Summer Institute in Statistical Genetics
University of Queensland, Brisbane
Population Genetics Module

A likelihood-based approach to allele count data from structured populations

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

Feb 10, 2017

1. Allele fractions in structured populations

2. Interpretations of F_{ST}

3. A sampling formula: the Beta-Binomial likelihood

4. Likelihood inference for the inbreeding coefficient

5. Likelihood inference for F_{ST}

6. F_{ST} estimates in worldwide human populations for forensics

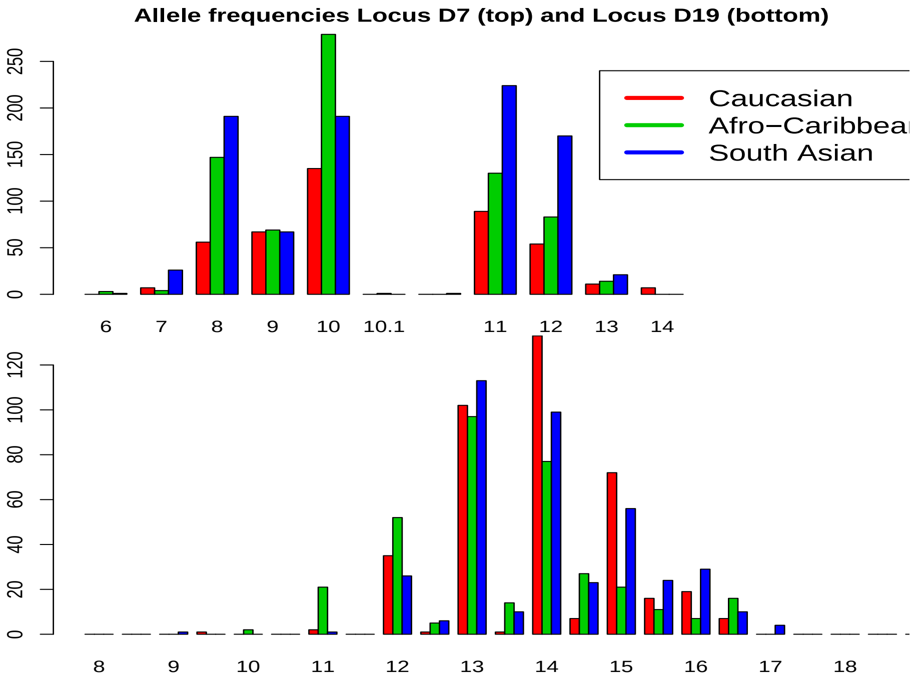
1. Allele fractions in structured populations
2. Interpretations of F_{ST}
3. A sampling formula: the Beta-Binomial likelihood
4. Likelihood inference for the inbreeding coefficient
5. Likelihood inference for F_{ST}
6. F_{ST} estimates in worldwide human populations for forensics

- Allele fractions vary across subpopulations due to a number of factors, such as
  - drift,
  - migration,
  - mutation,
  - selection.
- Here we do not focus on the causes but we wish to describe this variation.

In the figure on the next slide we see allele counts in samples drawn from three subpopulations of the UK human population.

- The counts are at two multi-allelic Short Tandem Repeat (STR) loci used in forensics.
- The allele labels indicate the number of copies of the (tetranucleotide) tandem repeat (not all allele labels are shown).
  - An allele label of the form x.y means x full copies of the repeat unit plus a y nucleotide fragment (usually $1 \leq y \leq 3$).
  - Soon full sequencing of STR alleles will distinguish all microvariants; to date allele classification is based only on the number of nucleotides.

Which locus has the greater variance in allele fractions across these three subpopulations?



Allele frequencies Locus D7 (top) and Locus D19 (bottom)

To define a variance of allele counts or fractions we need to have some probabilities. To keep things simple let's assume a diallelic locus, and focus on one allele, call it A.

- What are the probabilities for the number X of A alleles obtained when sampling n alleles at random in a given large subpopulation?

If we know the fraction π of A alleles in the subpopulation then the probabilities are given by the binomial distribution:

$$\Pr(X{=}m|\pi) = \binom{n}{m}\pi^m(1{-}\pi)^{n-m} \qquad \text{for } m = 0, 1, \ldots, n. \tag{1}$$

In particular, when $n{=}1$, the probability that the allele is A is

$$\Pr(X{=}1|\pi) = \pi.$$

Now suppose that we don't know π, but we assume that we know p, the population fraction of A alleles.

- A natural assumption is $\mathrm{E}(\pi) = p$; i.e. allele fractions in subpopulations vary about the population value.

Then the variance of π, denoted $\mathrm{V}(\pi)$, reaches its maximum value

$$\mathrm{V}(\pi) = p(1-p) \qquad \text{when} \qquad \Pr(\pi{=}1) = p \text{ and } \Pr(\pi{=}0) = 1-p.$$

It is therefore convenient to write

$$\mathrm{V}(\pi) = Fp(1-p)$$

where $F \in [0, 1]$.

Our goal is to estimate F, a parameter introduced by Wright (1951) who called it F_{ST}, where S is for subpopulation and T is for total population. It is also sometimes called θ (but θ has several other meanings in population genetics).

1. Allele fractions in structured populations
2. Interpretations of F_{ST}
3. A sampling formula: the Beta-Binomial likelihood
4. Likelihood inference for the inbreeding coefficient
5. Likelihood inference for F_{ST}
6. F_{ST} estimates in worldwide human populations for forensics

Wright (1951) interpreted F_{ST} as measuring the average progress of subpopulations towards fixation, and hence he called it a *fixation index*.

- $F_{ST} = 1$ implies that all subpopulations have reached fixation ($\pi = 1$ or 0) at the locus;
- $F_{ST} = 0$ implies that $\pi = p$ in all subpopulations, and so the population is homogeneous.

Wright also described F_{ST} as

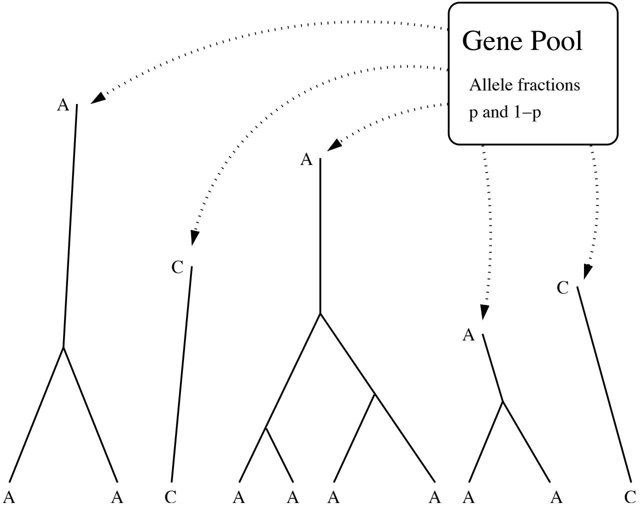
> *"the correlation between random gametes, drawn from the same subpopulation, relative to the total"*

This correlation is due to relatedness, and F_{ST} can also be interpreted as measuring the relatedness among individuals within sub-populations relative to the total population (Crow and Kimura, 1970).

- Thus it is often called a *coancestry coefficient*.
- More relatedness within subpopulations means higher F_{ST} and a greater variation in allele fractions across subpopulations.

In some simple models F_{ST} is the probability that two alleles drawn at random in the subpopulation are identical-by-descent (IBD) from an ancestral allele within the subpopulation (without any migration event).

In the model sketched here IBD $=$ correlation, but more generally they are not the same (correlations can be $-$ve).



The biggest factor affecting F_{ST} in most structured populations is migration, and in the past F_{ST} was estimated as an indirect way to estimate the migration rate, for example via the formula

$$F_{ST} = \frac{1}{1 + 2Nm}$$

(replace 2 with 4 for diploid populations, $N =$ population size) which holds in a simple island model assuming symmetric migration at rate m between all pairs of subpopulations.

In order to estimate F_{ST} we have to deal with the problem that we don't know p.

- At first, we solve this problem by simply pretending we know it.

However we still need to keep in mind what p represents, and there are at least two schools of thought:

1. p is the actual allele fraction among all individuals in all the subpopulations; this means that the largest subpopulations dominate the value of p.
2. p is the allele fraction in a hypothetical ancestral population from which all the observed subpopulations are descended.

p is unknown in either case, but can be estimated.

In general p can be any reference value, but its definition affects the value and interpretation of F_{ST}.

- In forensic applications p is the allele fraction in the population from which the frequency database was drawn. Then F_{ST} is defined in terms of the mean square error (MSE) of π about the given reference value,

$$\mathsf{MSE}[\pi, p] = \mathsf{E}\left((\pi - p)^2\right) = F_{ST}\, p(1-p).$$

There are many methods for estimating F_{ST}; we won't review them all.

- Some methods can be classified as "method of moments" estimation, which is based on equating sample moments (mean, variance, etc) to their expected values under the assumed probability model
  - it isn't necessary to specify a full probability distribution for π.
- Some of these methods are based on the idea of partitioning the variance in the sample allele counts into within- and between-subpopulation components of variance.
- Because there is little information in a single allele count, it is necessary to "share information" across different alleles at a multi-allelic locus, or across subpopulations or across loci.
  - This can require assuming that F_{ST} is constant.
  - In the past it was common to assume F_{ST} constant across populations, which is rarely true due to different N_e and demographic histories.
  - Now we usually have many markers genome-wide and F_{ST} can be estimated for individual populations or for pairs of populations by averaging over markers; however selection can cause some markers to have discrepant F_{ST} values.
- Bhatia et al. (2013) is a recent reference focussing on genome-wide human data and the effect of many rare variants on estimates of F_{ST}.

1. Allele fractions in structured populations
2. Interpretations of F_{ST}
3. A sampling formula: the Beta-Binomial likelihood
4. Likelihood inference for the inbreeding coefficient
5. Likelihood inference for F_{ST}
6. F_{ST} estimates in worldwide human populations for forensics

We will focus here on *likelihood-based* estimation of F_{ST}.

- I developed some of this approach, see particularly Balding (2003) and also my DNA forensics book Steele and Balding (2015).

To proceed, we need to specify a probability distribution for π (remember, we are assuming that p is known). For a diallelic locus, a natural candidate is the *beta* distribution, which has probability density function (pdf)[1]:

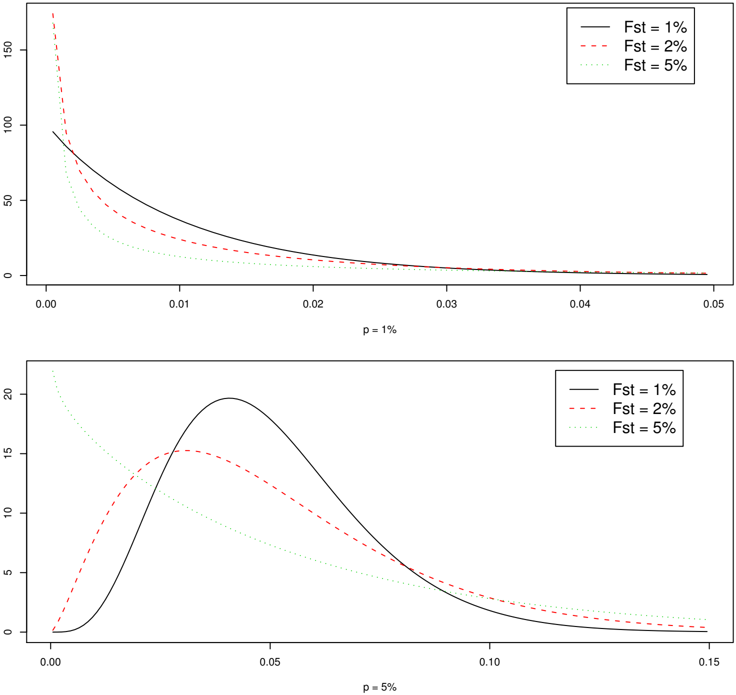
$$f(x) = cx^{\lambda p-1}(1-x)^{\lambda(1-p)-1}, \tag{2}$$

where c is a normalising constant whose value is known but not needed here, $0 \leq x \leq 1$, and

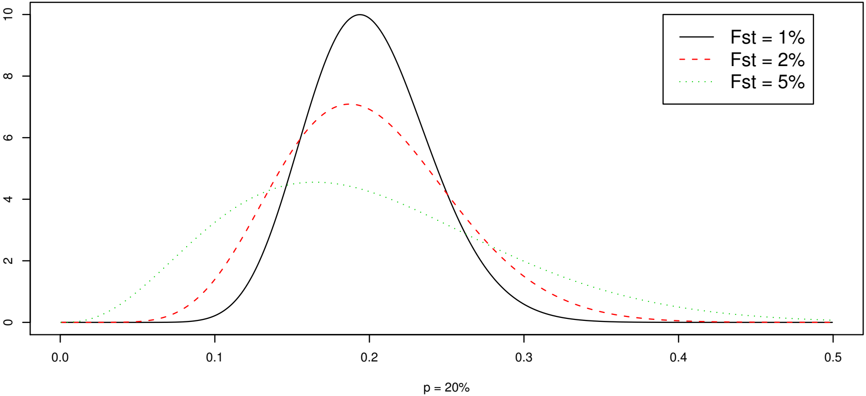
$$\lambda = \frac{1}{F_{ST}} - 1.$$

The expectation and variance of the beta are, respectively, p and $F_{ST}p(1-p)$.

[1] NB This parametrisation of the beta is not standard, the usual parametrisation has $\alpha = \lambda p$ and $\beta = \lambda(1-p)$. Then the mean is $\alpha/(\alpha+\beta) = p$ and variance is $\alpha\beta/((\alpha+\beta)^2(\alpha+\beta)+1) = p(1-p)F_{ST}$.

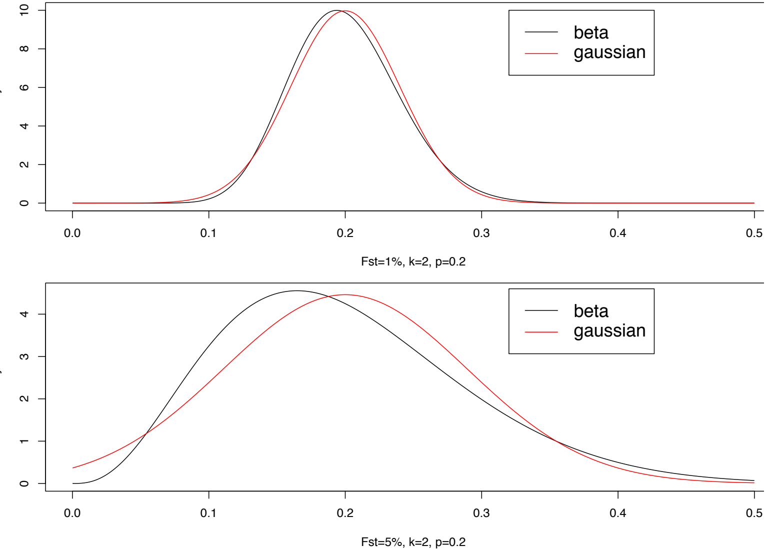


Areas under the curves represent the probabilities for different values of π given the value of p shown on the x-axis (see next slide for further details).



Beta pdf when $p = 0.01$ *(previous page, top),* $p = 0.05$ *(previous page, bottom),* $p = 0.20$ *(above), and* $\lambda =$ *99, 49, and 19, so that* $F_{ST} =$ *1%, 2%, 5%.*

The beta distribution applies exactly under some theoretical models, both of pure drift and of weak selection in large populations (Ewens, 2004). It allows the essential features of genetic differentiation to be modelled and estimated in actual populations.



An alternative to the beta (Nicholson et al., 2002) is the truncated Gaussian, with probability density outside (0,1) replaced with atoms of probability at 0 and 1.

To obtain the probability distribution for the allele count X in a subpopulation, we need to integrate the binomial probabilities (1) over the beta distribution for π. Remarkably, this integration can be done exactly and the result is the Beta-Binomial (BB) distribution, which we can represent schematically as

$$\mathrm{BB}(X) = \int \mathrm{binomial}(X|\pi)\mathrm{beta}(\pi)\mathrm{d}\pi$$

The BB is like the binomial but with higher variance, controlled by an additional parameter: BB variance is $np(1-p)(1+(n-1)F_{ST})$, which equals the binomial variance $np(1-p)$ when $F_{ST} = 0$ or $n = 1$.

Sampling formula: There is a simple recursive formula for the BB probabilities. Suppose that n alleles have been sampled in the subpopulation, of which m are A. Then the probability that the next allele sampled in the subpopulation is also A is:

$$\frac{mF_{ST} + (1-F_{ST})p_A}{1 + (n-1)F_{ST}}. \qquad (3)$$

When $m = n = 0$, we obtain probability p_A that the first allele drawn is A. The probability that the first two alleles drawn are both A is

$$p_A(F_{ST} + (1-F_{ST})p_A) = p_A^2 + F_{ST}\,p_A(1-p_A).$$

Increasing F_{ST} thus increases the probability of two A alleles, but decreases the probability of an A allele followed by a B, which is:

$$(1-F_{ST})p_A p_B. \tag{4}$$

The probability of an A and a B in an unordered sample of size two is obtained by multiplying (4) by two.

Non-recursive form of the sampling formula (3): the probability of an unordered sample of size n containing m copies of allele A is

$$\Pr(X=m) = \binom{n}{m} \frac{\Gamma(\lambda)}{\Gamma(n+\lambda)} \frac{\Gamma(m+\lambda p)}{\Gamma(\lambda p)} \frac{\Gamma(n-m+\lambda(1-p))}{\Gamma(\lambda(1-p))}, \tag{5}$$

where Γ is the gamma function, which satisfies $\Gamma(x+1) = x\Gamma(x)$. Replacing λ with $1/F_{ST}-1$ we obtain the likelihood formula for F_{ST}.

The multi-allelic case: Formula (3) still holds for a locus with > 2 alleles. The multivariate extension of the beta distribution is the Dirichlet, which has pdf:

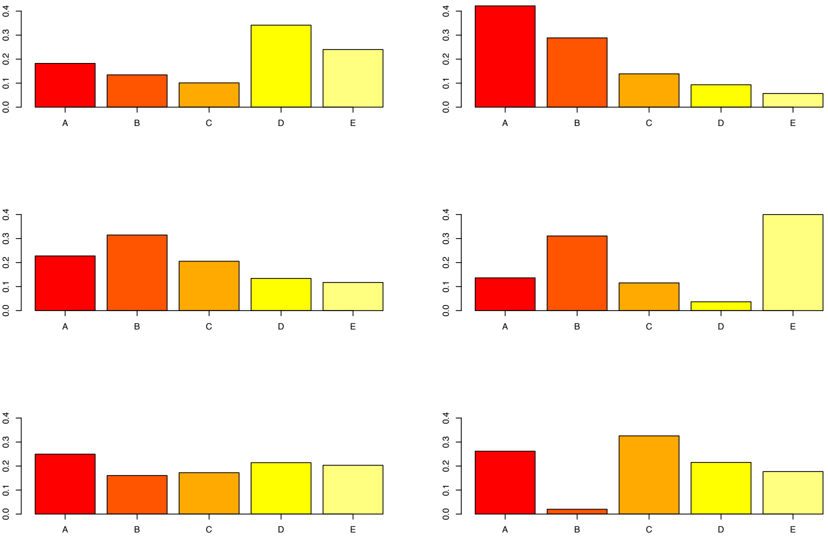
$$f(x_1, x_2, \ldots, x_K) = c \prod_{k=1}^{K} x_k^{\lambda p_k - 1}, \tag{6}$$

where c is a constant, $p_1, p_2, \ldots, p_K$ denote the population allele fractions, with $\sum_{k=1}^{K} p_k = 1$, and similarly the x_k are all positive and sum to one. If $K = 2$ then $p_2 = 1 - p_1$ and $x_2 = 1 - x_1$ and the beta pdf (2) is recovered.

The non-recursive form of the multinomial-Dirichlet is

$$\Pr(\mathbf{X} = \mathbf{m}) = \frac{n!\Gamma(\lambda)}{\Gamma(n+\lambda)} \prod_{k=1}^{K} \frac{\Gamma(m_k + \lambda p_k)}{m_k!\Gamma(\lambda p_k)}, \tag{7}$$

where $\mathbf{m} = (m_1, m_2, \ldots, m_K)$ denotes the sample count vector so that $n = \sum_{k=1}^{K} m_k$.



Allele fractions simulated under the multinomial-Dirichlet for a 5-allele locus in 3 subpopulations, 100 alleles sampled per subpopulation and $F_{ST} = 1\%$ (left) and 5% (right).

R code to simulate from the beta-binomial

Function BB generates a histogram of counts of A allele in samples of size nall from each of npop populations, for given Fst and p. The standard deviation of the allele count is also returned.

```
BB = function(Fst=0.01,npop=50,nall=10,p=0.2){
   if(Fst>0){
      lam = 1/Fst-1
      pi = rbeta(npop,lam*p,lam*(1-p))
   }
   else pi = rep(p,npop)
   dat = rbinom(npop,nall,pi)
   hist(dat,n=20)
   return(sd(dat))
}
```

Use BB to compare the distribution of the allele count for different values of F_{ST} when nall = 100 and p = 0.2.

Computing multinomial-Dirichlet probabilities under the sampling formula: Suppose that there are three alleles with population fractions p_1, p_2, and p_3, so that $p_1 + p_2 + p_3 = 1$. Using (3) repeatedly, or (7), the probability P(1,1,1) that an unordered sample of size three from the subpopulation consists of one copy of each allele is

$$\Pr(1,1,1) = \frac{6}{(1-F_{ST})(1+F_{ST})} \prod_{k=1}^{3} (1-F_{ST}) p_k = 6 p_1 p_2 p_3 \frac{(1-F_{ST})^2}{1+F_{ST}}.$$

Similarly,

$$\begin{aligned} \Pr(2,1,0) &= 3 p_1 p_2 (1-F_{ST}) \frac{(F_{ST}+(1-F_{ST})p_1)}{1+F_{ST}}, \\ \Pr(3,0,0) &= p_1 (F_{ST} + (1-F_{ST})p_1) \frac{(2F_{ST} + (1-F_{ST})p_1)}{1+F_{ST}}. \end{aligned}$$

These two formulas are the same whether the locus is diallelic or multi-allelic.

The multinomial-Dirichlet (or BB if diallelic) is not exact. Marchini et al. (2004) found that the BB provided an excellent fit for a genome-wide study of SNP markers. However STR mutant alleles usually differ from their parents by exactly one repeat unit, and this makes it unlikely that the Dirichlet assumption will be strictly valid.

- For inferences about variances it can be shown to give a good approximation.
- Alternative distributions are the multivariate Gaussian (Weir and Hill, 2002) and multivariate Gaussian log-ratios (Aitchison, 2003).

1. Allele fractions in structured populations
2. Interpretations of F_{ST}
3. A sampling formula: the Beta-Binomial likelihood
4. Likelihood inference for the inbreeding coefficient
5. Likelihood inference for F_{ST}
6. F_{ST} estimates in worldwide human populations for forensics

The multinomial-Dirichlet gives probabilities for a sample of alleles drawn from a subpopulation in terms of F_{ST} and the population allele fractions (the p). If we know the latter, this specifies a likelihood function for F_{ST}.

Likelihood-based inference has many advantages. We can start to illustrate these using observed diploid genotypes (samples of size two from the subpopulation) to infer the inbreeding coefficient f, or F_{IT} in Wright's notation. He also derived in a simple island model

$$1 - F_{IT} = (1 - F_{IS})(1 - F_{ST})$$

where F_{IS} is the inbreeding coefficient when the π are known.

Under the inbreeding model, the genotype probabilities are

$$\begin{aligned}
\Pr(\mathrm{AA}) &= p_A^2 + f p_B p_A \\
\Pr(\mathrm{AB}) &= 2(1-f) p_B p_A \\
\Pr(\mathrm{BB}) &= p_B^2 + f p_B p_A,
\end{aligned}$$

where $\max(-p_B/p_A, -p_A/p_B) \leq f \leq 1$.

Then the likelihood of a sample with genotype counts n_{AA}, n_{AB}, n_{BB} is

$$L(f) = c\,\Pr(\mathrm{AA})^{n_{AA}}\,\Pr(\mathrm{AB})^{n_{AB}}\,\Pr(\mathrm{BB})^{n_{BB}}\,,$$

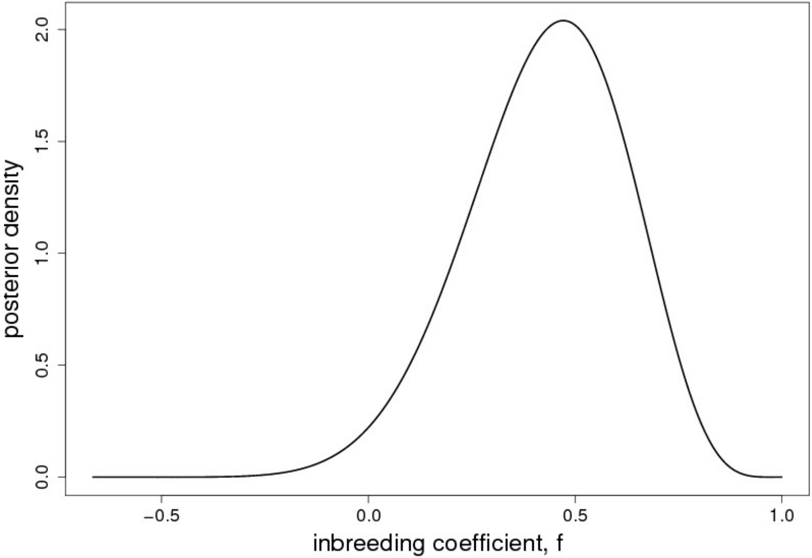
where c is a constant.

- We can choose c such that $L(f)$ takes value one at the HWE value $f = 0$, in which case we obtain, for $\max(-p_A/p_B, -p_B/p_A) < f < 1$,

  $$L(f) = (1 + fp_B/p_A)^{n_{AA}}(1-f)^{n_{AB}}(1 + fp_A/p_B)^{n_{BB}}.$$

  Maximising over f gives the Maximum Likelihood Estimator (MLE), but there is more information in the likelihood than just its maximum.
- If we choose c so that the integral over f is one, then the likelihood also specifies the posterior pdf for f given a uniform prior (see plot next slide for an illustration).

The uniform prior for f isn't appropriate when we have information about reasonable values (e.g. close to 0), but when n is large the choice of prior has little impact. We discuss choice of prior for F_{ST} below.



Posterior pdf for the inbreeding coefficient f given sample genotype counts $n_{AA} = 10$, $n_{AB} = 5$, and $n_{BB} = 5$, $p_A = 0.6$, $p_B = 0.4$, and a uniform prior.

R code for likelihood inference of f

inbflik computes a posterior pdf for *f* given genotype counts nAA, nAB, and nBB at a diallelic locus. If pA is unknown, enter a −ve value and it will be estimated from the genotype data:

```
inbflik = function(nAA=10,nAB=5,nBB=5,pA=-1){
   if(pA<0) pA = (nAA+nAB/2)/(nAA+nAB+nBB)
   pB = 1-pA
   minf = max(-pA/pB,-pB/pA)
   f=seq(minf,1,by=0.001)
   lik = (1 + f*pB /pA)^nAA * (1-f)^nAB * (1+f*pA/pB)^nBB
   plot(f,lik,ty="l",xlab="f",ylab="likelihood")
   return(f[which.max(lik)])
}
```

- Use inbflik to obtain the plot on previous page (NB *y* axis scale differs).
- Now repeat but assuming that pA is unknown.
- What are the approximate MLEs $\hat{f}$ in each case.
- Find a set of genotype counts that give a −ve value of $\hat{f}$.

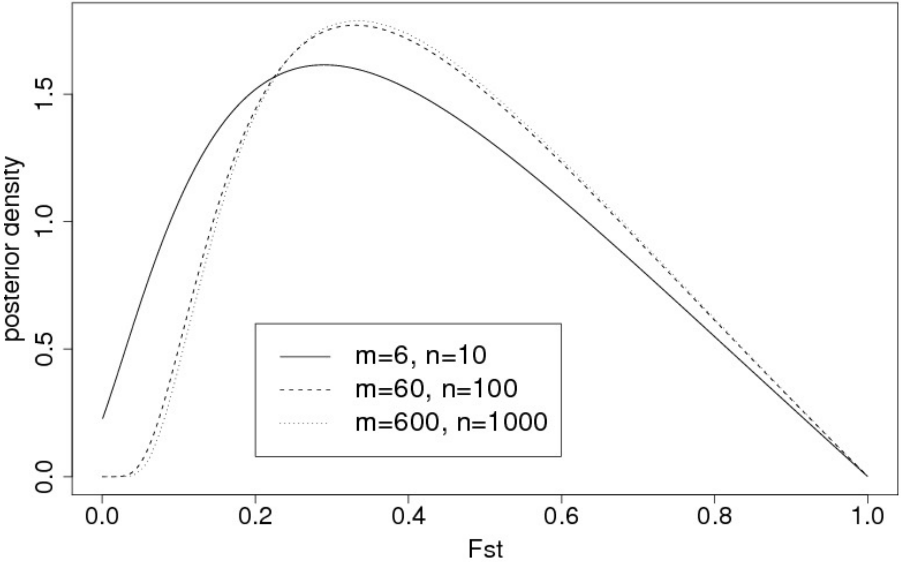
1. Allele fractions in structured populations
2. Interpretations of F_{ST}
3. A sampling formula: the Beta-Binomial likelihood
4. Likelihood inference for the inbreeding coefficient
5. Likelihood inference for F_{ST}
6. F_{ST} estimates in worldwide human populations for forensics

Consider a sample of 10 alleles, with $n_A = 6$ and $p_A = 0.2$.

- The sample proportion 6/10 is $\gg p_A$, suggesting that F_{ST} is large, but any inferences must be weak with so little data. To make this precise, from (7) we obtain:

$$L(F_{ST}) \;=\; \frac{\prod_{i=0}^{5}(iF_{ST} + (1-F_{ST})/5) \times \prod_{i=0}^{3}(iF_{ST} + 4(1-F_{ST})/5)}{(1-F_{ST})\prod_{i=1}^{8}(1 + iF_{ST})}$$

- This curve is plotted in next slide (solid line) scaled so that it can be interpreted as a posterior density given a uniform prior for F_{ST}.
- As expected, a wide range of F_{ST} values is supported: the 95% highest posterior density (hpd) interval for F_{ST} is (0.027,0.83).
- Increasing the sample size by a factor of 10, the 95% hpd interval is $0.099 \leq F_{ST} \leq 0.84$, excluding a larger interval near zero.
- Stepping up by a further factor of 10 (dotted curve), the pdf is almost unchanged. Once we get a good fix on π, there is no additional benefit from increasing the sample size for that subpopulation and locus.



Likelihood curves for F_{ST} given samples from one subpopulation, at a diallelic locus with $p = 0.2$. The curves have been scaled so that they can also be interpreted as posterior densities given a uniform prior for F_{ST}.

Cheating simulation study (p assumed known)

	True F_{ST}	0.2%	0.5%	1%	2%	4%	8%
$n = 100$	MoM	52	66	87	132	224	404
	MLE	36	53	74	114	184	317
$n = 200$	MoM	31	43	67	110	198	377
	MLE	23	37	57	91	159	288
$n = 400$	MoM	19	33	56	101	192	376
	MLE	16	28	46	83	150	276

Standard deviations ($\times 10^4$) of MoM and MLE estimators of F_{ST} **when the multinomial-Dirichlet assumption is valid**. There were 10^4 simulations of samples of size n from each of five subpopulations, typed at a locus with $K = 4$ alleles. The population allele fraction vector $\mathbf{p}$ was sampled uniformly randomly, independently for each simulation, and was regarded as known for the estimation of F_{ST}.

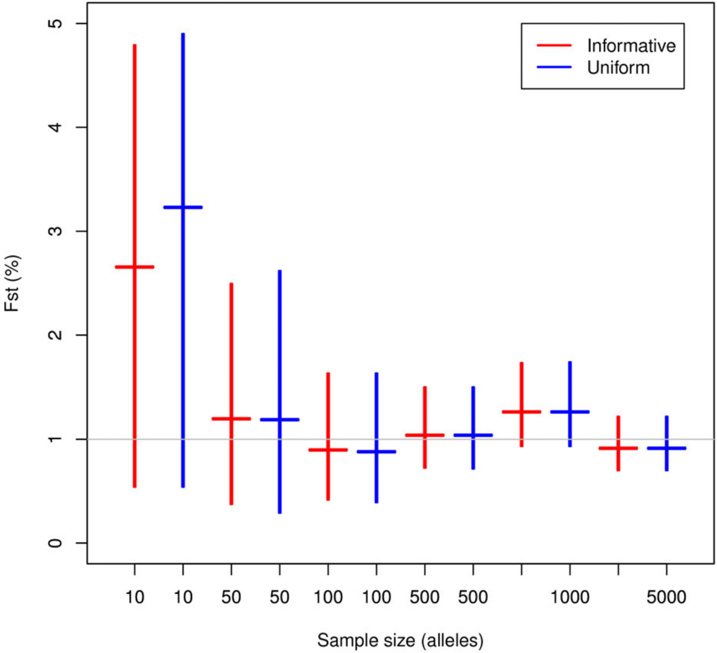
R code to simulate data and generate Fst likelihood curves

```
fstsim = function(Fst,nloc,nall,p=0.2){
   npop = length(nall);
   afcur = matrix(p,nloc,npop);
   if(Fst > 0){
      alph = (1/Fst-1)*p; beta = (1/Fst-1)*(1-p);
      for(i in 1:npop) afcur[,i] = rbeta(nloc,alph,beta)
   }
   ac = afcur;
   for(i in 1:npop) ac[,i] = rbinom(nloc,nall[i],afcur[,i]);
   return(list(nall,ac))
}
fstlik = function(nall=dat[[1]],ac=dat[[2]],p=0.2,fststep=100){
   nloc = nrow(ac);
   Fst = seq(0.001,0.99,len=fststep);
   loglik = rep(0,fststep);
   for(i in 1:nloc){
      alph = (1/Fst-1)*p;  beta = (1/Fst-1)*(1-p);
      loglik = loglik+lgamma(alph+ac[i,1])+lgamma(beta+nall[1]-ac[i,1])+lgamma(alph+beta);
      loglik = loglik-lgamma(alph)-lgamma(beta)-lgamma(alph+beta+nall[1]);
   }
   lik = exp(loglik-max(loglik));
   lik = lik*fststep/sum(lik);
   plot(Fst,lik,type="l",ylim=c(0,1.5*max(lik)));
   return(Fst[which.max(lik)])
}
```

`fstsim` simulates BB data at nloc loci in npop subpopulations with given `Fst`, then `fstlik` plots a likelihood curve and returns its approximate maxmimum. First run the following commands

```
f=0.1; dat=fstsim(f,20,c(20,20)); fstlik()
```

Then try other values for `f`, `nloc`, and `nall`.



Effect of prior, single locus, *p* known: F_{ST} posterior 95% interval using: (red) a beta prior with median 2.3% and 95% CI (0.26%, 8.0%)); (blue) the uniform prior. Data were simulated at a multiallelic locus with $F_{ST} = 1\%$. The vertical lines indicate the 95% equal-tailed CI, and medians are indicated with horizontal segments.

Effect of prior on F_{ST}, ten multi-allelic loci

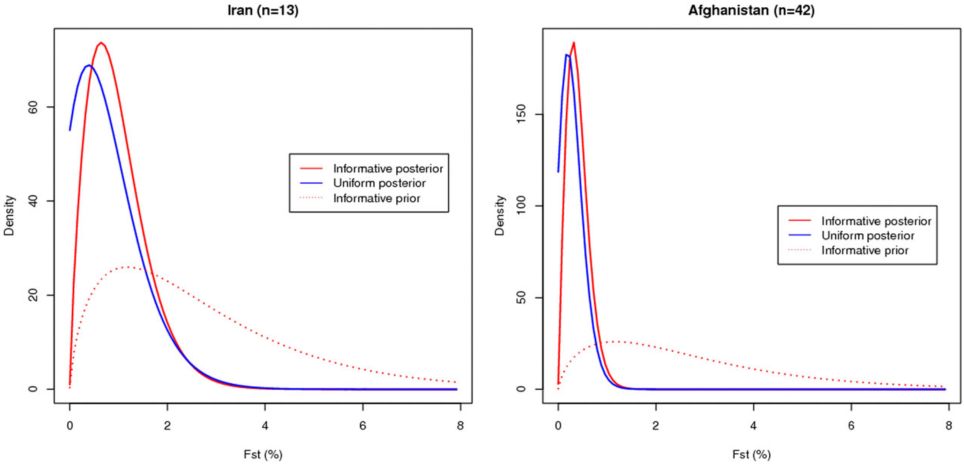


Figure 3 F_{ST} posterior densities (solid lines) using the direct method, given a uniform prior (blue) and an informative beta prior (red). Dotted red lines show the beta prior density. The subpopulations analysed are (left) Iran and (right) Afghanistan, with the reference populations being EA6 (Middle East/North Africa) and EA4 (South Asia), respectively.

- Most examples above use only a single population and a single locus.
  - It's an advantage of the likelihood approach that some inference is possible even with so little information.
- But in practice, to obtain better inferences about F_{ST} we need to combine information across loci and/or across subpopulations.
  - Another advantage of likelihood is that we can do this in many different ways using hierarchical models.
- With collaborators, I have developed the `BayesFST` software[2] which assumes the model

$$F_{ST}^{ij} = \frac{\exp(a_i+b_j)}{1+\exp(a_i+b_j)}, \tag{8}$$

  where i indicates the locus and j the population.

`BayesFST` also deals with the fact that we often we don't know the p:

- Integrate over the p with respect to a prior, by default uniform.
  - Different assumptions about p can have a big impact on inferences.
  - Uniform prior not appropriate for SNPs: a U-shaped distribution of allele fractions is common, and can be modelled by a beta prior.

[2] `www.reading.ac.uk/Statistics/genetics/software.html`

1. Allele fractions in structured populations

2. Interpretations of F_{ST}

3. A sampling formula: the Beta-Binomial likelihood

4. Likelihood inference for the inbreeding coefficient

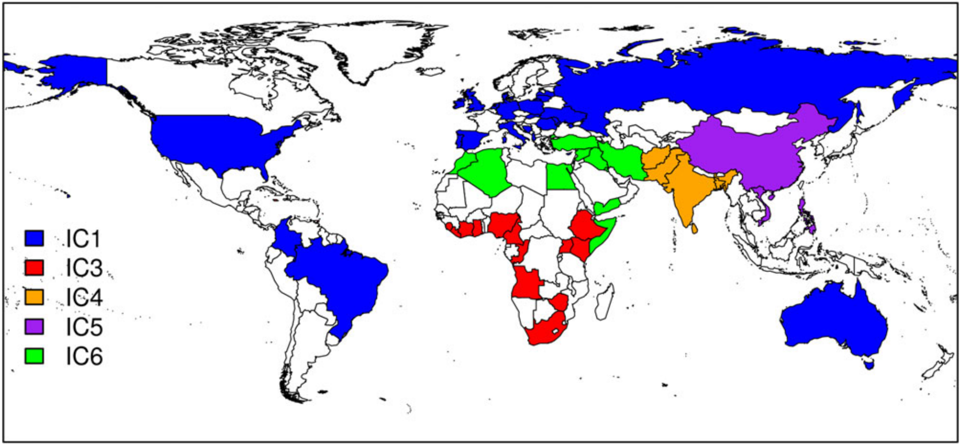
5. Likelihood inference for F_{ST}

6. F_{ST} estimates in worldwide human populations for forensics

- In forensic DNA analysis, the weight of evidence depends on the coancestry of alleged and alternative contributors of DNA to a sample.
- We can't know the right F_{ST} in any given case but we can use large values relative to the observed range.
- The reference population is that of the frequency database
    - this leads to larger values than typical population-genetics estimates, for which the reference population is either an ancestral population or a (weighted) mean of the subpopulations.
- Conversely, most human population genetic studies are of distinct populations, often geographically or socially isolated from other populations
    - this leads to higher F_{ST} estimates than for the heterogeneous, cosmopolitan populations that are often appropriate in forensic work.
- F_{ST} is also affected by the mutation rate.
    - We might guess that higher mutation causes greater divergence among populations.
    - In practice this doesn't seem to be true: F_{ST} at (high-mutation) STR loci tends to be lower than at (low-mutation) SNPs.

We obtained worldwide allele count data for 16 (multi-allelic) forensic STR loci from UK migration applicants.

- We grouped then into 5 continental-scale regions IC1 to IC6 (map).
- We assumed (after checking) that F_{ST} is constant over loci.



Results in following slides are a sample from Steele et al. (2014).

F_{ST} posterior median + 95% interval in Africa/Caribbean[3]

		Direct			Indirect		
IC3	n	2.5	50	97.5	2.5	50	97.5
Ghana	214	0.8	1.1	1.6	0.2	0.3	0.5
Jamaica	166	0.5	0.7	1.0	0.0	0.1	0.2
Kenya	51	0.7	1.2	1.9	0.8	1.3	1.9
Nigeria	444	0.9	1.2	1.5	0.2	0.3	0.3
Sierra Leone	41	0.7	1.3	2.2	0.1	0.3	0.8
Uganda	63	0.3	0.5	1.0	0.0	0.2	0.4

F_{ST} values are expressed in %.

Direct means relative to a forensic database;

Indirect reference population is a hypothetical ancestral population.

[3] A preliminary analysis indicated that Somalia fit better with Middle East/North Africa and is not shown here.

Inter-continental F_{ST}

Global	n	EA1	EA3	EA4	EA5	EA6	Indirect
IC1	3582	0.4	3.1	1.9	1.9	0.9	2.7
IC3	2032	1.7	0.7	1.7	1.4	1.1	1.0
IC4	285	1.4	3.1	0.7	1.3	0.8	2.3
IC5	304	3.1	4.2	2.4	0.5	2.0	3.3
IC6	604	1.8	1.7	1.9	1.7	0.9	1.4

EA1/IC1	European
EA3/IC3	Afro-Caribbean
EA4/IC4	South Asian
EA5/IC5	East Asian
EA6/IC6	Middle East/North African

EA refers to an older 10-locus database, IC is a newer 16-locus database. Above results use the 10 common loci.

Conclusions

- I've highlighted several advantages of a flexible likelihood-based approach to inference based on F_{ST}.
- The main disadvantage is that the assumed likelihood may not be exactly correct, but
  - it has been shown to fit well for SNP data;
  - validity of inferences can be checked by simulation.
- Another disadvantage of likelihood methods is computational speed for very large numbers of loci, but some calculations are still feasible.

References

Aitchison, J. 2003. *The Statistical Analysis of Compositional Data*. Blackburn Press.

Balding, D. J. 2003. Likelihood-based inference for genetic correlation coefficients. *Theoretical population biology*, 63(3):221–230.

Bhatia, G., Patterson, N., Sankararaman, S., and Price, A. L. 2013. Estimating and interpreting FST: The impact of rare variants. *Genome Research*, 23:1514–1521.

Crow, J. and Kimura, M. 1970. *An Introduction to Population Genetics Theory*. Harper and Row, New York.

Ewens, W. 2004. *Mathematical Population Genetics: I. Theoretical Introduction*. Springer, New York.

Marchini, J., Cardon, L. R., Phillips, M. S., and Donnelly, P. 2004. The effects of human population structure on large genetic association studies. *Nature genetics*, 36(5):512–517.

Nicholson, G., Smith, A., Jonsson, F., Gustafsson, O., Stefansson, K., and Donnelly, P. 2002. Assessing population differentiation and isolation from single-nucleotide polymorphism data. *Journal Of The Royal Statistical Society Series B — Statistical Methodology*, 64:695–715.

Steele, C. and Balding, D. 2015. *Weight of evidence for forensic DNA profiles*. Wiley, 2nd edition.

Steele, C., Syndercombe Court, D., and Balding, D. 2014. Worldwide FST estimates relative to five continental-scale populations. *Annals of Human Genetics*, 78(6):468–477.

Weir, B. S. and Hill, W. 2002. Estimating F-statistics. *Annual Review of Genetics*, 36(1):721–750.

Wright, S. 1951. The genetic structure of populations. *Ann. Eugen.*, 15:313–354.