Along the Genome - Crossing over, Recombination, Linkage Disequilibrium, and Effective Population Size

Henner Simianer
Department of Animal Sciences
Animal Breeding and Genetics Group
Georg-August-University Göttingen, Germany

Overview

- Metrics and maps
- Crossing over, recombination and interference
- Mapping functions

- Linkage disequilibrium
- Haplotype blocks

- Estimating Ne from LD

Physical map

Length of the nuclear DNA counted in
basepairs (bp), kilobasepairs (Kbp), megabasepairs (Mbp)

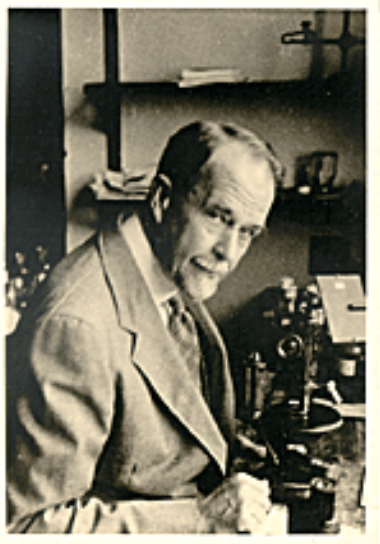
...GTCTTATTCTATATATATAGCCTTTACTATTCTTATGGGCTAGTGGTGCT...
...CAGAATAAGATATATATATCGGAAATGATAAGAATACCCGATCACCACGA...

Physical map length of the
- Mammalian genome: 3 x 10^9 bp = 3 mio kbp = 3000 mbp
- Chicken genome: 1.2 x 10^9
- Barley genome: 5.3 x 10^9 bp

Requires a high quality reference sequence and a conserved genome

Genetic map

In the early days of genetics the relative position of qualitative genes to each other could only be assessed by counting recombinations



Thomas Hunt Morgan (1866-1945)
Nobel prize for Medicine in 1933

1900 Re-discovery of Mendel‘s laws

1903 Sutton publishes the idea, that chromosomes are the physical carrieres of inheritance

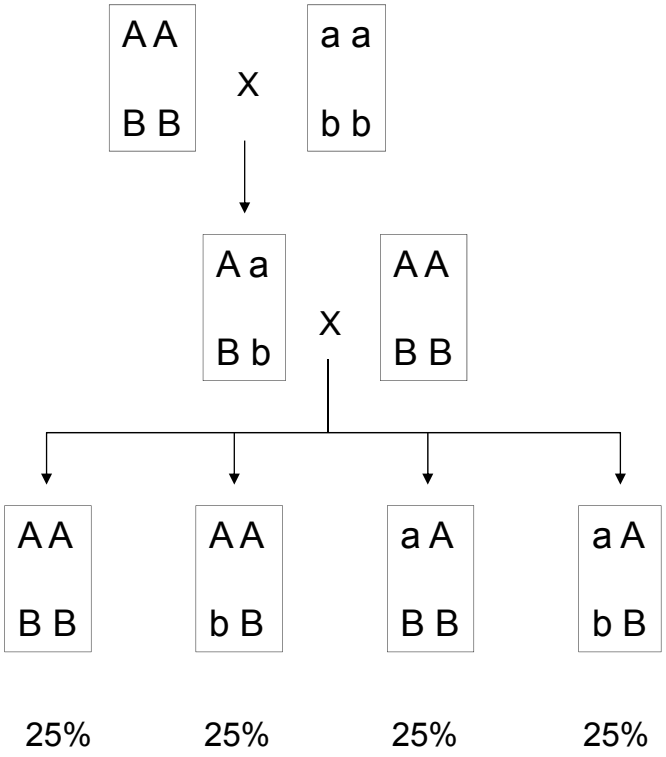
1905 The term ‚genetics‘ is coined for the new discipline by Sturtevant

1907 T.H. Morgan starts systematic experiments with *Drosophila melanogaster* confirming the theory of chromosomal inheritance

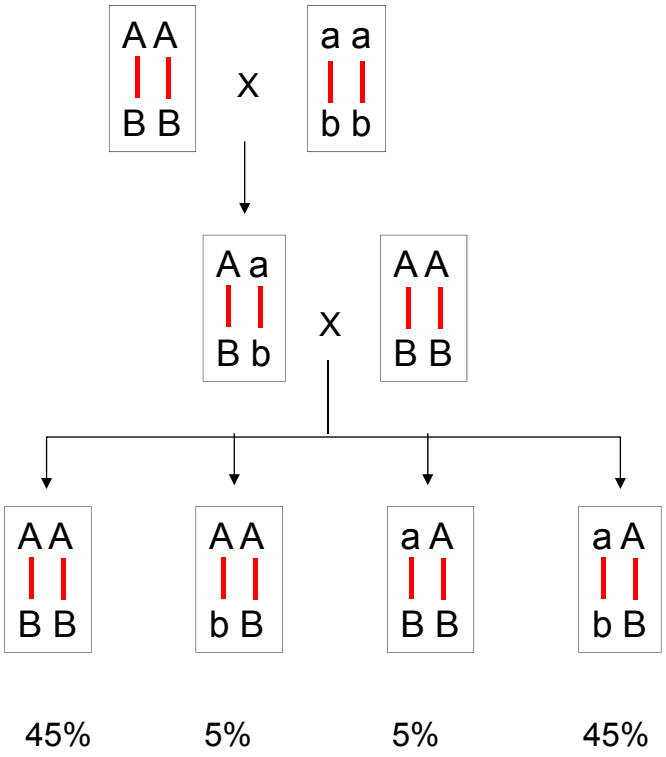
1911 First genetic map for *D. melanogaster*

1919 ‚The physical basis of Heredity‘ published

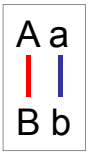
Independent segregation (according to Mendel‘s laws)



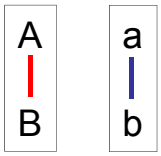
Non-independent segregation - linkage



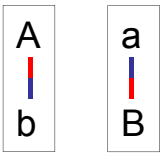
Non-independent segregation - linkage



non-recombined gametes



recombined gametes

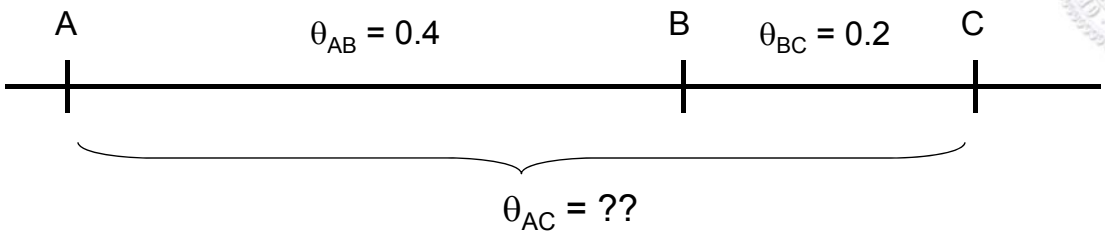


Recombination rate θ = $\frac{\text{number of recombined gametes}}{\text{total number of gametes}}$ $0 \le \theta \le 0.5$

In general:

- the recombination rate of two loci on different chromosomes is $\theta = 0.5$
- the recombintaion rate of two loci in the same chromosome is $0 \le \theta \le 0.5$
- the recombination rate of two loci on the same chromosome increases monotonically with distance on the chromosome
- Recombination rate is a probability, rules of probability calculus can be applied

But:



A and C are recombined if:

- A and B are recombined and B and C are not recombined
- A and B are not recombined and B and C are recombined

$$\theta_{AC} = \theta_{AB}(1-\theta_{BC}) + (1-\theta_{AB})\theta_{BC}$$
$$= 0.4 \times 0.8 + 0.6 \times 0.2$$
$$= 0.44$$

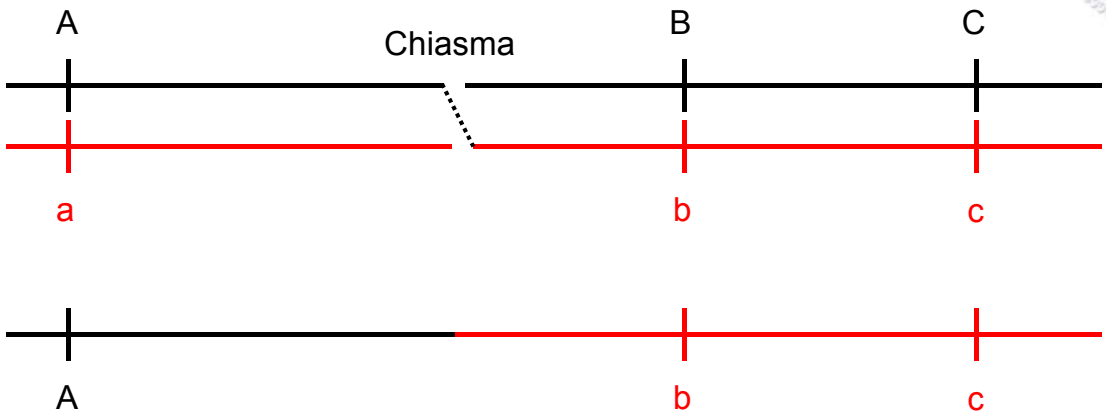
⇨ Recombination rates are not additive!

A and C are not recombined if:

- both A and B and B and C are not recombined
- both A and B and B and C are recombined

$$1-\theta_{AC} = \theta_{AB}\theta_{BC} + (1-\theta_{AB})(1-\theta_{BC})$$
$$= 0.4 \times 0.2 + 0.6 \times 0.8$$
$$= 0.56$$

Recombination and crossing over



- Two loci recombine, if between them an odd number of crossing-over events takes place
- Crossing-overs are usually not directly observable
- Crossing-over is not an ‚accident' in meiosis, but is a mechanism to generate and maintain variability, crossing-over activities have increased in domesticated species relative to the wild ancestors

Genetic map length – Morgan

Definition of the Morgan unit (named after T. H. Morgan)

For a chromosome segment of length one Morgan we expect one crossing over in one meiosis

1 Morgan (M) = 100 centiMorgan (cM)

Simple assumptions:

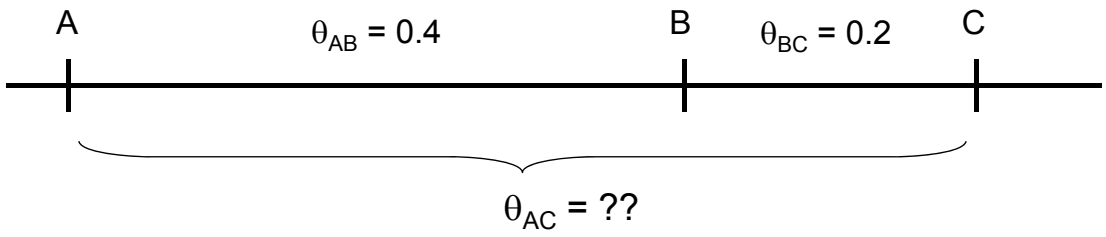
Crossing-overs are random events and follow a Poisson distribution

Crossing-over events on the same chromosome in the same meiosis are independent (no genetic interference)

Rule of thumb for mammals:

physical genome length:	3000 Mbp
genetic map length:	30M = 3000 cM
⇨ on average (!):	1 cM = 1 Mbp = 1% recombination ($\theta = 0.01$)

Genetic map length (in Morgan) is based on expected values and is additive across segments



⇨ ‚Translate' recombination rates of segments in Morgan

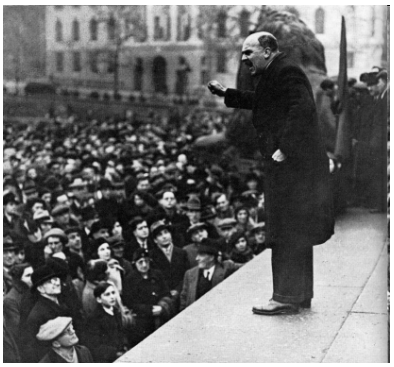
⇨ Add up the genetic length of all segments

⇨ ‚Translate back' from Morgan to recombination rate

Mapping functions ‚translate' between scales

The Haldane mapping function

Underlying assumptions:
⇨ crossing over events are Poisson distributed
⇨ no genetic interference



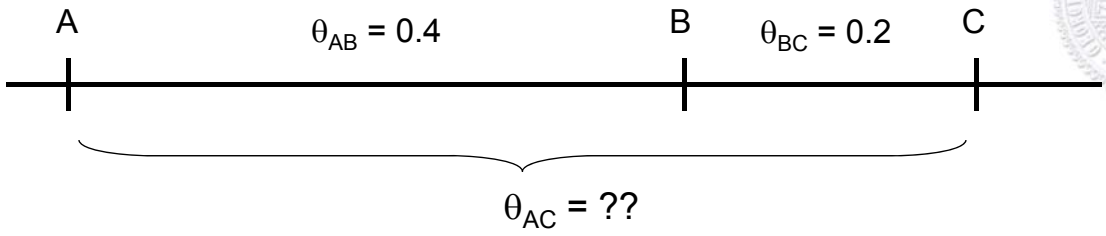
J.B.S. Haldane
1892 - 1964

Recombination rate θ → genetic length x

$$x = -\frac{1}{2}\ln(1-2\theta)$$

Genetic length x → recombination rate θ

$$\theta = \frac{1}{2}(1-e^{-2x})$$



θ_{AC} = ??

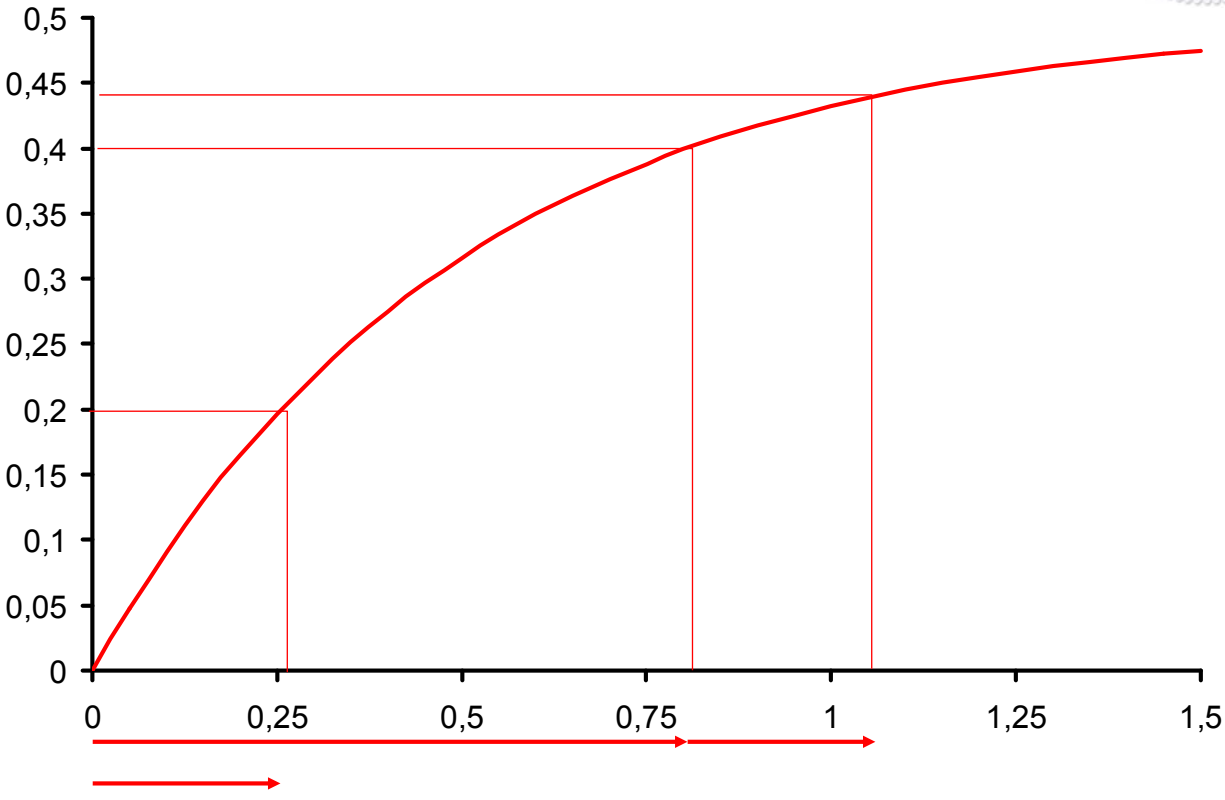
$$x_{AB} = -\frac{1}{2}\ln(1-2\theta_{AB}) = -\frac{1}{2}\ln(1-2\times 0.4) = 0.8047$$

$$x_{BC} = -\frac{1}{2}\ln(1-2\theta_{BC}) = -\frac{1}{2}\ln(1-2\times 0.2) = 0.2554$$

$$x_{AC} = x_{AB} + x_{BC} = 0.8047 + 0.2554 = 1.0601$$

$$\theta_{AC} = \frac{1}{2}(1-e^{-2x_{AC}}) = \frac{1}{2}(1-e^{-2\times 1.0601}) = 0.44$$

The Haldane mapping function



Some further aspects:

- ⇨ Usually the recombination activity in the homogametic sex (females in mammals, males in birds) is increased, therefore the respective map is on average longer (in humans female : male map length ~ 2 : 1)
- ⇨ Recombination activity varies along the genome (hot spots, cold spots) and is a complex process
- ⇨ Recombination activity is genetically regulated (family differences, selection experiments, domestication)
- ⇨ On short chromosomes (as e.g. the microchromosomes in chicken or the pseudoautosomal region on the heterosomes) one often assumes an ‚obligatory chiasma', so that the genetic length is much larger than the physical length

Linkage Disequilibrium (LD)
Gametic phase disequilibrium

2 biallelic loci (SNPs)

Locus A: $P(1) = p_A$ $P(0) = 1- p_A$
Locus B: $P(1) = p_B$ $P(0) = 1- p_B$

In equilibrium: $P(1 – 1) = p_A p_B$

If $P(1 – 1) \neq p_A p_B$ → linkage disequilibrium (LD)

Measure for LD: r^2 = squared correlation between 0/1 coded genotypes at two loci ($0 \leq r^2 \leq 1$)

Another look at r^2

Consider two loci A,B on one gamete, each with possible (random) realisation

- at locus A: x_A=1 (with probability p_A) and x_A=0 (with probability 1-p_A)
- at locus B: x_B=1 (with probability p_B) and x_B= 0 (with probability 1-p_B)

We estimate allele frequencies at both loci and the joint haplotype frequency from a sample of size N

$$\hat{p}_A = \frac{N_{(x_A=1)}}{N} = \frac{\sum x_A}{N}$$

$$\hat{p}_B = \frac{N_{(x_B=1)}}{N} = \frac{\sum x_B}{N}$$

$$\hat{p}_{AB} = \frac{N_{(x_A=1 \cup x_B=1)}}{N} = \frac{\sum x_A x_B}{N}$$

Another look at r^2

What is the correlation r between the realisations of the 0/1 random variable at the two loci?

Remember: $$r = \frac{Cov(x,y)}{\sqrt{Var(x)Var(y)}} = \frac{\sum xy - \frac{\sum x \sum y}{N}}{\sqrt{\left(\sum x^2 - \frac{\left(\sum x\right)^2}{N}\right)\left(\sum y^2 - \frac{\left(\sum y\right)^2}{N}\right)}}$$

$$\sum xy - \frac{\sum x \sum y}{N} = N\hat{p}_{AB} - \frac{N\hat{p}_A \times N\hat{p}_B}{N} = N(\hat{p}_{AB} - \hat{p}_A \times \hat{p}_B)$$

$$\sum x^2 - \frac{\left(\sum x\right)^2}{N} = N\hat{p}_A - \frac{(N\hat{p}_A)^2}{N} = N(\hat{p}_A - \hat{p}_A^2) = N(\hat{p}_A \times (1 - \hat{p}_A))$$

$$\sum y^2 - \frac{\left(\sum y\right)^2}{N} = N\hat{p}_B - \frac{(N\hat{p}_B)^2}{N} = N(\hat{p}_B - \hat{p}_B^2) = N(\hat{p}_B \times (1 - \hat{p}_B))$$

Another look at r^2

What is the correlation r between the realisations of the 0/1 random variable at the two loci?

$$r = \frac{Cov(x,y)}{\sqrt{Var(x)Var(y)}} = \frac{\sum xy - \frac{\sum x \sum y}{N}}{\sqrt{\left(\sum x^2 - \frac{\left(\sum x\right)^2}{N}\right)\left(\sum y^2 - \frac{\left(\sum y\right)^2}{N}\right)}}$$

$$= \frac{N(\hat{p}_{AB} - \hat{p}_A \times \hat{p}_B)}{\sqrt{N(\hat{p}_A \times (1-\hat{p}_A)) \times N(\hat{p}_B \times (1-\hat{p}_B))}}$$

$$= \frac{(\hat{p}_{AB} - \hat{p}_A \times \hat{p}_B)}{\sqrt{(\hat{p}_A \times (1-\hat{p}_A)) \times (\hat{p}_B \times (1-\hat{p}_B))}}$$

$$r^2 = \frac{(\hat{p}_{AB} - \hat{p}_A \times \hat{p}_B)^2}{\hat{p}_A \times (1-\hat{p}_A) \times \hat{p}_B \times (1-\hat{p}_B)}$$

- ⇨ r^2 is the squared correlation between allele realisations on the same gamete at two loci
- ⇨ this is also true for any other coding of alleles

A simple example

		A = 1	A = 0	
B	1	4	2	6
	0	1	3	4
		5	5	10

$$\hat{p}_A = 5/10 = 0.5$$
$$\hat{p}_B = 6/10 = 0.6$$
$$\hat{p}_{AB} = 4/10 = 0.4$$

$$r^2 = \frac{(\hat{p}_{AB} - \hat{p}_A \times \hat{p}_B)^2}{\hat{p}_A \times (1-\hat{p}_A) \times \hat{p}_B \times (1-\hat{p}_B)} = \frac{(0.4 - 0.5 \times 0.6)^2}{0.5 \times 0.5 \times 0.6 \times 0.4} = 0.167$$

A pipeline to estimate LD

Genotyping results (e.g. from Illumina SNP array)

↓ Quality control, filtering (e.g. with PLINK)

Filtered unphased genotypes

↓ Phasing and imputing (e.g. with Beagle)

↓ Recoding (e.g. with R)

Phased and imputed genotypes

↓ Correlations between allelic states (e.g. with R)

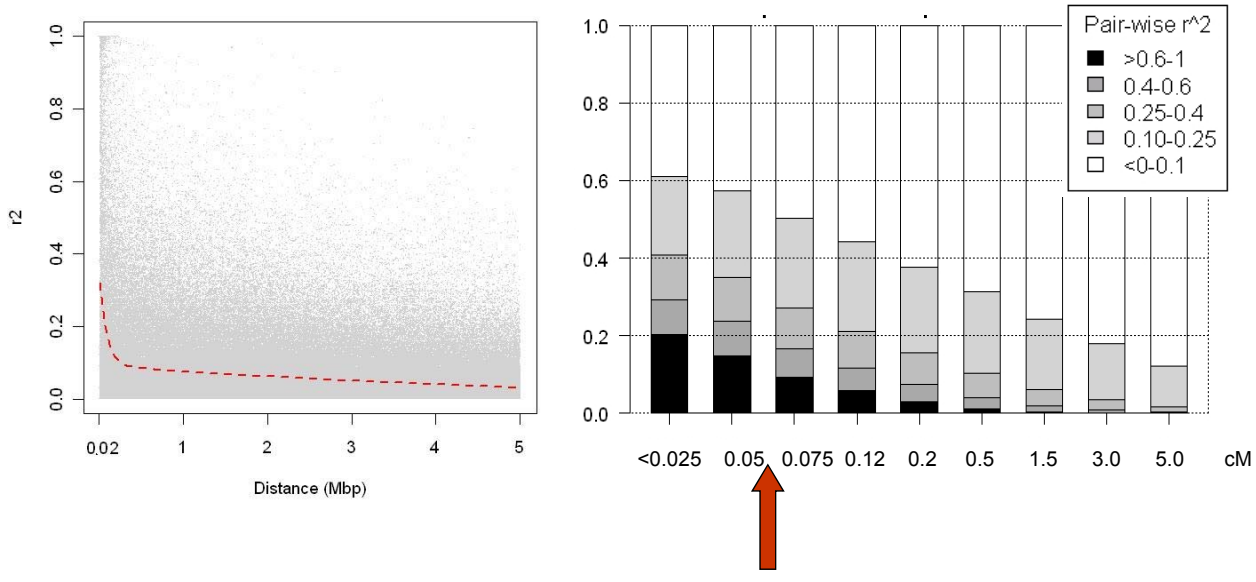
r^2 as a function of marker distance

↓

Summary statistics and visualisation

LD as a function of physical and genetic distance

(Holstein Friesian population, 56k SNP array, Qanbari et al., 2010)



~ average density with 50k SNPs

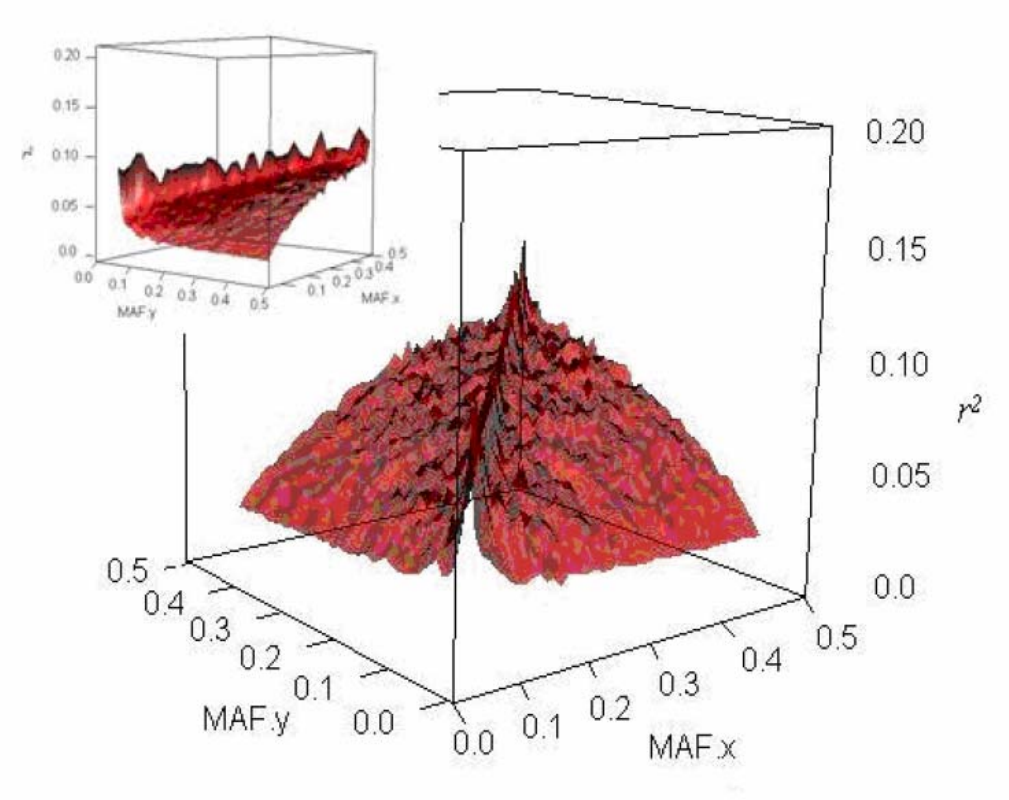
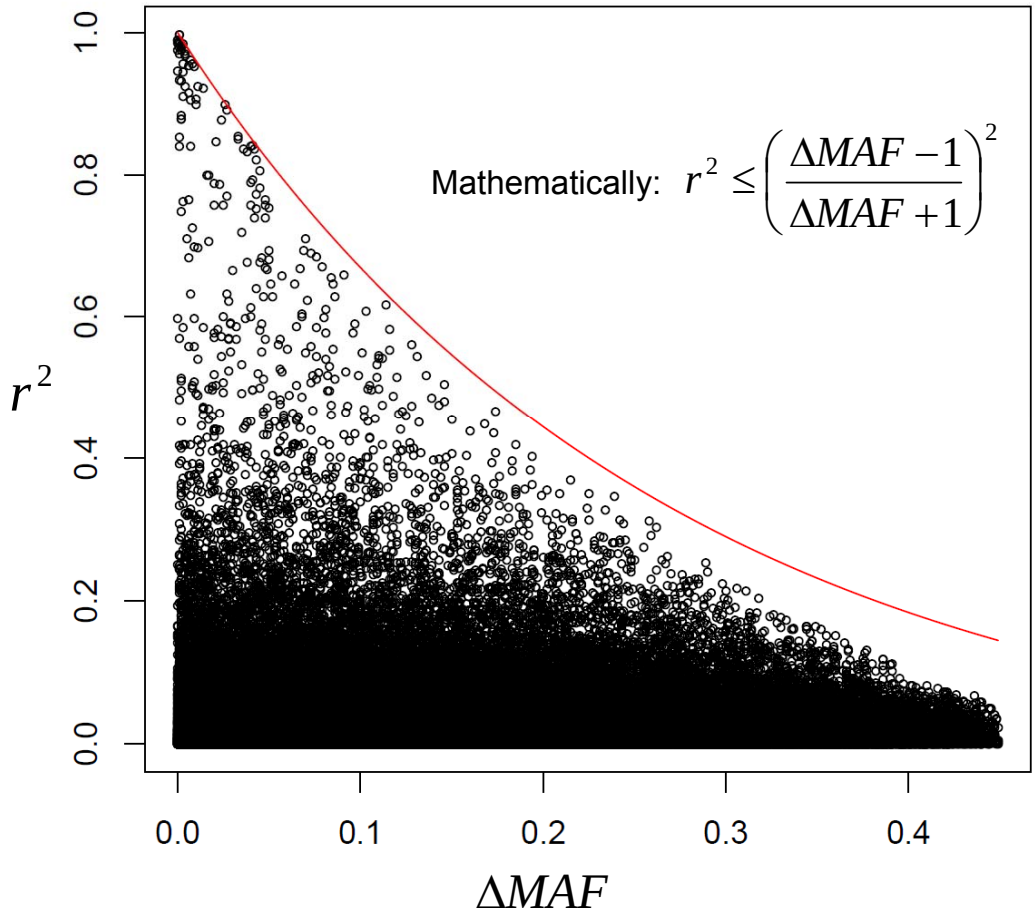


Figure 7. The prospective plot depicts the decay of LD with allele frequencies of SNP pairs. r^2 means were calculated for 45 bins of each 0.01 allele frequency.

Qanbari et al. (2010) [23]



For gene mappers:

- ⇨ Two loci can only be in high LD, if they have similar allele frequencies
- ⇨ To map rare disease alleles by association mapping you need rare marker alleles
- ⇨ A too strict filtering for MAF is therefore questionable in some applications (esp. GWAS)
- ⇨ one of the reasons for launching the ‚1000 genomes project'

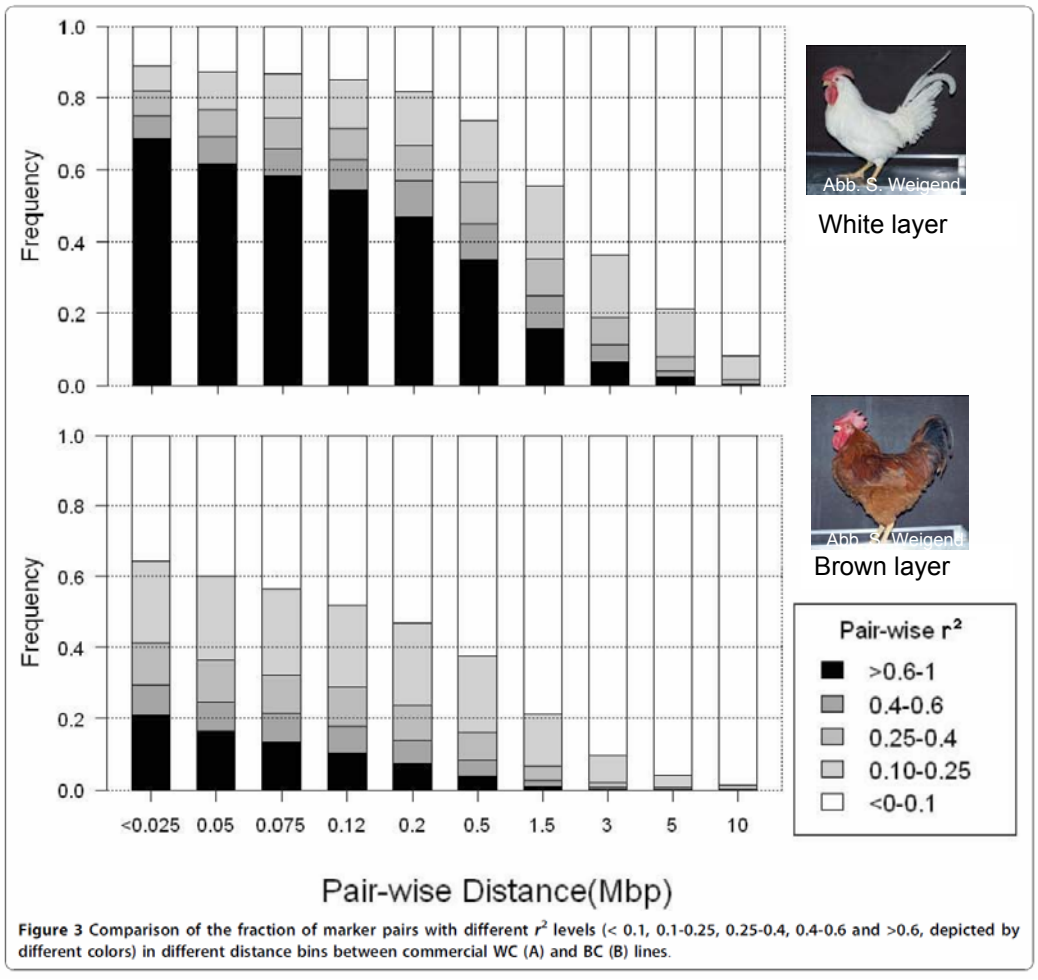
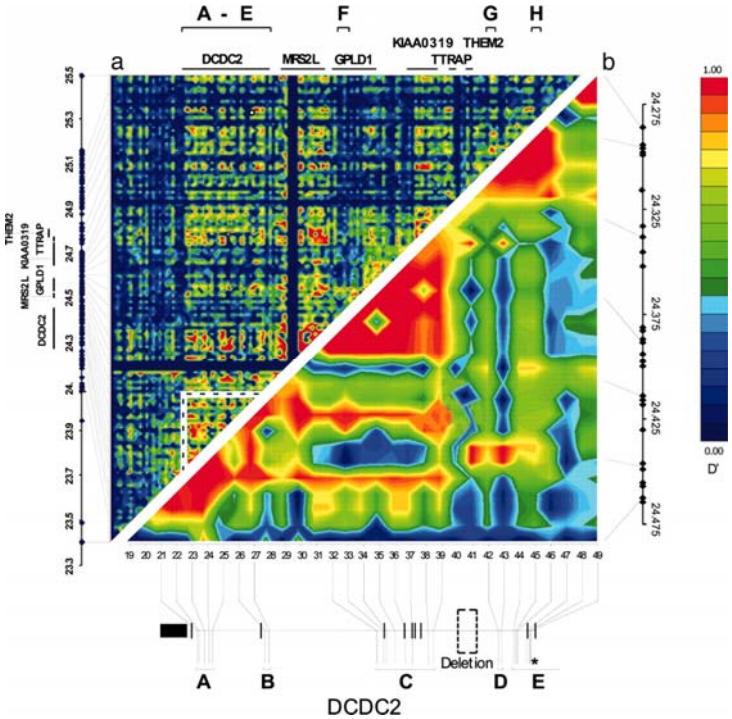
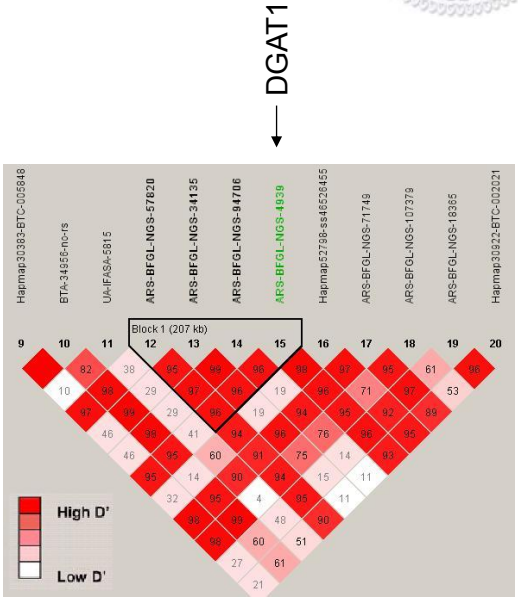


Figure 3 Comparison of the fraction of marker pairs with different r^2 levels (< 0.1, 0.1-0.25, 0.25-0.4, 0.4-0.6 and >0.6, depicted by different colors) in different distance bins between commercial WC (A) and BC (B) lines.

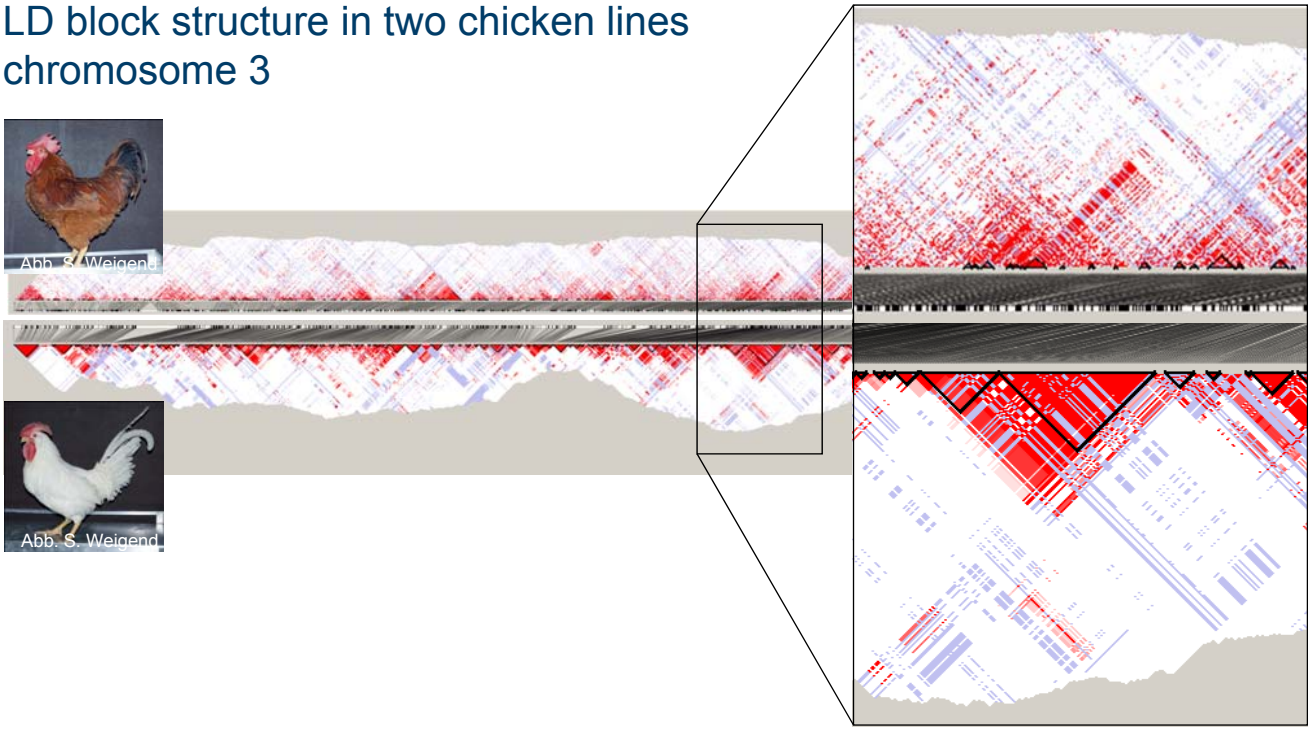
LD has a block structure





Meng H. et.al. PNAS 2005;102:17053-17058

LD block structure in two chicken lines chromosome 3

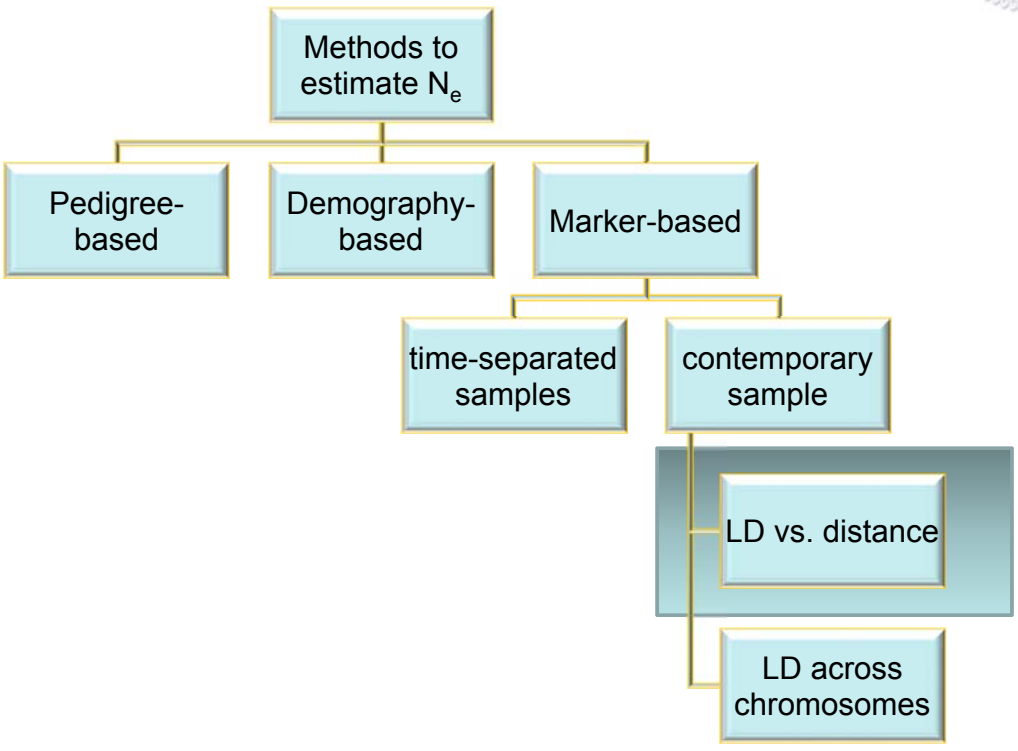


What are the mechanisms generating LD?

- ⇨ Linkage
- ⇨ Selection
- ⇨ Stratification/admixture
- ⇨ Genetic drift
- ⇨ Inbreeding

} ‚spurious LD'
within and across chromosomes

Estimation of effective population size from linkage disequilibrium



(Linkage disequilibrium) effective population size N_e

Definition: The (LD) effective population size N_e of a real population X with an observed LD for a given interval length is the size of a hypothetic ideal population that in an equilibrium state would display the same pattern of LD for the same interval length as observed in the real population

The underlying principle:

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

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

The equilibrium LD of two loci in a population of finite size

Sved (1971) $E(r^2) = \frac{1}{1+4Nc}$

where

r^2 is the squared correlation between gametic states at the two loci

c is the distance of loci in Morgan

N is the size of the population

Based on Sved's recursion formula:

Development of r^2 from generation T to $T+1$

$$E(r^2_{T+1}) = \left(1 - \frac{1}{2N}\right)(1-c)^2 E(r^2_T) + \frac{1}{2N}(1-c)^2 \xrightarrow{T \to \infty} E(r^2_\infty) = \frac{1}{1+4Nc}$$

A closer look at Sved's (1971) derivation

BAD NEWS...

No mathematically valid derivation for this recursion formula exists.

From John Sved's homepage: „This was all introduced in a very messy way, and was not understood by anyone, evidently including myself."

GOOD NEWS...

Simulation results indicate that the formula works reasonably well

Two problems with Sved's approach to estimation of N_e from LD

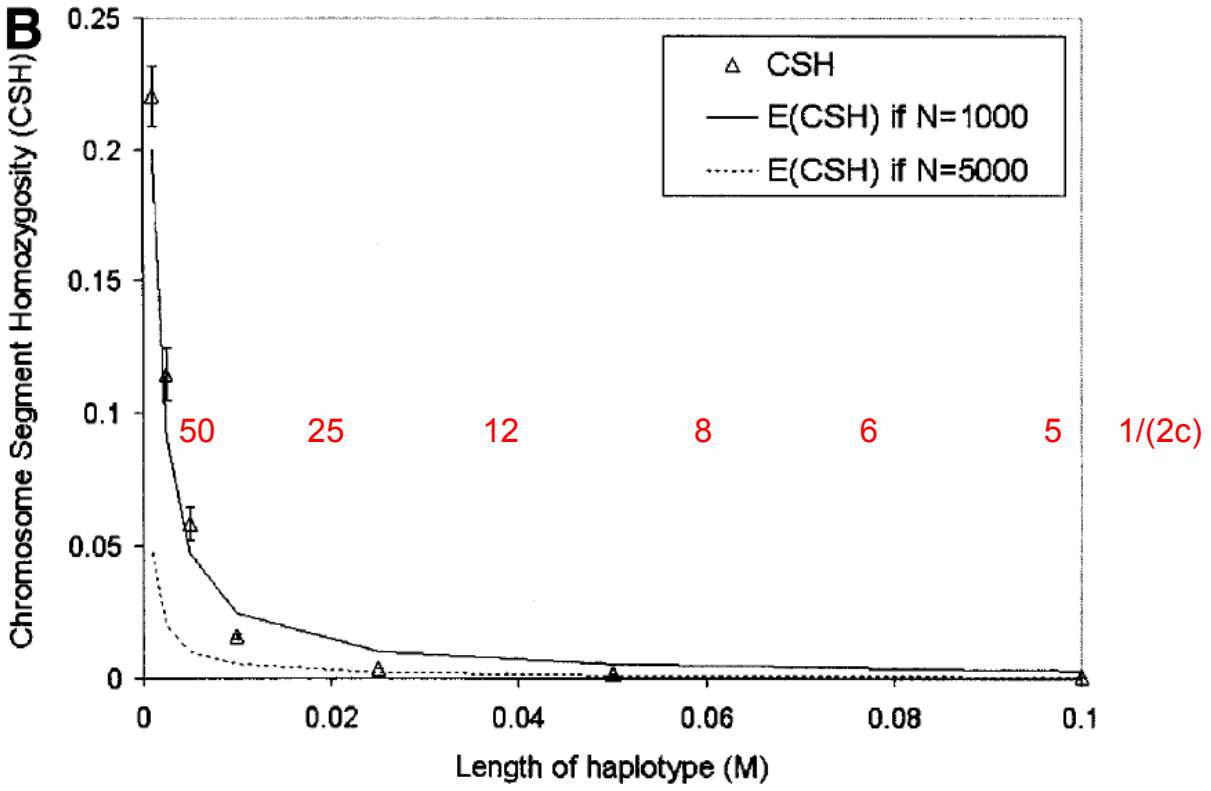
a) one obtains different estimates of N_e from different interval sizes

Hayes et al. (2003): The N_e estimated from the r^2 for a segment of length c (in Morgan) reflects the effective population size $(2c)^{-1}$ generations ago.

b) if $E(r) = 0$ → $E(r^2) \neq 0$

Simulation vs. expectation – linearly increasing population size
N = 1000 → N = 5000 in 50 generations

(Hayes et al., 2003)



Two problems with Sved's (or any other) approach to estimation of N_e from LD

a) one obtains different estimates of N_e from different interval sizes

Hayes et al. (2003): The N_e estimated from the r^2 for a segment of length c (in Morgan) reflects the effective population size $(2c)^{-1}$ generations ago.

b) if $E(r) = 0$ → $E(r^2) \neq 0$

Bishop et al. (1975): In the bivariate Bernoulli distribution with independent components and sample size n, nr^2 has an approximate χ_1^2 distribution so that $E(r^2) = 1/n$

(note: n is the number of sampled gametes!)

Estimating N_e in a contemporary sample from LD

a) from loci on the same chromosome

Sved (1971): $E(r^2) = \frac{1}{1+4N_e c}$ where c is the distance of loci in Morgan

→ $\hat{N}_{e,c} = \frac{1-\left(\overline{r_c^2} - \frac{1}{2n}\right)}{4c\left(\overline{r_c^2} - \frac{1}{2n}\right)}$ ← $E(r^2|LE)$ substracted

An example:

$c = .03M \qquad \overline{r_c^2} = 0.18 \qquad n = 50\ individuals\ sampled$

$$\hat{N}_e = \frac{1-\left(\overline{r_c^2} - \frac{1}{2n}\right)}{4c\left(\overline{r_c^2} - \frac{1}{2n}\right)} = \frac{1-\left(0.18 - \frac{1}{2\times 50}\right)}{4\times 0.03\times\left(0.18 - \frac{1}{2\times 50}\right)} = 40.7$$

1/0.06 = 17 generations ago

Estimating N_e in a contemporary sample from LD

a) from loci on the same chromosome

With SNP data:

1. many estimates of N_e for a distance bin → average, box plot
2. many (divergent) estimates of N_e for different distance bins
   → N_e calculated for a bin of length c is an estimate of the effective population size 1/(2c) generations ago.
3. different autosomes can be used as natural replications → distribution of results

N_e of Holstein cattle (Qanbari et al., 2009)

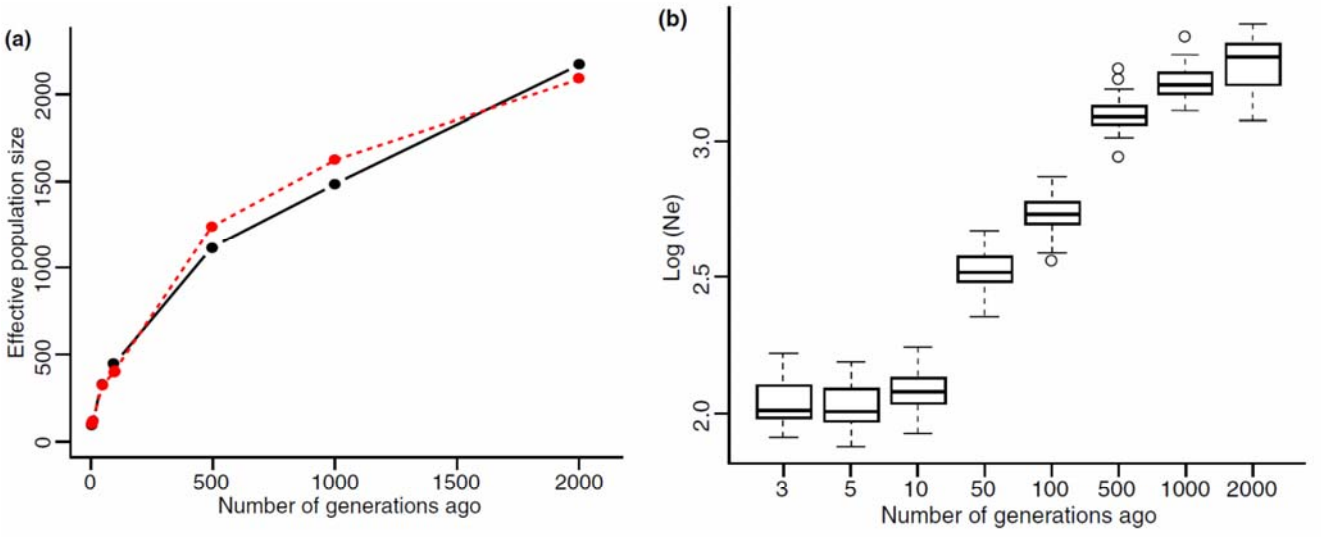
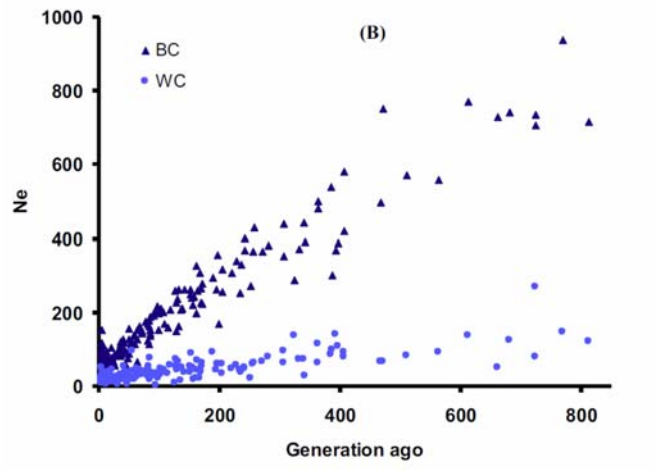
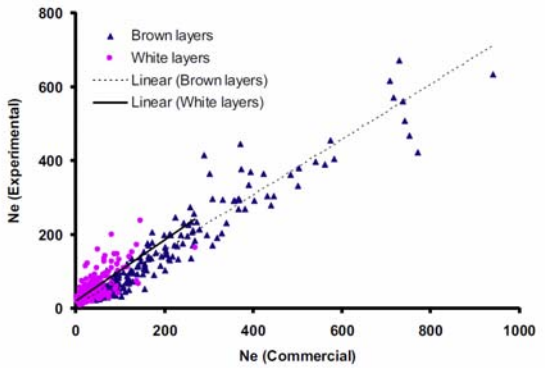


Figure 8 Estimated effective population size over the past generations from linkage disequilibrium data. (a) Dashed and solid lines represent N_e based on estimates of recombination rates and approximate linkage distances respectively. (b) Boxplot representing the trend of $\log_{10}(N_e)$ over time. The variability at each point of time reflects the variation of estimates between the 29 autosomes.

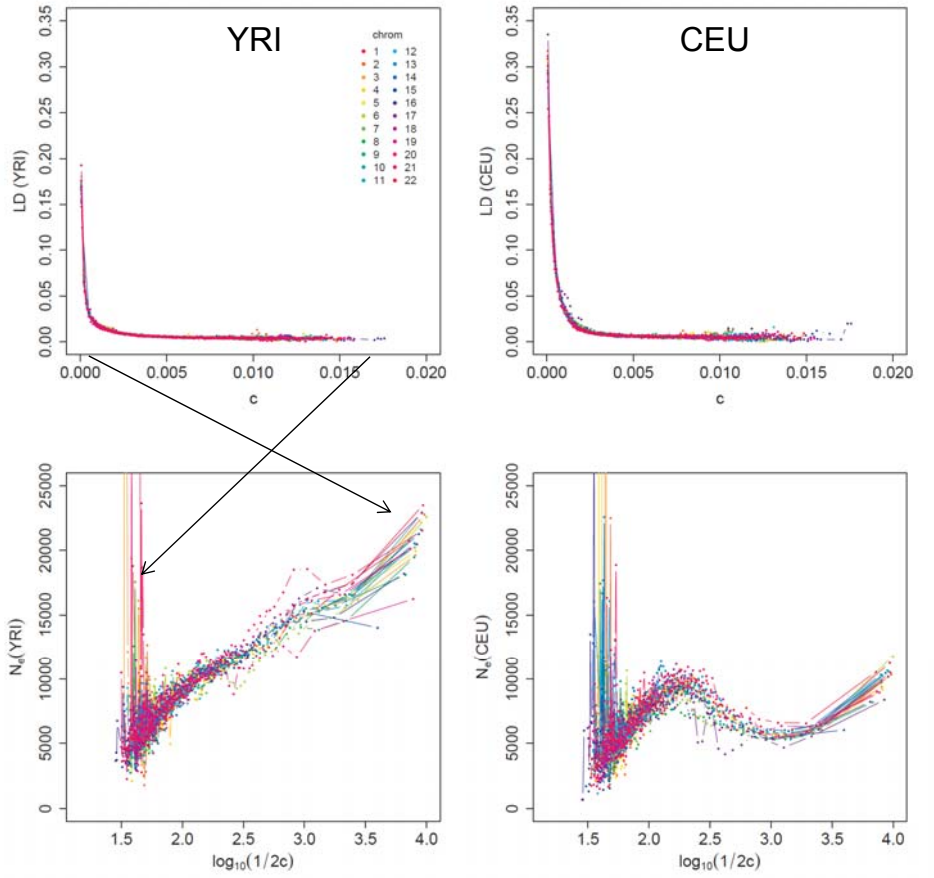
N_e of layers





Qanbari et al. (2010)

N_e for human populations (Ober et al. 2012)



Wrap up N_e

- Effective population size N_e is a relevant parameter in many areas of population and conservation genetics
- The LD effective population size is by definition different from other N_e's (c.f. the inbreeding N_e)
- With high density SNP genotypes N_e can be estimated from pairwise LD for different time points in the past
- Large sample sizes are needed if the more recent N_e is to be estimated
- The underlying recursion formula suggested by Sved (1971) is largely heuristic and lacks a sound mathematical justification, but empirically seems to work reasonably well
- … and yes, N_e remains „notoriously difficult to estimate“

Thanks to my coworkers

Malena Erbe

Ulrike Ober

Saber Qanbari

Animal Breeding and Genetics Group
Department of Animal Sciences
Georg-August-University Göttingen, Germany