,230)$ | (219, 235) | (217, 217) | (202, 205) |
| 39 | (151, 168) | $(227,236)$ | (222, 25) | (217, 220) | (199, 205) |
| 4) | (192, 16) | $(227,236)$ | (216, 225) | (217, 2231 | (196, 202) |
| 41 | $(192,192)$ | $(227,236)$ | (213, 231) | (184, 217) | (199, 202) |
| 42 | (192, 217) | (236,242) | (222, 23) | (220, 2231 | (199, 199) | 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 |
| - |  |
| $\underline{\square}$ | ACTGTCGCAA |
|  | position along ch |

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 $\mu$ per base-pair per generation ( $\odot$ ).

Wright (193I), Fisher (1930)
Consider sample genealogies (red).
Next: coalescent process. Model for sample genealogies of chromosomes in the limit of large $N$.

## Coalescent process

Wright-Fisher model. Probability $P_{2}$ that 2 individuals have distinct parents in previous generation: $P_{2}=1-N^{-1}$. Similarly: $P_{3}=\left(1-N^{-1}\right)\left(1-2 N^{-1}\right)$. In general:

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

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

$$
\left(1-P_{n}\right) P_{n}^{t} \approx \frac{\binom{n}{2}}{N} \mathrm{e}^{-\binom{n}{2} t / N} \quad \text { in the limit } \quad N \rightarrow \infty .
$$

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 $N \rightarrow \infty$ ).

## Coalescent process

Kingman (1982)

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


Sample size $n$. Independent random times $\tau_{n}, \tau_{n-1}, \ldots, \tau_{2}$ between coalescent events.
Distributed according to

$$
P\left(\tau_{j}\right)=\frac{\binom{j}{2}}{N} \mathrm{e}^{-\frac{\binom{j}{2}}{N} \tau_{j}} \text { so that }\left\langle\tau_{j}\right\rangle=N /\binom{j}{2}
$$

$$
n=5
$$

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

## Population-size fluctuations

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


Variable population size
$N(t)=N_{0} \sin (\omega t)$

## Effective population size

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

$$
N_{\mathrm{eff}}=\left(\lim _{T \rightarrow \infty} \frac{1}{T} \int_{0}^{T} \frac{\mathrm{~d} t}{N(t)}\right)^{-1} \begin{aligned}
& \text { Kaj \& Krone (2003) }
\end{aligned} \begin{aligned}
& N \\
& \begin{array}{l}
\text { Nordborge \& Krone (2003) } \\
\text { Jagers \& Sagitov (2004) }
\end{array}
\end{aligned}
$$

Harmonic average: effect of catastrophic events (bottlenecks $x \ll 1$ ).
Whether or not the effective population-size approximation works depends on frequency $\omega$ 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 $K(t)=K_{0} \sin (\omega t)$.

Distribution $\rho\left(T_{n}\right)$ of genealogical branch length $T_{n}=\sum_{j=2}^{n} j \tau_{j}$.


In a population of changing size, the times
$\tau_{j}$ are correlated. Zivkovic \& Wiehe (2008)

## Result for moments of $T_{n}$

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

$$
\begin{aligned}
\left\langle T_{n}^{k}\right\rangle= & k!\sum_{m_{1}=2}^{n} \sum_{m_{2}=2}^{m_{1}} \cdots \sum_{m_{k}=2}^{m_{k-1}} m_{1} \cdots m_{k} \\
& \int_{0}^{\infty} \mathrm{d} t_{1} f_{n m_{1}}\left(0, t_{1}\right) \cdots \int_{t_{k-1}}^{\infty} \mathrm{d} t_{k} f_{m_{k-1} m_{k}}\left(t_{k-1}, t_{k}\right) .
\end{aligned}
$$

where $f_{n m}\left(t_{1}, t_{2}\right)=g_{n m}\left(\Lambda\left(t_{2}\right)-\Lambda\left(t_{1}\right)\right)$ 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 $\Lambda(t)=N_{0} \int_{0}^{t} \mathrm{~d} s N^{-1}(s)$ and $g_{n m}\left(t_{2}-t_{1}\right)=\operatorname{Prob}\left(\ell\left(t_{2}\right)=m \mid \ell\left(t_{1}\right)=n\right)$ in a population of constant size.

## Example I

Population subject to time-changing carrying capacity $K(t)=K_{0} \sin (\omega t)$.

carrying capacity $K(t)$





$$
\begin{aligned}
\operatorname{var}\left(T_{n}\right) & =\left\langle T_{n}^{2}\right\rangle-\left\langle T_{n}\right\rangle^{2} \\
\operatorname{skew}\left(T_{n}\right) & =\frac{\left\langle\left(T_{n}-\left\langle T_{n}\right\rangle\right)^{3}\right\rangle}{\operatorname{var}^{3 / 2}\left(T_{n}\right)}
\end{aligned}
$$

Large frequencies $\omega$ : rigid shift of $\rho\left(T_{n}\right)$.

$$
\operatorname{kurt}\left(T_{n}\right)=\frac{\left\langle\left(T_{n}-\left\langle T_{n}\right\rangle\right)^{4}\right\rangle}{\operatorname{var}^{2}\left(T_{n}\right)}
$$

## Example 2

Population size randomly changing between two values, $N$ and $x N$ (for $0<x<1$ ) with rates $\lambda$ and $\lambda_{x}$.

Piecewise constant random process

$$
x(t)=N(t) / N
$$




Compute moments $\overline{\left\langle T_{n}^{k}\right\rangle}$ by averaging $\left\langle T_{n}^{k}\right\rangle$ over $x(t)$.
By way of example we have done this for $k=1,2$. For $k=1$ :

$$
\overline{\left\langle T_{n}\right\rangle}=\sum_{j=1}^{n} d_{n ; j} \frac{\lambda+\lambda_{x}+b_{j} / x}{b_{j}\left(\lambda / x+\lambda_{x}+b_{j} / x\right)}
$$

for $b_{j}=j(j-1) / 2$ and $d_{n ; j}=(2 j-1)\left(1+(-1)^{j}\right) \frac{\binom{2 n-1}{n-j}}{\binom{2 n-1}{n}}$.
In the limit $\lambda=\lambda_{x} \rightarrow \infty$ recover effective population-size approximation.

## Evolution punctuated by bottlenecks

Consider neutral evolution punctuated by severe bottlenecks



Example of $\Xi$-coalescent (simultaneous multiple mergers)
as opposed to Kingman coalescent:


Sagitov, Rafajlovic, Eriksson \& Mehlig (2010)

## 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 $R=2 \mathrm{Nr}$ where $r$ is recombination rate per generation per chromosome between loci in question.


Recombination causes decorrelation of times $\tau_{x(i j)}$ and $\tau_{y(i j)}$ 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 $R=2 N r$.

In a population of constant size

$$
\operatorname{cov}\left[\tau_{x(i j)}, \tau_{y(i j)}\right]=\frac{R+18}{R^{2}+13 R+18}
$$

Griffiths (1981), Hudson (1983),
Hudson, in Oxford Surveys of Evolutionary Biology (1990)
Measure of linkage disequilibrium.
Other measures $\hat{r}^{2}, \sigma_{\mathrm{d}}^{2}$.
McVean, Genetics 162 (2002) 987
0.6

Jakobsson et al., Nature 45 I (2008) 998

For humans the covariance has been estimated from SNP data.

Jakobsson et al., Nature 45 I (2008) 998

Influence of demographic history?

## Fluctuating population size

Covariance of gene histories in population of fluctuating size?
Effective population-size approximation. Replace $N$ by $N_{\text {eff }}$

$$
N_{\mathrm{eff}}=\left(\lim _{T \rightarrow \infty} \frac{1}{T} \int_{0}^{T} \frac{\mathrm{~d} t}{N(t)}\right)^{-1}
$$

and $R$ by $R_{\text {eff }}=2 N_{\text {eff }} r$. Can this work? Time-scale separation?
Investigate covariance of gene-histories in bottleneck model.


Parameters: $R, \lambda, \lambda_{x}$, and $x$.
Time scales: $\tau=\lambda^{-1}$ and $\tau_{x}=\lambda_{x}^{-1}$.

## 'Multiple mergers'

Weak versus strong bottlenecks


Period of low population-size

- Ancestral line

8 Chromosome

- Locus $a$
- MRCA of locus $a$
- Locus $b$
- MRCA of locus $b$

Material not ancestral to loci $a$ and $b$

## 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 45 I (2008) 998



## 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.

mainland islands (carrying capacity $K$ )

Table 1. The number of sires

|  |  |  |  |  |
| :--- | :--- | :--- | :--- | :--- |
| 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 | 73 | 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

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 $c$

$x$
mainland islands (carrying capacity $K$ )


Steady state with migration
Rimark \& Rafajlovic (2010)

## Waves of genetic variation

Study time-dependence of population heterozygosity $H_{2}$ in steady state.

$x$
mainland
islands (carrying capacity $K$ )

high variation
low variation

## Conclusions

Single-locus sample genealogies in populations of fluctuating size

- how quickly (slowly) are the 'effective population-size' limits are attained when $\omega \rightarrow 0$ and $\omega \rightarrow \infty$ ?
- 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?

