Elements of R for Genetics & Bioinformatics
Session 7: Large data files

Before you start: download and install the `RNetCDF` package from CRAN – one way to do this is via the GUI menus.

Next: Download the file `hapmap6.nc` from the module website (using your web browser, not in R)

1. Tabulate the genotypes for each of the first ten SNPs (they are coded 1,2,3 for the AA, Aa, aa genotypes)

2. Compute the proportion of heterozygous genotypes for each individual

3. The file `SEAflights.db` is an SQLite database with the same flights data you saw in Session 3. Read in the arrival and departure delays for all flights from SFO. Check that the mean delay times match what you found in that session.

4. Database file `hapmap6.db` contains the same data as `hapmap6.nc`. For the task in Q1, compare the speed using the SQL vs nCDF versions.

5. (For keen people) Load the full dataset from `hapmap6.nc,` and compute principal components of ancestry;

    a. First, use `scale()` to scale the data so each SNP has the same variance

    b. Remove any SNPs with missing data (in real life, one might impute these)

    c. Use `svd(yourmatrix, nu=10, nv=0)` to get 10 principal components, and plot them – you could use base R's `pairs()` function or `parcoord()` in the MASS package.