The program fdist2 has been modified from the original fdist in two main
respects:

a) You can specify the number of islands (demes) that can be occupied (originally a model with 100 islands was used, as an approximation to the
infinite-island case). This number must always be greater than or equal to
the number of samples that you have.

b) Since computers are so fast these days, I have simplified the program so
that the user doesn't have to specify a particular mutation rate. The program
now attempts to get a (very!) roughly uniform distribution of heterozygosities
by first generating a heterozygosity uniformly on (0,1) and then treating this
as an 'expected value' and generating a theta by inverting the known relationships for the expected heterozygosity as a function of theta for the
stepwise and infinite allele mutation models.

The program cplot has been modified to allow the user to specify the level of
the equal-tail probability interval, which was previously just fixed at 0.95.

Important note, especially relevant for human and microsatellite data.

These programs carry out analyses described in Beaumont and Nichols (1996) Proc
Roy. Soc. Lond. B. 263: 1619-1626. This paper demonstrated that the distribution of Fst with heterozygosity was generally very robust to many
demographic assumptions. However it was not appreciated in this paper, and only
became apparent with extra results reported in Flint et al. (1999) Human Genet
105: 567-576, that much of the robustness in Beaumont and Nichols (1996) stems
from the assumption that the number of populations is large - i.e. that we are
dealing with something close to the infinite island model. With smaller
populations, the mutation rate, mutation model and the demographic history
(i.e. migration model) may have a larger effect on the distribution of Fst - in
particular depressing it at high heterozygosities.

It looks as though much of the genetic variability in humans can be well
partitioned into 3 continental groups - African, Asian, and European. So if you
are going to analyse human data with fdist you are probably going to want to
lump your samples into 3, corresponding to these groupings, and set the model
to have a maximum of 3 islands (demes). If this is done then the

following
observations are pertinent:

a) the theoretical distribution of Fst is quite broad - roughly, the distribution narrows slightly with decreasing number of islands, but narrows more strongly with increasing number of samples.

b) there is a depression of Fst at high heterozygosities.

c) stepwise and infinite-allele models have similar distributions of Fst up
until heterozygosities of around 0.8. Thereafter, the infinite allele model
predicts a sharper decline in Fst with heterozygosity than the stepwise model.

d) Beaumont and Nichols (1996) found that equilibrium and non-equilibrium (e.g.
populations diverging from a common ancestral population) models of population structure had very similar distribution of Fst. This assumed a large
number of demes for the equilibrium model, and we haven't tested whether this
holds up with a small number of demes.

e) Given the way that average Fst varies with heterozygosity, I suggest a
modification to the original approach of Beaumont and Nichols (1996) whereby
the mean Fst calculated using datacal is plugged into fdist, and then this is
iteratively modified as outlying loci is removed. A more robust approach may be
to do a more exploratory analysis where you first use the mean of your data to
get the fdist output, as before, but you then vary it slightly to try to get
half the data points above and 1/2 the data points below the median line that
is output by cplot. You could then do an iterative fitting, as suggested in the
paper, where you remove the outlying points and then redo the analysis,
however, since you are now matching up the medians, it won't matter very much
because removing the odd outlying point will only slightly alter the medians
(assuming you have an extensive number of loci).

This distribution contains the files
README_fdist2
INTFILE
INTFILE_orig
fdist_params2.dat
infile
datacal.exe
cplot.exe
pv.exe
datacal.exe
fdist2.exe
as100.c
as99.c

cplot.c
datacal.c
fdist2.c
pv.c

The .exe files are pc binaries, which is only what most people want. I
include the source file in case anyone wants to play with them. In this case
compilation is e.g.:

```
gcc -o fdist2 -O fdist.c -lm
gcc -o cplot -O cplot.c as100.c as99.c -lm
gcc -o pv -O pv.c as100.c as99.c -lm
gcc -o datacal -O datacal.c -lm
```

For queries, complaints, etc please email m.a.beaumont@reading.ac.uk

These programs originated very much as research tools, they
are not particularly user friendly, but should work if used as
described.

The file INTFILE contains the state of the random number generator, and
copies of it should live in the same directory as the data you are analysing.
INTFILE_orig is just a copy, in case you lose/corrupt INTFILE.

infile and fdist_params2.dat contain the Singh and Rhomberg drosophila data and
a suitable parameter file for fdist. You should be able to use them to recreate
figure 1 in the paper.

fdist2 - this is the simulation program. It reads a file
"fdist_params2.dat", which consists of 6 lines. This file must only contain
numbers in the format described:
Total number of demes (100 max).
No of populations sampled (must be less than or equal to total number of demes).
Expected Fst for infinite allele, infinite island model.
Sample size (assumed the same in all populations)
Indicator - 1 for stepwise mutations; 0 for infinite alleles
No of realizations (loci)

The program takes around 10 minutes to carry out 20,000 realizations with the
parameter file here on a 500Mhz Pentium. **I strongly recommend running more
realizations than suggested for the original fdist, now that the method of
varying heterozygosity has been changed - do a minimum of 20,000. The more
points you do the smoother the contours become.**

Output in "out.dat". 2 columns: (1-hat(f1)) ; (hat(f0) - hat(f1))/(1-hat(f1)).
If a stepwise mutation model is used, or a small number of demes, the mean Fst
of the simulated data will generally be less than the expected value, and will
vary with heterozygosity so it is worth playing around to get the value you

want.

datacal - this program will take input from a file "infile" and give you the
heterozygosity and fst estimates for each locus in a file "data_fst_outfile",
which can then be compared with the simulation output. infile should have the
following format
1/0 indicator. alleles by rows in data matrix (1), or pops by rows (0).
No of pops
no of loci
no of alleles at locus 1
matrix of data at locus 1 either with each row corresponding to the same
allele or to the same population.
no of alleles at locus 2
matrix...
.
.
.
etc.
datacal also gives you the mean fst, and the median sample size, which can be
input as parameters to fdist in an initial analysis (see note above). It will
also produce pairwise mean fsts for pairs of populations in a file "pdist.dat".
These can then be ordinated. pdist.dat consists of one column of pairwise Fsts
corresponding to the upper triangle of a symmetric matrix in row order. The
data matrices can contain populations for which a locus was not genotyped,
these should be indicated by zero entries (see the infile). Fst is undefined
when populations are homozygous for an allele. These are indicated as -100.0 in
any output, but it should only be a problem in pairwise calculations when all
loci are monomorphic in a pair. Undefined Fsts are not included in any
averages. I have not included an ordination program. Many programs will carry
out the ordination - ie cmdscal in R or Splus; there are ordination routines in
NTSYS.

datacal and fdist calculate Fst in the way described in the paper.

cplot - This will calculate the quantiles as described in the paper from the
output of fdist. It will prompt you for input file (e.g "out.dat") and output
file (any name you want). It will then prompt you for the probability level.
Call this pval.
The structure of the output is
Heterozygosity, 0.5(1-pval) quantile, median, 0.5(1+pval) quantile

pvj - This will give approximate p-values for each data point. p-value means
the probability of getting values as small as or smaller than the

data point.
It will prompt you for data file (e.g. "data_fst_outfile"), output file (any
name you want), and simulated data file(e.g. "out.dat").
The structure of the output is
sample het, sample fst, test statistic, P(simulated Fst < sample Fst)

To run example data
(1) Double click on datacal - This will produce the observed Fst and heterozygosities in "data_fst_outfile")
(1) Double click on fdist. Wait for it to stop.
(2) Double click on cplot. At prompt enter "out.dat quant.dat"
(3) You can then visualize quant.dat in your favourite plotting program.
(4) Double click on pv. At prompt enter "data_fst_outfile probs.dat out.dat"
(5) Inspect the p-values in probs.dat.

Note - from a unix background I've tended to put .dat on the end of data files,
but these have some special meaning in windows, so you may want to name or rename
files to e.g. .txt or something more suitable for automatic opening, otherwise you
tend to get warning messages.

Another unix vs. windows problem is that the window disappears the instant a program stops. I've put in an ad hoc prompt in datacal and fdist2 to stop this happening if the program runs normally, because these programs type out informative messages. However if the program dies
with an error message you won't see the message. In which case it may be better
to run the program from the dos prompt.