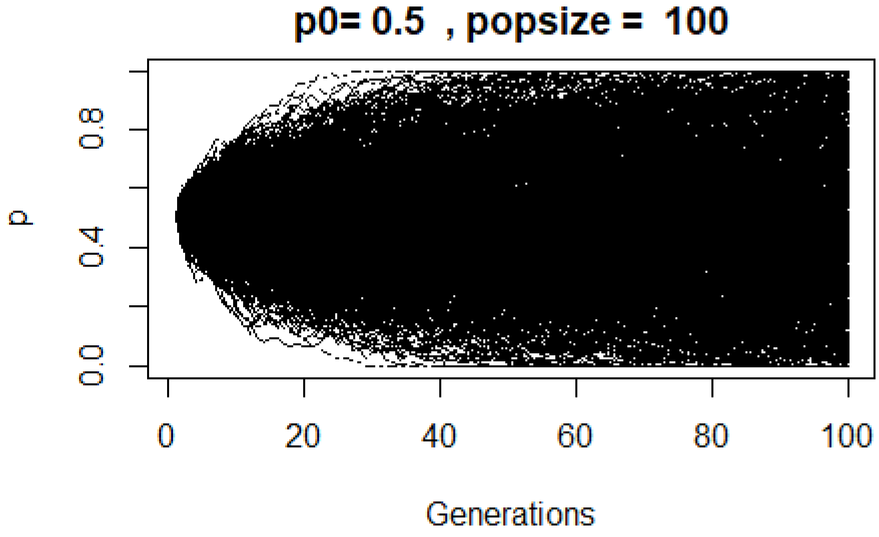
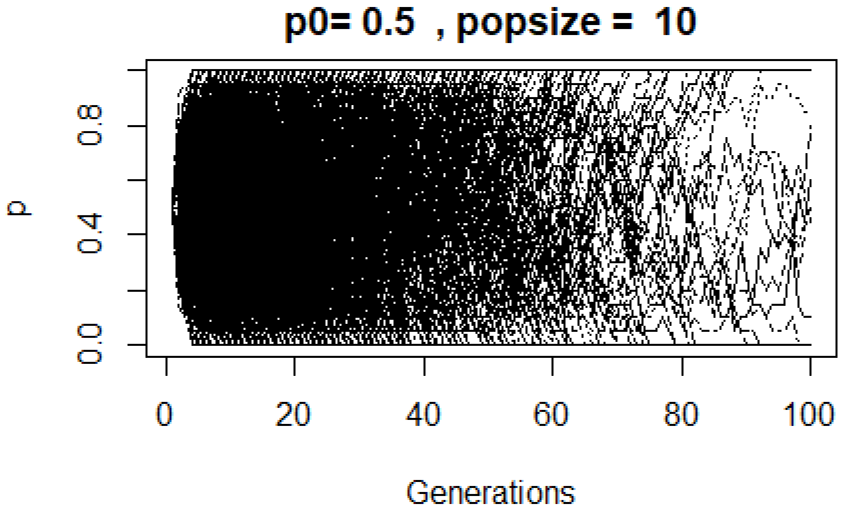
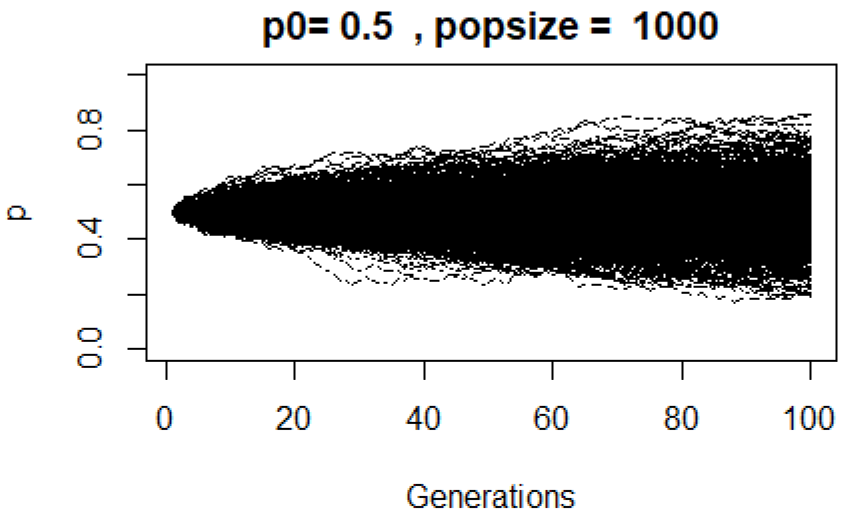
1. We will start by doing in silico Buri’s drosophila cages experiment. We will generate using the JGTeach::drift function the allelic frequency trajectories for a number of replicates, corresponding to the different cages:
    a. **Describe the figure produced**

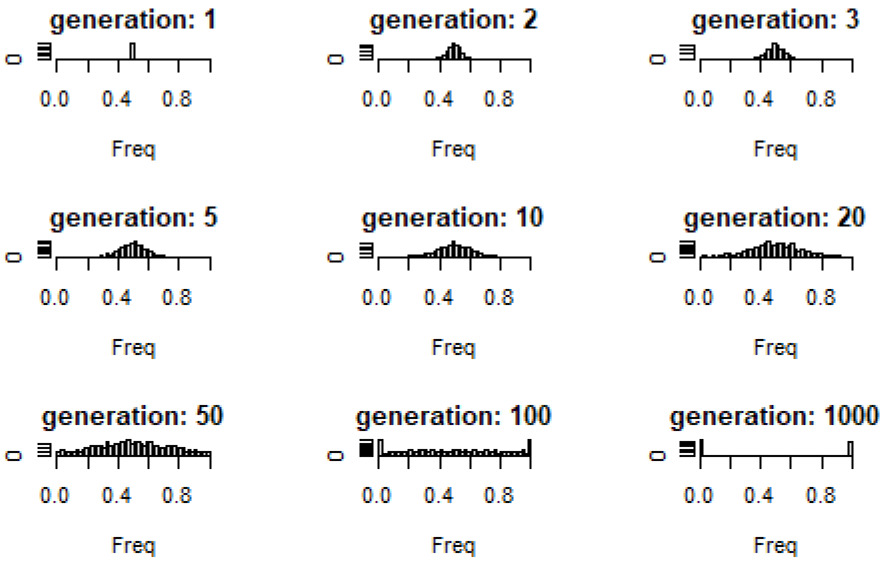


    b. set the number of individuals to 10; to 1000. **How does population size affects the drifting process?**





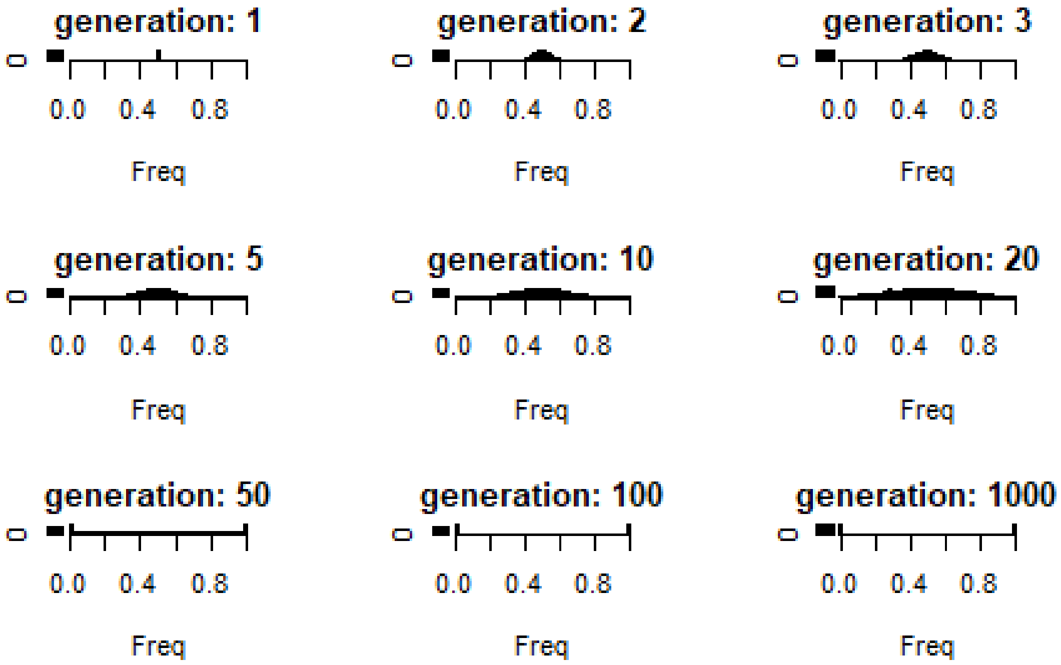
c. Set the number of individuals to 100 and the number of generation to 1000. Take time slices (generations 1, 2, 3, 5, 10, 50, 100,1000) to **look at the distribution of the replicates allele frequencies over time**.



d. **Describe the resulting figure and compare it to Buri's results**

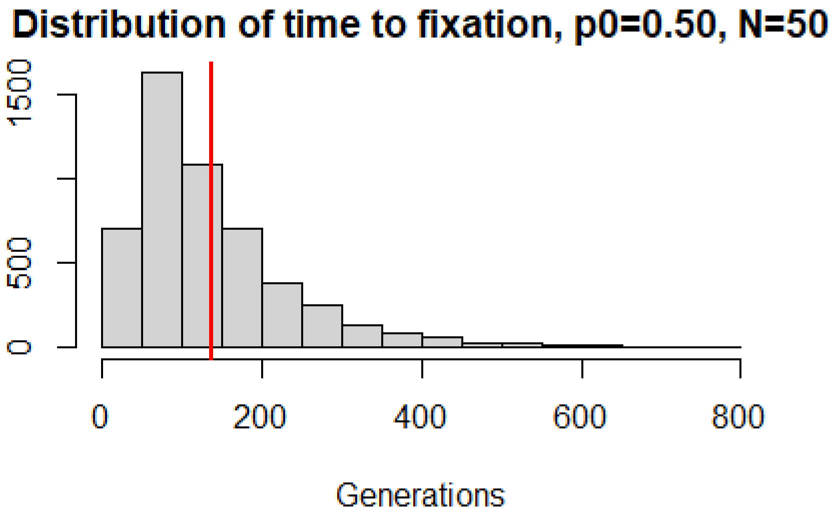
2. **[optional]** We will use the drift function to explore some properties of drift, namely the probability and time to fixation of an allele.

   a. **[optional]** simulate 10,000 replicates of a population of 50 diploid individuals starting with a frequency p0=0.5. The expected probability of fixation is 0.5, and it should take on average 135 generations to achieve fixation.



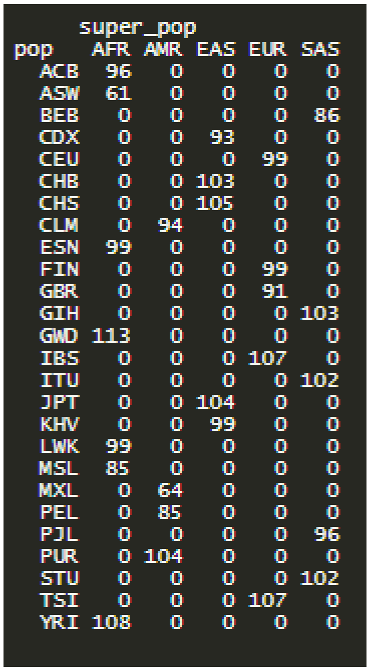
b. **[optional]** extract the time to fixation of the different replicates, from which you will extract both the probability and mean time to fixation.

c. **[optional]** Make a histogram of the time to fixation and **describe it'**



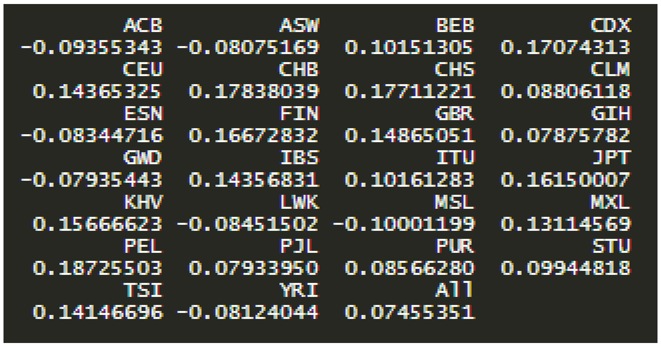
3. Simulate using hierfstat::sim.genot.t 3 populations evolving independently one from another, in order to check the approximation.
    a. **Check whether the approximation holds after 200 generations**
    b. **[optional] Imagine a way to do this simulation with ms.**
4. **[optional]** Have a look at the help page, and run the examples (fig 1 of the paper)

5. **[optional]** Next you will show that the estimates matches the expected values. For this, we will make use of the hierfstat function sim.genot.metapop.t to generate genetic data according to this model
6. Load chromosome 22 fragment VCF file from the 1000 genome into R (see practical 1) and the samples description
    a. **Describe object samp.desc:** Print a table of the number of samples per super_pop and pop

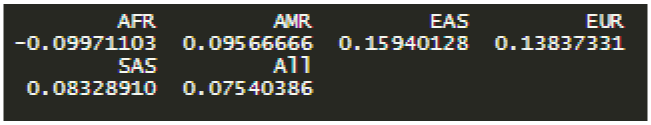


```
     super_pop
pop   AFR AMR EAS EUR SAS
  ACB  96   0   0   0   0
  ASW  61   0   0   0   0
  BEB   0   0   0   0  86
  CDX   0   0  93   0   0
  CEU   0   0   0  99   0
  CHB   0   0 103   0   0
  CHS   0   0 105   0   0
  CLM   0  94   0   0   0
  ESN  99   0   0   0   0
  FIN   0   0   0  99   0
  GBR   0   0   0  91   0
  GIH   0   0   0   0 103
  GWD 113   0   0   0   0
  IBS   0   0   0 107   0
  ITU   0   0   0   0 102
  JPT   0   0 104   0   0
  KHV   0   0  99   0   0
  LWK  99   0   0   0   0
  MSL  85   0   0   0   0
  MXL   0  64   0   0   0
  PEL   0  85   0   0   0
  PJL   0   0   0   0  96
  PUR   0 104   0   0   0
  STU   0   0   0   0 102
  TSI   0   0   0 107   0
  YRI 108   0   0   0   0
```

    b. Verify that samples are the same and in the same order in ch22 and samp.desc.
    c. Estimate population and continent FSTs from this dataset (it takes some time). You will have to download the Allele sharing (matching) file first



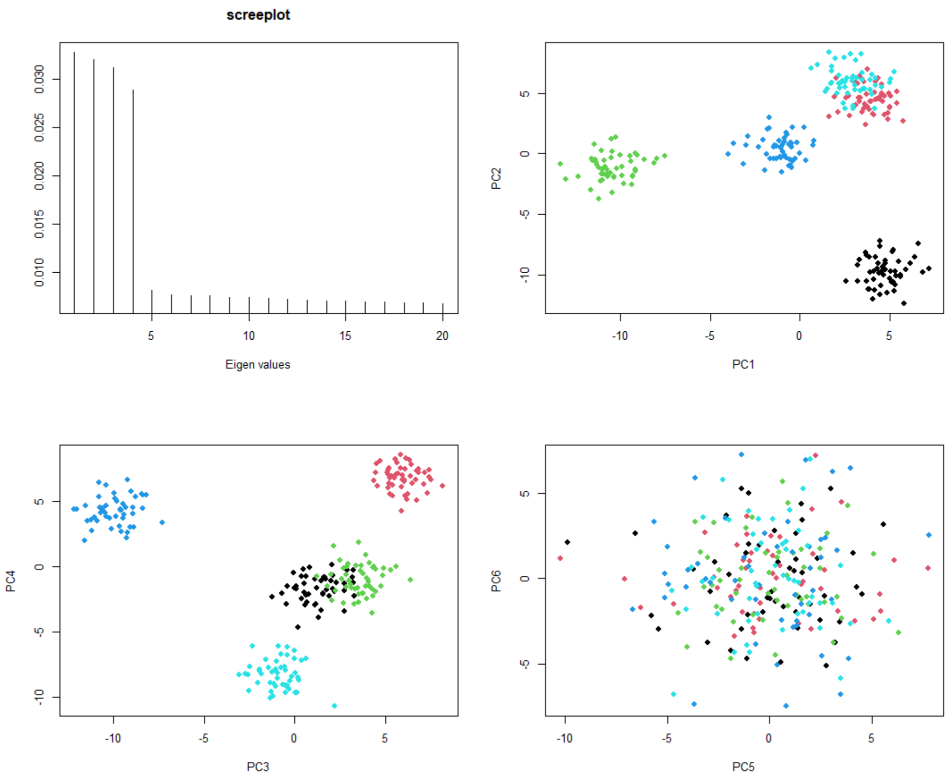
```
        ACB         ASW         BEB         CDX 
-0.09355343 -0.08075169  0.10151305  0.17074313 
        CEU         CHB         CHS         CLM 
 0.14365325  0.17838039  0.17711221  0.08806118 
        ESN         FIN         GBR         GIH 
-0.08344716  0.16672832  0.14865051  0.07875782 
        GWD         IBS         ITU         JPT 
-0.07935443  0.14356831  0.10161283  0.16150007 
        KHV         LWK         MSL         MXL 
 0.15666623 -0.08451502 -0.10001199  0.13114569 
        PEL         PJL         PUR         STU 
 0.18725503  0.07933950  0.08566280  0.09944818 
        TSI         YRI         All 
 0.14146696 -0.08124044  0.07455351 
```



```
        AFR         AMR         EAS         EUR 
-0.09971103  0.09566666  0.15940128  0.13837331 
        SAS         All 
 0.08328910  0.07540386 
```

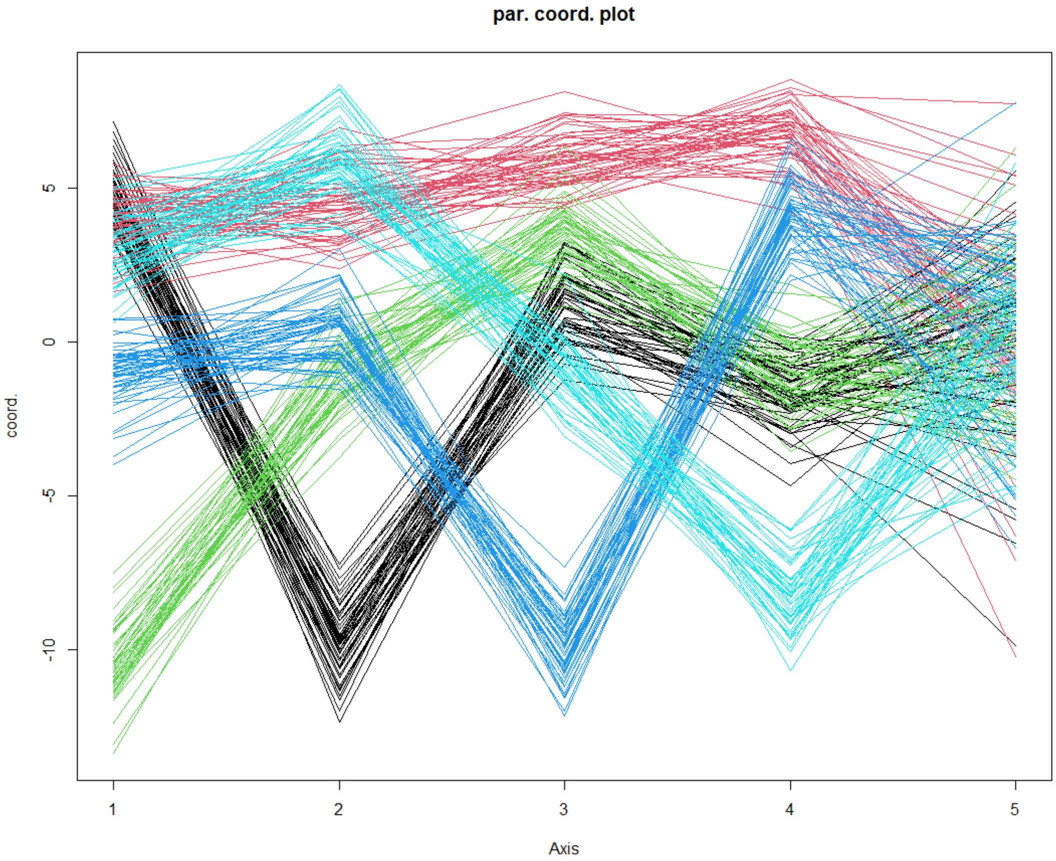
    d. There is a plot function associated to fs.dosage. Produce the corresponding plot after having ordered the population by continent and discuss the results in the light of what you know about human demographic history: the top left panel show individual inbreeding coefficients per population, relative to the mean kinship in their population. The right column panels show FXYST (or βXY) on top and population specific FST at the bottom. Last, the bottom left panel shows pairwise FST. **Do you see a similarity between FXYST and the heatmap of KAS kinship?**

7. In this part, we will learn how to conduct Principal Component Analyses (PCA) on a genomic data set, **[optionally]** look at the different flavors of PCA, and get a feel for how to interpret the results
    a. Start by simulating the genotypes of 250 individuals, 50 from each of 5 populations, at 1000 bi-allelic loci, assuming an island model at equilibrium between migration, mutation and drift. Assume each island is made of 1000 individuals, with m=0.003 between them. For this, use the sim.genot function of the hierfstatpackage.
    b. Convert the data into dosage format using the biall2dos function
    c. Transform the dosage matrix into matrix X, as done in Patterson et al. (2006). You may find the scale function useful for this.
    d. Calculate XX′ using function tcrossprod
    e. Obtain the eigen value decomposition of XX′ using the eigen function
    f. Obtain the individual's coordinates UV1/2
    g. **Compare the results you obtain to what is obtained using the prcomp function on matrix X**
    h. In a four panel graphic windows, produce the following 4 plots:
        i. A scree plot of the first 20 eigen values, expressed as proportion of the sum of all eigenvalues
        ii. A plot of PC2 against PC1, using a different color for each population
        iii. A plot of PC4 against PC3
        iv. A plot of PC6 against PC5



i. **Describe what you see**: How many eigen values stand out? are samples grouped according to their population of origin? If so along which axes? How many PCs are necessary to describe the structure of this dataset?

j. Produce a parallel coordinate plot, where the x-axis correspond to the first 5 PCA axes, and the y axis shows the coordinates of each individuals along the 5 axes, drawn as a line with color corresponding to the population in which the individual has been sampled



8. **[optional].** Redo this analysis on the matrix of dosages centered but not scale to sd=1. **Conclusions?**
9. **[optional].** Instead of using XX′ for eigenvalue decomposition, use the kinship matrix KAS obtained from the allele sharing matrix, you might want to use the beta.dosage function for this. **Conclusions?**
10. **[optional].** Simulate a new genotype data set, this time with only three populations. Rerun step 2-9 above. **Conclusions?**
11. **[optional].** You might want to experiment with higher migration rate in the simulated data set, and observe the effect it has. **Conclusions?**
12. **[optional].** Rather than an island model of population structure, you might want to simulate a stepping stone in one dimension, using the sim.genot.metapop.t function. Assume 8 demes, connected by migration with nearest neighbors only, and with 5% migration between neighbors. Redo the PCA on the data set. **Interpret the figure.**