Molecular population genetics

Hedrick 2005, chapter 8, pp. 452-462, 428-449

- Demography and selection effects on coalescent trees
- Detecting selection
  - Tajima's *D*
  - Compare divergence and polymorphism
  - Selective sweeps and diversity

Original slides by Outi Savolainen

Coalescent trees and demography, Hedrick 2005

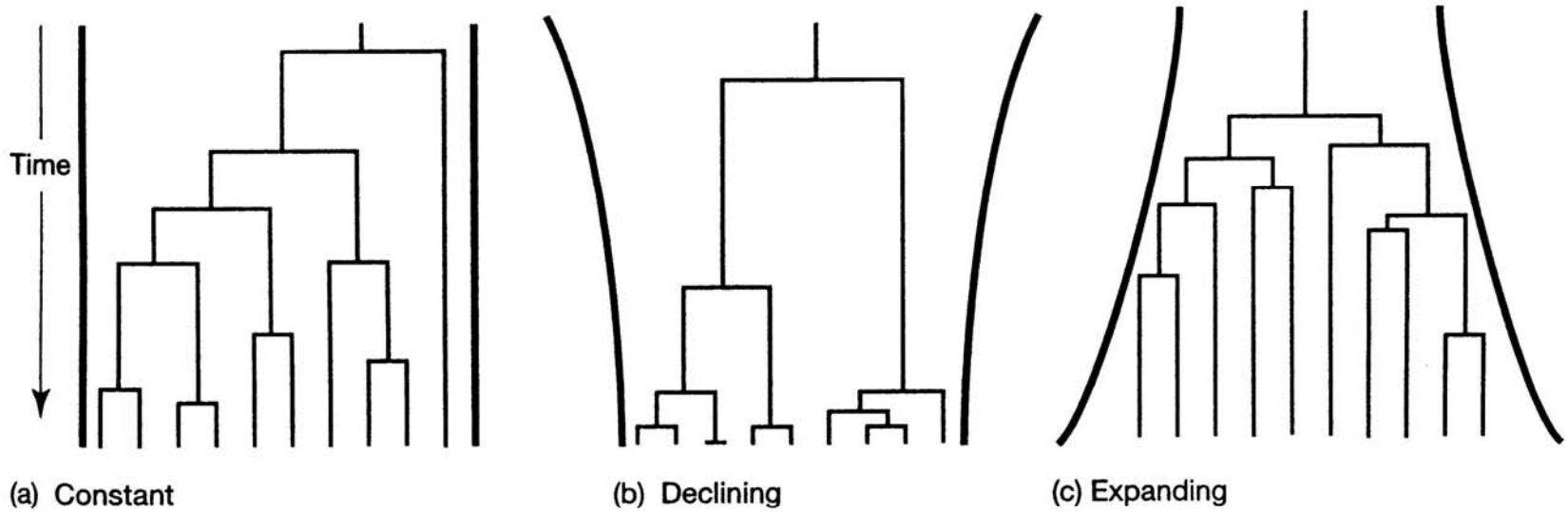


Figure 8.18. The theoretical distributions of coalescent times in genealogies with 10 contemporary samples under three scenarios of historical population change: (a) constant population size, (b) declining population size, and (c) increasing population size (Garrigan *et al.*, 2002). The distance between the thick lines indicates the relative population sizes at different times, and the genealogies reflect coalescent events in periods of

Extensions of the coalescent

- Mating system
- Population size changes
  - exponential, logistic, random, bottlenecks
- Population structure
  - island models, models with geographical structure, continuous models
- Selection
  - balancing - two-allelic and multi-allelic
  - directional – adaptive and deleterious
- Recombination

Important references

- Hein, J. Schierup, M. Wiuf, C. 2004 Gene genealogies, variation and evolution. Oxford Univ. Press.
- Stumpf, M and McVean, G. A. T. 2003. Estimating recombination from population genetic data. Nature Reviews Genetics 4: 959-968.
- Rosenberg, N. and Nordborg, M. 2002. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nature Reviews Genetics 3: 380-390.
- Kingman, J. F. C. 1982. The coalescent. Stochastic Proc. Appl. 13: 1461-1463.
- Hudson, R. R. 1983. Testing the constant-rate neutral allele model with protein sequence data. Evolution 37: 203-217.
- Hartl, D. H. and Clark, A. G. 2007. Principles of Population Genetics. 4th edition. Sinauer.
- www.coalescent.dk

	899 ins G	904 T	908 A	908/09 del —	919 G	1098 T	1114 G	1279 Phe/Ile T	1303 Val/Leu G	1377/93 del TAGAATTCTAATTTT
Landsberg	*	A	C	*	A	*	A	*	*	*
Al-0	*	A	C	*	A	*	A	*	*	*
Bs-1	*	A	C	*	A	*	A	*	*	*
Mt-0	*	A	C	*	A	*	A	*	*	*
Yo-0	*	A	C	*	A	*	A	*	*	*
Shokei	*	A	C	*	A	G	A	*	*	*
Bl-1	*	*	*	*	*	*	*	*	*	*
Gr-1	*	*	*	*	*	*	*	*	*	*
Col-0	*	A	C	*	A	*	*	*	*	*
Es-0	*	*	*	*	*	*	*	*	*	*
Chi-0	*	*	*	*	*	*	*	*	*	*
Pog-0	*	*	*	*	*	*	*	*	*	*
Aa-0	—	*	*	T	*	*	*	*	*	*
Ita-0	*	*	*	*	*	*	*	*	C	—
Ci-0	*	*	*	*	*	*	*	*	C	—
Bla-10	*	*	*	*	*	*	*	*	C	—
Hiroshima	*	*	C	*	*	*	A	A	C	—

Innan et al. 1996 Genetics143:1761. Adh in ecotypes of Arabidopsis thaliana nt 899-1398 ~500 nt

Sites	112	139	243	427	628	762	915	1156	1456	1687	2029	2038
Consensu	C	T	C	C	T	T	A	C	G	C	T	T
Bro1	.	.	.	.	.	.	.	.	.	.	.	G
Bro2	.	.	.	.	.	.	.	.	A	.	.	.
Bro3	.	.	.	.	C	.	.	.	.	T	.	G
Bro4	.	.	.	T	.	.	.	.	.	.	.	.
Bro5	.	.	.	.	.	.	.	.	.	T	G	.
Kol1	.	.	.	.	C	.	.	G	.	.	.	G
Kol2	.	.	.	.	.	.	.	.	.	T	.	.
Kol3	.	.	T	.	.	.	.	.	.	T	.	.
Kol4	.	.	T	.	.	.	.	G	.	.	.	.
Kol5	.	.	.	.	C	.	.	.	.	.	.	.
Kir19	.	C	.	.	.	G	.	.	.	.	.	.
Kir21	.	.	.	T	.	.	.	.	.	.	.	.
Kir23	T	.	.	.	.	.	.	.	.	.	.	.
Kir27	.	C	.	.	.	.	.	G	.	.	.	.
Kir30	T	.	.	.	.	.	T	.	.	T	.	G
Lillo1	.	.	.	.	C	.	.	.	.	T	.	G
Lillo3	.	.	.	.	C	.	.	.	.	T	.	G
Lillo17	.	.	.	.	C	.	.	.	.	.	.	.
Lillo19	.	.	.	.	C	.	.	.	.	.	.	.
Lillo22	.	.	.	.	C	.	.	.	.	.	.	.

Finding departures from the neutral equilibrium model

Adh variation in *A. thaliana* (Innan *et al*. 1996) 500 nt, π **=0.019**

Nucleotide polymorphism at 2043 sites of *pal1* (Dvornyk *et al*. 2002)

Detecting departures from the standard neutral model

- Tajima's *D*
- HKA-test
- McDonald Kreitman

- in some methods, deviations from neutral model could be either due to demography or selection

Tajima's test

- Comparing estimates of θ based on segregating sites (θ_w) and nucleotide diversity (θ_T)
  - θ_w number of segregating sites, also rare ones are counted
  - θ_T pairwise sequence comparison, mostly influenced by intermediate frequency alleles
- Difference is due to departure from the neutral equilibrium model
- Demography or selection?

Tajima's D

.

$$D = \frac{\Theta_T - \Theta_W}{\mathrm{Std}(\Theta_T - \Theta_W)}$$

A large value indicates shortened terminal branches

A small value indicates shortened deep branches

➔deviations from the shape of the coalescent tree may be detected by Tajima's *D*

Rough rule: $D>2$ or $D<-2$ suggests a significant deviation

Tajima 1989 Genetics 123: 585-595

Tajima's *D* and HLA, Garrigan and Hedrick 2003

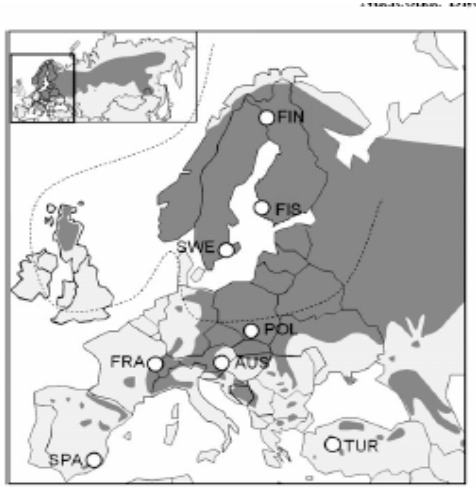
TABLE 8. Homozygosity (*F*) and Tajima's *D* statistics calculated for a global sample of *HLA-A* sequences. A total of 537 sites were examined. Simulations of the neutral genealogical process, conditioned on the observed number of segregating sites, were used to assess significance of Tajima's *D* statistic (significant values shown in bold). Data were obtained from the listed reference.

Population	2*N*	Alleles	*F*	Prob(*F*)	Tajima *D*	Prob(*D*)	Reference
Ainu	100	9	0.182	0.06	3.143	**<0.01**	Bannai et al. 2000
Australian	367	7	0.327	0.24	3.736	**<0.01**	Gao et al. 2000
Chinese	298	22	0.140	0.48	2.356	**0.01**	Middleton et al. 2000
French	248	20	0.162	0.60	2.440	**<0.01**	Bugawan et al. 2000
Havasupai	244	3	0.401	**0.03**	4.049	**<0.01**	Markow et al. 1993
Kagui	100	4	0.356	0.09	3.254	**<0.01**	Middleton et al. 2000
Molucca	52	7	0.214	0.09	2.514	**<0.01**	Bugawan et al. 1999
Omani	236	27	0.100	0.42	2.498	**<0.01**	Middleton et al. 2000
PNG-Lowland	94	6	0.336	0.32	2.447	**<0.01**	Bugawan et al. 1999
PNG-Highland	188	5	0.634	0.75	1.013	0.13	Bugawan et al. 1999
Zapotec	137	6	0.235	**0.02**	1.976	**0.01**	Hollenbach et al. 2001
Zulu	200	23	0.069	**<0.01**	2.441	**<0.01**	Middleton et al. 2000
Global	2164	45	0.126	0.73	3.691	**<0.01**	

Notice positive Tajima's *D* – too few rare alleles, selection maintains allele frequencies more even than in neutral situation.

Scots pine demography

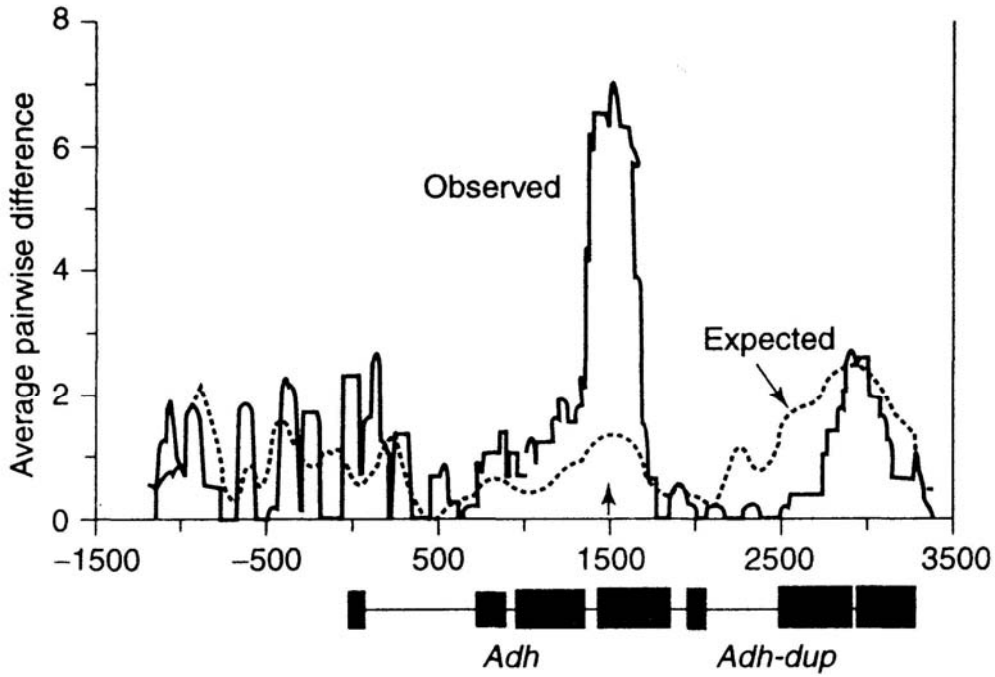
(Pyhäjärvi *et al*. 2007 Genetics 177: 1713–1724)



Tajima's *D* is negative in central and northern populations – ancient bottleneck and expansion

	θ^b	Tajima's D^a	
		Total	Silent
North	0.0056	−0.63 (1.06)**	−0.53*
Central	0.0050	−0.32 (0.99)	−0.51*
Spain	0.0047	−0.16 (1.09)	0.12
Turkey	0.0057	−0.24 (0.88)	0.21

HKA-test, *Drosophila* Adh
Compare divergence between species to polymorphism



ıre 8.8. A sliding window of the observed and expected genetic variation over 5′ flanking region, the *Adh* gene, and the *Adh-dup* gene in *Drosophila* (after

Expectation in neutrality: divergence and poly-morphism correlated, notice aberration in exon 4

Divergence μ
Polymorphism 4Nμ

McDonald-Kreitman test

- Compare patterns of divergence and polymorphism at synonymous and nonsynonymous nucleotide positions

	Div	Polym
Syn		
Nonsyn		

McDonald Kreitman test

from Hedrick 2005

TABLE 8.9 The number of nonsynonymous (replacement) and synonymous substitutions for fixed differences between species and polymorphism within species (a) in general, (b) for *Adh* in three *Drosophila* species (McDonald and Kreitman, 1991), and (c) for *G6pd* in *D. melanogaster and D. simulans* (Eanes *et al.*, 1993). Below, data from humans and chimpanzees are given for (d) mtDNA gene *ND3* (Nachman *et al.*, 1996), (e) *G6pd* (Verrelli *et al.*, 2002), and (f) *HLA-B* (Garrigan and Hedrick, 2003).

	(a) *General*		(b) *Adh*		(c) *G6pd*	
	Fixed	*Polymorphic*	*Fixed*	*Polymorphic*	*Fixed*	*Polymorphic*
Nonsynonymous	N_F	N_P	7	2	21	2
Synonymous	S_F	S_P	17	42	26	36
Ratio	$N_F/S_F = N_P/S_P$		0.41	0.05	0.81	0.06
	(c) *ND3*		(d) *G6pd*		(e) *HLA-B*	
Nonsynonymous	4	8	0	5	0	76
Synonymous	31	10	44	23	0	49
Ratio	0.13	0.8	0.0	0.28	—	1.61

McDonald and Kreitman (1991) applied this test to data from coding region of *Adh* from *Drosophila* species *D. melanogaster, D. simulans*, and *D. yakuba*. Table 8.10 gives the nucleotide sequence for the nonsynonymous (replacement) differences for the three species and the status as far as fixed differences between species (seven sites) and polymorphic variation within species (two sites). For example, position 781 is considered a fixed difference

Detecting selection

TABLE 1. The tests of selection and their evolutionary parameters for detecting selection on the three different timescales.

Timescale	Test	Parameters
Current generation	Hardy-Weinberg	Selection coefficient, number of alleles
	Mendelian proportions	Selection coefficient
	Random mating	Selection coefficient
	Fitness associations	Selection coefficient
Recent past	Ewens-Watterson	Population size, number of alleles, selection coefficient and duration
	Linkage disequilibrium	Population size, recombination rate, selection coefficient and duration
	Geographical congruency	Population size, population structure, selection coefficient and duration
Distant past	Nonsynonymous/synonymous substitution rate	Selection coefficient and duration, mutation rate
	McDonald–Kreitman	Selection coefficient and duration, mutation rate, species divergence times
	Tajima's *D*	Population size, population structure, mutation rate, selection coefficient and duration
	Transspecies polymorphism	Selection coefficient, mutation rate, species divergence times

Selective sweep by hitchhiking

- Directional selection reduces variation
- Example: chloroquine resistance locus of malaria parasite *Plasmodium falciparum*

  -medicine introduced about 45 years ago

  -resistance mediated by mutations in *pfcrt* gene

  -less variation close to resistance locus

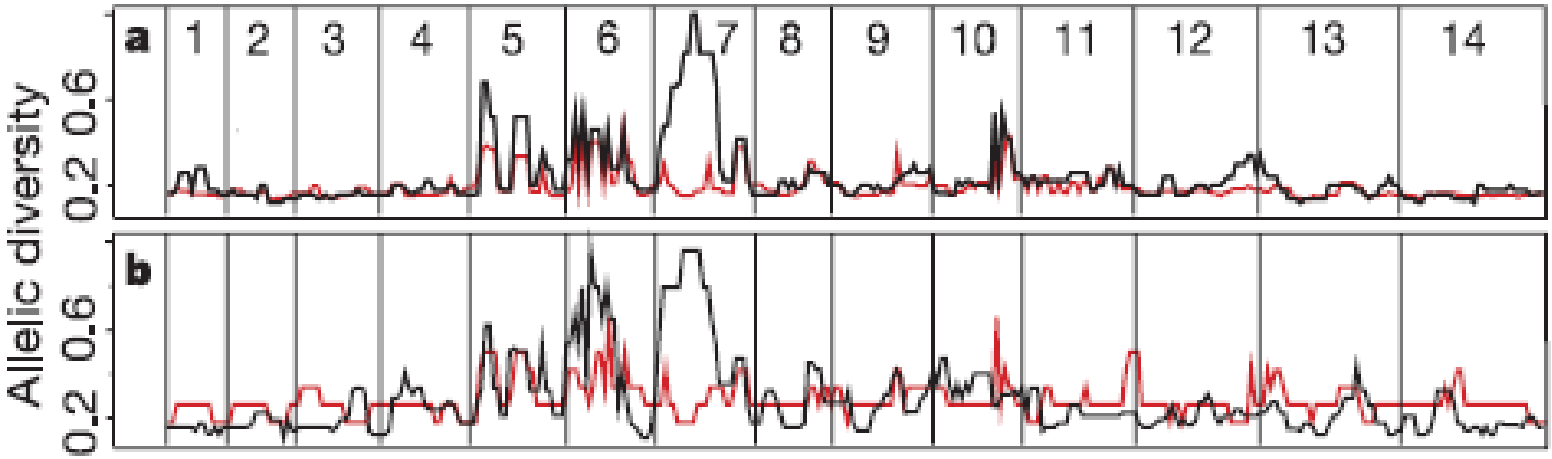


Figure 3 Genome-wide scans for loci of reduced diversity and association of reduced diversity with the CQR phenotype. The numbered panels represent the 14 chromosomes covered by 342 markers. **a**, **b**, Allelic diversity, plotted as (1 − median adjusted allelic diversity of five contiguous markers), for CQS (red) and CQR (black) isolates from Africa (**a**) and Asia (**b**, excluding CQR PNG isolates). Peaks represent regions with reduced diversity.

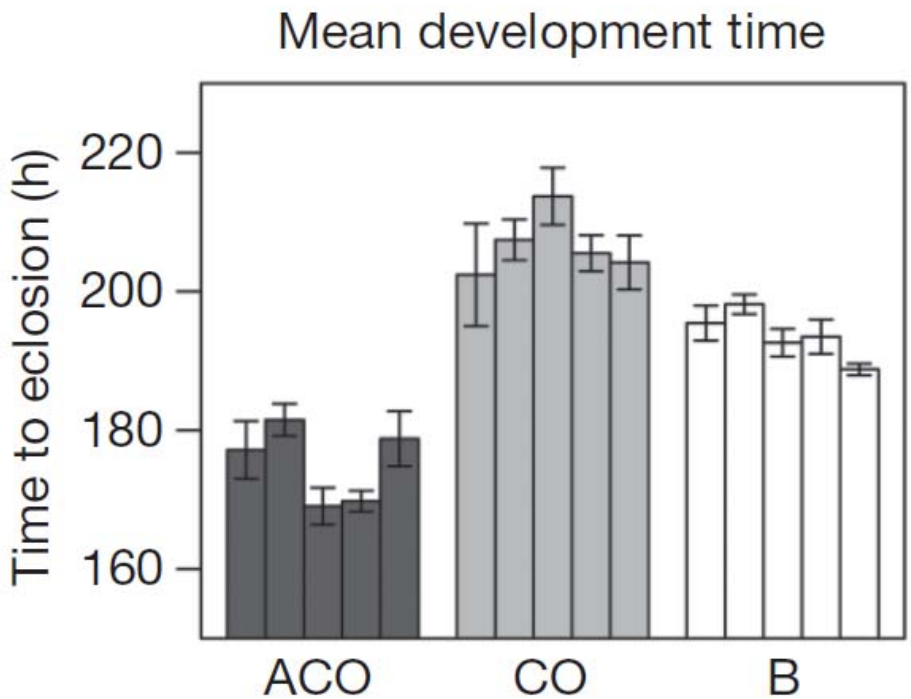
RED – sensitive, BLACK – resistant strains

Notice: Allelic diversity reduced especially in chromosome 7, where resistance locus is located, important to compare to the genomic background (e.g. demography effects)

Wootton *et al*. 2002 Nature 418, 320-323.

Experimental evolution in *Drosophila* – soft sweeps, Burke *et al*. 2010 Nature 467: 587–590.

- select lines of *D. melanogaster* for 600 generations for rapid development
- examine genetic changes
- complete fixation of new mutations in different lines OR
- selection from existing variation



ACO - selection for rapid development
CO - control
B - base population

Results, Burke *et al*. 2010

Complete selective sweep based on a new mutation: each replicate line has its own mutations that get fixed – monomorphic areas that differ between lines (likely)

Response based on existing variation – same genetic areas in different lines, perhaps not completely monomorphic

Soft sweeps, Burke *et al*. 2010

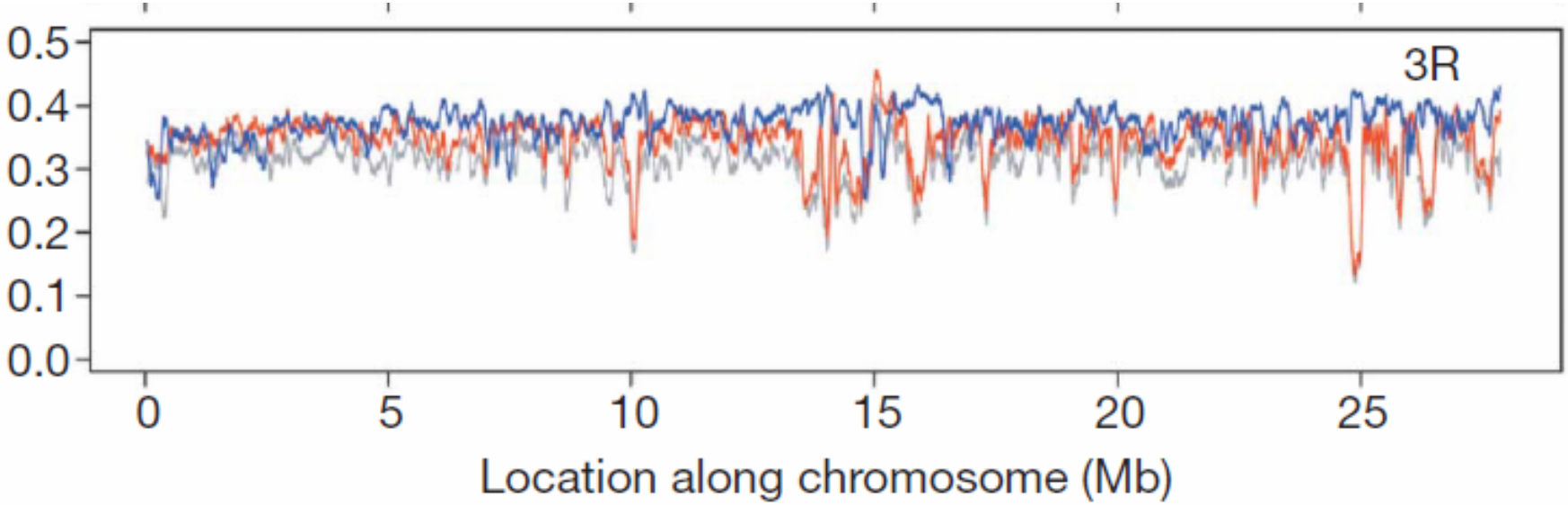


Figure 3 | Heterozygosity throughout the genome. Sliding-window analysis (100 kb) of heterozygosity in the CO pool (blue), the ACO pool (red) and ACO_1 (grey), with a 2-kb step size. The panels show the five major chromosome arms of *D. melanogaster*.

- local losses of heterozygosity
- regions of reduced heterozygosity associated with regions of differentiated allele frequency in selected vs. control lines
- soft sweeps: heterozygosities not near zero (as would be in classic sweeps)

Selection for development time, Burke *et al*. 2010

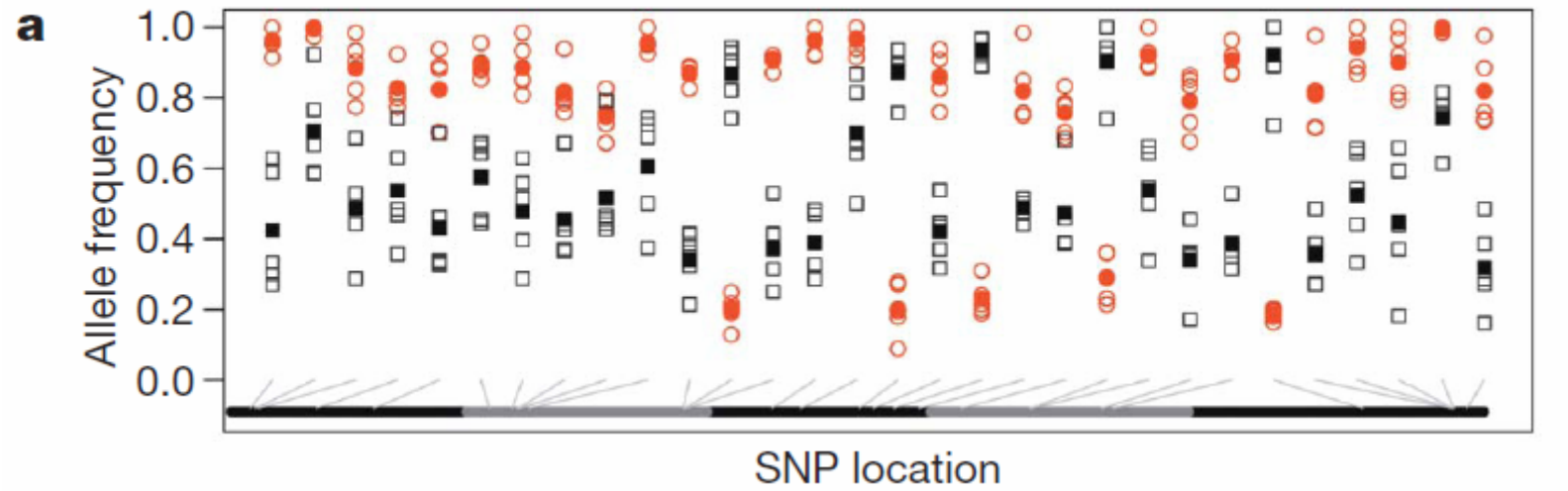


Figure 4 | Analysis of individual genotypes, measured by cleaved amplified polymorphic sequence (CAPS) techniques. **a**, Allele frequency estimates of the most common allele at 30 SNPs genotyped in 35 females per replicate population. Red circles represent ACO estimates and grey squares represent CO estimates. Open symbols are allele frequencies for ACO_1–ACO_5 and CO_1–CO_5, and filled symbols represent treatment means. Alternating black and grey bars designate the X, 2L, 2R, 3L, and 3R arms, respectively, with grey lines

- convergence of allele frequencies and H levels between replicates
- selection is acting on on the same intermediate-frequency variants in each population