Bioinformatics approaches in animal breeding

AGRONOMSKI FAKULTET ZAGREB MCMXIX

UNIVERSITAS STUDIORUM ZAGRABIENSIS MDCLXIX

BF
UNIVERZA V LJUBLJANI
Biotehniška fakulteta

VF
UNIVERZA V LJUBLJANI
Veterinarska fakulteta
NACIONALNI VETERINARSKI INŠTITUT

angen.agr.hr

https://bioinfo2025.splet.arnes.si

Čurik Ino

Effective population size

University of Zagreb Faculty of Agriculture

✉ icurik@agr.hr

Hungarian University of Agriculture and Life Sciences

Summer school 2025 / Thursday 10.7.2025 / 9:00 to 10:30 /

University of Zagreb Faculty of Agriculture

Random genetic drift / genetic drift

Genetic drift describes random **fluctuations** / **changes** in **allele frequencies**, inbreeding level, gametic and linkages disequilibrium, loss of heterozygosity, etc., in populations.

Genetic drift arises from the **stochastic nature** (random sampling errors) of allele transmission during inheritance in finite populations.

Idealised (Wright-Fisher) **population**

✓ **Infinite size of population – no genetic drift**

- **No selection**
- **No mutation**
- **No migration**
- **Mating at random**
- Equal sex ratio
- Variance of the family size - Poisson distribution
- Constant population size
- Discrete generations (no overlapping)

No change in allele and genotype frequencies (HWE), inbreeding level, LD, GD, etc.

Idealised (Wright-Fisher**) population**

✓ **Finite size of population – genetic drift**

- **No selection**
- **No mutation**
- **No migration**
- **Mating at random**
- Equal sex ratio
- Variance of the family size - Poisson distribution
- Constant population size
- Discrete generations (no overlapping)

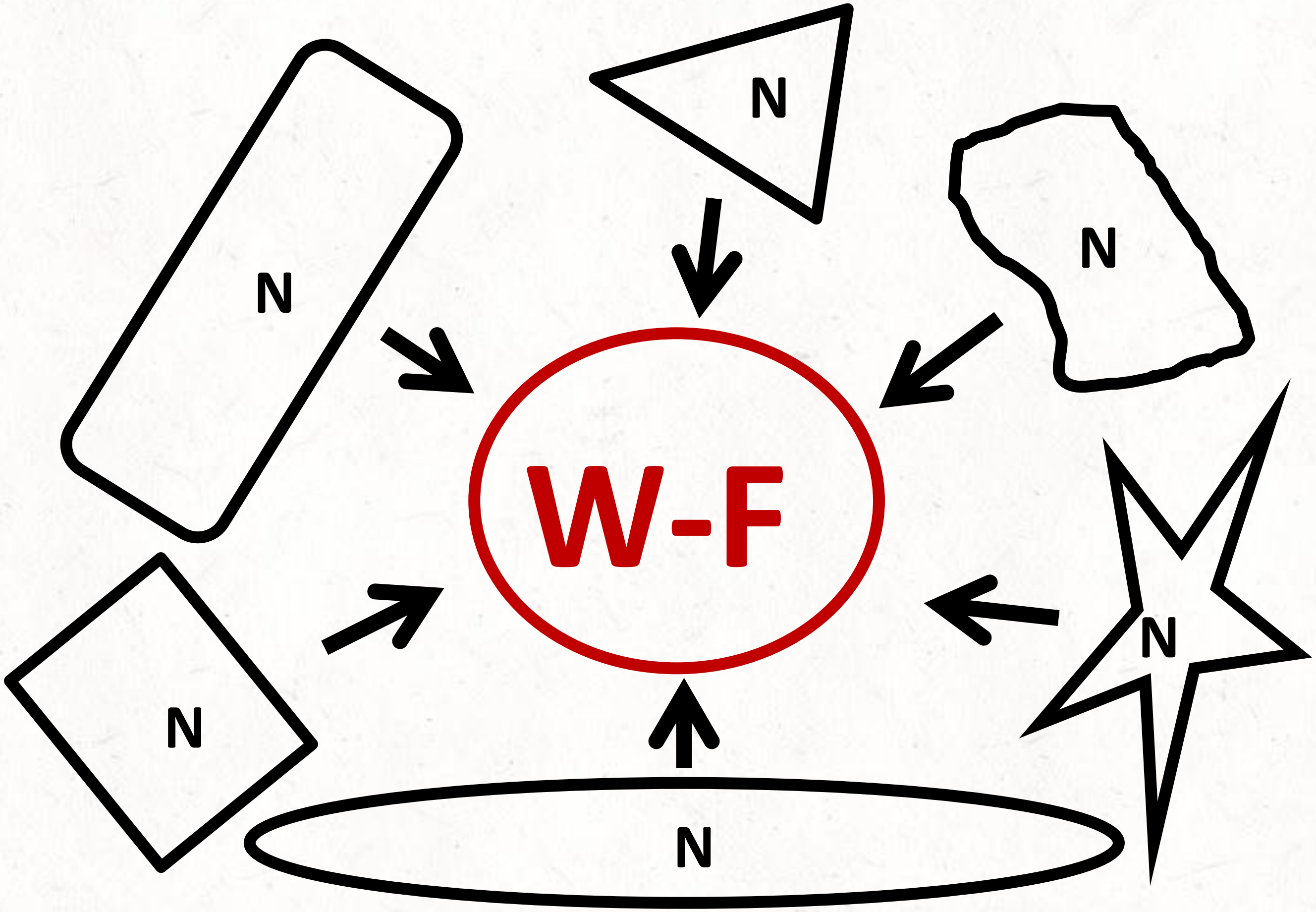
Theoretical derivations defining expected genetic drift

Effective population size (Ne)

The main reason to define **effective population size** (Ne) in population genetics is to quantify the genetic diversity and evolutionary dynamics of a population by accounting for deviations from an idealized population.

The concept of effective population size arises from the need to define population size in a way that **connects census size (N**), or the simple count of individuals, **to the genetic dynamics of the population**.

Required to predict genetic drift in future generations of 'natural' populations !



The concept of **effective population size (Ne)** was introduced by the S. Wright (1931, 1938).

The effective population size - Ne

The most important parameter in population genetics

The effective population size (Ne) of a **real population** X is the size of a **hypothetical ideal population** (Wright-Fisher) that will result in the - same amount of genetic drift **-** as in the **real (actual) population considered**.

Variance effective population size (N_{eV}):

- ✓ same change in allele frequencies -

Inbreeding effective population size (N_{eI}):

- ✓ same change in inbreeding level -

Eigenvalue effective population size (N_{eE}):

- ✓ same long-term rate at which genetic variants are lost -

Linkage disequilibrium effective population size (N_{eLD}):

- ✓ same change in gametic phase/linkage disequilibrium -

N_{eLD}, N_{eV}, N_{eI} and N_{eE} are not always equal (same) by definition !

Unequal sex-ratio

The combined probability of uniting gametes coming from the same grandparent is:

$$\frac{1}{N_e} = \frac{1}{4N_f} + \frac{1}{4N_m}$$

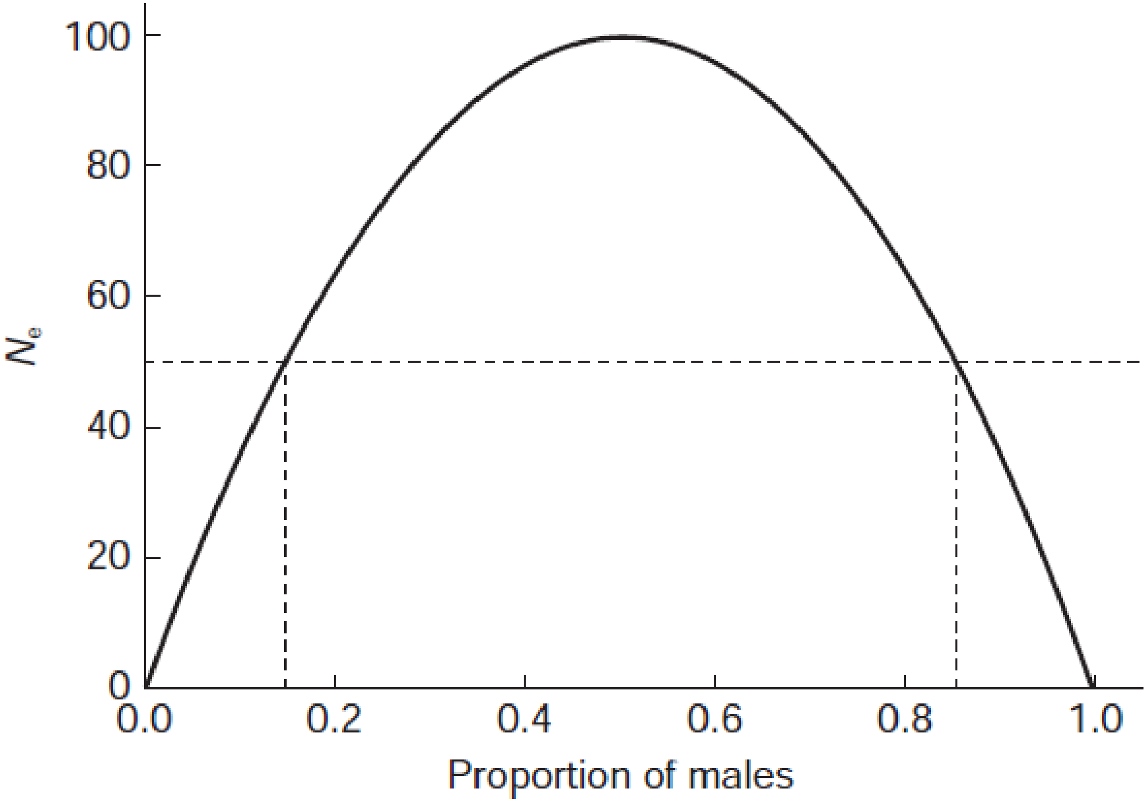
$$\boxed{N_e = \frac{4N_f N_m}{N_f + N_m}}$$

N_f is the number of breeding females

N_m is the number of breeding males

100 males & females

$$N_e = 4\frac{(21)(550)}{21+550} = 80.91$$



Unequal sex-ratios reduce the effective size of the population towards the number in the sex with the fewest breeding individuals.

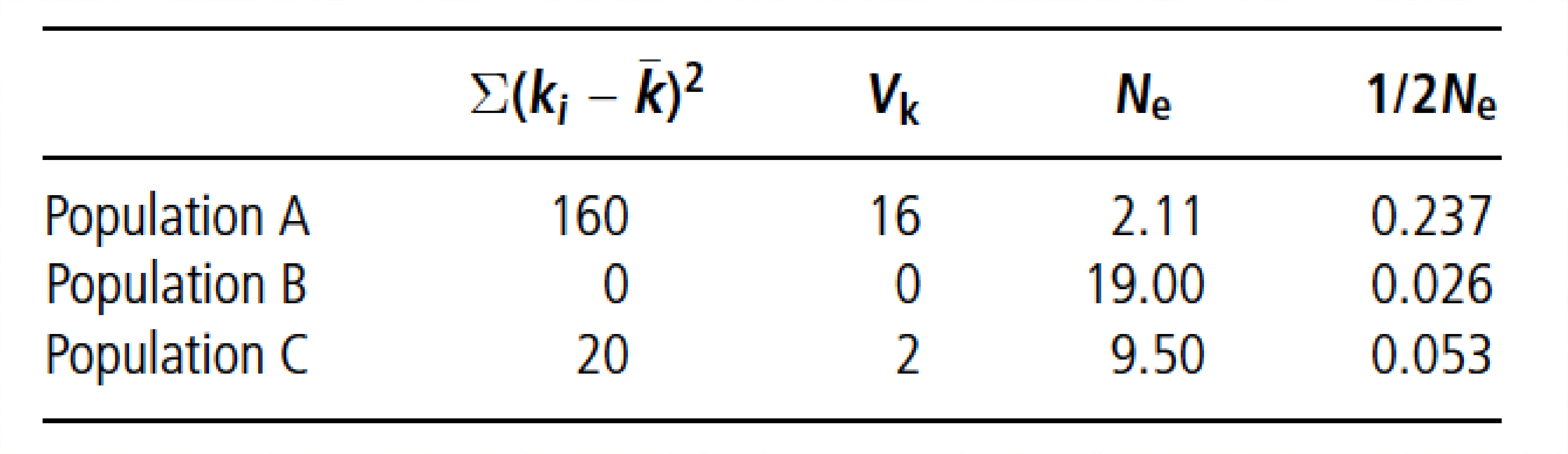
X-chromosome

$$\boxed{N_e = \frac{9N_mN_f}{4N_m+2N_f}}$$

Non-Random Contribution of Parents to Offspring

$$\boxed{N_e = \frac{4N - 2}{2 + V_k}}$$

A Poisson distribution has a mean equal to the variance, thus, Vk = k = 2 & Ne = N.



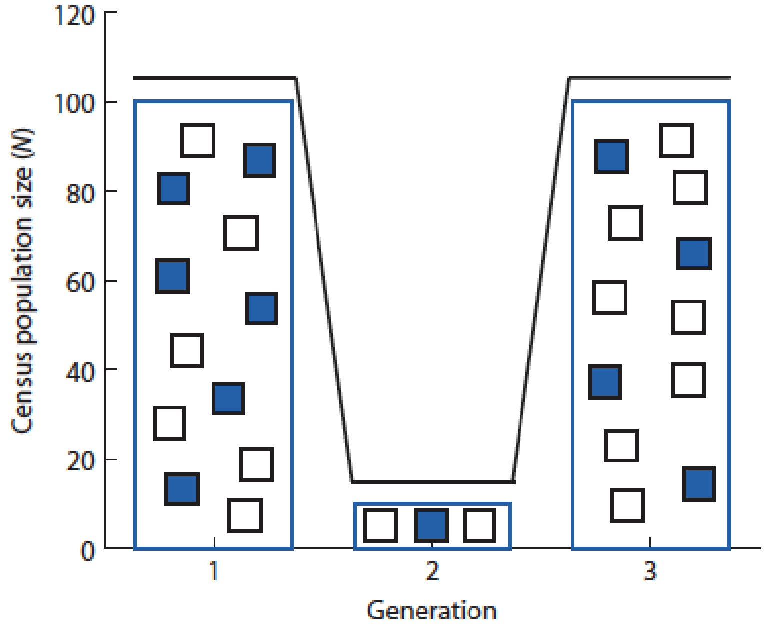
	$\Sigma(k_i - \bar{k})^2$	V_k	N_e	$1/2N_e$
Population A	160	16	2.11	0.237
Population B	0	0	19.00	0.026
Population C	20	2	9.50	0.053

In animal breeding the variance of the parents to offspring contribution depends on the sex.

$$\boxed{N_e = \frac{8N - 4}{V_{km} + V_{kf} + 4}}$$

Note that effective population size (Ne) can be defined as reciprocal of the probability that two gametes come from the same parent.

Variable Population Size across Generations



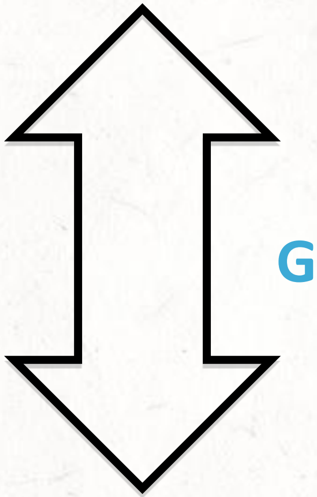
$$\frac{1}{N_e} = \frac{1}{t}\left[\frac{1}{N_{e(t=1)}} + \frac{1}{N_{e(t=2)}} + \cdots + \frac{1}{N_{e(t)}}\right]$$

$$\frac{1}{N_e} = \frac{1}{3}\left[\frac{1}{100} + \frac{1}{10} + \frac{1}{100}\right] \quad N_e = {}^{1}/0.04 = 25$$

Linkage disequilibrium effective population size (N_{eLD})

The effective population size (Ne) of a **real population** X is the size of a **hypothetical ideal population** (Wright-Fisher) that will result in the **- same change in gametic phase/linkage disequilibrium -** as in the **real (actual) population considered**.

In an ideal population of infinite size that has reached an equilibrium state, all loci are in linkage equilibrium.



Genetic drift generates LD between loci

In an ideal population of finite size that has reached an equilibrium state, loci are in linkage disequilibrium while the amount of LD is a function of the genetic distance of the considered loci and the size of population.

Please, note the problems in terminology;

The linkage disequilibrium effective population size reflects the association of alleles at loci that are physically linked and influenced by recombination rates.

The gametic phase disequilibrium effective population size captures allele associations across loci, or those on different chromosomes.

In practice N_{eLD} cannot be measured while LD can.

Use of functional relation between r_{LD}^2, N_{eLD} and recombination rate c is the basis for the estimation of N_{eLD} $[E(r_{LD}^2) \approx f(c, N_{eLD})]$

Sved J.A. (1971) Linkage disequilibrium and homozygosity of chromosome segments in finite populations. Theor. Popul. Biol., 2, 125–141.

$$r_{LD}^2 = \frac{1}{1 + 4N_{eLD}c}$$

where;

r_{LD}^2 is linkage disequilibrium coefficient defined as squared correlation between gametic states at the two loci

N_{eLD} is linkage disequilibrium effective population size

C is genetic distance in Morgans

- Mathematically valid derivation does not exist as even Sved commented; "*This was all introduced in a very messy way, and was not understood by anyone, evidently including myself*".

- Simulation results indicate that the formula works reasonably well.

Ober et al. have reported a new recursion formula that better fitts the function **$E(r_{LD}{}^{2}) \approx f(c, N_{eLD})$**, although their formula was not broadly accepted in scientific community, probably as being more mathematically complex.

Weir B.S., Hill W.G. (1980) Effect of mating structure on variation in linkage disequilibrium. Genetics, 95, 477–488.

$$r_{LD}^2 = \frac{1}{1 + 4N_{eLD}c} + \frac{1}{n_g}$$

Adjusted formula for the sampling effect introduced by the Var(r)=1/n_g , relation where n_g is sample size of gametes i.e. 2*number of sampled individuals.

This and previous formula does not account for mutation!

McVean (2002) -> Tenesa et al., (2007) =>

$$r_{LD}^2 = \frac{1}{\alpha + kN_{eLD}c} + \frac{1}{n_g}$$

where;

- α is correction for the presence of mutations, α = 1 -> no mutations or α = 2 (Tenesa et al., 2007) or α > 2 (2.2 according to Corbin et al., 2012) for the model assuming mutations
- k is correction for the inheritance so k=4 for autosomes and k=2 for X chromosomes

Effective linkage disequilibrium population size

$$Ne_{LD} = \frac{1}{kc}\left[\frac{1}{r_{LD}^2 - \frac{1}{n_g}} - \alpha\right]$$

- c is genetic distance in Morgans
- r_{LD}^2 is linkage disequilibrium coefficient
- α is correction for the presence of mutations, α = 1 -> no mutations or α = 2 or α > 2 (2.2) for the model with mutations
- k is correction for the inheritance, k=4 for autosomes & k=2 for X chromosomes
- n_g is sample size of gametes (2*number of sampled individuals)

Historical effective linkage disequilibrium population size

Hayes et al., (2003) introduced concept of time i.e. historical effective population size by linking the distance between markers with past time related to N_{eLD}.

$$Ne_{LDf(c)\rightarrow T} = \frac{1}{k[f(c) \rightarrow T]}\left[\frac{1}{r_{LD}^{2} - \frac{1}{n_g}} - \alpha\right]$$

$$Ne_{LDf(c)\rightarrow T} = \frac{1}{4[f(c) \rightarrow T]}\left[\frac{1}{r_{LD}^2} - 1\right]$$

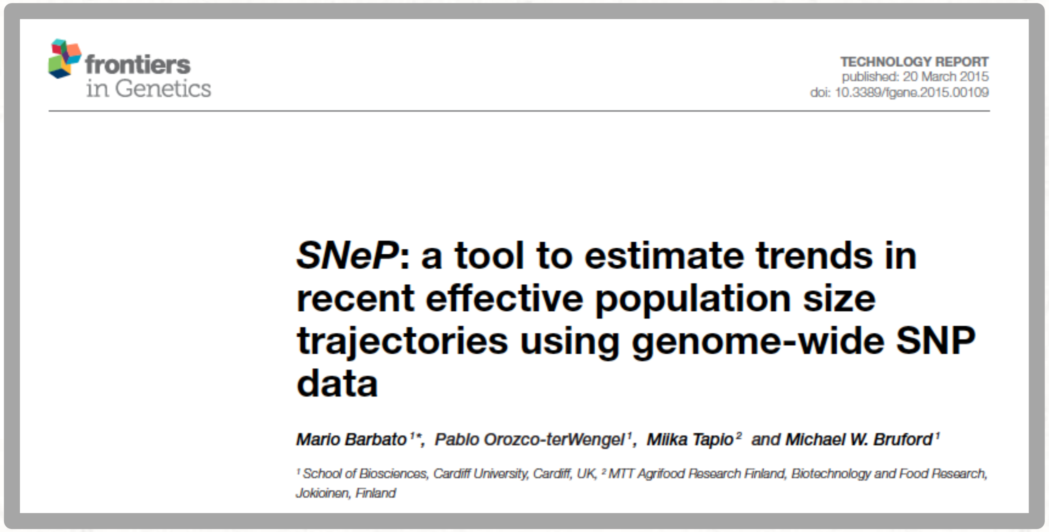
Domestic animal geneticists -> dominated method

A key assumption stated by Hayes et al. (2003) is constant linear growth of N_e with T.

SNeP - Linkage Disequilibrium Ne

Barbato et al., 2015 FG*,

$$Ne_{LD} = \frac{1}{kc}\left[\frac{1}{r_{LD}^2 - \frac{1}{n_g}} - \alpha\right]$$



frontiers in Genetics

TECHNOLOGY REPORT
published: 20 March 2015
doi: 10.3389/fgene.2015.00109

SNeP: a tool to estimate trends in recent effective population size trajectories using genome-wide SNP data

Mario Barbato[1]*, Pablo Orozco-terWengel[1], Miika Tapio[2] and Michael W. Bruford[1]

[1] School of Biosciences, Cardiff University, Cardiff, UK, [2] MTT Agrifood Research Finland, Biotechnology and Food Research, Jokioinen, Finland

Practical issues

- f(c)->T, distance bins from 50 kb to 2000 kb by 50 kb (Flury et al., 2010)

- SNPs with adjacent r_{LD}^2 values [0.01, 0.99] -> Uimari & Tapio (2011)

- MAF issues -> mutation rate

Assumes constant linear growth of N_e with *T (Hayes et al., 2003)*

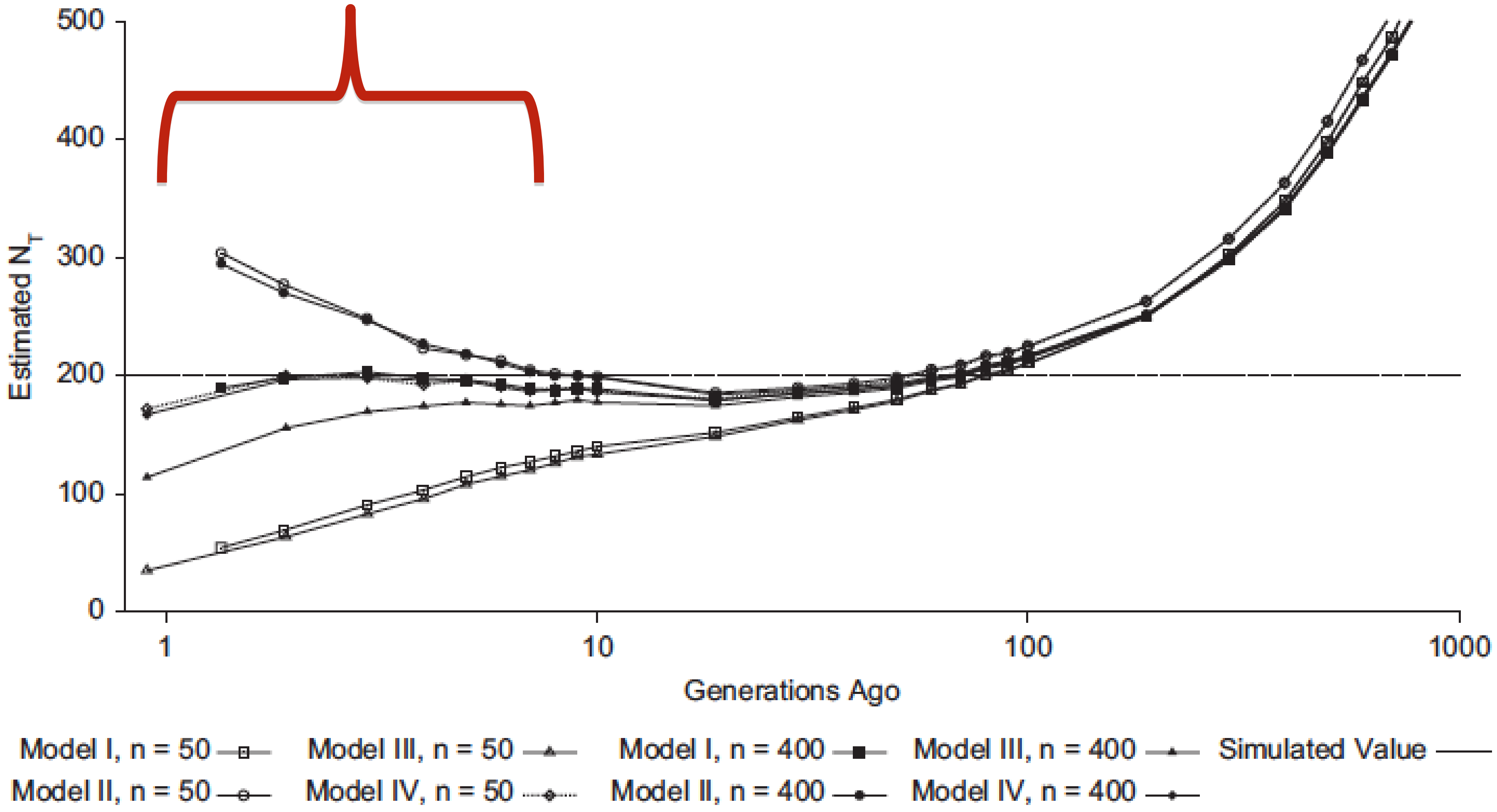
J. Anim. Breed. Genet. ISSN 0931-2668; 2012

ORIGINAL ARTICLE

Estimation of historical effective population size using linkage disequilibria with marker data

L.J. Corbin, A.Y.H. Liu, S.C. Bishop & J.A. Woolliams

The Roslin Institute & Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Midlothian, UK



MOLECULAR ECOLOGY RESOURCES

Molecular Ecology Resources (2014) 14, 209–214

NeEstimator V2

NeEstimator v2: re-implementation of software for the estimation of contemporary effective population size (N_e) from genetic data

C. DO,* R. S. WAPLES,† D. PEEL,‡ G. M. MACBETH,§ B. J. TILLETT¶ and J. R. OVENDEN**

Effective gametic phase disequilibrium population size

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

England et al., (2006): "estimates are biased when sample size (S) is small and S < N_{eGD}"

$$Ne_{GD} = \frac{1}{3\left(r^2_{LD} - \frac{1}{S}\right)}$$

- r_{LD}^2 is linkage disequilibrium coefficient
- S is number of sampled individuals
- c is equal to 0.5 as the loci are unlinked

Many combinations of loci on different chromosomes are possible !

If r_{LD}^2=0.025 & S = 200

$$Ne_{GD} = \frac{1}{3\left(0.025 - \frac{1}{200}\right)} = 50$$

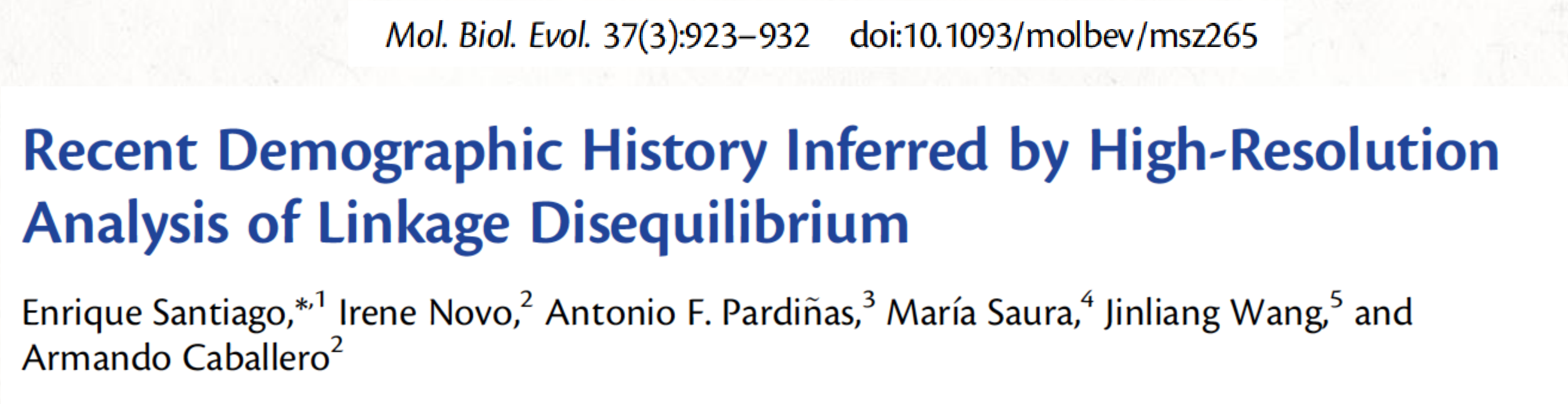
Wildlife animal geneticists -> dominated method

Waples, R. S., 2005 Genetic estimates of contemporary effective population size: To what time periods do the estimates apply? Mol. Ecol. 14: 3335–3352.

Waples, R. S., 2006 A bias correction for estimates of effective population size based on linkage disequilibrium at unlinked gene loci. Conserv. Genet. 7: 167–184.

Waples, R. S., and C. Do, 2008 LdNe: a program for estimating effective population size from data on linkage disequilibrium. Mol. Ecol. Res. 8: 753–756.

Waples, R. S., and C. Do, 2010 Linkage disequilibrium estimates of contemporary Ne using highly variable genetic markers: a largely untapped resource for applied conservation and evolution. Evol. Appl. 3: 244–262.



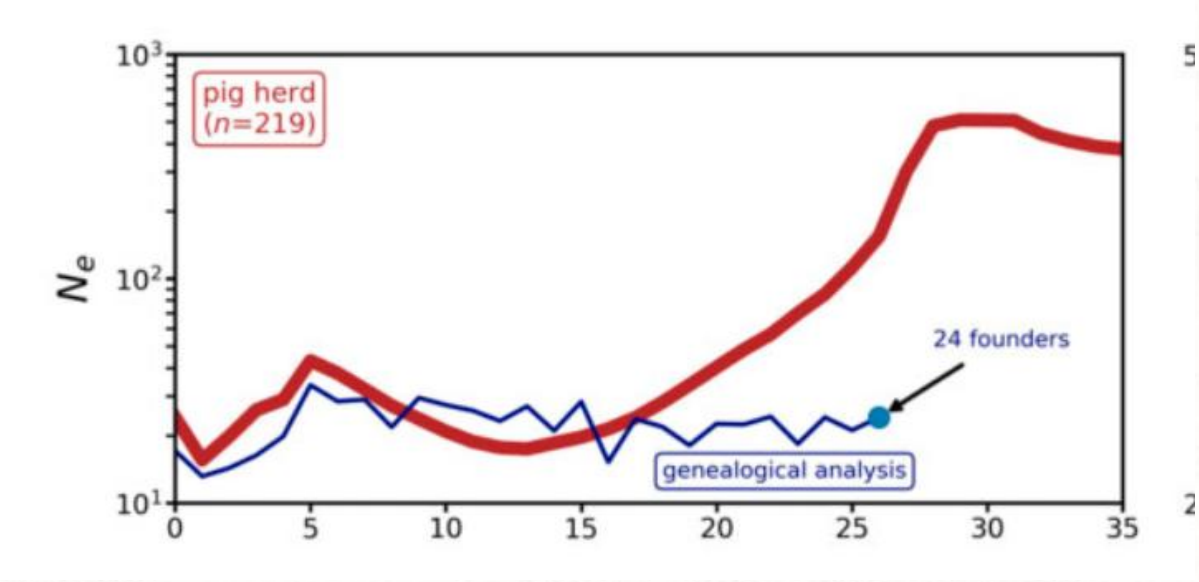
Mol. Biol. Evol. 37(3):923–932 doi:10.1093/molbev/msz265

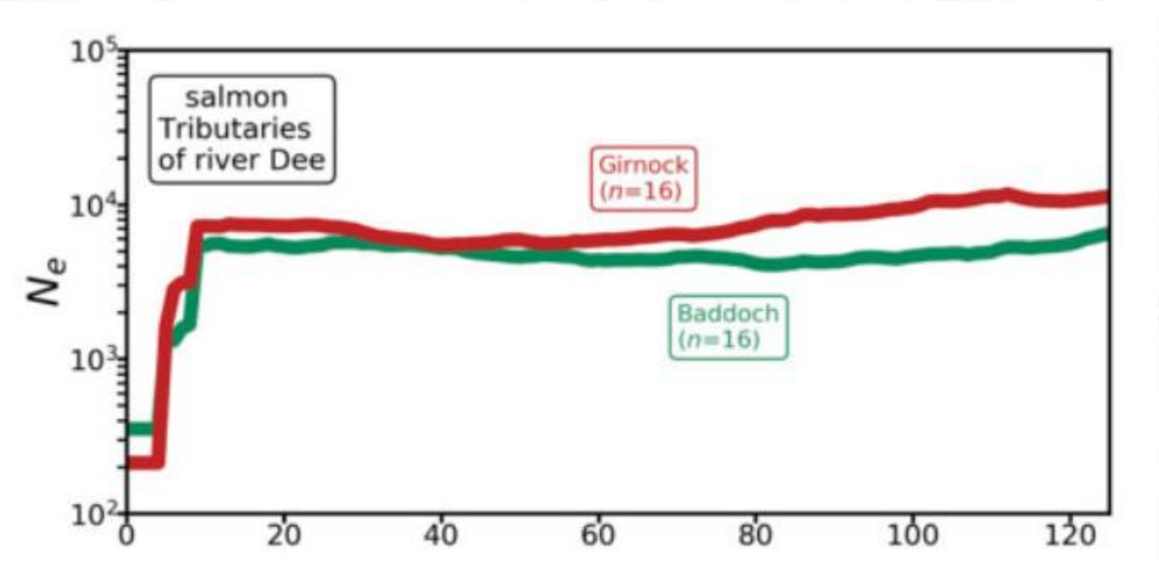
Recent Demographic History Inferred by High-Resolution Analysis of Linkage Disequilibrium

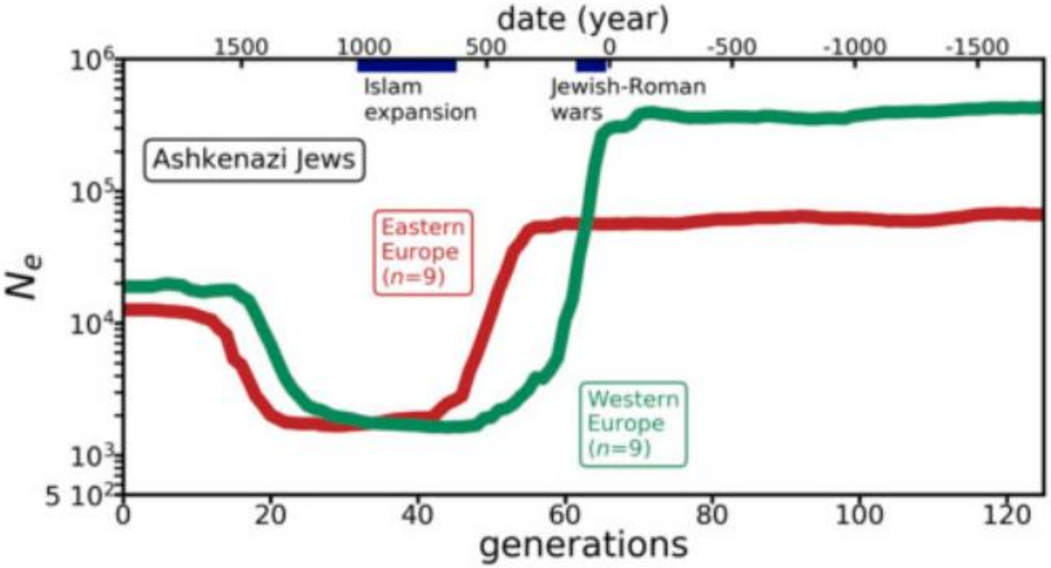
Enrique Santiago,*[1] Irene Novo,[2] Antonio F. Pardiñas,[3] María Saura,[4] Jinliang Wang,[5] and Armando Caballero[2]

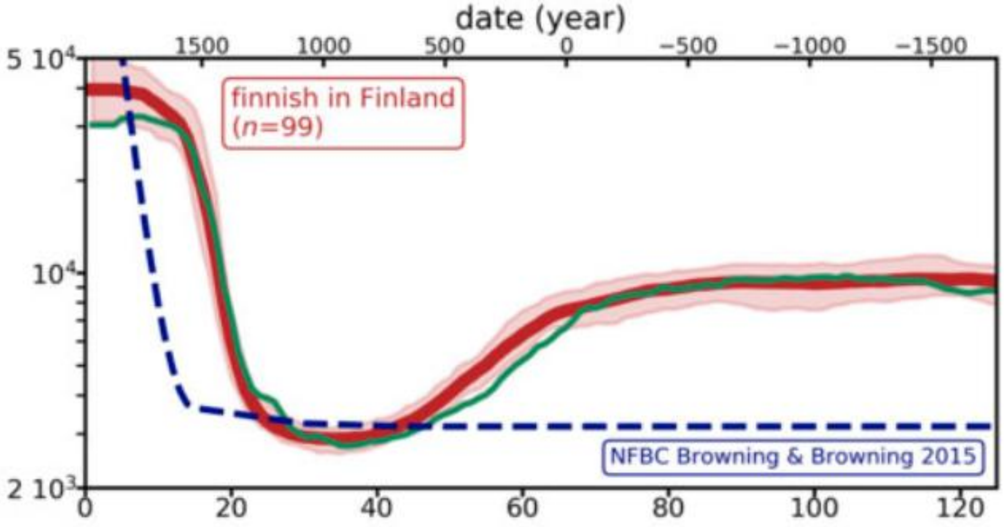
- Genotyping or sequencing data
- Low sample sizes : 10, 20, …. even 2 (?)
- Phased or unphased data
- Pseudo-haploid data (low-coverage ancient DNA sequencing)
- No need to apply MAF

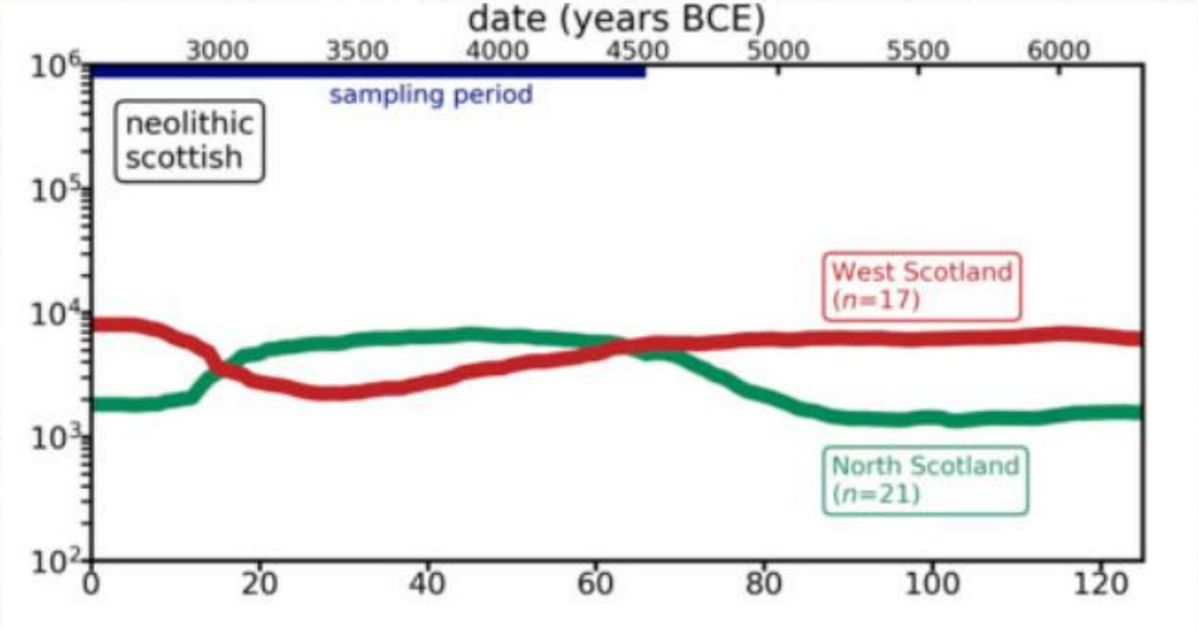
Software GONE https://github.com/esrud/GONE

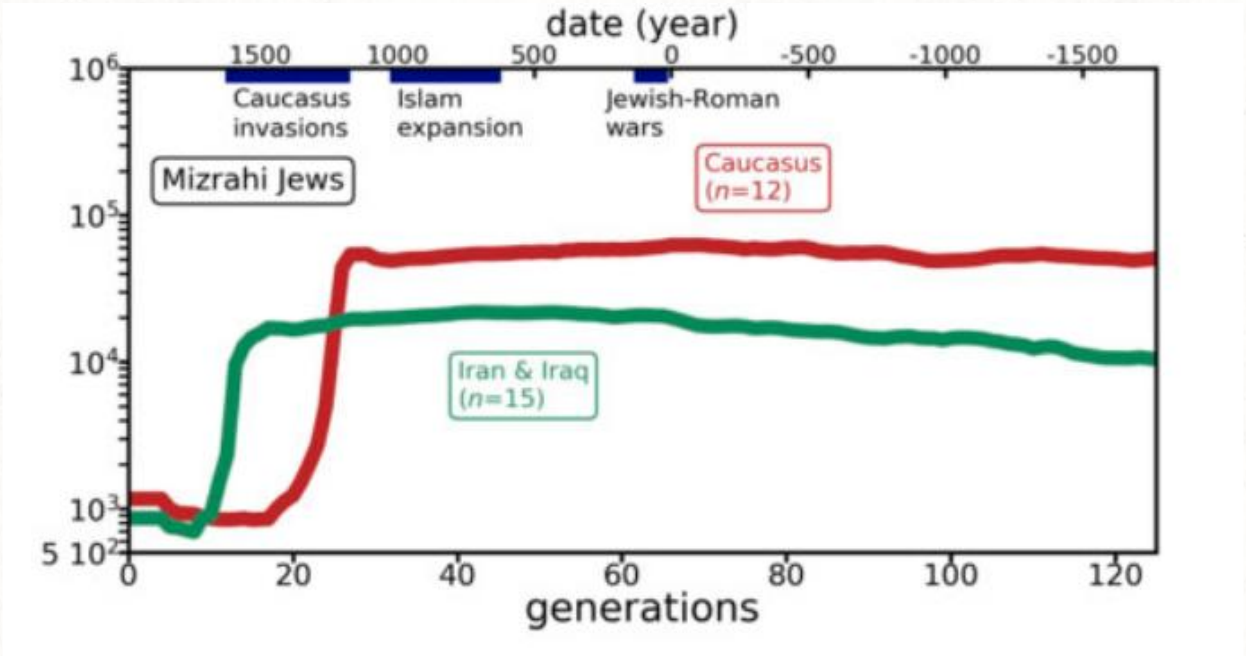

pig herd
(n=219)
24 founders
genealogical analysis
N_e
10^3
10^2
10^1
0
5
10
15
20
25
30
35
5
2


salmon
Tributaries
of river Dee
Girnock
(n=16)
Baddoch
(n=16)
N_e
10^5
10^4
10^3
10^2
0
20
40
60
80
100
120

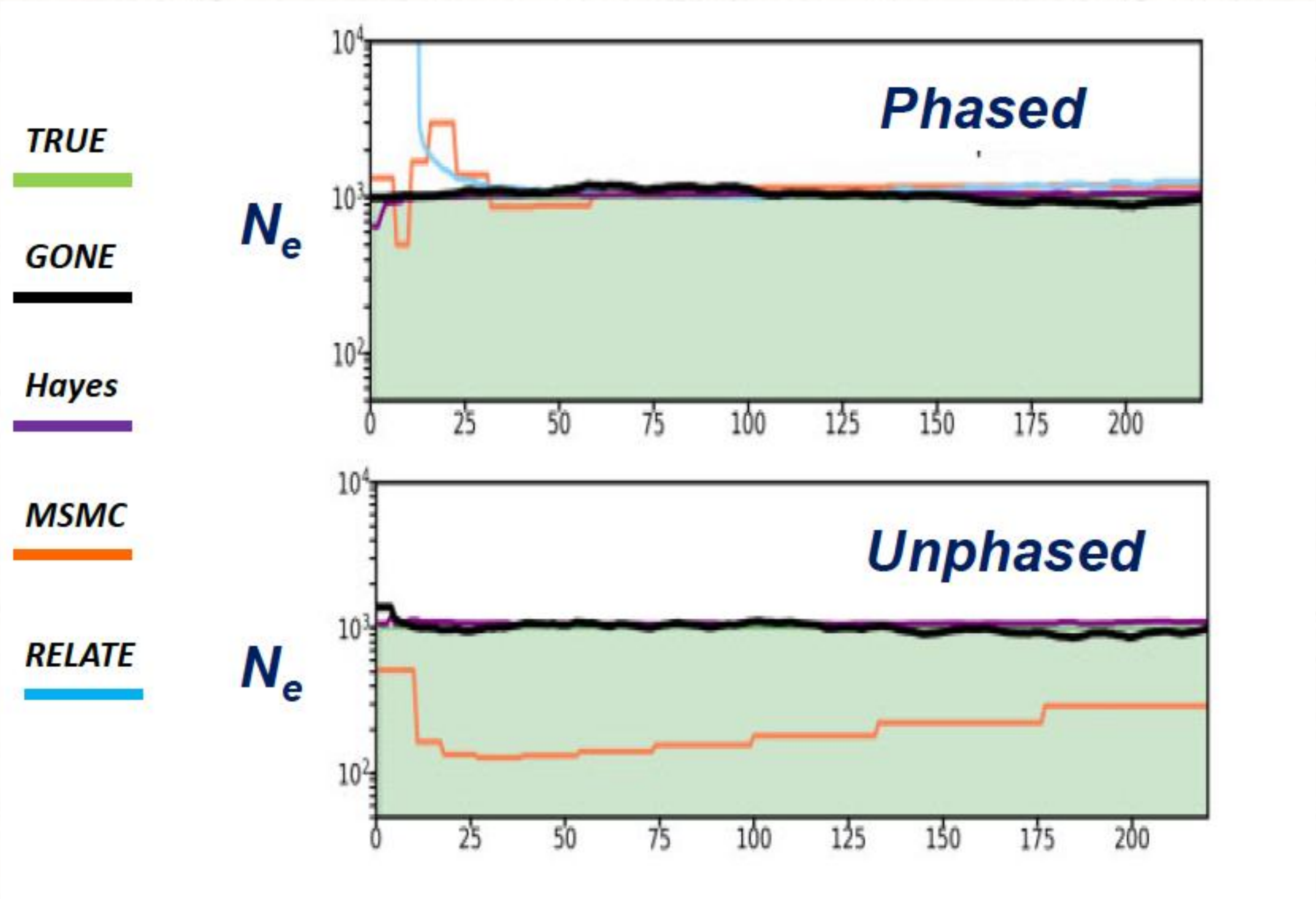

date (year)
1500
1000
500
0
-500
-1000
-1500
Islam expansion
Jewish-Roman wars
Ashkenazi Jews
Eastern Europe (n=9)
Western Europe (n=9)
N_e
10^6
10^5
10^4
10^3
$5\ 10^2$
0
20
40
60
80
100
120
generations


date (year)
1500
1000
500
0
−500
−1000
−1500
$5\ 10^4$
10^4
$2\ 10^3$
finnish in Finland (n=99)
NFBC Browning & Browning 2015
0
20
40
60
80
100
120

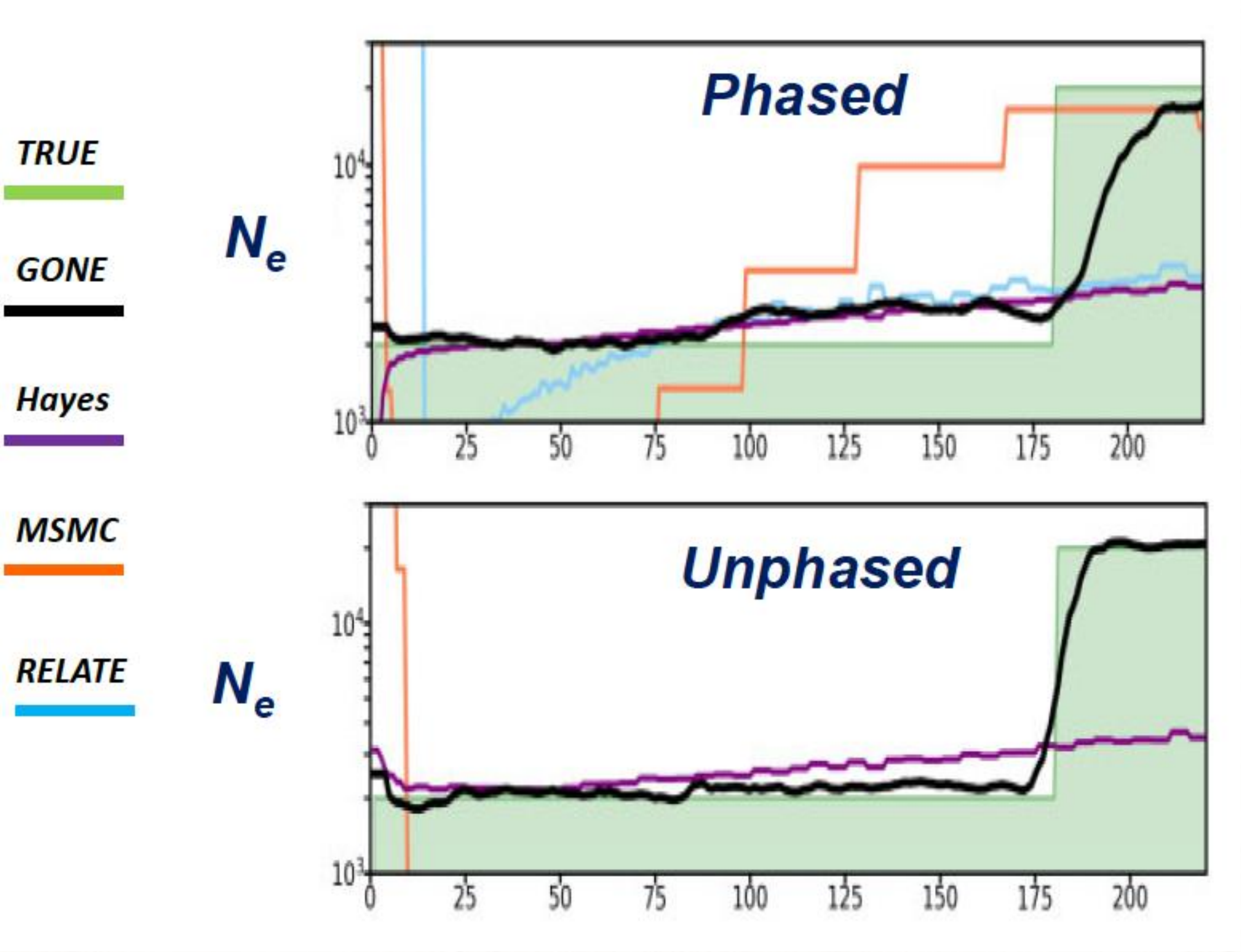

date (years BCE)
3000
3500
4000
4500
5000
5500
6000
sampling period
neolithic scottish
West Scotland (n=17)
North Scotland (n=21)
10^6
10^5
10^4
10^3
10^2
0
20
40
60
80
100
120


date (year)
1500
1000
500
0
-500
-1000
-1500
Caucasus invasions
Islam expansion
Jewish-Roman wars
Mizrahi Jews
Caucasus (n=12)
Iran & Iraq (n=15)
10^6
10^5
10^4
10^3
5 10^2
0
20
40
60
80
100
120
generations

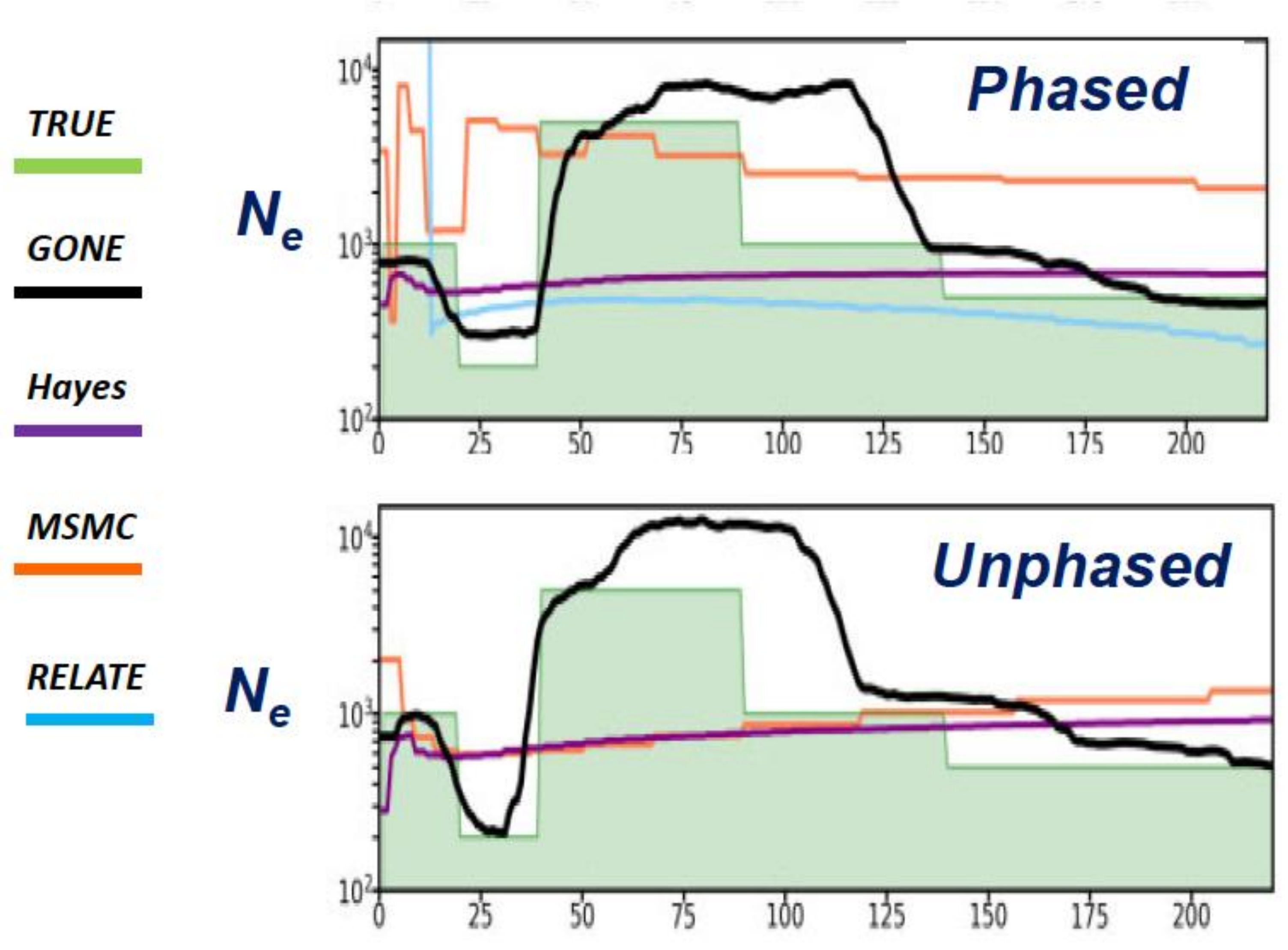
- MSMC (Schiffels& Durbin, Nat. Genet. 2014)
- RELATE (Speidelet al., Nat. Genet. 2019)



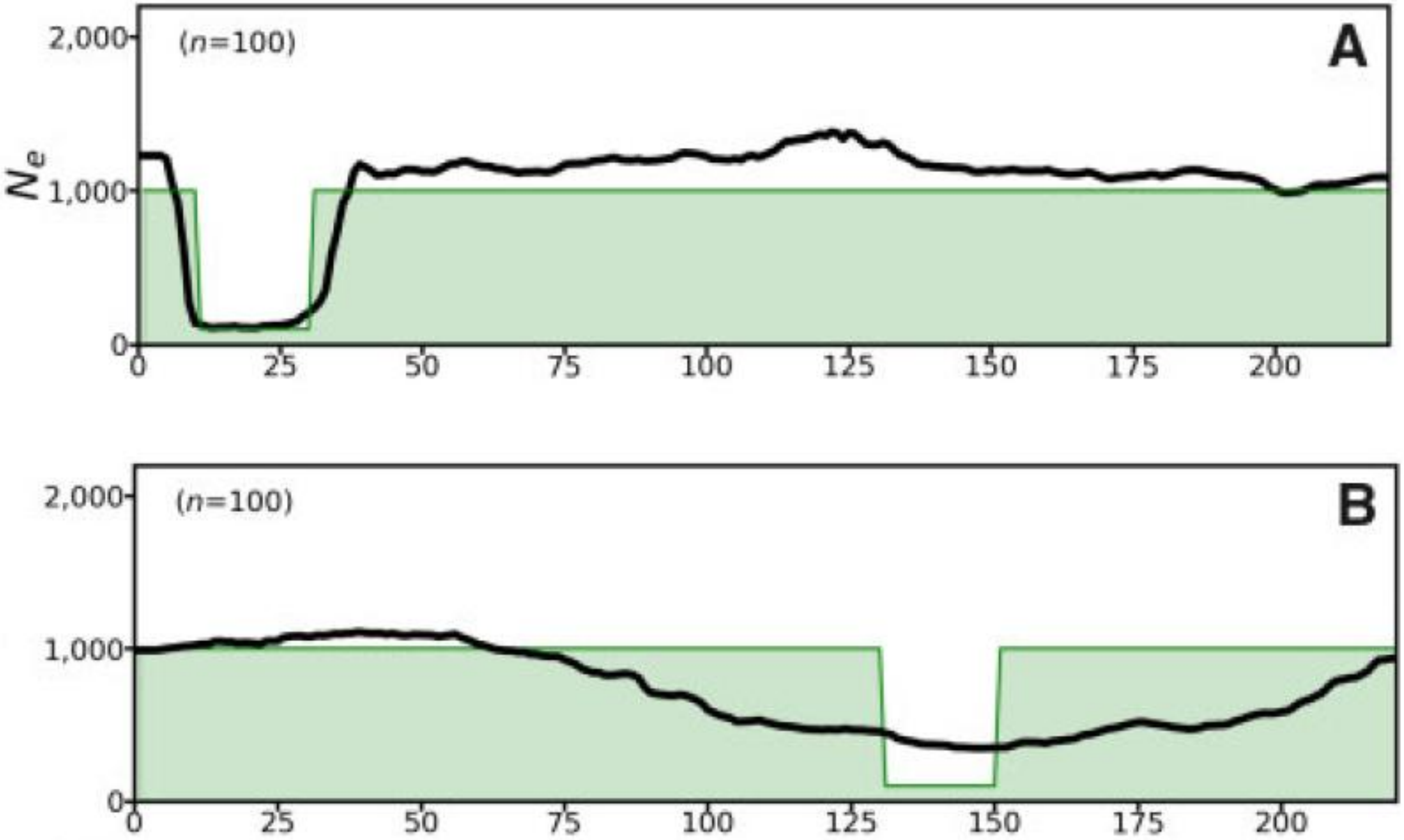
From the presentation: Armando Caballero – Enrique Santiago



From the presentation: Armando Caballero – Enrique Santiago

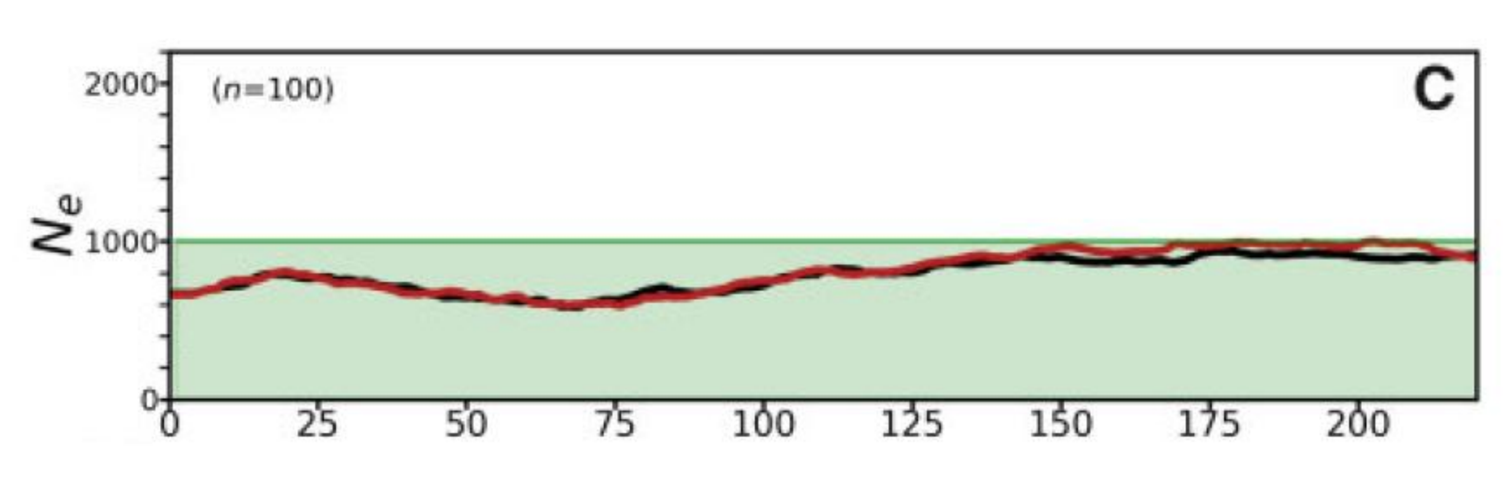


From the presentation: Armando Caballero – Enrique Santiago

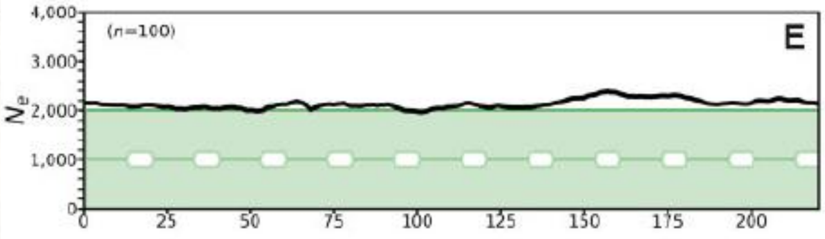


The method is more precise for recent changes than for old ones !

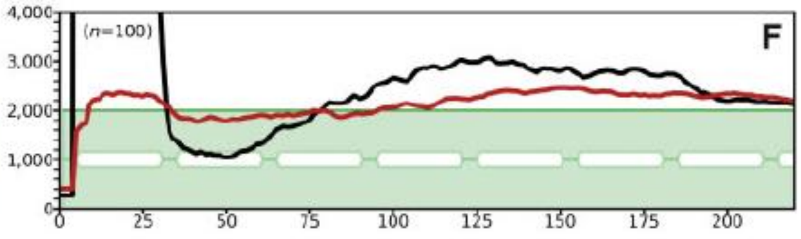
From the presentation: Armando Caballero – Enrique Santiago



Overlapping generations may cause a substantial bias !



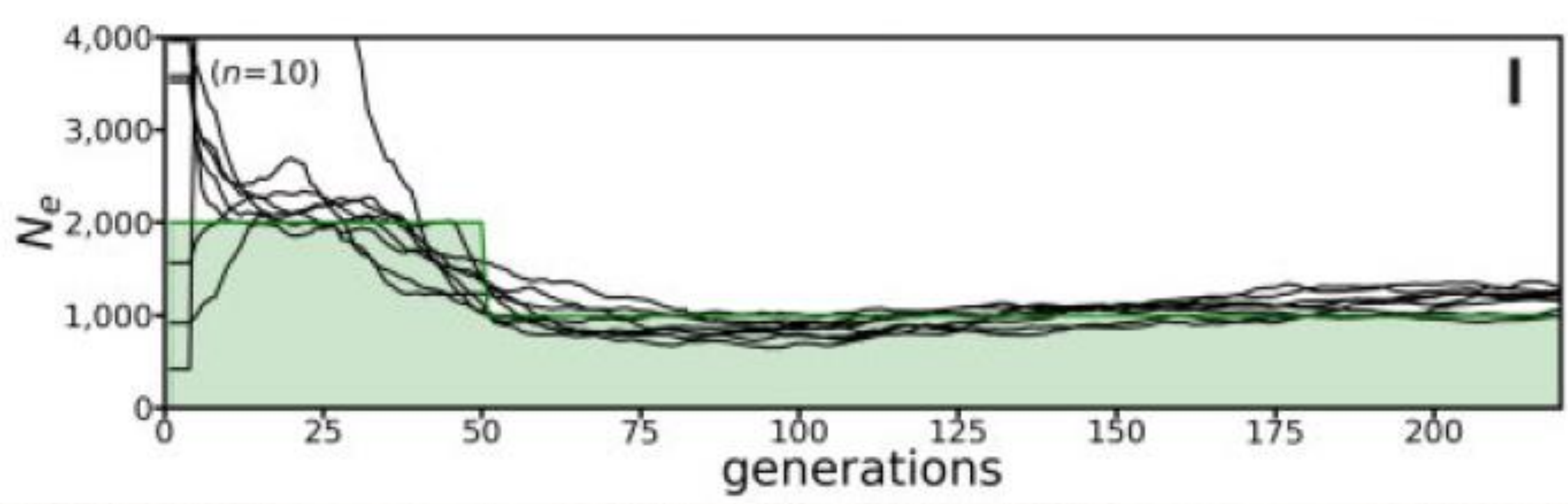
2 % migration between two subpops of *N* = 1000

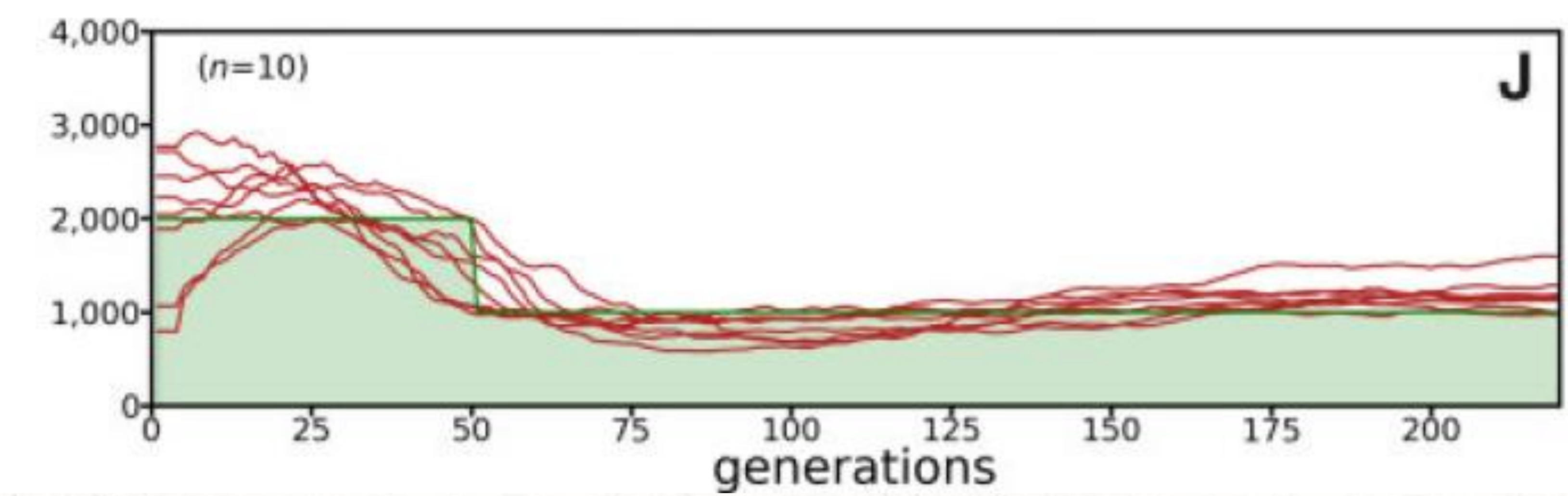


0.2 % migration between two subpops of *N* = 1000

Migration may produce substantial artefacts in the recent past !

From the presentation: Armando Caballero – Enrique Santiago





A low sample size causes a large noise in the estimates, e.g. *n*= 10 !

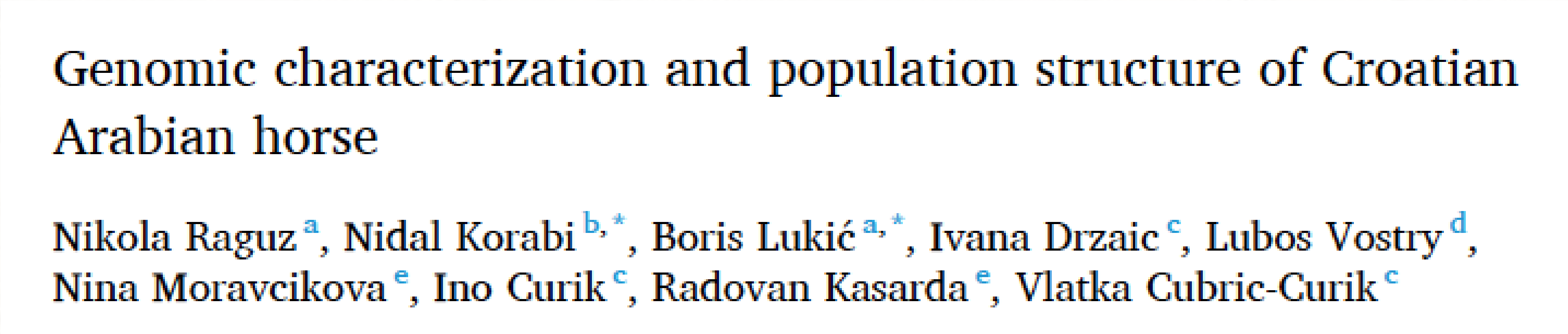
Conclusions: Armando Caballero – Enrique Santiago

- ✓ The short-term trajectory (about 100 generations) of *Ne* can be estimated from linkage disequilibrium among SNPs
- ✓ The method accounts for phased, unphased and pseudo-haploid data
- ✓ A good genetic map, absence of migration or recent admixture and large random samples are the best scenarios to obtain reliable estimates
- ✓ The estimates of historical *Ne* provided by GONE may be substantially biased when there has been a recent mixture of populations that were previously separated for a long period of time

- ✓ Biases may occur when the **rate of continued migration between populations is low,** or when chromosomal inversions are present at high frequencies.

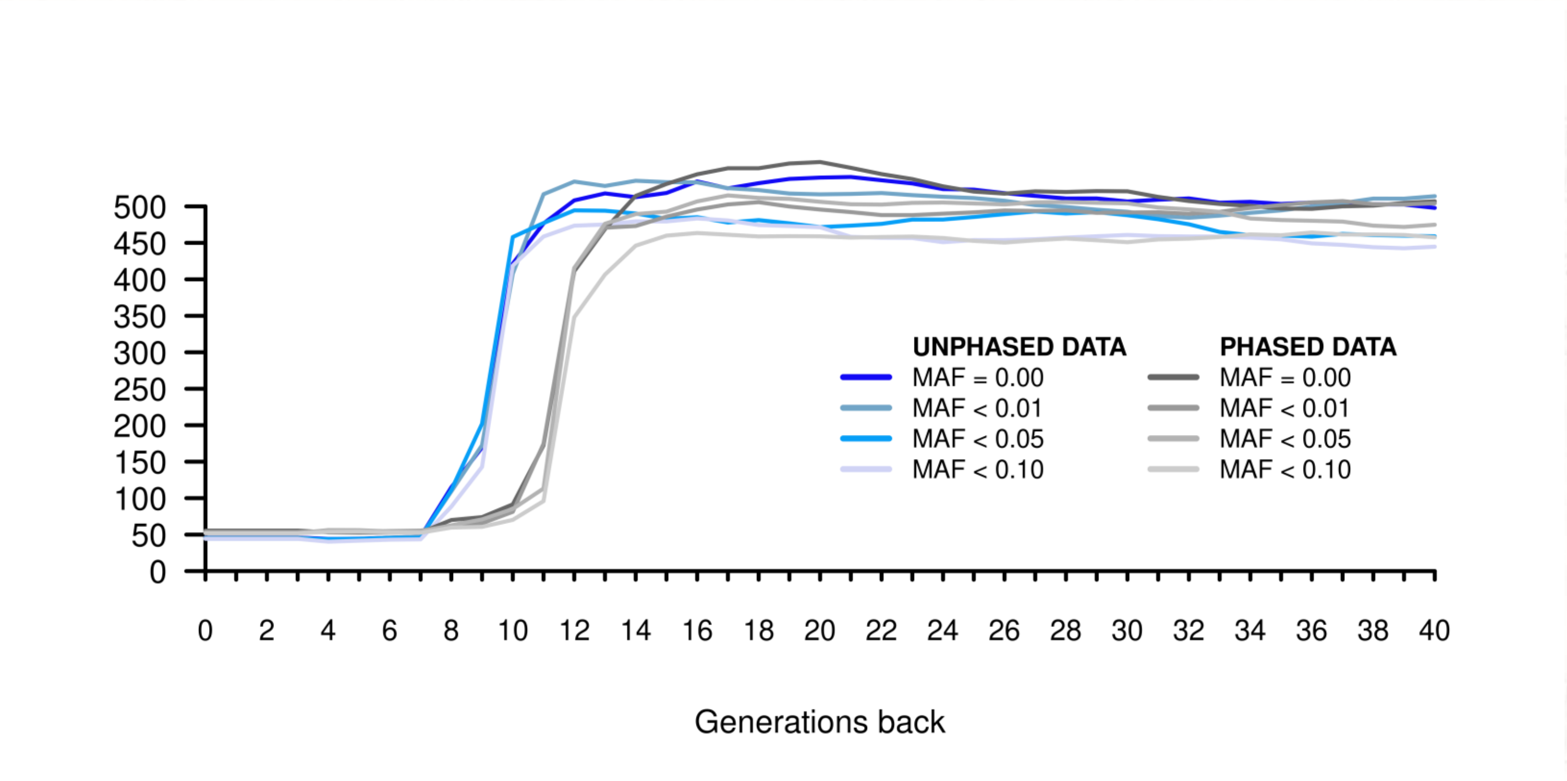
- ✓ However, some biases due to population structuring **can be eliminated by conducting population structure analyses** and restricting the estimation to the differentiated groups.

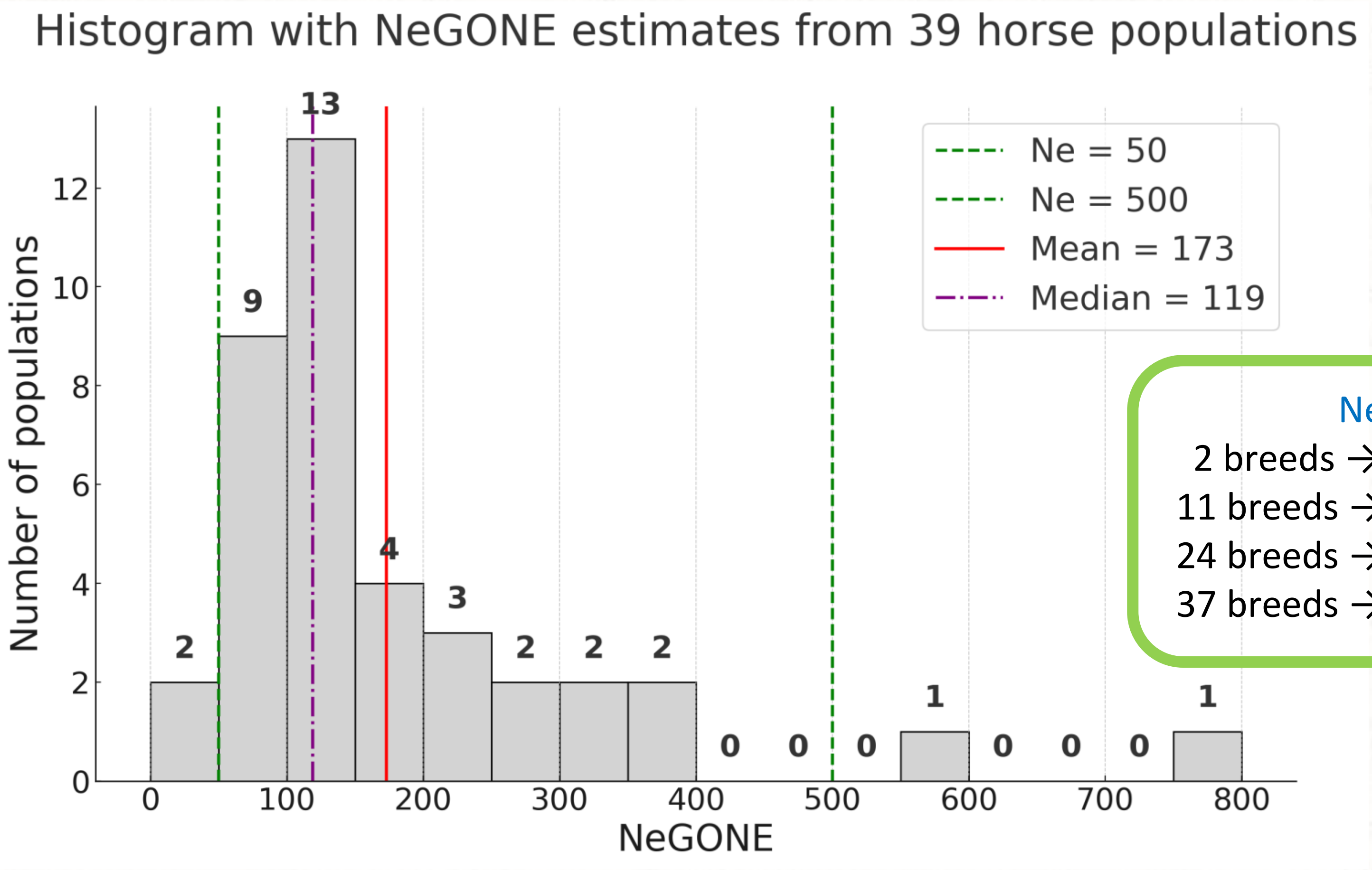
- ✓ In addition, disregarding the genomic regions that are involved in inversions can also remove biases in the estimates of *Ne*.



Genomic characterization and population structure of Croatian Arabian horse

Nikola Raguz [a], Nidal Korabi [b,*], Boris Lukić [a,*], Ivana Drzaic [c], Lubos Vostry [d], Nina Moravcikova [e], Ino Curik [c], Radovan Kasarda [e], Vlatka Cubric-Curik [c]




Histogram with NeGONE estimates from 39 horse populations

NeGONE

2 breeds → Ne < 50 (5%)

11 breeds → Ne < 100 (28%)

24 breeds → Ne < 150 (62%)

37 breeds → Ne < 500 (95%)

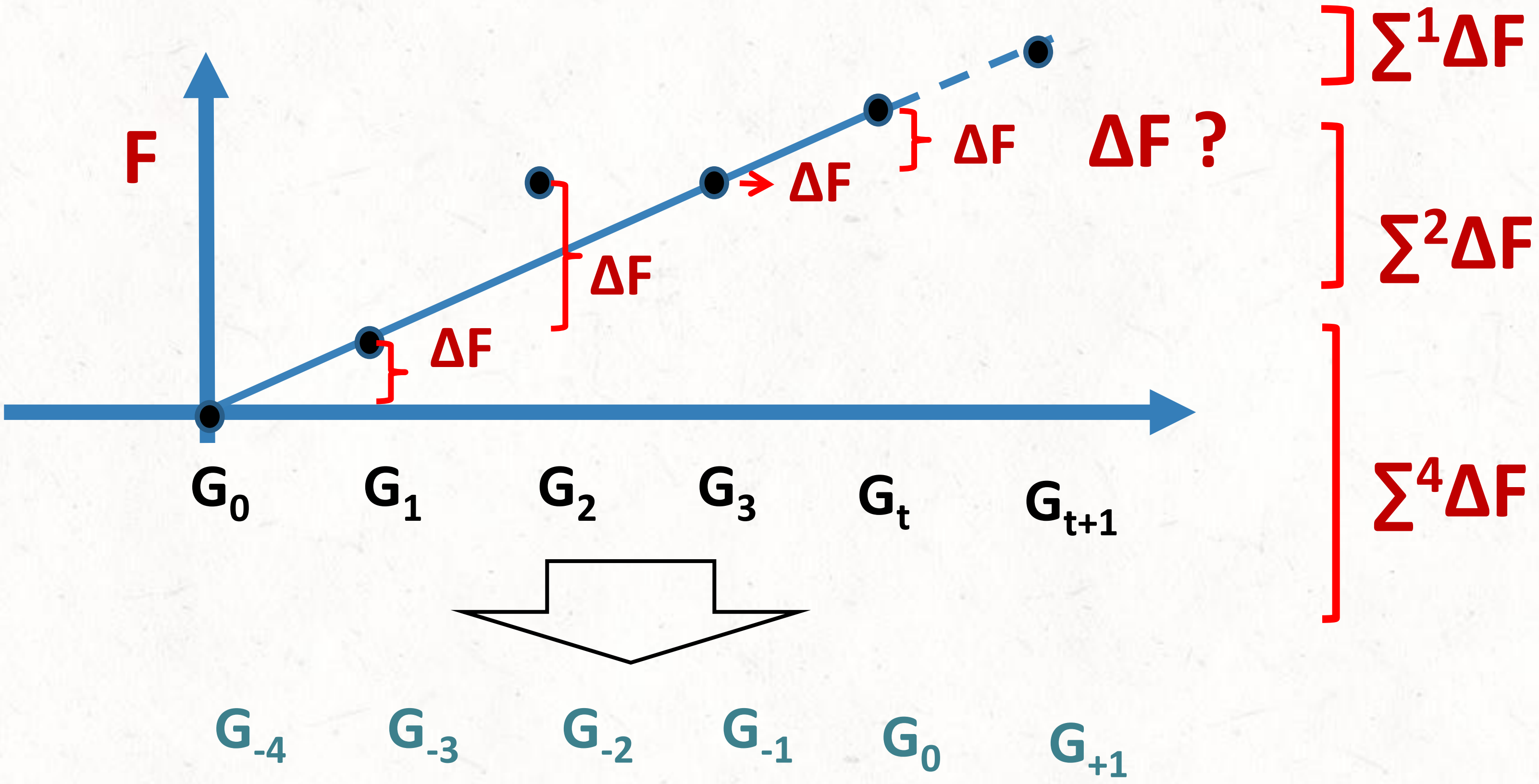
Effective inbreeding population size (N_{eI} or N_{eF}):

The effective population size (Ne) of a **real population** X is the size of a **hypothetical ideal population** (Wright-Fisher) that will result in the - **same amount of change in inbreeding level -** as in the **real (actual) population considered**.

NeI is based on the rate of inbreeding in a finite population and is defined as the size of an idealized population that would experience the same rate of increase in the probability of autozygosity per generation.

$$\Delta F = 1/(2Ne) \rightarrow \quad Ne = 1/(2\Delta F)$$

$\Delta F = 1/(2Ne)$ → $Ne = 1/(2\Delta F)$



F

ΔF ΔF ΔF ΔF ΔF ?

$\Sigma^1 \Delta F$

$\Sigma^2 \Delta F$

$\Sigma^4 \Delta F$

G_0 G_1 G_2 G_3 G_t G_{t+1}

G_{-4} G_{-3} G_{-2} G_{-1} G_0 G_{+1}

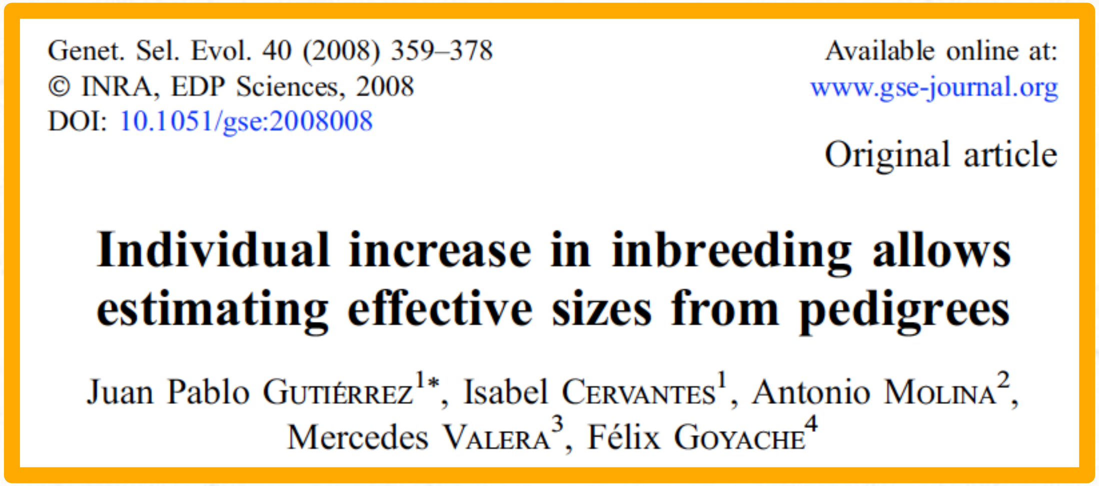
$$\Delta F = \frac{1}{2N_e} \Longrightarrow N_{el} = 1/2\Delta F$$

$$F_t = \overset{\Delta F}{\boxed{\frac{1}{2N_e}}} + \left(1 - \overset{\Delta F}{\boxed{\frac{1}{2N_e}}}\right) F_{t-1} \Longrightarrow \boxed{\Delta F = \frac{1}{2N_e} = \frac{F_t - F_{t-1}}{1 - F_{t-1}}}$$

$$F_t = 1 - (1 - \Delta F)^{t\,-1} \Longrightarrow \Delta F_i = 1 - \sqrt[t-1]{1 - F_i}$$

t - 1 to correct for no-selfing by delay of one generation

Gutiérrez et al., 2008 GSE, 2009 JABG



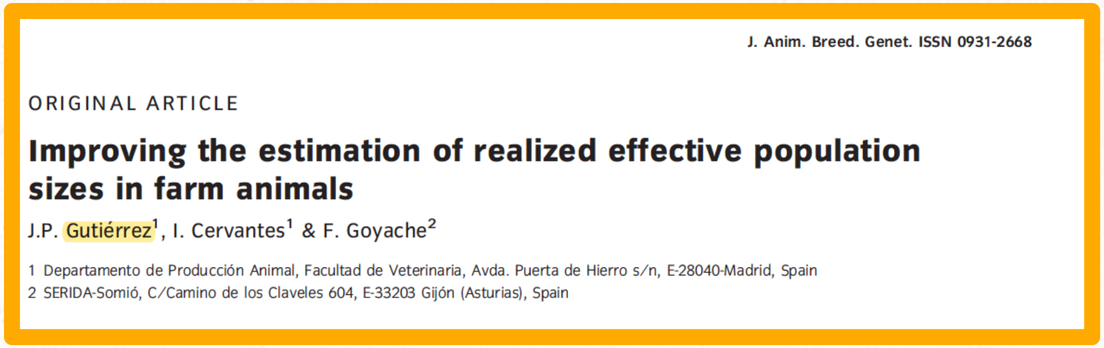
Genet. Sel. Evol. 40 (2008) 359–378
© INRA, EDP Sciences, 2008
DOI: 10.1051/gse:2008008

Available online at:
www.gse-journal.org

Original article

Individual increase in inbreeding allows estimating effective sizes from pedigrees

Juan Pablo Gutiérrez[1*], Isabel Cervantes[1], Antonio Molina[2], Mercedes Valera[3], Félix Goyache[4]



J. Anim. Breed. Genet. ISSN 0931-2668

ORIGINAL ARTICLE

Improving the estimation of realized effective population sizes in farm animals

J.P. Gutiérrez[1], I. Cervantes[1] & F. Goyache[2]

1 Departamento de Producción Animal, Facultad de Veterinaria, Avda. Puerta de Hierro s/n, E-28040-Madrid, Spain
2 SERIDA-Somió, C/Camino de los Claveles 604, E-33203 Gijón (Asturias), Spain

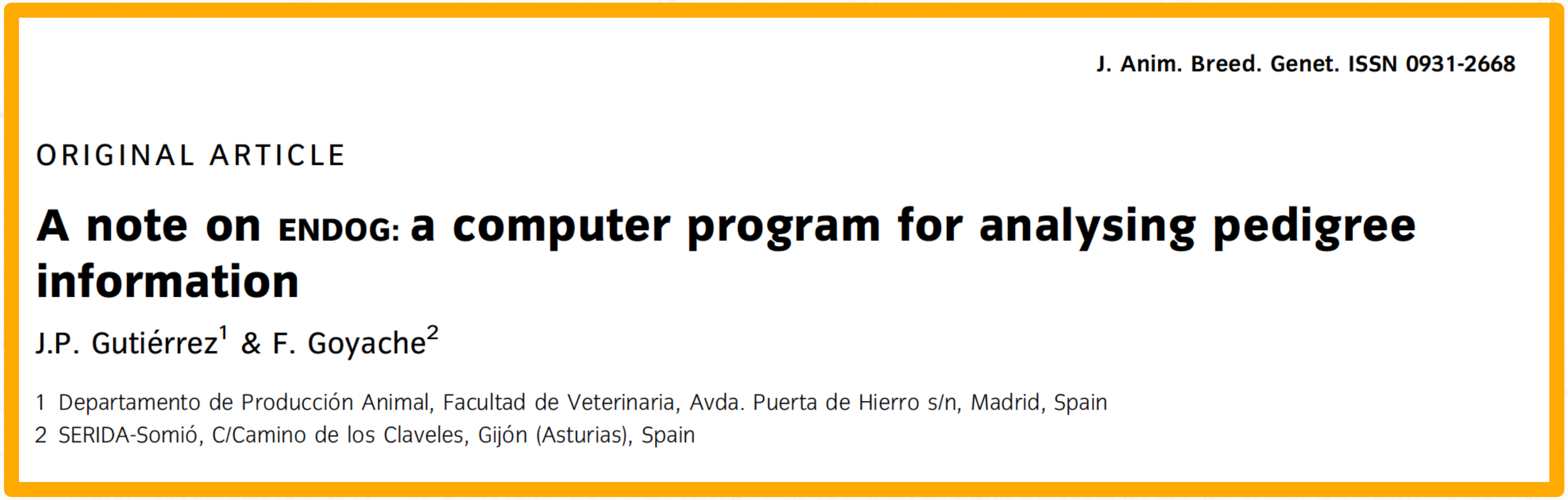
$$F_t = 1 - (1 - \Delta F)^t$$

$$\Delta F_i = 1 - \sqrt[t]{1 - F_i}$$

$$\overline{N}_e = \frac{1}{2\overline{\Delta F}}$$

$$\Delta F_i^* = 1 - \sqrt[t_i - 1]{1 - F_i}$$

$$\overline{N_e^*} = \frac{1}{2\overline{\Delta F^*}}$$



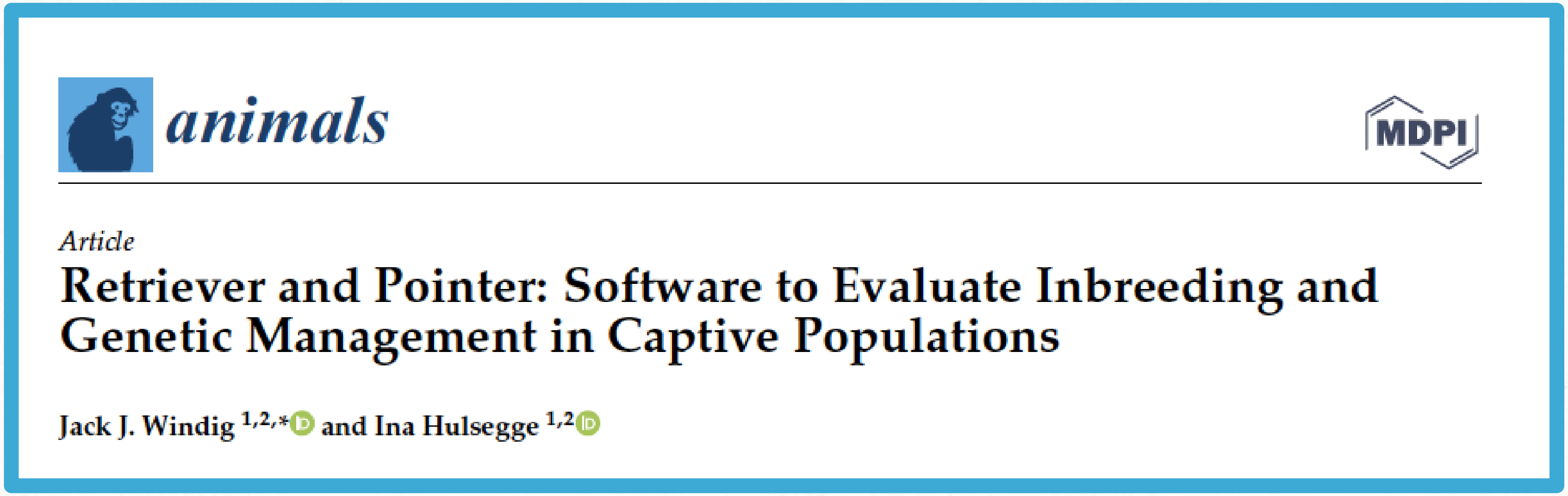
J. Anim. Breed. Genet. ISSN 0931-2668

ORIGINAL ARTICLE

A note on ENDOG: a computer program for analysing pedigree information

J.P. Gutiérrez[1] & F. Goyache[2]

1 Departamento de Producción Animal, Facultad de Veterinaria, Avda. Puerta de Hierro s/n, Madrid, Spain
2 SERIDA-Somió, C/Camino de los Claveles, Gijón (Asturias), Spain



animals

MDPI

Article

Retriever and Pointer: Software to Evaluate Inbreeding and Genetic Management in Captive Populations

Jack J. Windig [1,2,*] and Ina Hulsegge [1,2]

ROH inbreeding effective population size, Ne_{FROH}

Gutiérrez et al., 2008 GSE, 2009 JABG

$$\Delta F_i^* = 1 - \sqrt[t_i - 1]{1 - F_i}$$

$$\overline{N_e^*} = \frac{1}{2\overline{\Delta F^*}}$$

$$\Delta F = \frac{F_t - F_{t-1}}{1 - F_{t-1}}$$

$$** \; Ne_{FROH>LMb} \Rightarrow 1 - \sqrt[\#generations\;as\;f(LROH)]{1 - F_{ROH}}$$

Empirical population with informative pedigree

The Old Kladruber horse (OKH)

Informative pedigree

Individuals in pedigree: 9288 (1729 – 2018)
Tracing up to 55 generations

The **Kladruber** (Czech *Starokladrubský kůň*) is the oldest Czech horse breed and one of the world's oldest horse breeds, bred for more than 400 years.

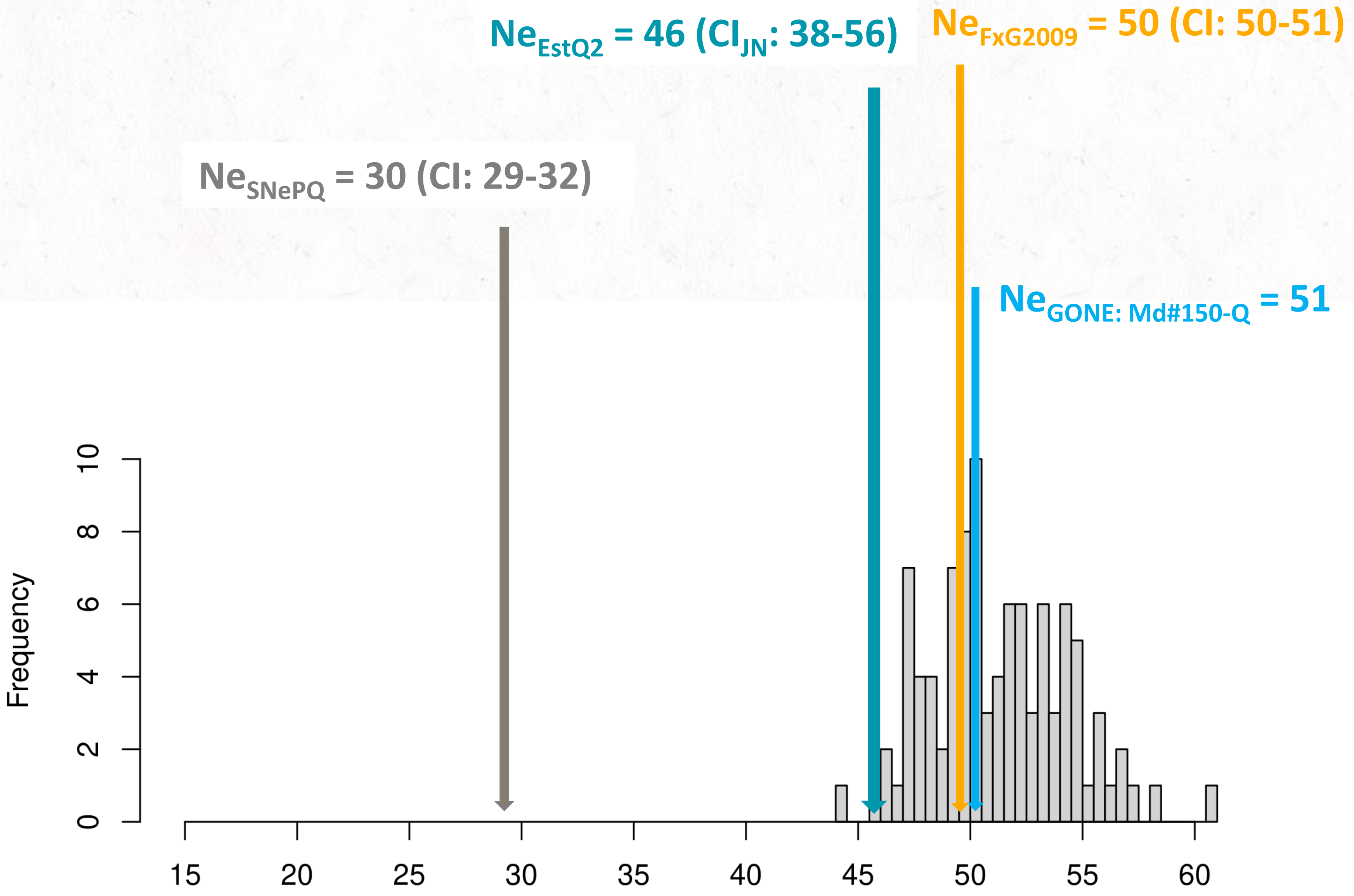
Calculation based data set

Genotyped animals (n = 215; 1994 - 2014)

Q: **60645** autosomal SNPs

# ECG: 15.9 (13.5 – 17.4)

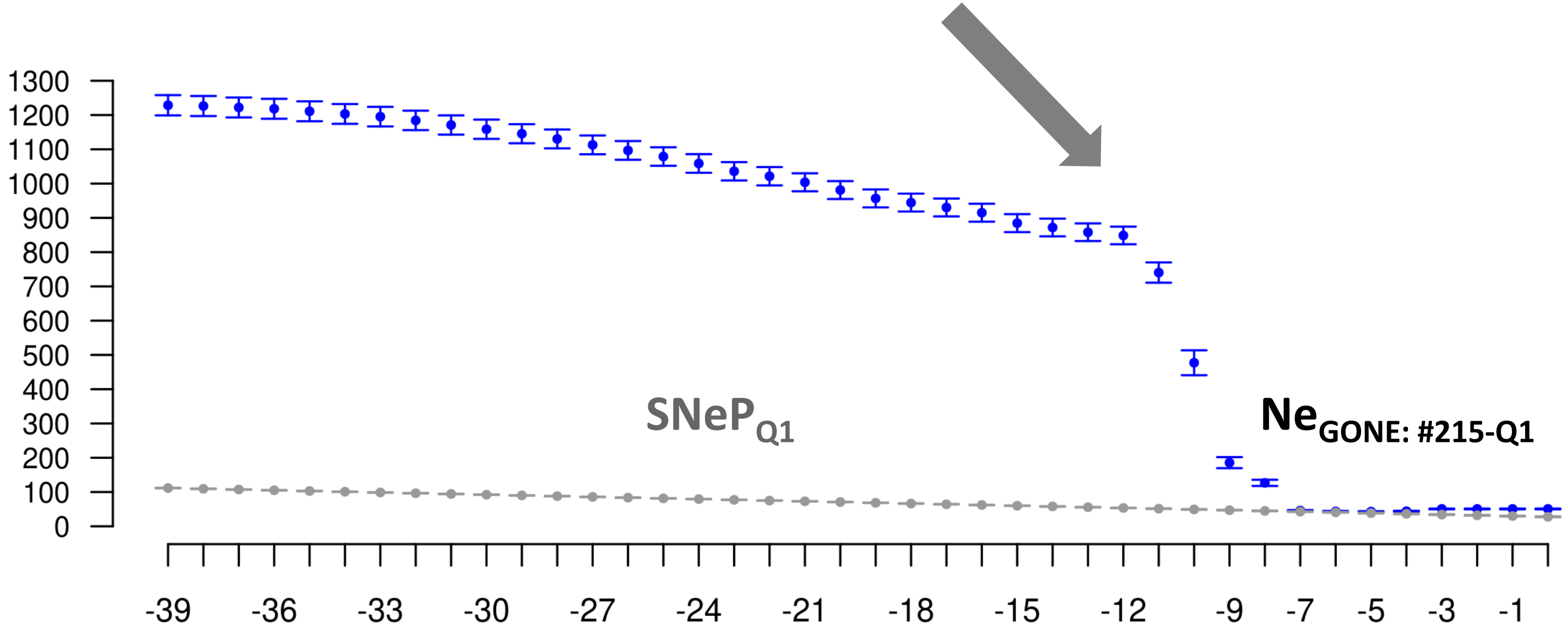
Results of the current effective population size: OKH



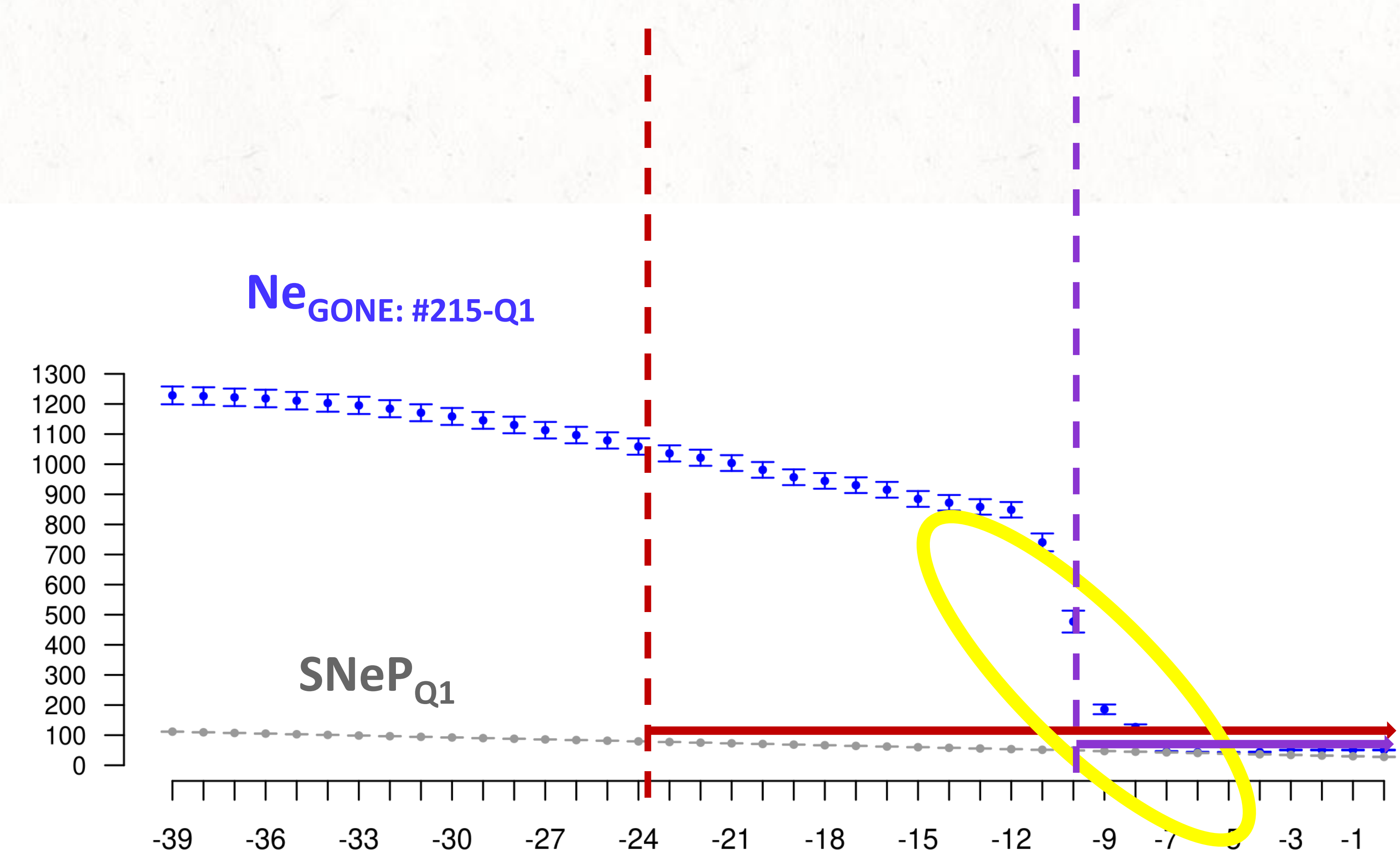
Results of the historical effective population size: OKH

Generation interval: 10 to 11 years

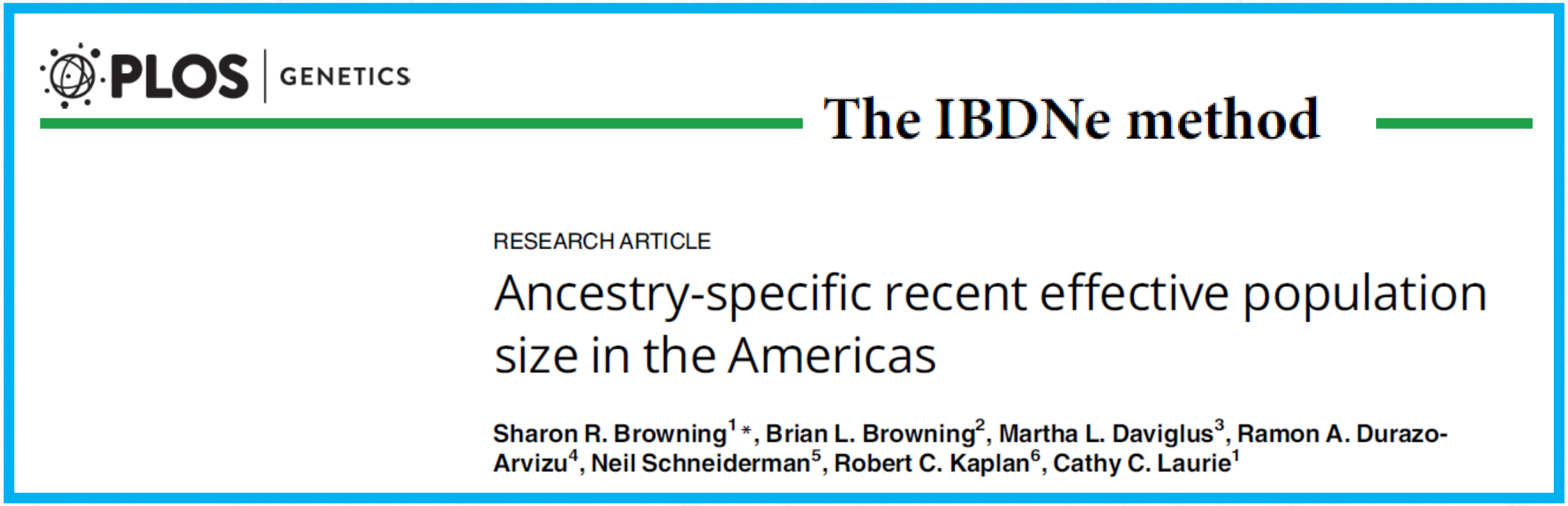
≈ 100 years back / fall of Austro-Hungarian Monarchy / drop in the number of horses



WARNING: Sensitivity of Ne estimation in declining population



The IBDNe method



PLOS | GENETICS

RESEARCH ARTICLE

Ancestry-specific recent effective population size in the Americas

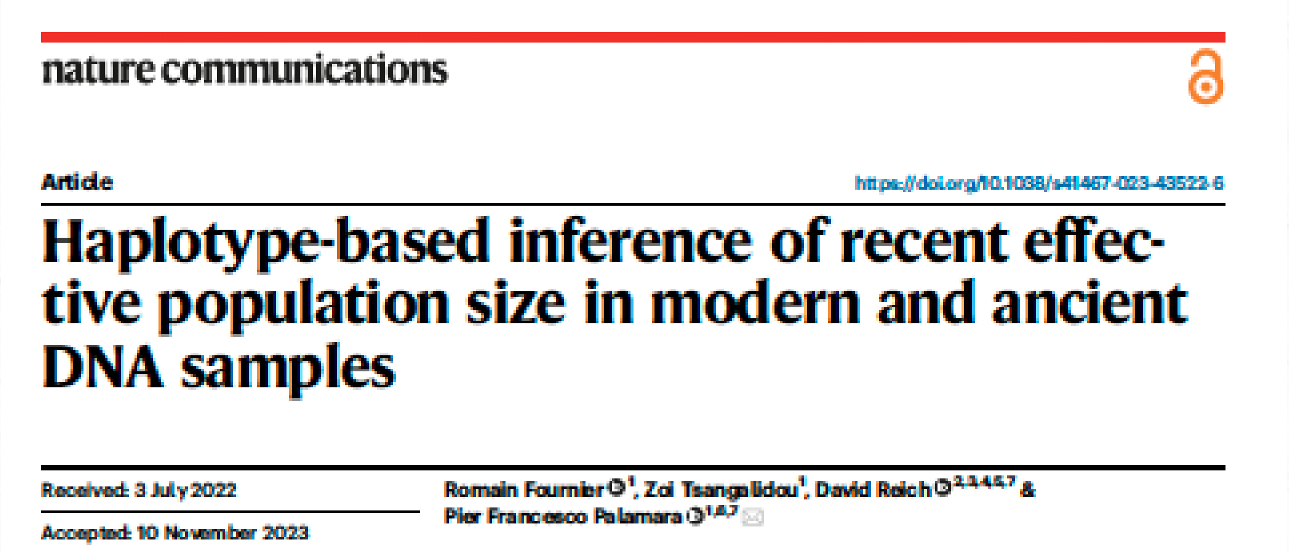
Sharon R. Browning[1] *, Brian L. Browning[2], Martha L. Daviglus[3], Ramon A. Durazo-Arvizu[4], Neil Schneiderman[5], Robert C. Kaplan[6], Cathy C. Laurie[1]

Some methods that allow for the estimation of the recent trajectory of *Ne* are based on identity by descent (IBD) segments.

Analogously to the relationship of the recombination fraction between SNPs and time in the LD method, longer IBD fragments give information on the *Ne* in more recent generations, as mentioned above.

Both SNP array or sequencing data can be used by these methods.

The software **IBDNe**, developed by Browning and Browning (2015), implements this method and the proper generation range **to be applied is between generations 4 and 200 in the past**.



nature communications

Article

https://doi.org/10.1038/s41467-023-43522-6

Haplotype-based inference of recent effective population size in modern and ancient DNA samples

Received: 3 July 2022

Accepted: 10 November 2023

Romain Fournier[1], Zoi Tsangalidou[1], David Reich[2,3,4,5,7] & Pier Francesco Palamara[1,6,7]

HapNe

HapNe (Fournier et al. 2023) implements both an IBD method and a LD method, enabling the use of either **phased or unphased data**, even though the IBD method is encouraged to be used with phase information for accurate IBD identification.

One novelty of this method is the **capacity to use**, at the same time, **individuals from different time points**, which can be frequent in ancient DNA samples.

Up to date the method has only been tested and applied to human data!

**Thank you
for your
attention!**