# Historic effective population size

# Effective Population size

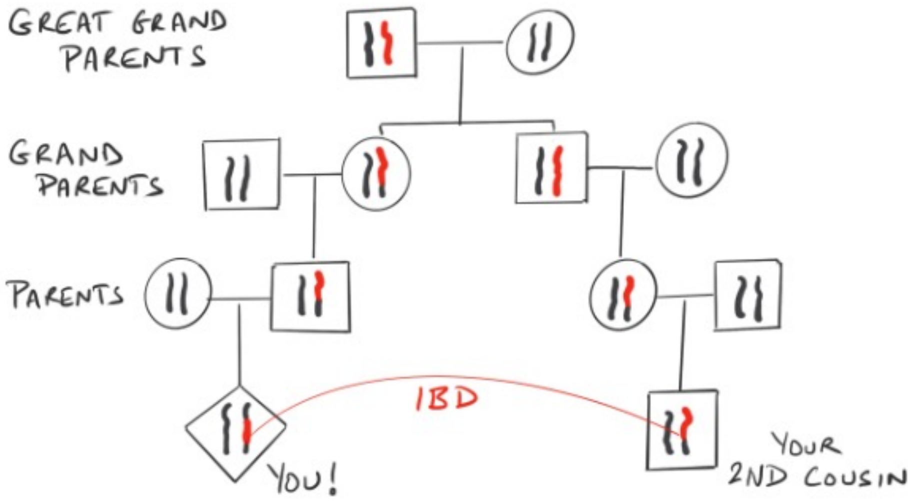
- Effective population size:
  - the size of a population that shows the same amount of dispersion of allele frequencies under drift as an ideal population
- Effective population size is important for the conservation of endangered species
  - indicates the effective number of breeders in a population
  - Often lower than census population size Ne=0.2*Nc
  - If Ne is low there is a high chance of inbreeding, reducing Ne further (-> extinction vortex)
- Genetics allow us to estimate Ne with a single sample for current population sizes (NeEstimator)
- Genomics allow us to estimate Ne with a single sample for historic population sizes (<- but how good are they)

# Methods

- MSMC, PMSC
  - Need whole genome sequences for each individual
- Stairways2, Epos, Snep (using the site frequency spectrum)
  - Using SNPs only
  - Based on the Coalescent approach
- Gone, LinkNe
  - Using linkage disequilibrium
  - Using SNPs and a linkage map (needs a good reference genome)

# Coalescent (Kingsman 1980)

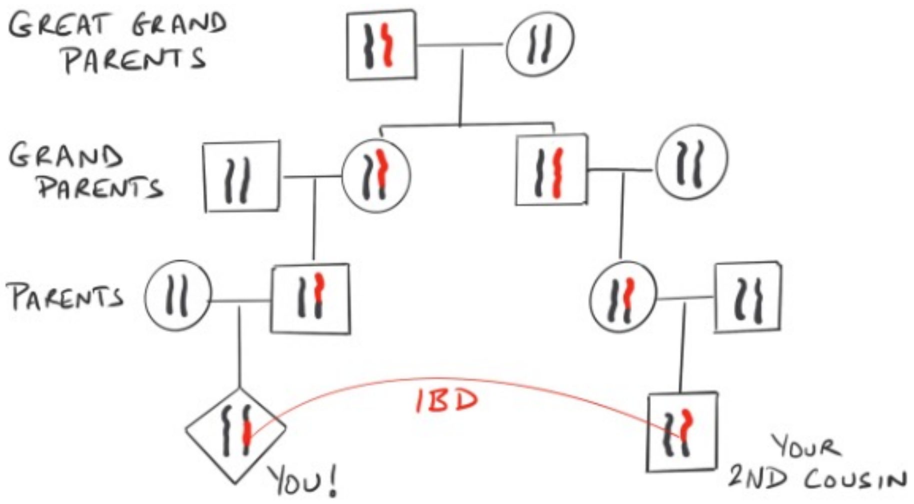
- IBD (identical by descent) [not isolation by distance!!!]
- Inheritance of a chromosome from a great parent
- For the two cousins the overlapping segment is said to have coalesced in their great-grandfather. (IBD)



(from Pritchards new book)
[https://web.stanford.edu/group/pritchardlab/HGbook/]

# Coalescent (Kingsman 1980)

- IBD (identical by descent) [not isolation by distance!!!]
- Inheritance of a chromosome from a great parent
- For the two cousins the overlapping segment is said to have coalesced in their great-grandfather. (IBD)

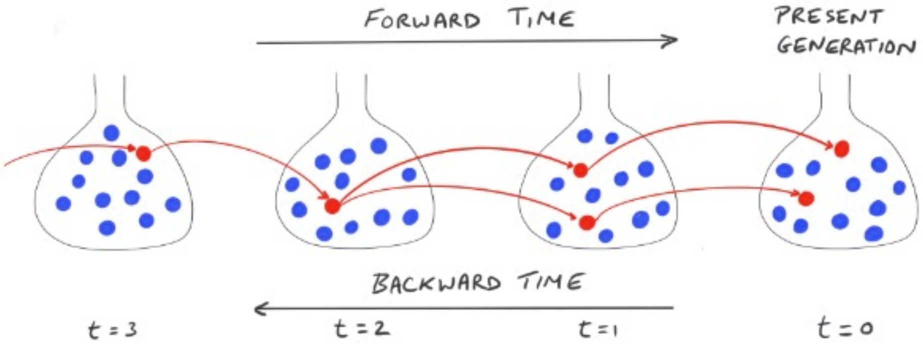


(from Pritchards new book)
[https://web.stanford.edu/group/pritchardlab/HGbook/]

# Coalescent (Kingsman 1980)

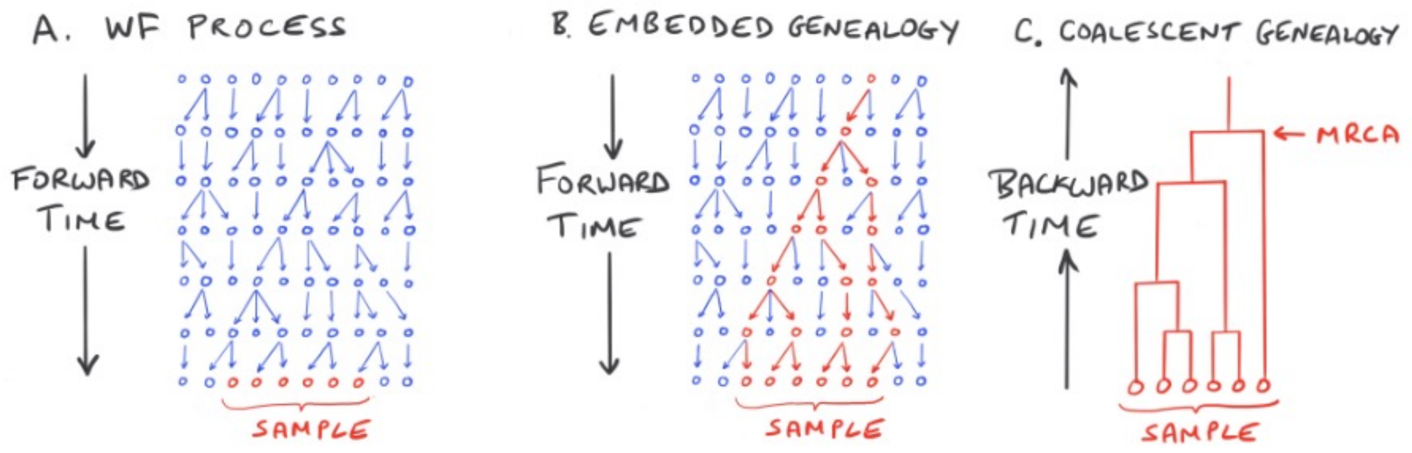
## Backwards in time

- Two copies of a locus in the present generation are marked by red balls.
- These descend from a common ancestor (i.e. they coalesce) two generations ago. In coalescent models it is most natural to measure time backward from the present.



(from Pritchards new book)
[https://web.stanford.edu/group/pritchardlab/HGbook/]

# Coalescent (Kingsman 1980)



WF genealogy for a small population.

six chromosomes sampled at the present day, in red.

Red circles and arrows indicate the ancestors of the sampled chromosomes

The coalescent genealogy abstracts away all irrelevant details of the WF process, showing only the ancestral relationships of the 6 samples and the coalescent times.

(from Pritchards new book) [https://web.stanford.edu/group/pritchardlab/HGbook/]

# Coalescent (Kingsman 1980)

Lots of calculation can be done (probability to coalesce in t generation) $(1-\frac{1}{2N})^t.$

Most recent common ancestor is: 4Ne

**We now add mutations in the mix**

# Coalescent (Kingsman 1980)

*Annual Review of Ecology, Evolution, and Systematics*

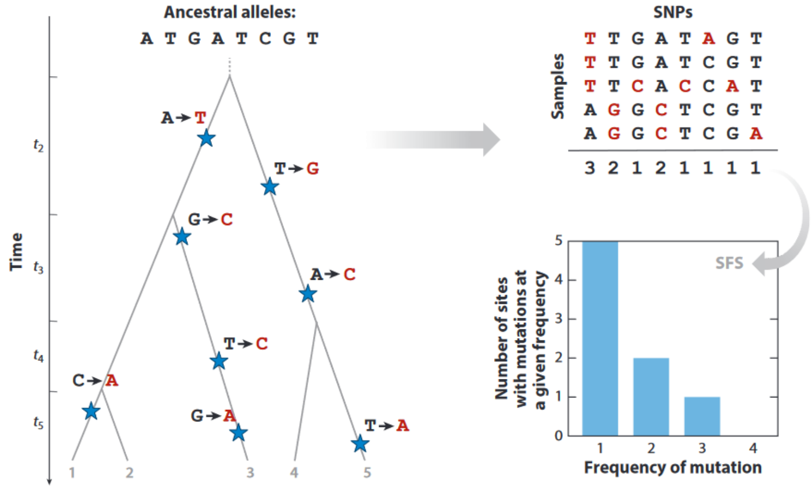
Using Genomic Data to Infer Historic Population Dynamics of Nonmodel Organisms

Annabel C. Beichman,[1] Emilia Huerta-Sanchez,[2,3] and Kirk E. Lohmueller[1,4]

Lots of calculation can be done (probability to coalesce in t generation) $(1 - \frac{1}{2N})^t.$

Most recent common ancestor is: 4Ne

**We now add mutations in the mix SFS**



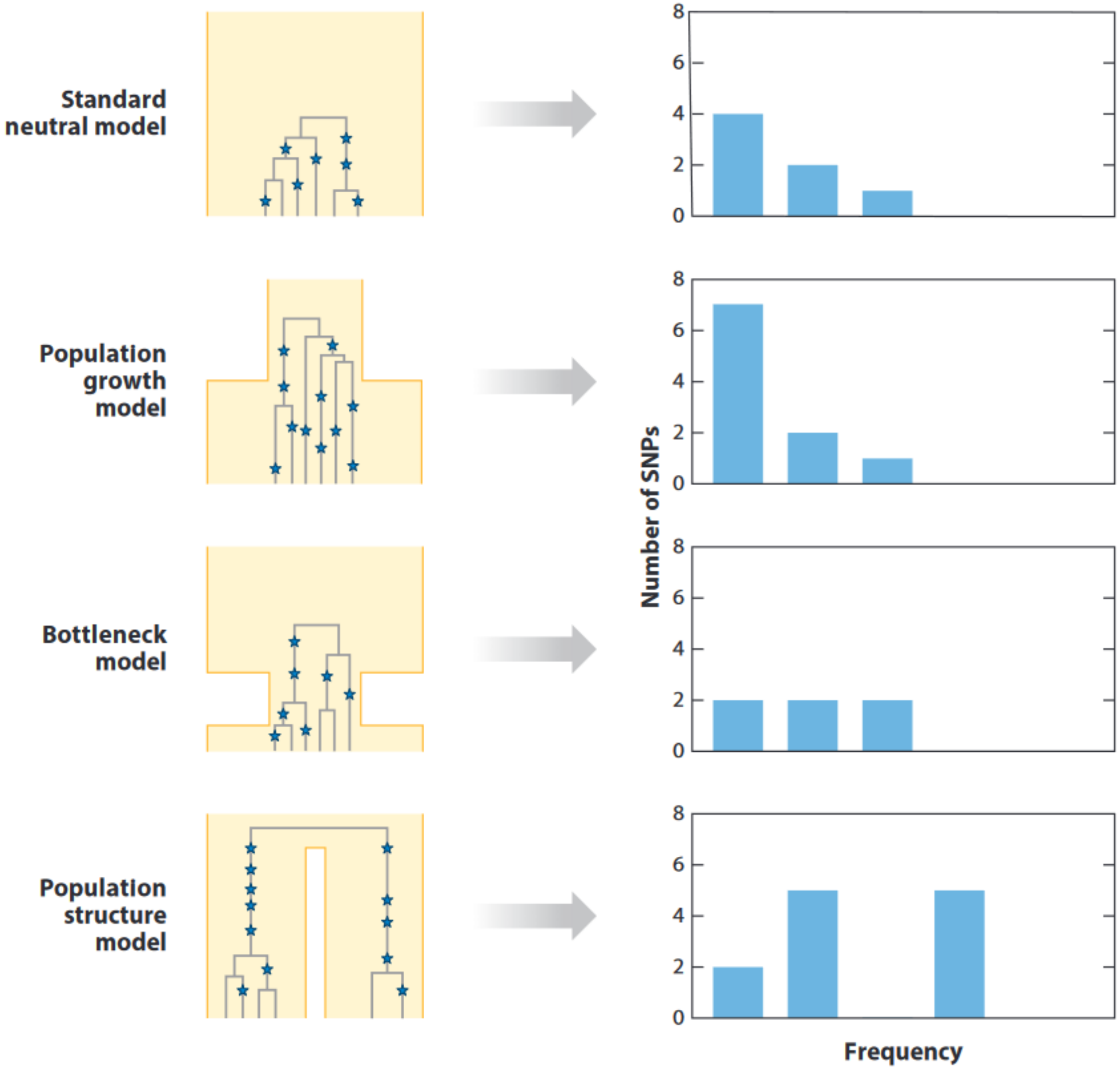
| | | | | | | | |
|---|---|---|---|---|---|---|---|
| T | T | G | A | T | A | G | T |
| T | T | G | A | T | C | G | T |
| T | T | C | A | C | C | A | T |
| A | G | G | C | T | C | G | T |
| A | G | G | C | T | C | G | A |
| 3 | 2 | 1 | 2 | 1 | 1 | 1 | 1 |

**Figure 1**

The locations of mutations on the coalescent genealogy (*left*) give rise to patterns of genetic variation data (*top right*). The SFS depicts the mutational patterns seen in the genetic variation data. Abbreviations: SFS, site frequency spectrum; SNPs, single nucleotide polymorphisms.

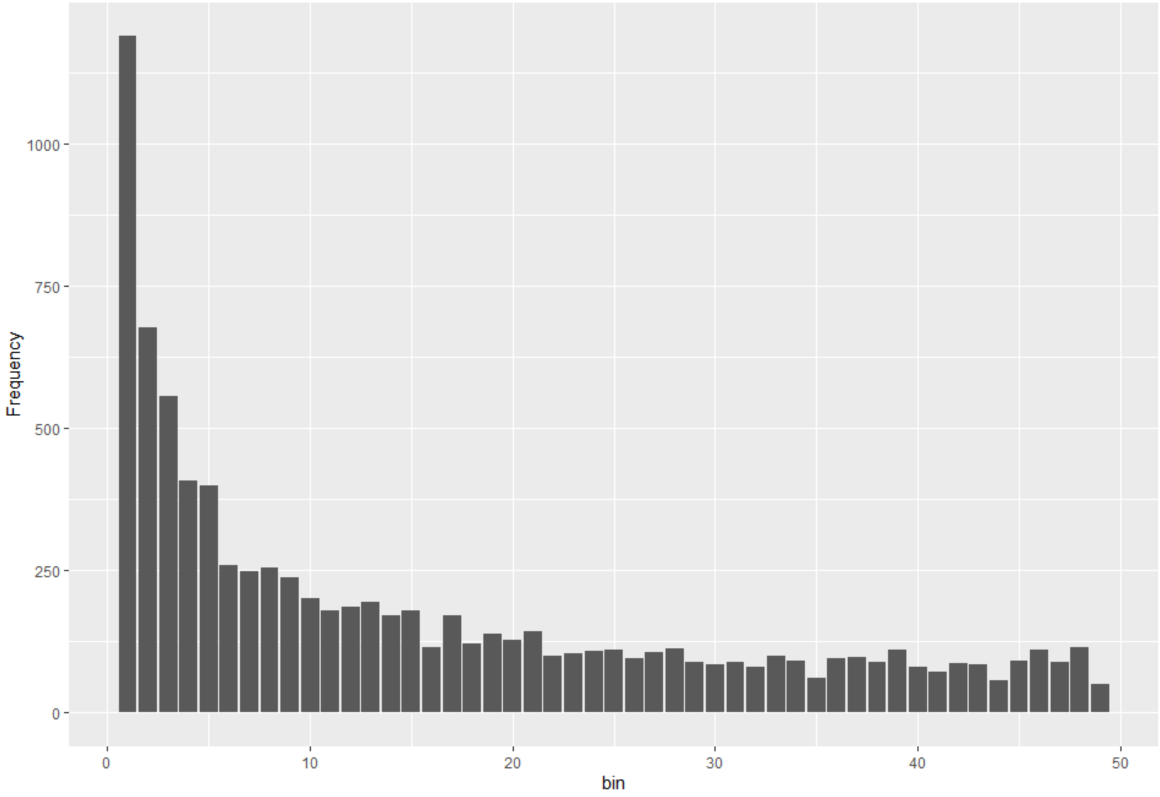
# Site frequency spectrum

- Population history influences the shape of genealogies and the SFS.
- The yellow shaded areas on the left denote the history of each population.
- These demographic histories give rise to the genealogies shown within each model. Blue stars denote mutations that occur on the genealogies.
- The histograms denote the SFS for each model that is generated from the mutational pattern that occurred on the genealogy.



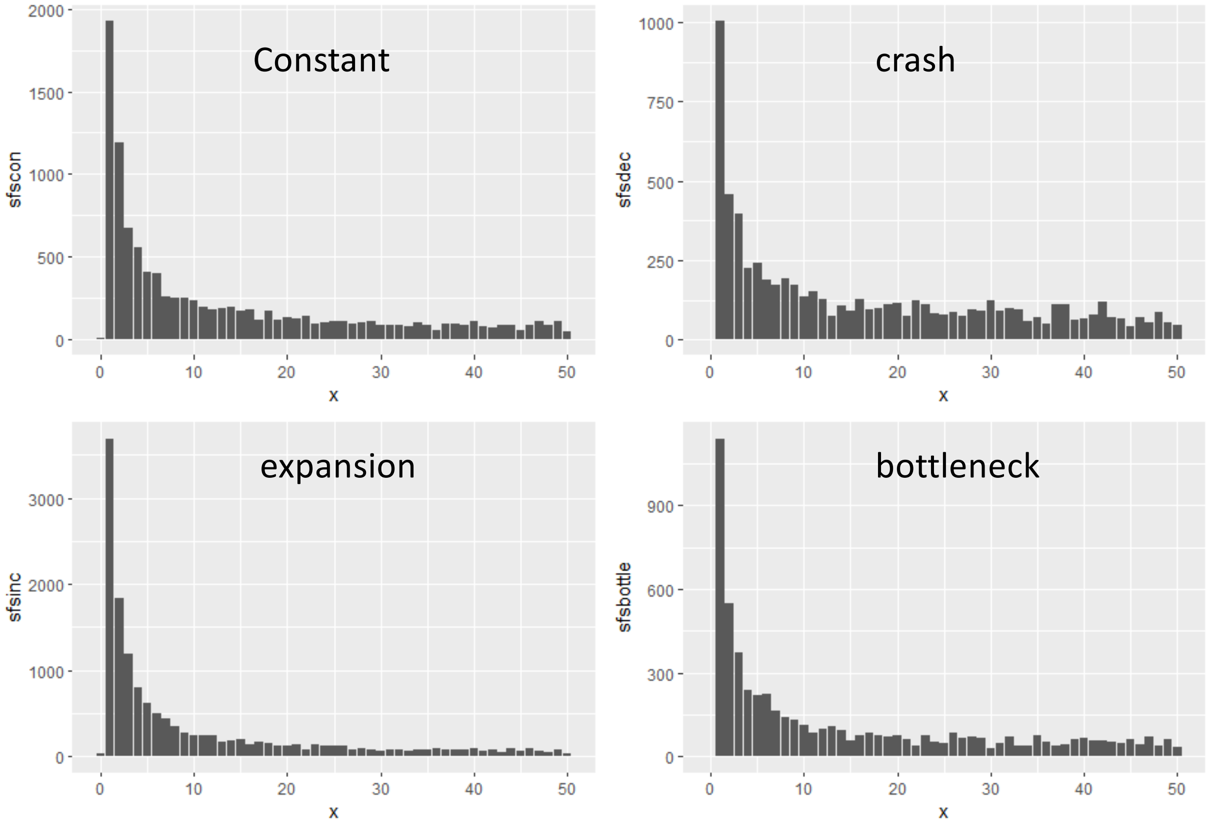
# Site frequency spectrum

- A more realistic SFS
- (Ne=100, constant last 400 years)



# Site frequency spectrum

- SFS for different population trajectories



# Epos and Stairways

- Aim to find the historic population trajectories by fitting the SFS
- Basically all possible SFS are created and compared to the actual one

- Methods that simulate certain histories (Fastsimcoal, dadi) [testing certain hypothesis]
- Methods that do not need a proposal history: Epos/Stairways[/Gone]

# Input for Epos/Stairways2

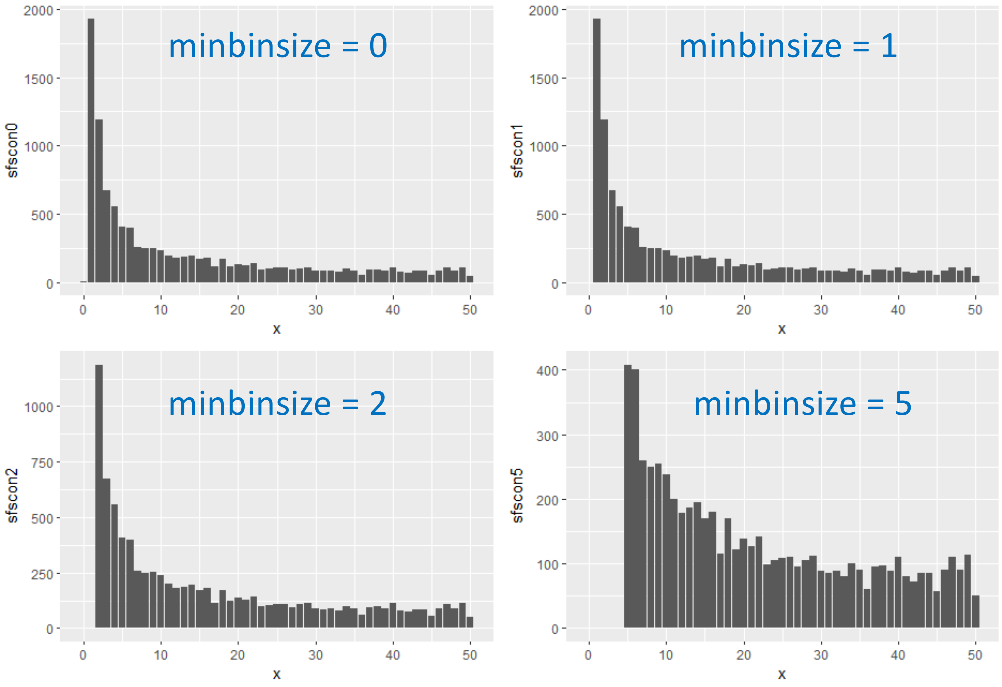
- SNPs (genlight object)
- L, mu and SFS (minbinsize)
- gl.sfs creates a SFS based on dartR/genlight object

- `> gl.sfs( gl, minbinsize = 1, folded=TRUE, singlepop=TRUE)`

- Exercise 1: Parameter in gl.sfs
- Exercise 2: Compare sfs for different data sets
  (glcon, glinc, gldec, glbottle)

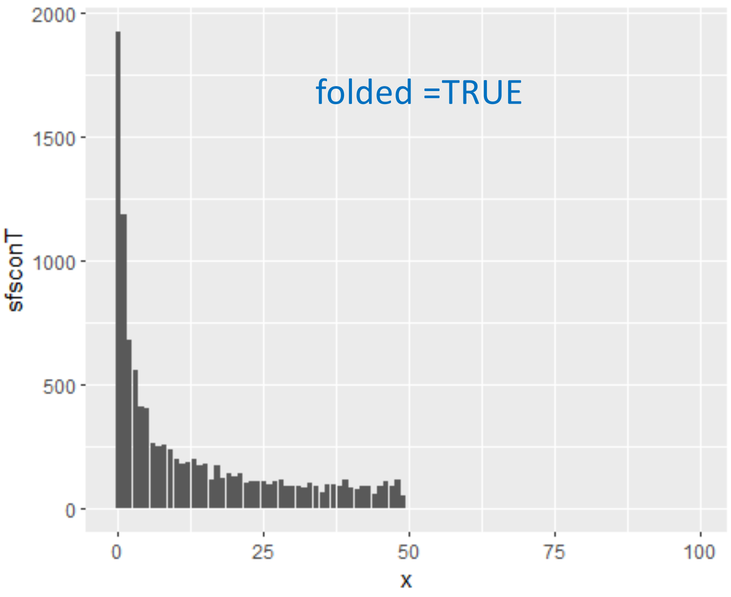
# Input for Epos/Stairways2

- > gl.sfs( gl, minbinsize = X, folded=TRUE, singlepop=TRUE)

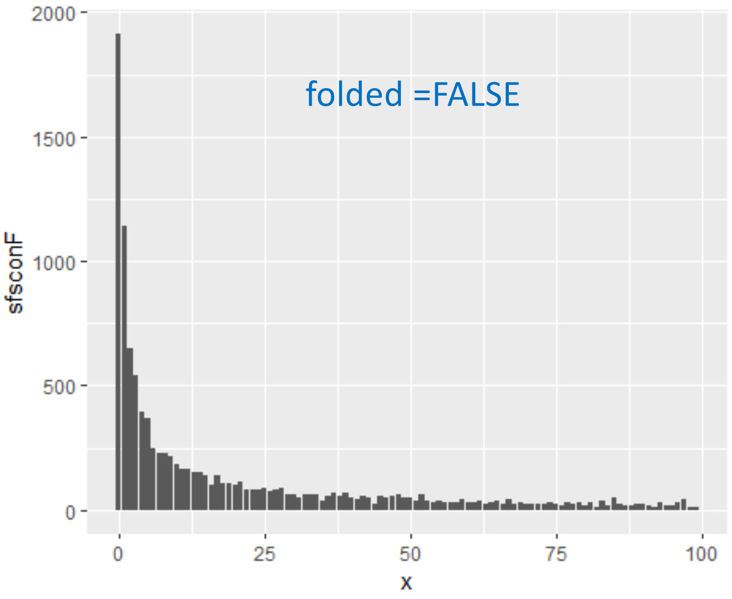


# Input for Epos/Stairways2

- > gl.sfs( gl, minbinsize = 1, folded=TRUE/FALSE)



folded =TRUE



folded =FALSE

# Input for Epos/Stairways2

- L: the length of all combined sequences (genome length)
- mu: mutation rate

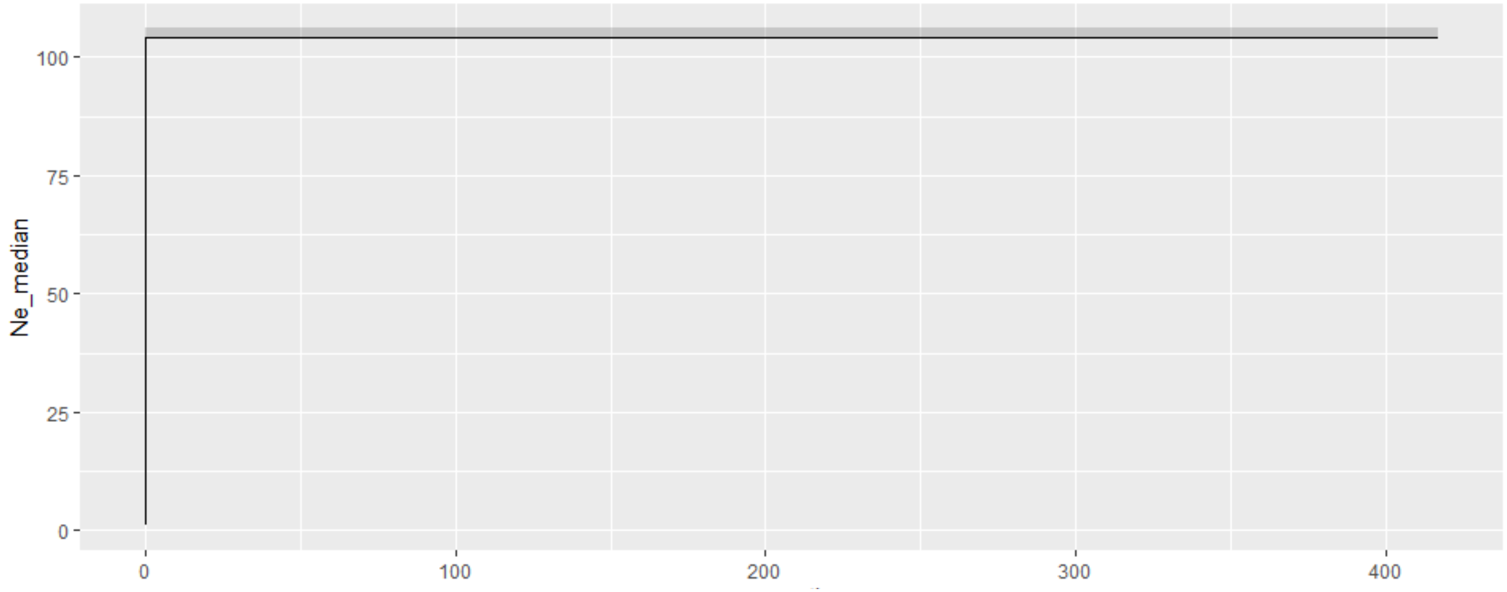
$$x = mu \times L \times 2 \times Ne$$

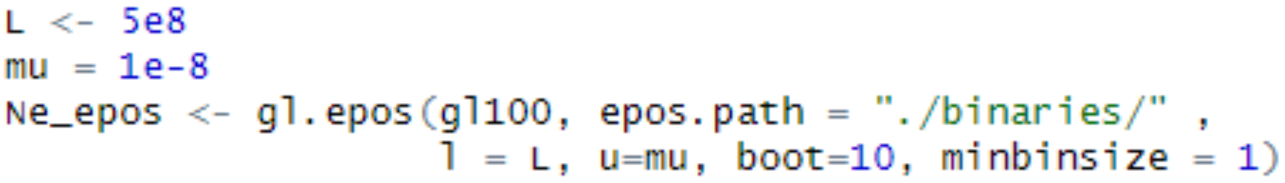
Principle: x: number of mutations in a generation

So the “trajectory” is calibrated for this product

- Exercise 3: Run gl.epos for different settings of L and mu

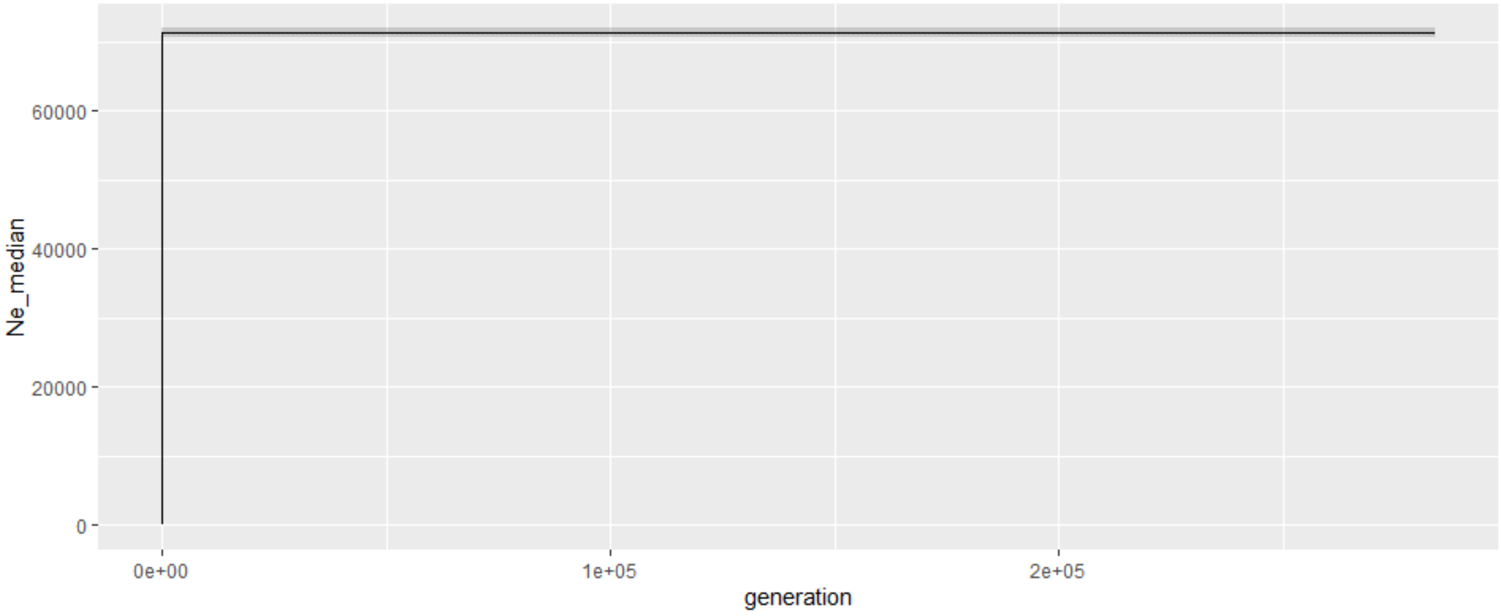
# Epos

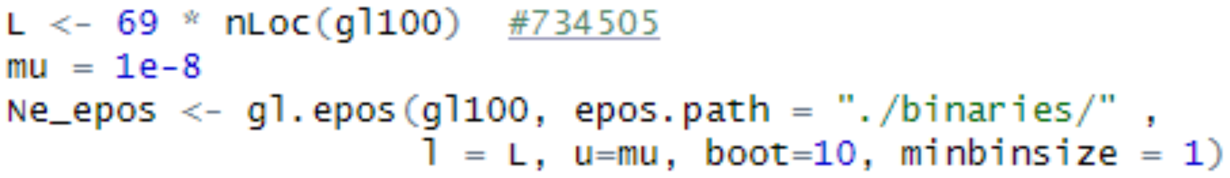




```
L <- 5e8
mu = 1e-8
Ne_epos <- gl.epos(gl100, epos.path = "./binaries/" ,
                   l = L, u=mu, boot=10, minbinsize = 1)
```

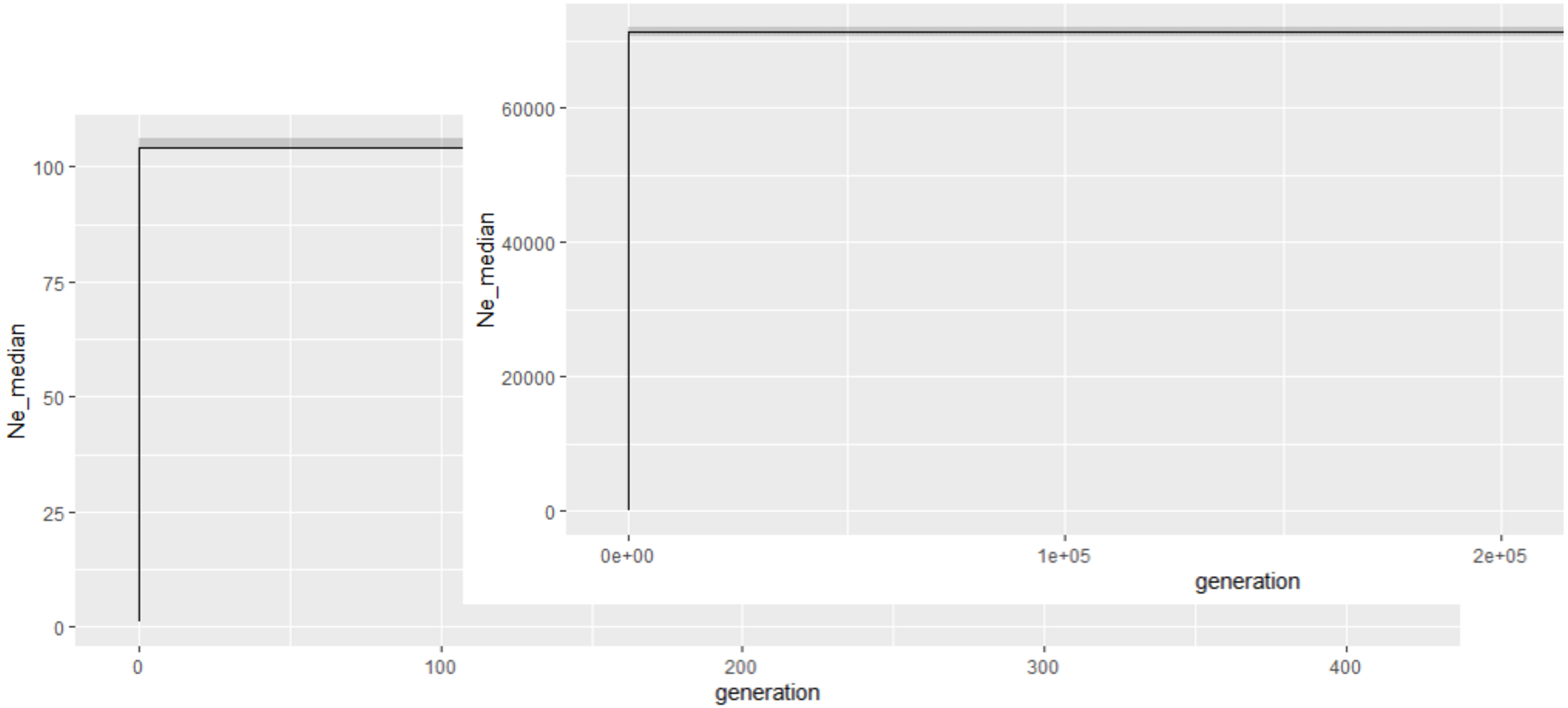
# Epos





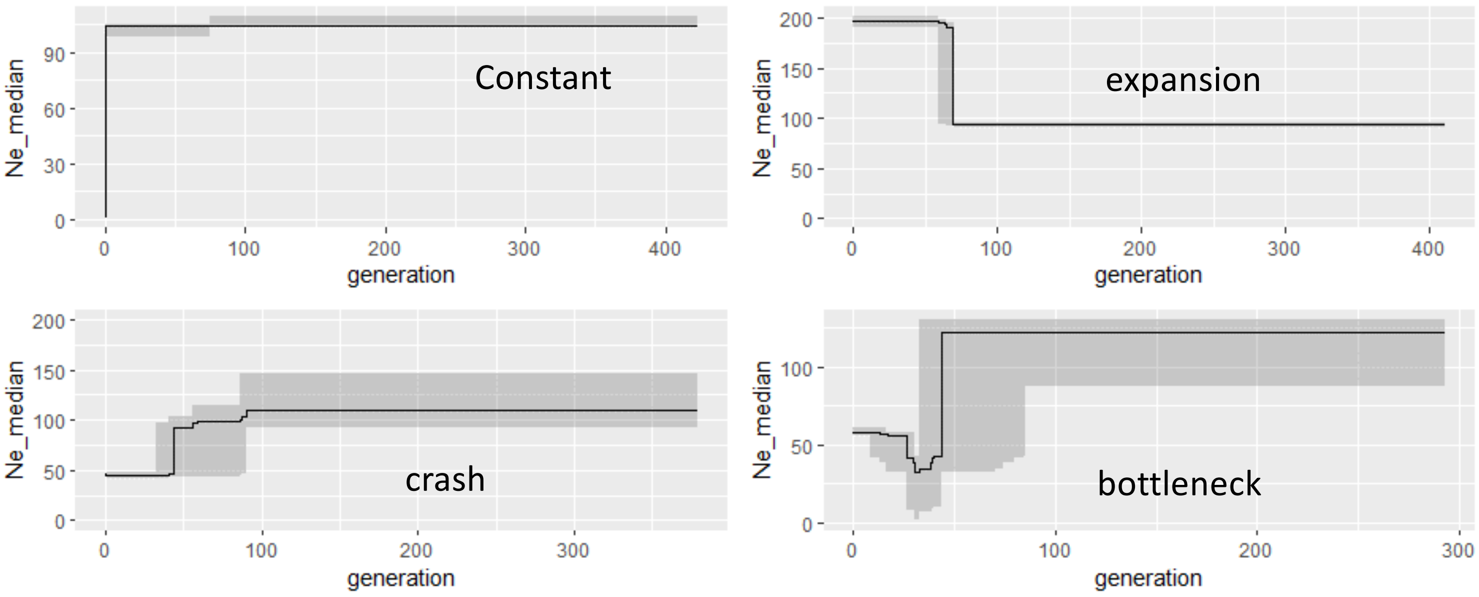
```
L <- 69 * nLoc(gl100)  #734505
mu = 1e-8
Ne_epos <- gl.epos(gl100, epos.path = "./binaries/" ,
                   l = L, u=mu, boot=10, minbinsize = 1)
```

# Epos


Ne_median
100
75
50
25
0
0
100
200
300
400
generation
Ne_median
60000
40000
20000
0
0e+00
1e+05
2e+05
generation

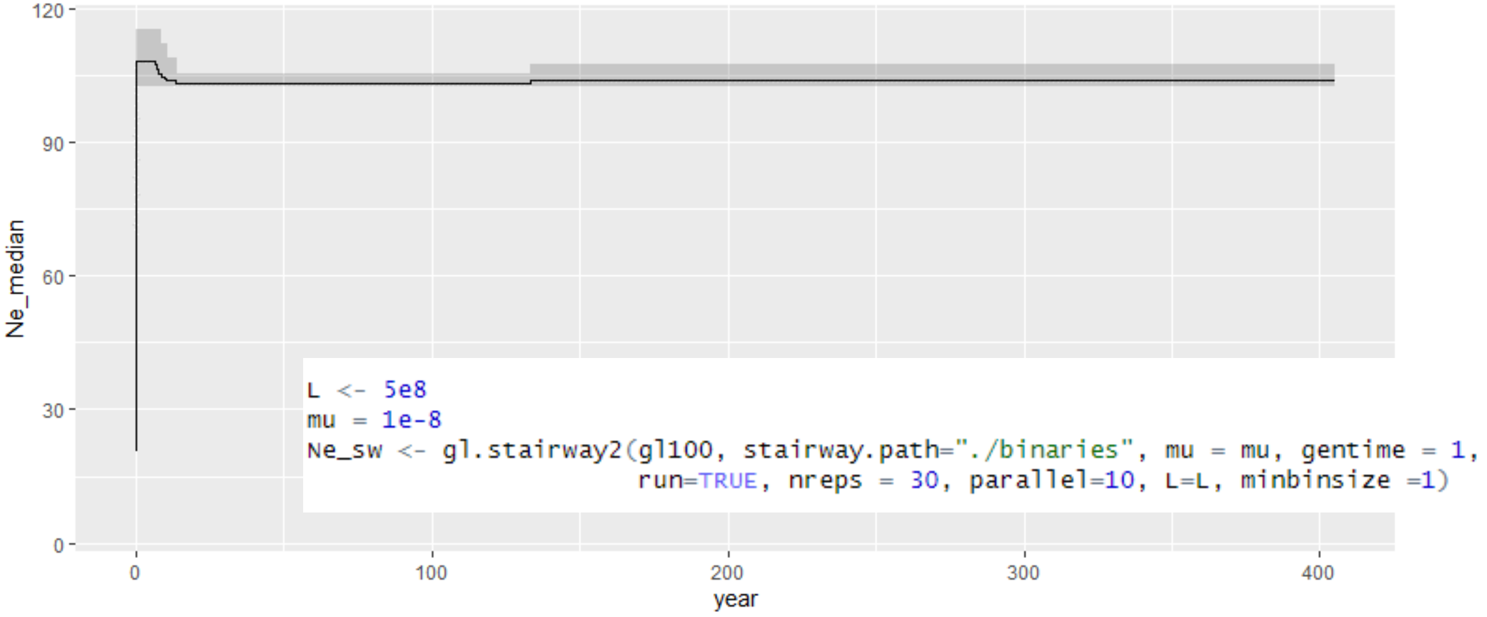
# Epos

- Exercise 4: Run Epos for all four data sets



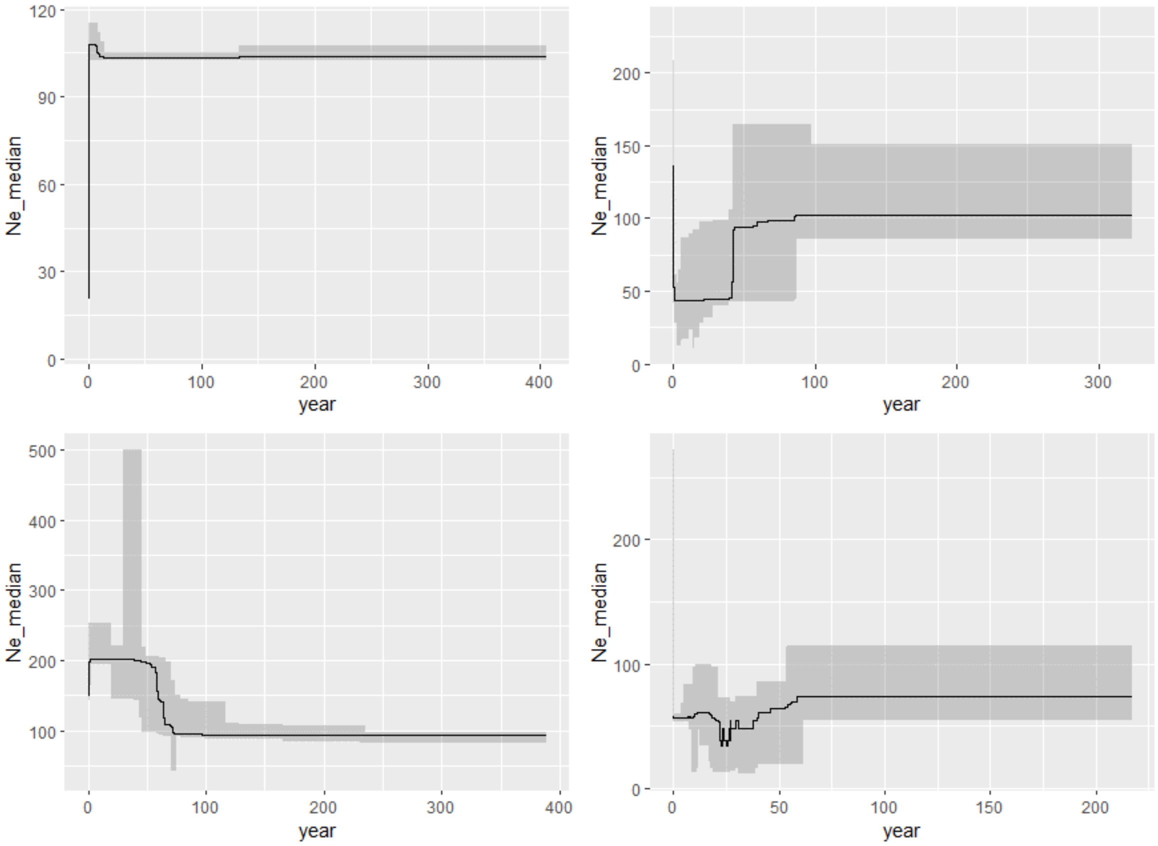
# Stairways2

- Stairways2 (has the same input), but takes >50 times longer or so you can repeat the same exercise for all four scenarios.



```
L <- 5e8
mu = 1e-8
Ne_sw <- gl.stairway2(gl100, stairway.path="./binaries", mu = mu, gentime = 1,
                      run=TRUE, nreps = 30, parallel=10, L=L, minbinsize =1)
```

# Stairways [all four scenarios]



# Isobel Walcott (Honours) Demographic Inference in a Conservation Context

- Standard cases – decline, expansion, stable
  - 586 scenario combinations, 30 replicates
  - 6 levels of loci subsampling for each
  - 3 methods tested
  - 316,440 runs (using NCI infrastructure)
- Bottleneck cases – based on known decline-recovery cases
  - Saltwater crocodiles, southern right whales, fur seals, Fleay's barred frog

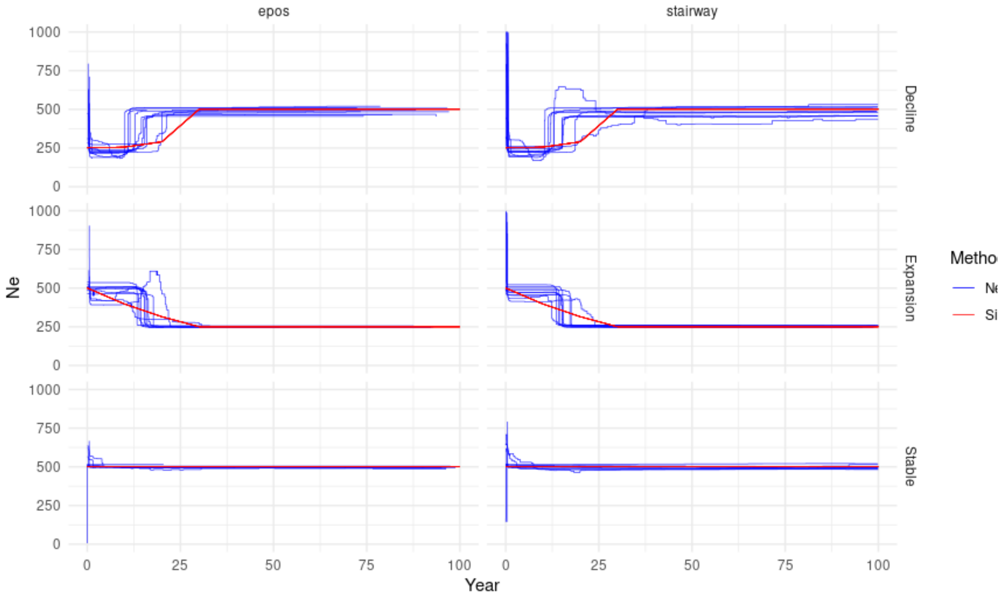
# BEER ACCOUNT

- Bernd Gruber
- 062 924
- 10173614

- If you  only drink soft drinks it is $10
- $20 for alcoholic drinks
- Talk to Jason in case you want something special

- Board games at 7:00

# Stairway2 and Epos Example runs

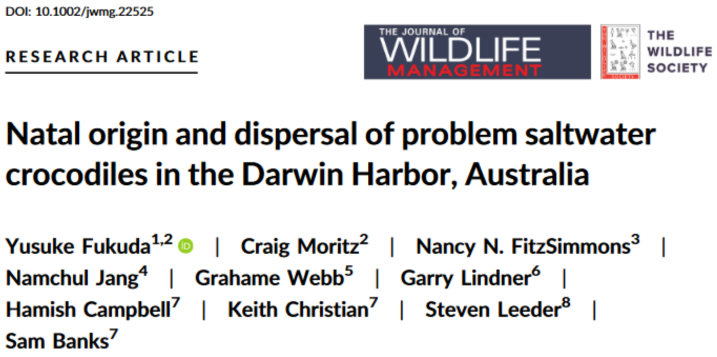


- 200 individuals
- 20k loci
- Decline: Population crash from 500 to 250 over 30 years
- Expansion: Population increase from 250 to 500 over 30 years
- Stable: Population of 500

# Crocodiles

Population history

- Up until the mid-20th century, saltwater crocodiles were abundant across northern Australia.
- unregulated hunting and habitat destruction led to a dramatic decline in their numbers (~1950 – 1970)
- The saltwater crocodile was fully protected under Australian law by 1971.
- conservation efforts included habitat protection, regulated farming, and sustainable use initiatives, which have been pivotal in the species' recovery (>100.000)
- length of genome: l=2.123e9 (or
- length of mu = 7.9e-9  (slowest found in reptiles)



DOI: 10.1002/jwmg.22525

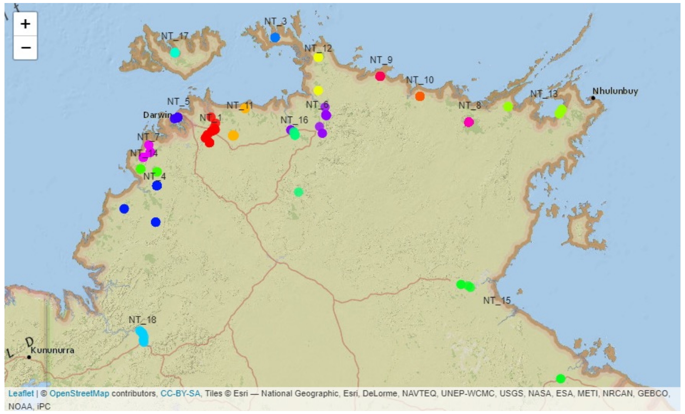
RESEARCH ARTICLE

THE JOURNAL OF WILDLIFE MANAGEMENT

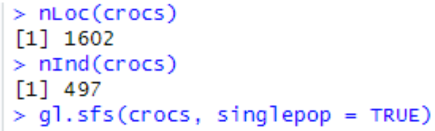
THE WILDLIFE SOCIETY

**Natal origin and dispersal of problem saltwater crocodiles in the Darwin Harbor, Australia**

Yusuke Fukuda[1,2] | Craig Moritz[2] | Nancy N. FitzSimmons[3] | Namchul Jang[4] | Grahame Webb[5] | Garry Lindner[6] | Hamish Campbell[7] | Keith Christian[7] | Steven Leeder[8] | Sam Banks[7]



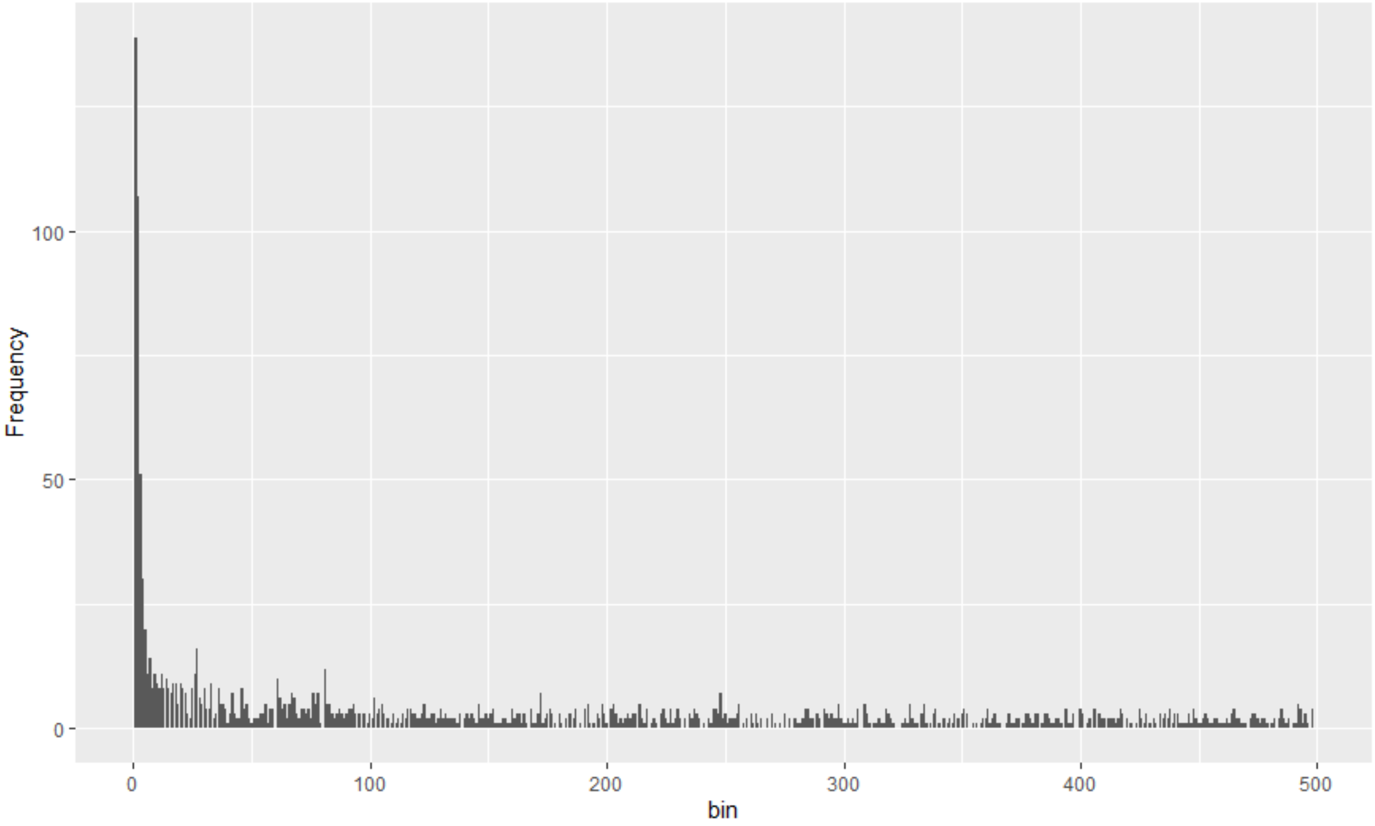
Leaflet | © OpenStreetMap contributors, CC-BY-SA, Tiles © Esri — National Geographic, Esri, DeLorme, NAVTEQ, UNEP-WCMC, USGS, NASA, ESA, METI, NRCAN, GEBCO, NOAA, iPC



```
> nLoc(crocs)
[1] 1602
> nInd(crocs)
[1] 497
> gl.sfs(crocs, singlepop = TRUE)
```

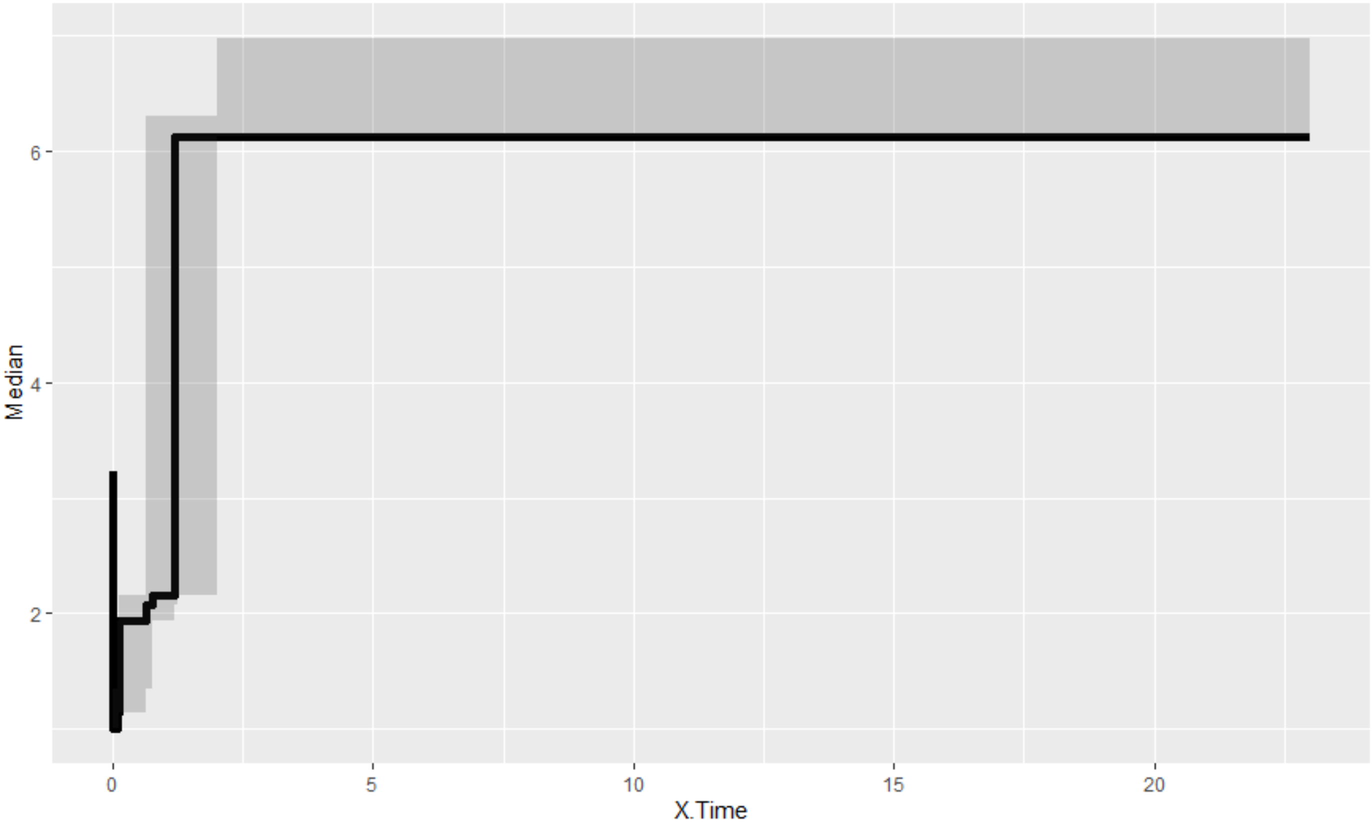
# Crocodiles

- sfs



# Crocodiles

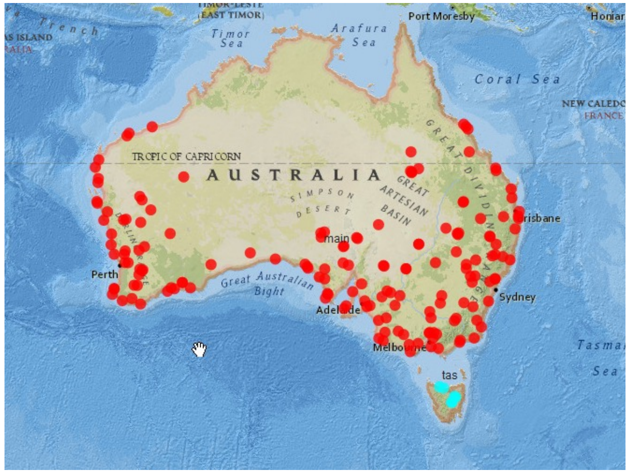
- trajectory

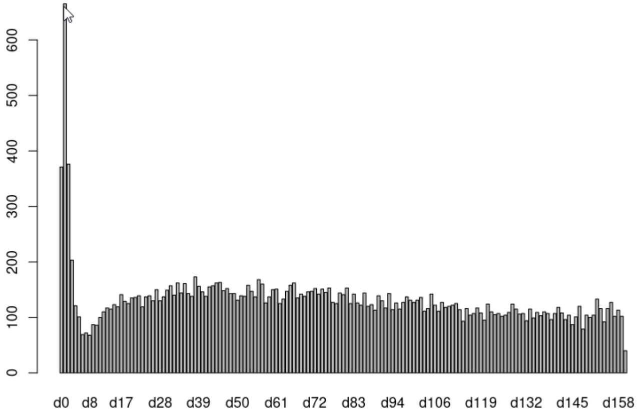


# Foxes

- 160 foxes, 21259 loci

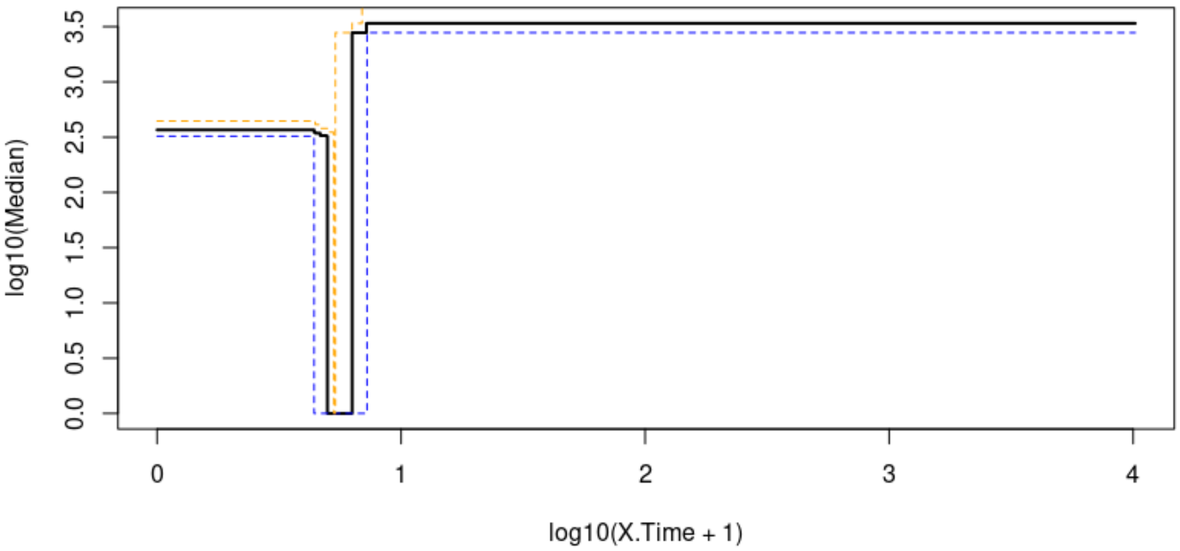
- SFS





# Foxes

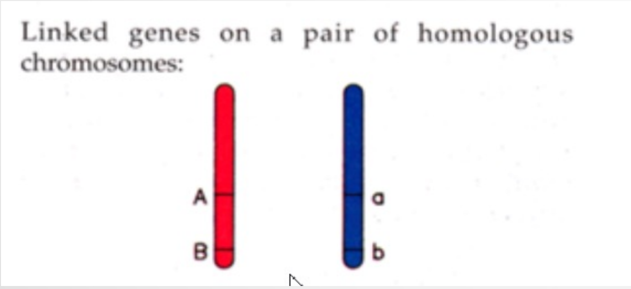
- 160 foxes, 21259 loci
- SFS



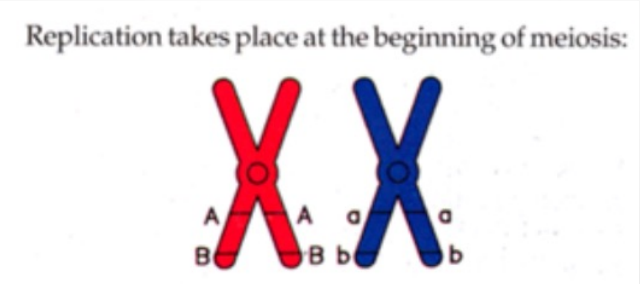
# LD methods

- Linkage disequilibrium
- Hayes 2003 (CSH phased) and LD (for phased and unphased data)
- LD (CSH) along; LD = $(1+4N_e x\ c)$
- LD changes over distances along the chromosomes and number of recombination are determined by effective population size. LD over distance
- Recombination rate c is unknown (using distance between as a proxy)

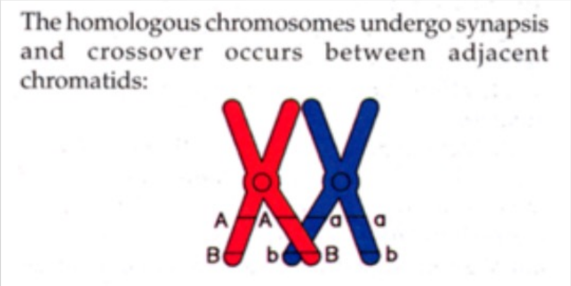
# Linkage (Dis)equilibrium



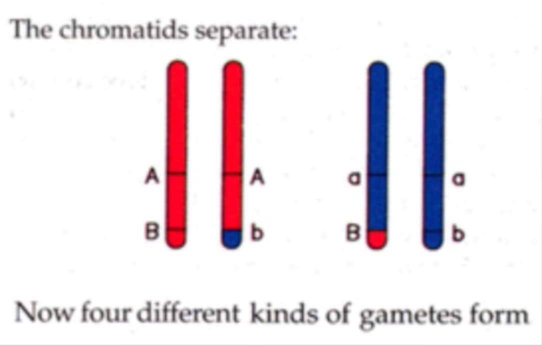
Linked genes on a pair of homologous chromosomes:



Replication takes place at the beginning of meiosis:



The homologous chromosomes undergo synapsis and crossover occurs between adjacent chromatids:



The chromatids separate:

Now four different kinds of gametes form

# Ne and LD

- ***Ne*****:** the number of breeding individuals in an idealized population that would show the same amount of dispersion of allele frequencies under random genetic drift
- $E(\hat{r}^2)$: LD [Linkage disequilibrium]
- ***S*** : sample size (number if individuals)
- ***c*** : recombination rate (for all pairs of SNPs)

$$E(\hat{r}^2_\Delta) \approx \frac{(1-c)^2 + c^2}{2N_e c(2-c)} + \frac{1}{S},$$

# Ne and LD

- ***Ne*:** the number of breeding individuals in an idealized population that would show the same amount of dispersion of allele frequencies under random genetic drift
- $E(\hat{r}^2)$: LD [Linkage disequilibrium]
- ***S*** : sample size (number if individuals)
- ***c*** : recombination rate (for all pairs of SNPs)

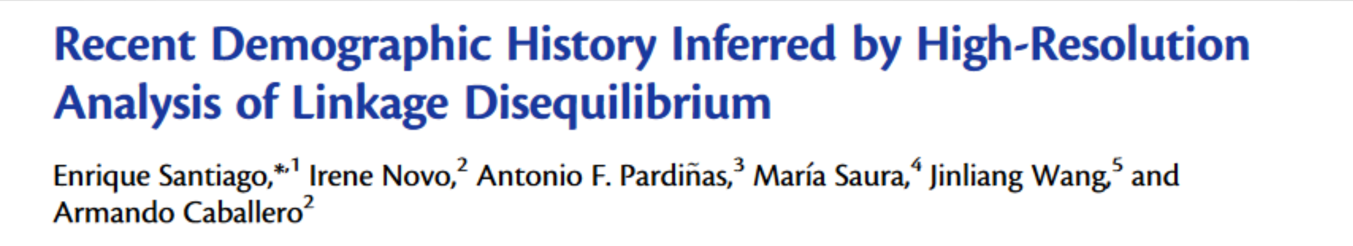
$$E(\hat{r}_{\Delta}^{2}) \approx \frac{(1-c)^2 + c^2}{2N_{\mathrm{e}}c(2-c)} + \frac{1}{S},$$

If c=0.5:

$$E(\hat{r}_{\Delta}^{2}) \approx \frac{1}{3N_{\mathrm{e}}} + \frac{1}{S}$$

# LD

- LD between pairs of SNP at binned distances  provide information on population size at times in the past
- Originally only for expanding and declining population
- New methods eg GONE using phased, unphased genotypes)
  - Input SNPs + linkage map (position of SNPs on the chromosomes)



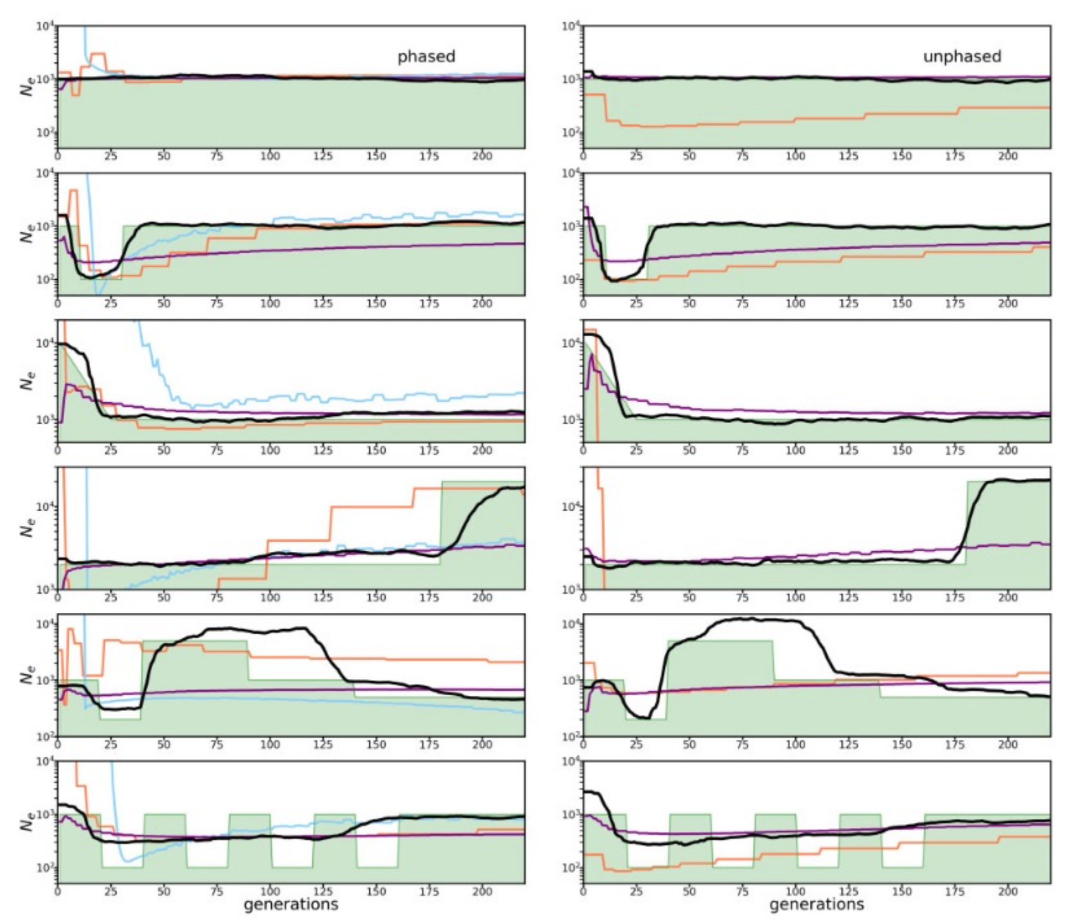
**Recent Demographic History Inferred by High-Resolution Analysis of Linkage Disequilibrium**

Enrique Santiago,*[1] Irene Novo,[2] Antonio F. Pardiñas,[3] María Saura,[4] Jinliang Wang,[5] and Armando Caballero[2]

# Gone

Gone —
MSMC —
Relate —
LinkNe —

- Comparison to other methods
- Number of SNPs: 255,000 or 450,000
- Number of samples: 20 (4 MSMC)



# Gone

- Very data hungry
- Ne=1000 and 100 individuals sampled.

- Implemented to work with dartR, but you need a linkage map

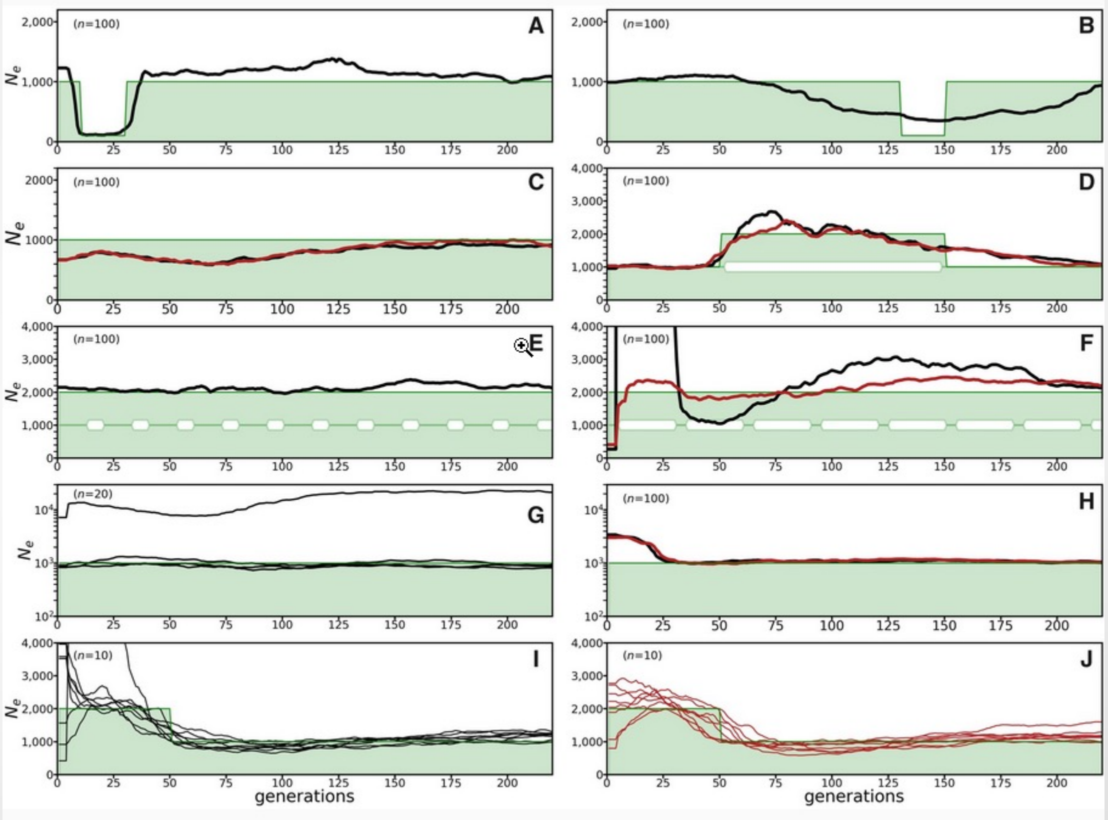


FIG. 2. Estimates of temporal $N_e$ by GONE under different simulated demographic scenarios from present (generation 0) to 220 generations in the past. The true population size is the green shadowed area and *n* is the sample size of individuals for analysis. For all panels, the black lines refer to an analysis where all recombination bins from $c = 0.001$ up to $c = 0.5$ are considered (option $hc = 0.5$), whereas the red lines refer to analyses with rate bins from $c = 0.001$ up to only 0.05 ($hc = 0.05$). (*A*) and (*B*) Detection of bottlenecks occurring at different times. (*C*) Scenario with overlapping generations with three cohorts per generation and mixed-cohort sampling. (*D*) A population $N_e = 1{,}000$ was divided into two populations $N_e =$ 1,000 each, which were isolated for 100 generations and then mixed 50 generations ago into a single population with $N_e = 1{,}000$. (*E*) and (*F*) Metapopulation composed of two subpopulations $N_e = 1{,}000$ each with 2% and 0.2% of migration, respectively, between them. (*G*) Estimations under different base-calling error rates. From top to bottom, 10%, 1%, 0.1% and 0%, the latter two being indistinguishable. (*H*) A hundred individuals were sampled from the population over a period of 100 consecutive generations at a rate of one sampled individual per generation. (*I*) and (*J*) Eight small samples ($n = 10$ each) were taken from the same population at the same time.

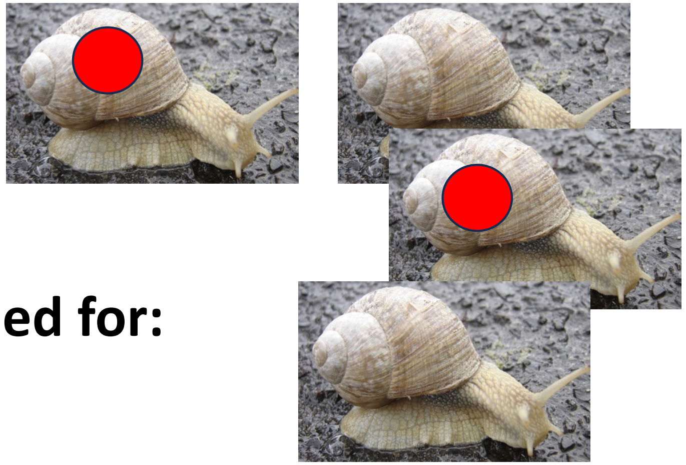
# DnaDot - Fixing Ecology and Evolution’s Blind Spot, Population size $N_c$

Ecological Indicators 2024

WB Sherwin EERC UNSW-Sydney

W.Sherwin@unsw.edu.au

# $N_c$ from: Mark Recapture MR;
# Close Kin Mark Recapture CKMR



**$N_c$ is CRUCIAL – abundance needed for:**

- ± 10% accuracy needed for any forecasting, eg IUCN listing.
- Biodiversity measures with narrow confidence limits (Simpson, Shannon, NOT species lists)
- In most Ecology texts, ~75% of chapters rely heavily on $\mathbf{N_c}$ , for mechanics or outcomes of competition, predation, etc.

Snail: Geierunited commons.wikimedia.org/w/index.php?curid=95926

# $N_c$ from: Mark Recapture MR; Close Kin Mark Recapture CKMR

## IMPERFECT MR:

- Sample sizes (for ±10% precision, need 80% true $\mathbf{N_c}$ !!!)
- 2+ samples
- Assume: no birth death immigration emigration

## IMPERFECT CKMR:

- Sample size $\sqrt{true\ N_c}$ only gives ± 30% accuracy – not ± 10%
- Individual or kin identification, often from worst possible DNA!
- Assume: family size mean and variance known independently

# DnaDot

HOW:

- Based on MR, but no marking
- Pre-existing polymorphisms ‘mark’ separate groups in population (eg, at this SNP, these ones have ‘T’, these ones have ‘C’)

ADVANTAGES:

- Single sample
- No need for independent knowledge of birth, death, immigration, emigration, family size mean and variance, etc
- Genotyping only good enough to estimate allele proportions, NOT to identify individuals and kin

# DnaDot  performs well  (accurate & precise)

| | Bias<br>Estimate =±10% of true $N_c$<br>or better | Variability<br>CL Confidence limits =<br>±10%  or  better |
|---|---|---|
| **MR** | **No** (1/3) Maybe (1/3) **Yes** (1/3) | **No** (4/5) Maybe (1/5) |
| **CKMR** | **No** (2/2) | **No** (4/4) |
| **DnaDot** | **No** (6/36) **Yes** (30/36) | **Yes** (36/36) |

# DnaDot  Summary

Compared to older genetic and non-genetic measures:

- DnaDot avoids most pitfalls of previous estimates of population size $N_c$
- DnaDot is sufficiently accurate and precise for most uses of $N_c$
- Article has link to App requiring Excel input and output, no programming:  Sherwin WB. 2024. DnaDot - Fixing Ecology and Evolution's Blind Spot, Population size. Ecological Indicators *******
- WB Sherwin EERC UNSW-Sydney W.Sherwin@unsw.edu.au

# Coalescent effective population size

- **Estimating Historical Population Sizes**: Coalescent models can use genetic data to estimate changes in Ne over time, providing insights into past population dynamics.
- In coalescent models, Ne influences the expected time until two lineages coalesce.

- Show different SFS (based on different trajectories)
- Explain the coalescent in sketches
- Explain Ne (in this sense)
- Run different scenarios (explain impact of L, mu)
- Show some examples and how users can use there data (preparation)

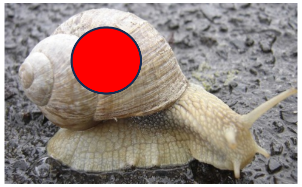
# DnaDot - Fixing Ecology and Evolution’s Blind Spot, Population size $\mathbf{N_c}$

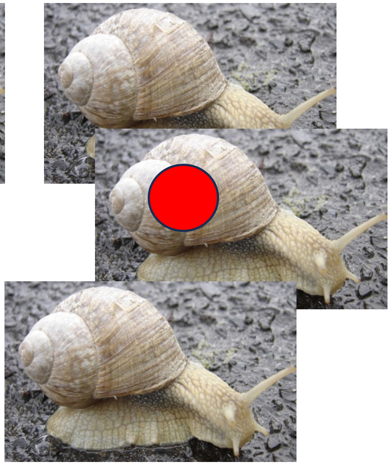
Ecological Indicators 2024

WB Sherwin EERC UNSW-Sydney

W.Sherwin@unsw.edu.au

# $N_c$ from: Mark Recapture MR; Close Kin Mark Recapture CKMR





**$N_c$ is CRUCIAL – abundance needed for:**

- ± 10% accuracy needed for any forecasting, eg IUCN listing.
- Biodiversity measures with narrow confidence limits (Simpson, Shannon, NOT species lists)
- In most Ecology texts, ~75% of chapters rely heavily on $\mathbf{N_c}$ , for mechanics or outcomes of competition, predation, etc.

Snail: Geierunited commons.wikimedia.org/w/index.php?curid=95926

# $N_c$ from: Mark Recapture MR; Close Kin Mark Recapture CKMR

### IMPERFECT MR:

- Sample sizes (for ±10% precision, need 80% true $\mathbf{N_c}$ !!!)
- 2+ samples
- Assume: no birth death immigration emigration

### IMPERFECT CKMR:

- Sample size $\sqrt{true\ N_c}$ only gives ± 30% accuracy – not ± 10%
- Individual or kin identification, often from worst possible DNA!
- Assume: family size mean and variance known independently

# DnaDot

HOW:

- Based on MR, but no marking
- Pre-existing polymorphisms ‘mark’ separate groups in population (eg, at this SNP, these ones have ‘T’, these ones have ‘C’)

ADVANTAGES:

- Single sample
- No need for independent knowledge of birth, death, immigration, emigration, family size mean and variance, etc
- Genotyping only good enough to estimate allele proportions, NOT to identify individuals and kin

# DnaDot performs well (accurate & precise)

| | **Bias**<br>**Estimate =±10% of true $N_c$ or better** | **Variability**<br>**CL Confidence limits = ±10% or better** |
|---|---|---|
| **MR** | **No** (1/3) Maybe (1/3) **Yes** (1/3) | **No** (4/5) Maybe (1/5) |
| **CKMR** | **No** (2/2) | **No** (4/4) |
| **DnaDot** | **No** (6/36) **Yes** (30/36) | **Yes** (36/36) |

# DnaDot Summary

Compared to older genetic and non-genetic measures:

- DnaDot avoids most pitfalls of previous estimates of population size $N_c$
- DnaDot is sufficiently accurate and precise for most uses of $N_c$
- Article has link to App requiring Excel input and output, no programming: Sherwin WB. 2024. DnaDot - Fixing Ecology and Evolution's Blind Spot, Population size. Ecological Indicators *******
- WB Sherwin EERC UNSW-Sydney W.Sherwin@unsw.edu.au