Summer Institute in Statistical Genetics
University of Queensland, Brisbane
Module 5: Population Genetic Data Analysis

The Wright-Fisher Model

David Balding
Professor of Statistical Genetics
University of Melbourne, and
University College London

Feb 9, 2017

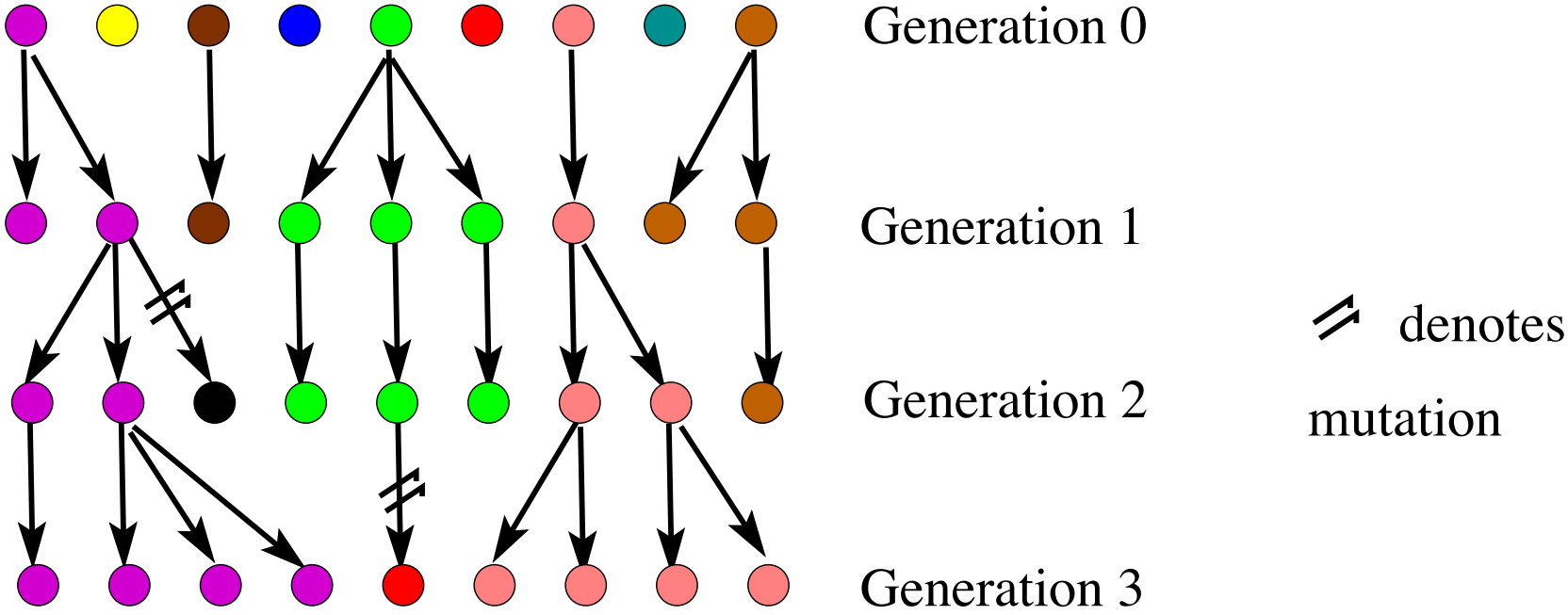
Contents:

Wright-Fisher model in haploid populations

Two-locus Wright-Fisher and linkage disequilibrium

Genetic Drift: single locus

- Drift is the stochastic process by which some alleles are lost from the population, while others increase in frequency, because of the randomness of reproductive success.
- We can investigate the effects of drift using a simple mathematical model: the **Wright-Fisher (W-F) Model** of a constant-size, random mating population at a neutral locus.
- Although real populations don't satisfy these assumptions, the W-F model can give important insights into how gene frequencies change over time in real populations.
- It's easy to extend to a diploid model, but to keep things simple we will consider a haploid model here, in which case an individual corresponds to one allele copy;
  - haploid models apply to humans at mtDNA and Y chromosome.



Wright–Fisher model 1 locus, haploid, n=9;

- The population size is N alleles (constant).
- Each allele has a type reflecting its DNA sequence.
- Each allele in generation k is chosen at random in gen. $k-1$;
  - the new allele is the same type as its parent (probability $1-\mu$), or is altered by mutation (probability μ).

Each allele has a binomial(N,1/N) number of offspring (mean $= 1$; variance ≈ 1), but the offspring numbers of different alleles are not independent.

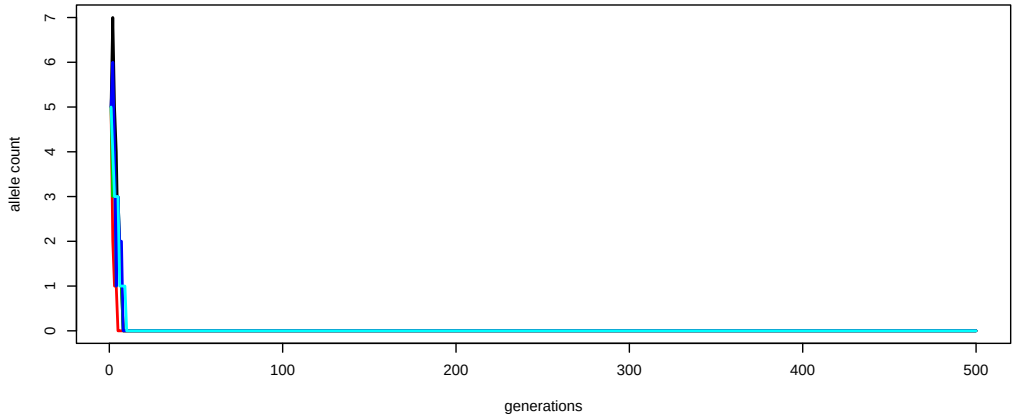
When $\mu = 0$, the probability F_k that two alleles in generation k are identical by descent (IBD) from an allele in generation 0 satisfies:

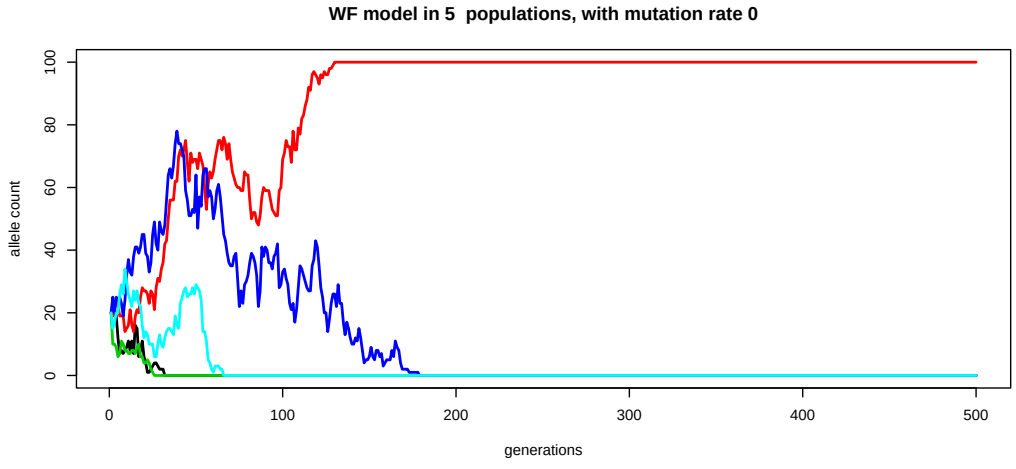
$$F_k = \frac{1}{N} + \left(1 - \frac{1}{N}\right) F_{k-1}.$$

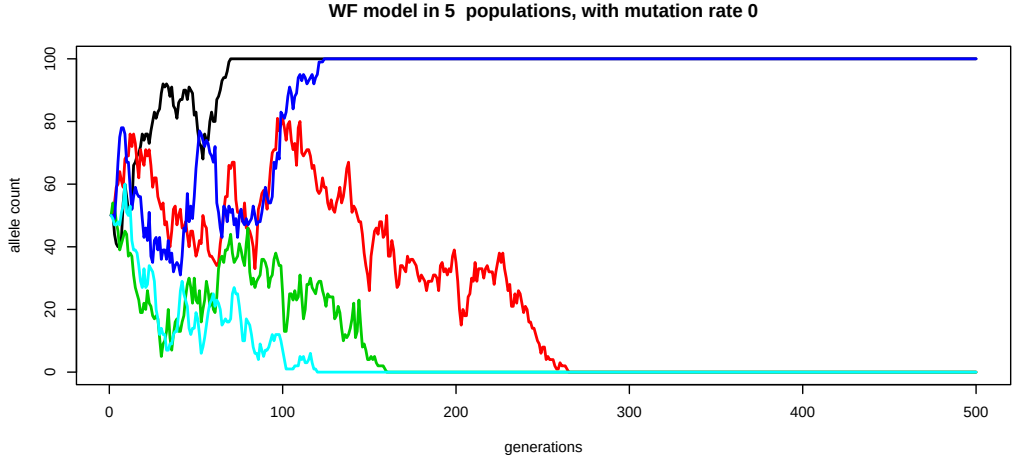
Whatever the value of F_0, we have $F_k \to 1$ as $k \to \infty$, i.e. eventually all alleles are descendants of one ancestral allele, whose allelic type becomes *fixed*. **Questions:**

1. What is the probability for an allelic type to become fixed if initially there are x alleles of that type?
2. How long does this take?

It is possible to make some progress theoretically, and much more can be inferred from simulation.







Each plot shows five haploid W-F simulations over 500 generations, with $N = 100$, $\mu = 0$ and initial frequencies of a particular allele:

- top: 1;
- middle: 20;
- bottom: 50.

R code for WF simulations

```
wf = function(npop=5,ngen=500,nall=100,init=1,mu=0)
{
freq = matrix(init,npop,ngen)
for(i in 1:(ngen-1))
     freq[,i+1] = rbinom(npop,nall,mu+(1-2*mu)*freq[,i]/nall)
matplot(t(freq),ty="l",lty=1,lwd=2,xlab="generations",ylab="allele \
    count",main=paste("WF model in",npop," populations,mu = ",mu))
}
```

After entering the function above into R, plots similar to those on the previous page can be obtained by entering

```
par(mfrow=c(3,1)); wf(); wf(init=20); wf(init=50)
```

Try varying some of the other parameters (npop, ngen and nall). For example you could try:

```
par(mfrow=c(2,1)); wf(ngen=20,nall=10,init=5); wf(ngen=2000,nall
```

- How does the probability of fixation vary with init?
- How does the mean time to fixation or loss vary with init?

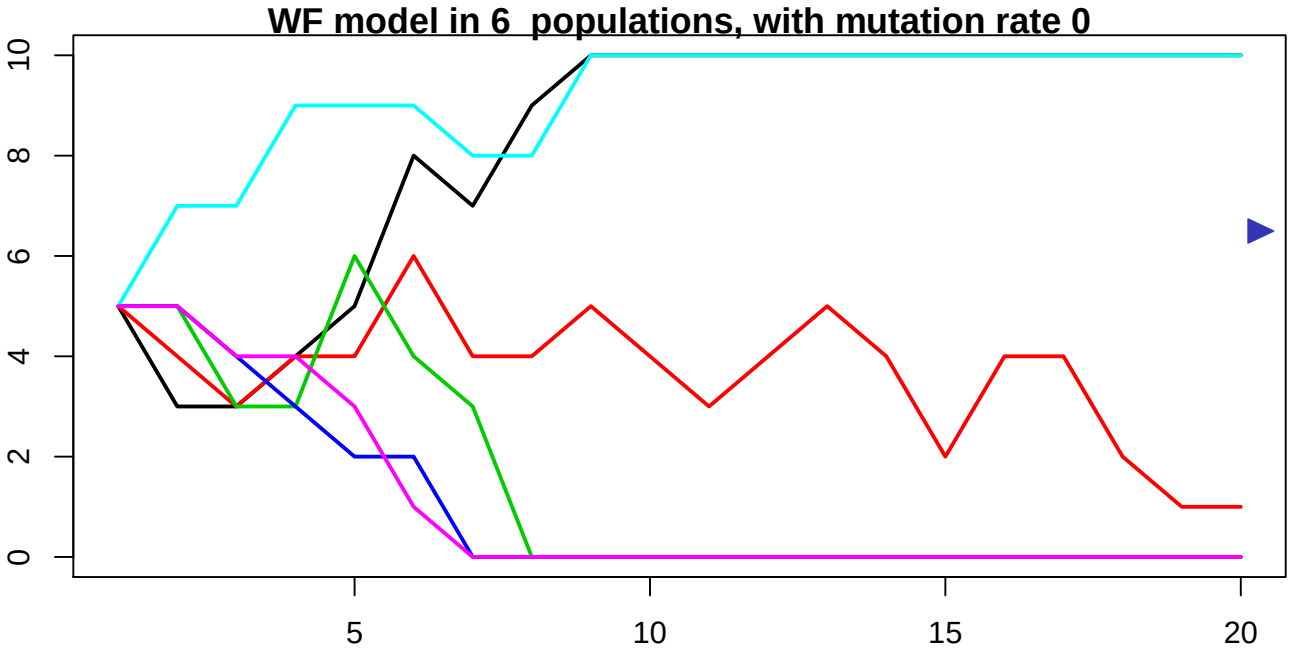
Time Scaling and Effective Population Size

The behaviour of the W-F model is similar for different N, but at a "speed" that is proportional to 1/N. Population geneticists often work with time scaled in units of N generations: then the same results fit all population sizes. Much of classical population genetics theory uses the "diffusion limit", $N \to \infty$.

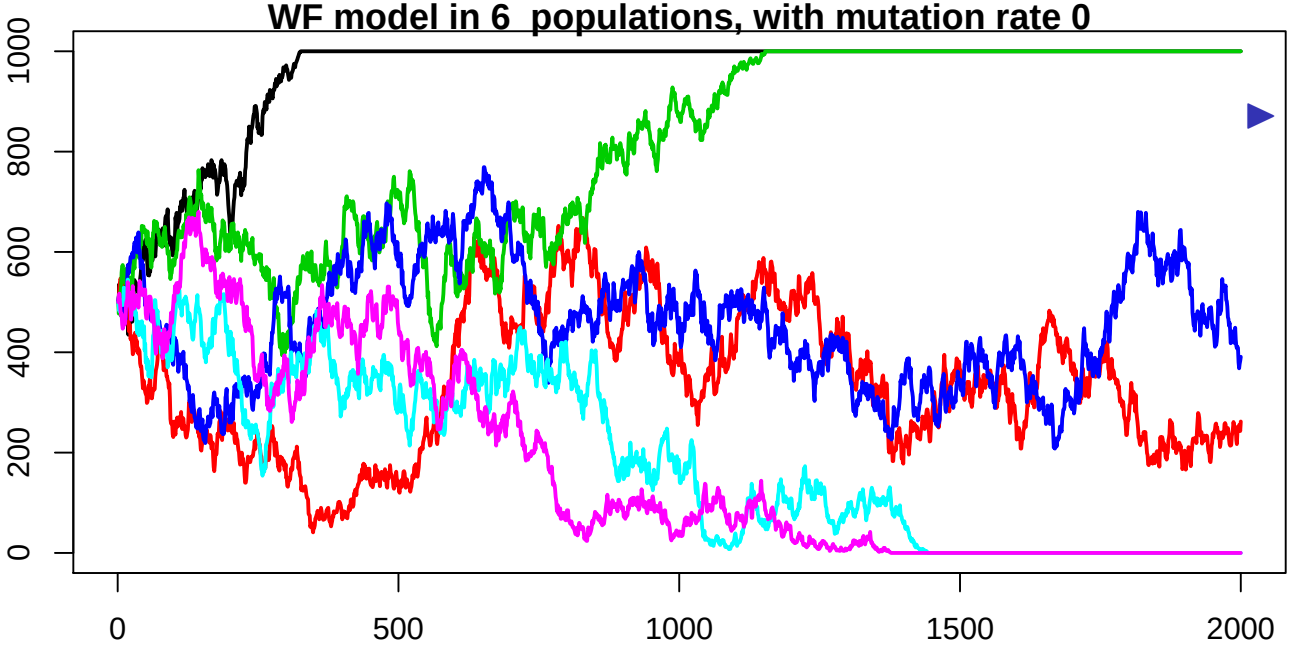
A population may be closely approximated by a W-F population of a different size N_e, the *effective population size*. Examples:

- If variance in offspring number $\sigma^2 \neq 1$ then the W-F model can be used with $N_e = N/\sigma^2$.
- If a (diploid and dimorphic) population has sex ratio $p/(1-p)$, we have $N_e = 4p(1-p)N$.

In real populations N_e is often estimated as the value of N that provides a best fit to the W-F model. Although powerful, this is not always satisfactory, e.g. population growth can require a non-linear change in time-scale and no W-F model gives a good fit.



Top: six haploid, $\mu = 0$, W-F simulations over 20 generations with $N = 10$ and initial allele count 5;



Bottom: six haploid, $\mu = 0$, W-F simulations over 2000 generations with $N = 1000$ and initial allele count 500.

For humans, $N_e \approx 10\,000$ breeding adults, surprisingly small compared with census size of ≈ 4 billion (of breeding age), which reflects the relatively low genetic diversity of our species. This in turn is due to many factors, including some or all of:

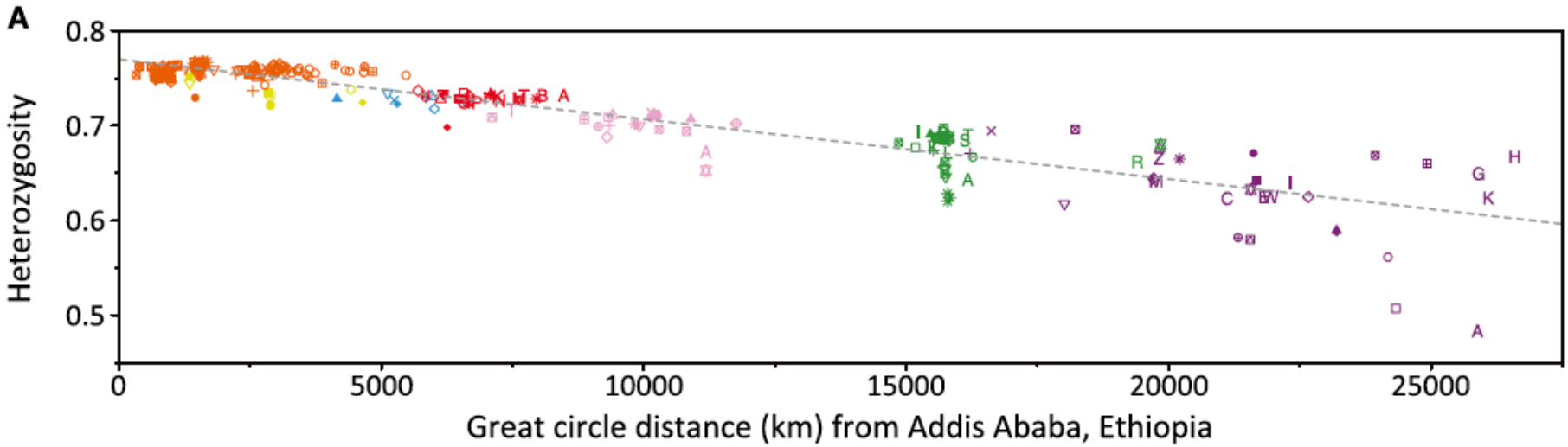
- geographic dispersal over wide areas;
- high between-male variance in reproductive success;
- war and disease;
- bottlenecks/intense selection in changing environments.

NB This does not mean that there was an ancestral human population of size 10 000.

The greatest human genetic variation occurs in Africa. Ignoring recent mass migrations, e.g. to the "New World", the genetic variation within human populations declines roughly linearly with their distance from Addis Abbaba. This reflects the relatively recent (perhaps 60-70 KYBP) migration out of Africa of the ancestors of most modern humans.

For further details see: "Human Evolutionary Genetics" by Jobling et al. (2004).

The greatest human genetic variation occurs in Africa. Ignoring recent mass migrations, e.g. to the "New World", the genetic variation within human populations was found to decline with their distance from Addis Ababa. This reflects the relatively recent (perhaps 70 KYBP) migration of modern humans out of Africa to found the populations of other continents.



From: Pemberton et al. G3 Vol 3 pp 891–907, 2013.

Mutation-drift equilibrium

For $\mu > 0$ in the W-F model we have

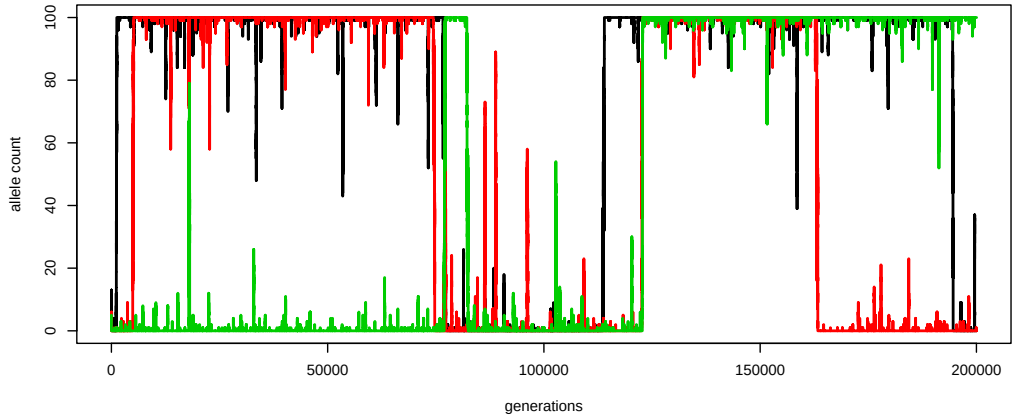
$$F_k = (1-\mu)^2 \left[\frac{1}{N} + \left(1 - \frac{1}{N}\right) F_{k-1} \right].$$

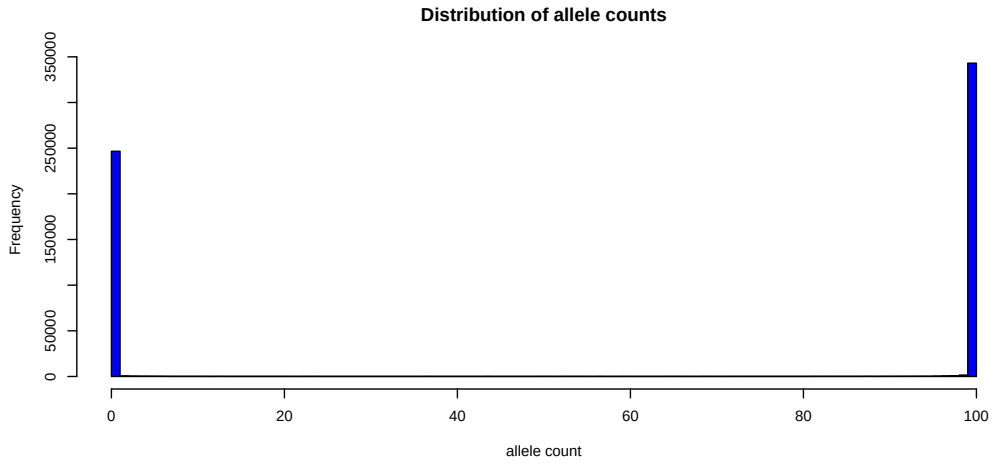
For any value of F_0, the value of F_k approaches (approximately):

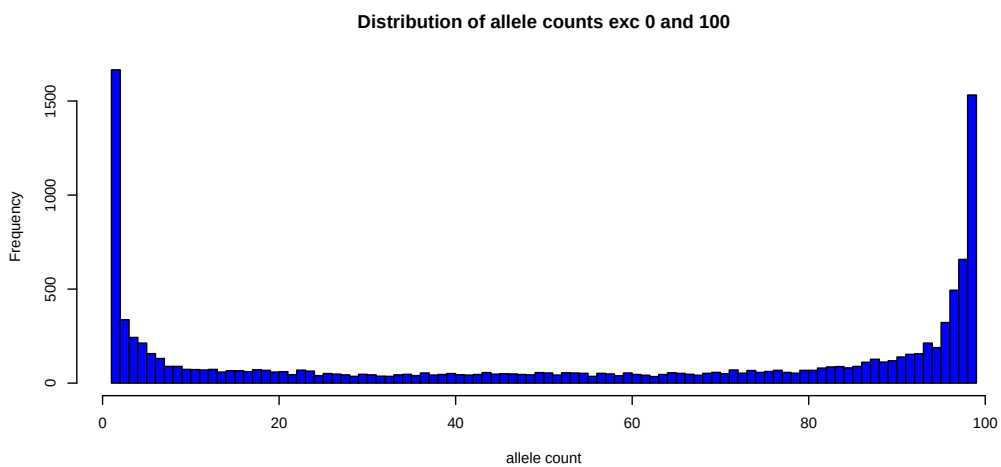
$$F_\infty = \frac{1}{1 + 2N\mu}.$$

The value of F_∞ measures a balance, known as *mutation-drift equilibrium*, between loss of variation due to drift and creation of variation through new mutants.

At equilibrium when $0 < \mu \ll 1/N$, there is a U-shaped distribution of allele frequencies: most allelic variants are rare. This holds for humans and many other species. The allele frequency distribution for markers used in genetic epidemiology may not be U-shaped, because of a bias favouring common alleles.







Top: Three haploid W-F simulations over 2×10^5 generations with $N = 100$ and $\mu = 2 \times 10^{-5}$.

Middle: Allele count distribution over the whole simulation. From this we can compute $F \approx 0{\cdot}9962$, compared with a theoretical value of $1/(1+2N\mu) = 0{\cdot}9960$.

Bottom: same as middle except that the counts at 0 and 100 are removed to better see the other values.

The histograms should become symmetric about 50 as the number of generations increases.

R simulations for mutation-drift equilibrium

- The genome-wide average mutation rate per site per generation is about $\mu = 10^{-8}$ (one per hundred million).
- Since a haploid human genome has about 3×10^9 sites, that means about 30 mutations per meiosis per genome.
- The rate is too low for efficient simulation in class, but time scaling can help: $\mu = 10^{-8}$ with $N_e = 2 \times 10^4$ generates a similar pattern of diversity to $\mu = 10^{-6}$ with $N_e = 200$ (it's the product μN_e that matters).
- The simulation on previous slide used a higher μ to generate more diversity:

```
wf(npop=3,nall=100,mu=2*10^-5,ngen=2*10^5)
```

To generate the histograms, we need to run `wf` for large values of `npop` (at least 10^3, preferably 10^4) in which case we don't want the plots, so you should comment out (with #) the `matplot` command and add a final line `return(freq[,ngen])` which returns the vector of allele counts in each population. Then run

```
res = wf(npop=10^3,nall=100,mu=2*10^-5,ngen=2*10^5)
hist(res,n=100)
hist(res[(res>0)&(res<100)],n=100)
```

Finally, to find F, the probability that two alleles drawn from a random population are the same, we compute:

```
mean((res/100)^2 + (1-res/100)^2)
```

For our simulations is about $2N\mu = 0.002$, which is about an order of magnitude larger than for humans and consequently the genetic diversity is much greater in the simulations than in human genomes.

Subpopulation differentiation and migration - F_{ST}

In a subdivided population with no migration, mutation-selection equilibrium may be established but with different alleles in different subpopulations. If μ is small, a given allele may be near fixation in some subpopulations and nearly lost in others.

Although migration does not generate new variation, from a local perspective in geographically-structured populations, new migrants can be the most important source of genetic variation. Globally, migration tends to reduce inter-population differences in allele fractions.

The parameter F_{ST} measures the variation in allele fractions between subpopulations. If $\tilde{p}$ denotes the subpopulation fraction of an allele, and p denotes a reference ("global") value, then

$$F_{ST} = \frac{\text{Var}[\tilde{p}]}{p(1-p)}.$$

- if $F_{ST} = 1$ all subpopulations have reached fixation (low/no migration),
- if $F_{ST} = 0$ allele fractions are the same in all subpopulations (high migration).

Can also estimate F_{ST} as $1 - H_S/H_T$, where H_S is the observed heterozygosity in a subpopulation (or averaged over several subpopulations) and H_T is the expected heterozygosity, given the allele fractions, in a panmictic population.

In simple models of a large population divided into k subpopulations each of size N/k, with m migrants in/out of each island in each generation, the equilibrium value of F_{ST} is

$$F_{ST} \approx \frac{1}{1 + 2N(m+\mu)}.$$

(the 2 becomes a 4 for diploids).

F_{ST} depends on local population size, migration and mutation. Typical SNP-based value is $\leq 1\%$ for comparisons of (large) European populations, 10-15% for intercontinental comparisons.

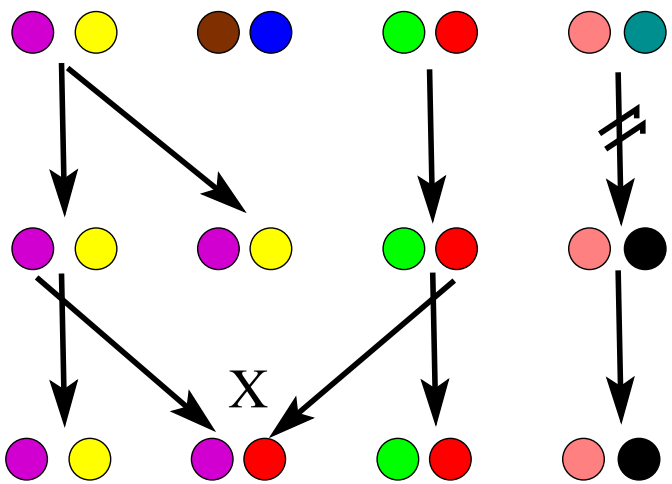
Wright-Fisher model in haploid populations

Two-locus Wright-Fisher and linkage disequilibrium

2-locus W-F model

The basic entity is a haplotype consisting of two loci A and B, with possible alleles A, a, and B, b. Each individual has two haplotypes, for example AB and aB, which may be written AB/aB.

With probability ρ, a haplotype in the new generation is obtained by choosing alleles independently from those in the previous generation at the corresponding locus, otherwise it is a copy of a haplotype from the previous generation. In either case, each allele is subject to mutation occurring independently with probability μ.



Linkage disequilibrium (LD)

Linkage equilibrium (LE), also called "gametic equilibrium", is the statistical independence of the alleles at two polymorphic loci on the same chromosome (or in the same gamete).

Let p_{AB} denote the population fraction of AB haplotypes, while p_A and p_B denote the A and B allele fractions. If $D_{AB} = p_{AB} - p_A p_B$ then LE corresponds to $D_{AB} = 0$, otherwise we have Linkage *Disequilibrium* (LD).

The range of values of D_{AB} (or just D) depends on the allele fractions: a value that arises for one pair of loci may be impossible at another pair given their allele fractions, making cross-locus comparisons difficult. To overcome this problem we define

$$\begin{aligned} D^- &\equiv \min\{p_a p_b, p_A p_B\} \\ D^+ &\equiv \min\{p_a p_B, p_A p_b)\}. \end{aligned}$$

and

Measures of LD

$$D' = \begin{cases} D/D^+ & \text{if } D \geq 0 \\ D/D^- & \text{if } D < 0 \end{cases},$$

which ranges between -1 and 1. Another measure of LD is the squared correlation coefficient:

$$r^2 = \frac{D^2}{p_A p_a p_B p_b}.$$

The population allele and haplotype fractions must be estimated via sample counts, denoted n. For example, r^2 can be estimated from

$$\hat{r}^2 = \frac{(n_{AB} n_{ab} - n_{Ab} n_{aB})^2}{n_A n_B n_a n_b}$$

Although haplotype counts such as n_{AB} are not usually directly observed, if LD is high they can be inferred reliably from multilocus genotype data.

Example: haplotype counts from a sample of size 1 000:

	A	a
B	400	150
b	300	150

The maximum likelihood estimators (MLEs) of the haplotype and allele fractions are $\hat{p}_{AB} = 0{\cdot}4$, $\hat{p}_{Ab} = 0{\cdot}3$, $\hat{p}_{aB} = 0{\cdot}15$, $\hat{p}_{ab} = 0{\cdot}15$, $\hat{p}_A = 0{\cdot}7$, $\hat{p}_a = 0{\cdot}3$, $\hat{p}_B = 0{\cdot}55$, $\hat{p}_b = 0{\cdot}45$, and:

- $\hat{D}_{AB} = 0{\cdot}4 - 0{\cdot}7 \times 0{\cdot}55 = 0{\cdot}015$.
- $\hat{D}^+ = \min\{0{\cdot}3 \times 0{\cdot}55, 0{\cdot}7 \times 0{\cdot}45\} = 0{\cdot}165$;
- $\hat{D}' = 15/165 \approx 0{\cdot}091$;
- $\hat{r}^2 = \dfrac{(400 \times 150 - 300 \times 150)^2}{550 \times 450 \times 700 \times 300} \approx 0{\cdot}0043$.

All these measures suggest a weak positive association between the 0 alleles at the two loci.

D' versus r^2

- D'. Advantage: sensitive to few recombinations between the loci since the most recent mutation at one of them. Disadvantage: we can have $|D'| = 1$ when one of the alleles is extremely rare, which is usually of little practical interest.
- r^2 is only large when the LD is likely to be statistically detectable, since r^2 is the Pearson's test statistic for independence in a 2×2 contingency table. Thus, the larger is r^2 the more likely we are able to detect LD using this, or a related statistical test. Put another way, every increase in r^2 permits a corresponding decrease in sample size n required to detect the statistical association.

Factors that affect LD: recombination

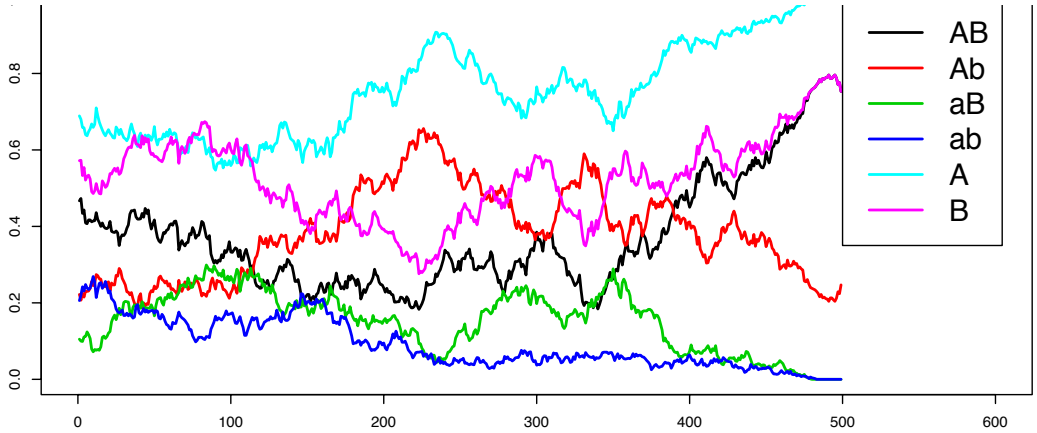
In the 2-locus WF model, we expect in every generation a fraction $1-\rho$ of haplotypes to avoid recombination, while the recombinant haplotypes are generated in LE proportions. Thus, LD reduces over time under random mating. Population genetics textbooks often derive the relationship

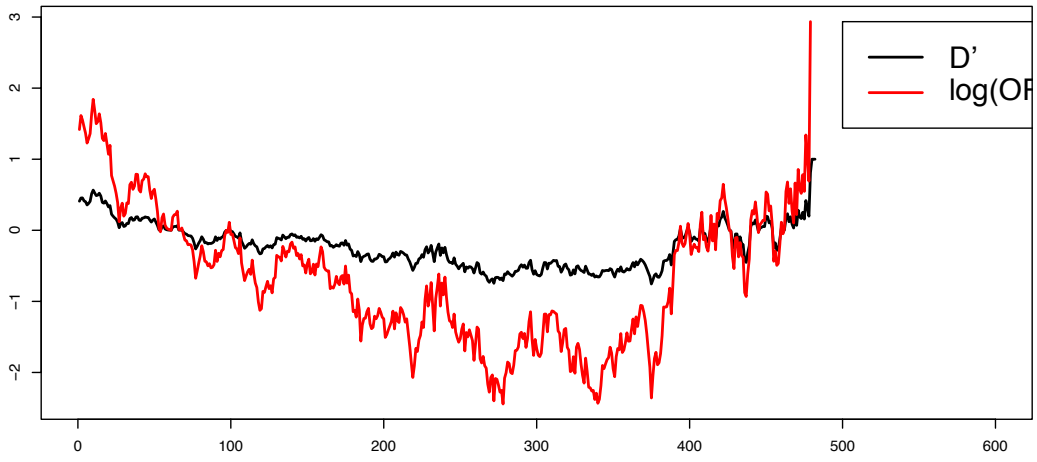
$$D^k = (1-\rho)^k D^0$$

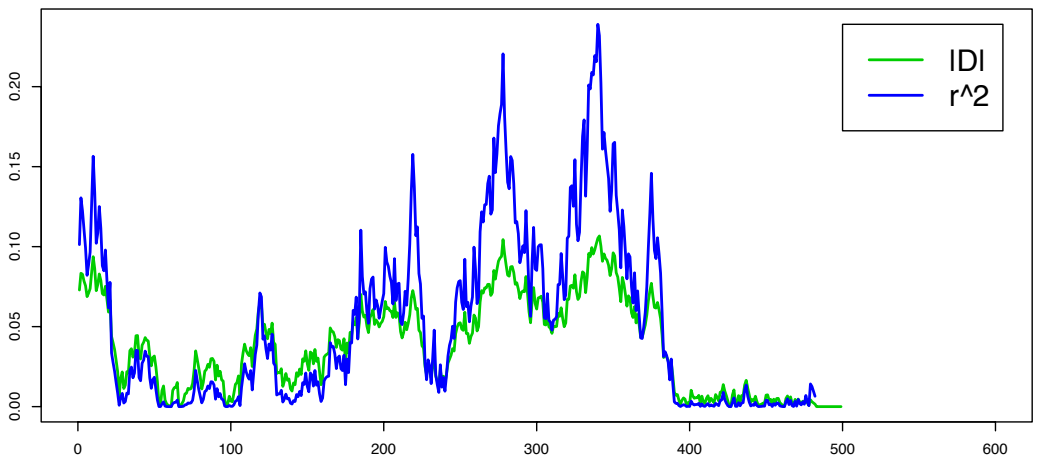
where the superscript indicates generation number.

In practice this formula is of little use unless the population size is very large. Then, any LD decays to approximate LE after roughly $1/\rho$ generations. If the LD is initially high then for a short period D and D' decay approximately exponentially.

Measures of the breakdown of LD along a genome can be used to estimate recombination rates: high LD between markers suggests low recombination between them. However such inference can be confounded by other factors described below.







Measures of LD in a simulation of the 2-locus WF model, with $\mu = 0$. The top plot shows the allele and haplotype fractions evolving over 500 generations. Near the end of the simulation the a allele is lost and this has different effects on the different measures (middle and bottom plots).

Factors that affect LD: mutation

Suppose that initially everyone carries the A allele, but then a mutant a arises in a gamete that also carries the B allele, to create the following haplotype counts:

	A	a
B	599	1
b	400	0

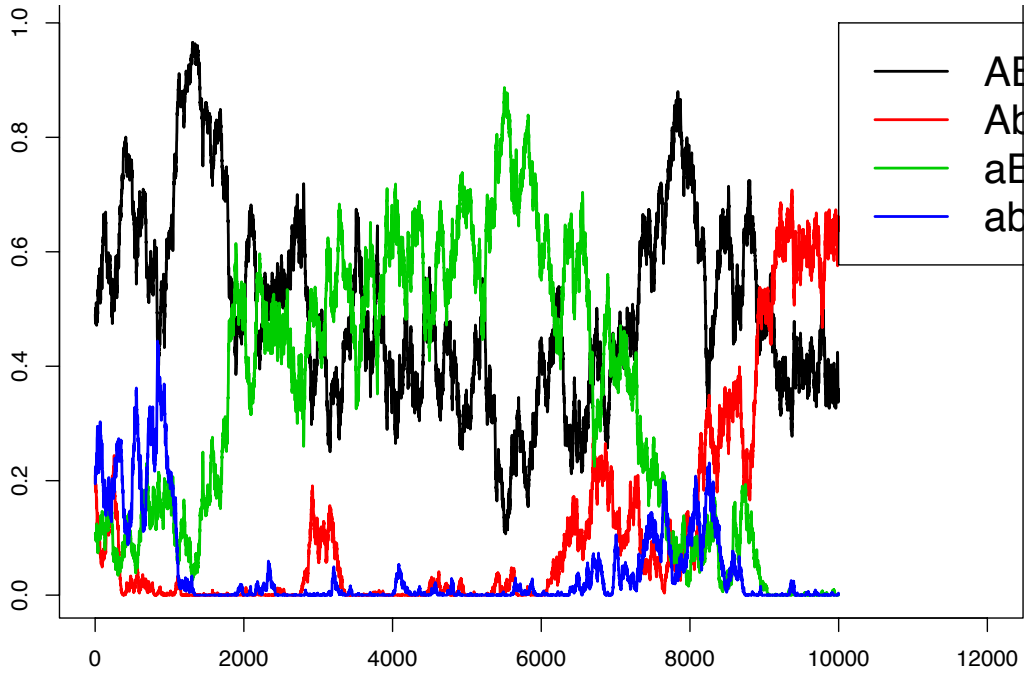
Then

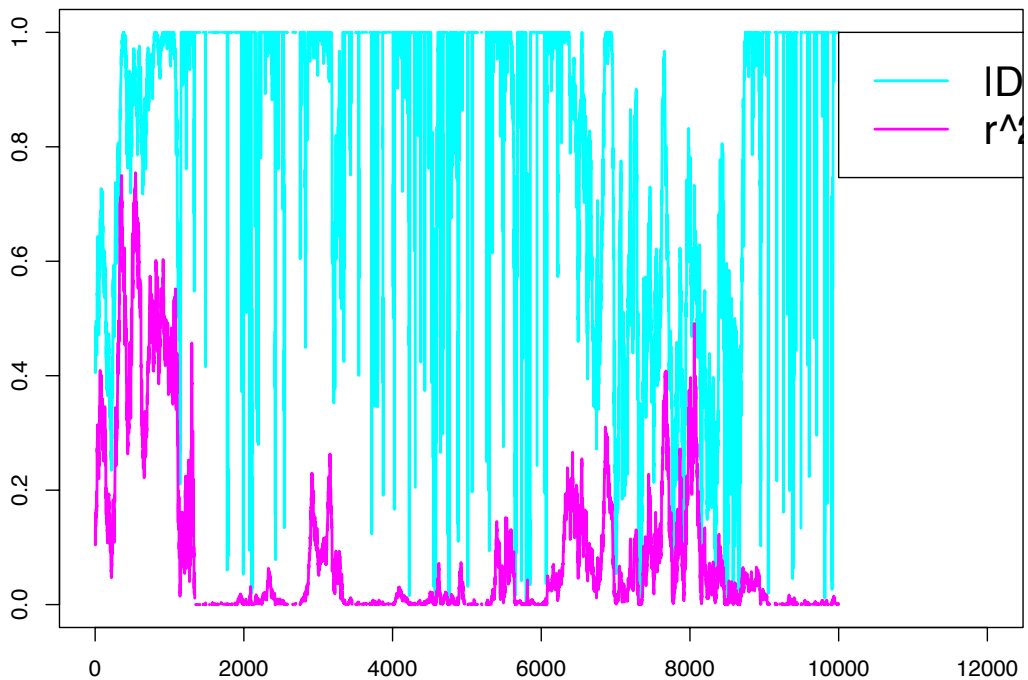
$$D' = \frac{0{\cdot}599 - 0{\cdot}999 \times 0{\cdot}6}{\min\{0{\cdot}001 \times 0{\cdot}4, 0{\cdot}999 \times 0{\cdot}6\}} = -1.$$

In fact, $|D'| = 1$ whenever any haplotype is absent from the population while both loci are polymorphic. However,

$$r^2 = \frac{(599 \times 0 - 400 \times 1)^2}{600 \times 400 \times 1 \times 999} = \frac{2}{3 \times 999} \approx 0.$$

Thus, according D' we have maximal LD, but according to r^2 we have almost no LD.





The 2-locus WF model, with $\mu > 0$ and low recombination. **Top:** the haplotype fractions evolving over 10K generations; **Bottom:** corresponding values of $|D'|$ and r^2. New mutants move $|D'|$ towards one, from which it declines towards zero; r^2 also increases but less so, depending on the frequency of the haplotype on which the new mutant arises.

- If the novel mutant increases in frequency, D' will quickly decline to zero unless the recombination fraction is small.
- In that case D' can remain high for many generations, and meanwhile r^2 can become large.
- Perfect LD in the sense of $r^2 = 1$ can only arise if the two loci are perfectly correlated, which means that only two haplotypes exist in the population, AB and ab.
- This can arise if the founding mutation events occurred on the same branch of the coalescent tree (and so occurred at about the same time).

Factors that affect LD: population structure

Two equally-large populations with:

	p_A	p_B	p_{AB}	D'	r^2
Pop. 1	0·5	0·1	0·05	0	0
Pop. 2	0·2	0·6	0·12	0	0

Both pops are in LE for these loci, but in the combined pop:

p_A	p_B	p_{AB}	D'	r^2
0·35	0·35	0·085	-0·306	0·027

Intuition: the observation of a haplotype bearing, say, the A allele, suggests that the haplotype originates from population 1, in which case the haplotype is likely to also carry the B allele.
LD due to population structure is "spurious" if we are interested in LD as a means to detect linkage.

2-locus WF linkage disequilibrium simulation code

```
ldsim = function(ngen=10000,nhap=2000,init=c(5,2,1,2),rho=0.5,mu=0){
  hp = matrix(0,ngen,4); p = rep(0,4); hp[1,] = init/sum(init);
  ldstat = matrix(0,ngen,2);
  p1 = hp[1,1]+hp[1,2]; p2 = hp[1,1]+hp[1,3];
  for(i in 2:ngen){
# haplotype proportions in next generation after recombination:
pp = hp[i-1,]*(1-rho)+rho*c(p1*p2,p1*(1-p2),(1-p1)*p2,(1-p1)*(1-p2));
# now let's have some mutation:
p[1] = sum(pp*c((1-mu)^2,mu*(1-mu),mu*(1-mu),mu^2));
p[2] = sum(pp*c(mu*(1-mu),(1-mu)^2,mu^2,mu*(1-mu)));
p[3] = sum(pp*c(mu*(1-mu),mu^2,(1-mu)^2,mu*(1-mu)));
p[4] = sum(pp*c(mu^2,mu*(1-mu),mu*(1-mu),(1-mu)^2));
# sample haplotypes in next generation and record counts
tmp = sample(1:4,nhap,repl=T,prob=p);
hp[i,] = hist(tmp,br=seq(0.5,4.5,1),plot=F)$c/nhap;
p1 = hp[i,1]+hp[i,2];  p2 = hp[i,1]+hp[i,3];  # allele prop at loc 1 and 2
# compute D'_00
D00 = hp[i,1]-p1*p2;
if(D00>0) ldstat[i,1] = D00/min(p1*(1-p2),p2*(1-p1))
        else ldstat[i,1] = -D00/min(p1*p2,(1-p2)*(1-p1));
# compute r^2_00
ldstat[i,2] = (hp[i,1]*hp[i,4]-hp[i,2]*hp[i,3])^2/p1/p2/(1-p1)/(1-p2);
  }
  cbind(hp[-1,],ldstat[-1,])
}
plotld = function(tmp,ngen=10000,nhap=2000,rho=0.5,mu=0){
  par(mfrow=c(2,1),mar=c(3,2,2,1));
  matplot(tmp[,1:4],type="l",xlim=c(0,ngen*1.2),ylim=c(0,1),lty=1,lwd=2,xlab="",ylab="haplotype \
   proportion",main=paste("LD sim: pop size=", nhap," rho=",rho,", mu=",mu));
  legend(ngen,1,leg=c("AB","Ab","aB","ab"),lty=1,col=1:4,lwd=2,cex=1.4);
  matplot(tmp[,5:6],type="l",xlim=c(0,ngen*1.2),ylim=c(0,1),lty=1,lwd=2,col=5:6,xlab="",ylab="");
  legend(ngen,1,leg=c("|D'|","r^2"),lty=1,col=5:6,lwd=2,cex=1.4)
}
```

After reading the `ldsim` and `plotld` functions into `R`, you can perform a run with default settings using:

```
res = ldsim()
plotld(res)
```

By default, `rho=0.5` corresponding to unlinked loci, and `mu=0` (no mutation). Since there is also no population structure and no selection in these simulations, levels of LD are low, varying only due to fluctuations in haplotype fractions due to finite population size. In most default simulations one allele will reach fixation and D' is then undefined.

Run some further simulations varying `rho`, `mu` and `nhap` to investigate their effects on D' and r^2.

- Any parameters set in `ldsim` should also be set to the same values in `plotld`, but they are only used for the plot legends.