Peter J. Green (*University of Bristol and University of Technology, Sydney*)
For several reasons, I regret not being able to come to the meeting, including that I understand that there is some connection between what I write here and the discussion by Dr Torben Tvedebrink.

Since this paper was completed, I have with Julia Mortera been exploring the effects of uncertainty in the allele frequencies $q=(q_a)_{a=1}^A$. In earlier work (Green and Mortera, 2009) addressing cases where the DNA traces are of discrete allele presence indicators rather than continous peak heights, such questions were explored under an (idealized) Dirichlet model—this leads to a Pólya urn scheme which is readily implementable in a Bayes net formulation for the inference. More precisely, $q|\rho \sim \text{Dirichlet}\{(M\rho_a)_{a=1}^A\}$, where q are the true, unknown, allele frequencies, $\rho=(\rho_a)_{a=1}^A$ the database frequencies and M the database size; this is typically only a few hundred in practice, so there is considerable uncertainty. We write $\alpha_a = M\rho_a$.

Combining this Dirichlet prior on q with the authors' set-up, Dirichlet–multinomial conjugacy then gives the joint distribution for the allele counts n_{ia}, recognizing this uncertainty. Recall that n_{ia} is the number of a alleles for the ith individual, $a=1,2,\ldots,A$, $i=1,2,\ldots,I$. Conditional on allele frequencies $\{q_a\}$, the vectors $n_{i.}=(n_{ia})_{a=1}^A$ are independent and identically distributed multinomial$\{2,(q_a)_{a=1}^A\}$. Then

$$n_{1.} \sim \text{DM}\{2,(\alpha_a)_{a=1}^A\}$$

where DM denotes the Dirichlet–multinomial distribution: $X \sim \text{DM}\{n,(\alpha_a)_{a=1}^A\}$ means

$$P(X=x)=\int \frac{n}{\prod_a x_a}\prod q_a^{x_a}\frac{\Gamma(\sum_a \alpha_a)}{\prod_a \Gamma(\alpha_a)}\prod_a q_a^{\alpha_a-1}\,\mathrm{d}_q=\left(\frac{n!}{\prod_a x_a}\right)\left\{\prod_a \frac{\Gamma(\alpha_a+x_a)}{\Gamma(\alpha_a)}\right\}\frac{\Gamma(\sum_a \alpha_a)}{\Gamma(\sum_a \alpha_a+n)},$$

so long as $\Sigma_a x_a = n$. Furthermore, again by conjugacy, for $i=2,3,\ldots,I$,

$$n_{i.}|(n_{j.})_{j=1}^{i-1} \sim \text{DM}\{2,(\alpha_a+T_{i-1,a})_{a=1}^A\}$$

where $T_{i-1,a}=\Sigma_{j=1}^{i-1} n_{ja}$.

Factorizing these distributions over alleles, we find that individual allele counts have beta–binomial conditional distributions:

$$n_{ia}|\{n_{jb}, j<i, \forall b\},\{n_{ib}, b<a\} \sim \text{BB}(2-S_{i,a-1}, \alpha_a+T_{i-1,a}, \beta_a+U_{i-1,a}) \qquad (9)$$

Here BB is the beta–binomial distribution: $\text{BB}(n,\alpha,\beta)$ is the same as $\text{DM}\{n,(\alpha,\beta)\}$, $\beta_a=\Sigma_{b>a}\alpha_b$, $S_{ia}=\Sigma_{b=1}^a n_{ib}$ as in the paper and $U_{i-1,a}=\Sigma_{b>a}T_{i-1,b}$. Note that $\text{BB}(1,\alpha,\beta)$ is just Bernoulli$\{\alpha/(\alpha+\beta)\}$. Equation (9) exhibits association among the n_{ia} that is positive across i and negative across a, as would be expected.

In the large database limit, $\alpha_a \to \infty$ but $\alpha_a/\Sigma_a \alpha_a \to q_a$, and the beta–binomial conditional probabilities (9) become

$$n_{ia}|\{n_{jb}, j<i, \forall b\},\{n_{ib}, b<a\} \sim \text{binomial}(2-S_{i,a-1}, q_a/\sum_{b \geqslant a} q_b) \qquad (10)$$

as in Section 2.4.1.

Graversen's (2013) R package `DNAmixtures` can readily be amended to use distribution (9) instead of (10) in a Bayes net computation to sum the terms in equation (8). The corresponding directed acyclic graph is now considerably more complex, owing to the presence of the additional nodes T_{ia} and U_{ia}, and the computation runs much more slowly. (Therese Graversen showed us how to amend our amendment to her code to use a more efficient elimination order, and this improved the times.)

Our limited numerical experiments with casework data using this code reveal a curiously mixed picture: uncertainty in allele frequencies may either increase or decrease the weight of evidence $\log_{10}(\text{LR})$, depending on the example. This is in contrast with all our earlier examples, with either allele presence indicator traces (in Green and Mortera (2009)) or with the model of Cowell *et al.* (2007), in which this uncertainty always reduced the weight of evidence. This needs further study, but we surmise that the difference might be attributable to maximizing out of parameters, in contrast with a more fully Bayesian approach.

In the literature, other phenomena causing dependence among DNA profiles, such as identity by descent, have been modelled in a way leading to the same probabilistic dependence as in the analysis above.

Han Liu and Junwei Lu (*Princeton University*)
We congratulate the authors for making an interesting contribution to the problem of analysing DNA mixtures. We first describe a protein identification problem which shows a resemblance to the DNA mix-