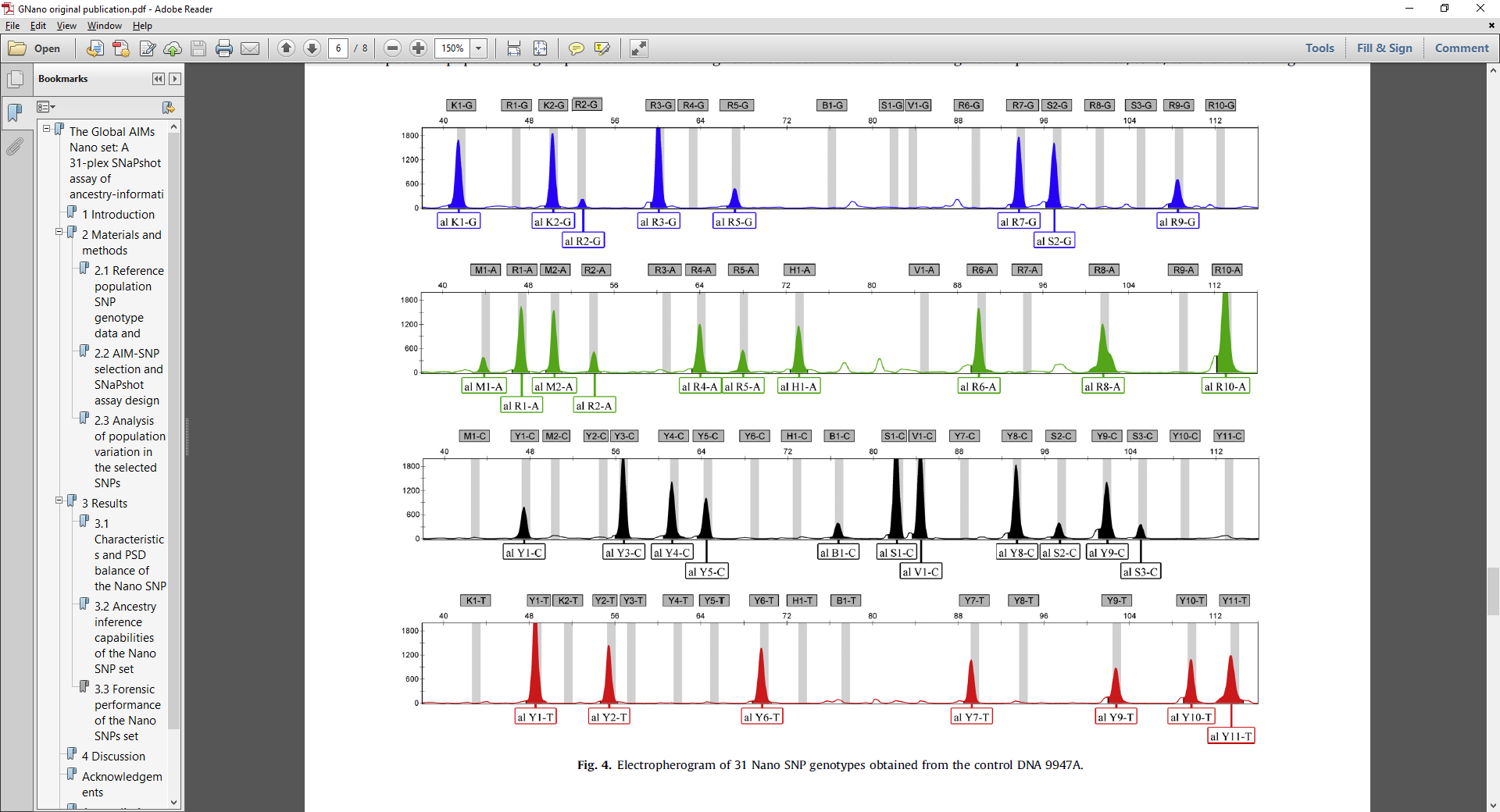
**INTRO**

An example of a DNA profile produced by the GNano identity SNP panel is shown in Figure 1.



*Figure 1: DNA profile produced by the GNano identity SNP panel*

We seek information regarding the performance of the GNano SNP ancestry kit, specifically to answer questions:

* What is the level of imbalance seen between heterozygous SNP peaks?
* What are the relative amplification efficiencies of the loci?
* Is there a dye effect that would affect heterozygous balance?

To answer these questions, we use a hierarchical Bayesian Model (HBM) that borrows from our knowledge of autosomal DNA profiles [1, 2].

**THE MODEL**

We consider that there are *L* loci, which possess two alleles, *a*, to produce a genotype, *G*:

.

We consider that to describe the observed fluorescence we need:

* A template DNA amount, *T*, that has prior , where *S* is the saturation of the capillary electrophoresis instrument
* A locus amplification efficiency, *Al*, for each of the *L* loci. The geometric mean of the *Al* values is constrained to 1, i.e. . We model 
* A dye effect, *D*, for each of the *F* fluorophores, *Df*, that has prior 

In previous models for DNA profile data [1] there was also a term used for the amount of degradation the profile exhibited. A term for degradation is not included in this model for two reasons:

1. The amplicons are very short and less prone to degradation
2. The DNA being used to generate a SNP profile is pristine. If the GNano panel were to start being used routinely on degraded samples then a degradation term may be present. At this stage, any degradation that is present would be accounted for by the locus amplification efficiency parameters

We use these parameters to calculate the expected height of allele *a*, at locus *l*, for dye *f* by:



Where *X* is a dose term that doubles the expected height when both alleles in the genotype are the same:



When a profile’s alleles are also known, the parameters *T*, *A* and *D* allow the calculation of expected peak heights. We expect that the efficiency of loci to amplify is based largely on their primer sequence. While every effort is made to achieve some level of peak height balance across a DNA profile with multiple loci, there are inevitable differences that will exist. We therefore expect that profiles will all generally exhibit a similar level of locus amplification efficacy for the same locus across multiple profiles. In professionally produced autosomal STR amplification kits the balance of the loci has been fine-tuned to the point where we can assume , however for the GNano system this same assumption cannot be made. We also know that locus amplification efficiency can have profile to profile effects that adds noise to the population-wide expectations. These profile, specific effects are most prominently by the presence of chemicals in the PCR reaction that can affect the amplification. Hence, we need to model profile specific locus amplification efficiencies as well as the population means for each locus.

Another difference between the model presented here and the model in [1] is the incorporation of a parameter for dye. Again, in professionally produced autosomal profiling systems the performance of the dyes is assumed to be approximately equal. In addition, even if there are indeed small differences in the performance of the dyes then this will be subsumed within the locus amplification efficiency terms. Such a subsumation cannot occur within the locus amplification efficiency terms for the GNanao DNA profiles, as the different alleles within a locus are labelled with different dyes. Therefore, in a manner similar to the locus amplification efficiency terms, our model has profile specific dye amplification efficiencies as well as population level means.

The observed peak heights will depend on the expected height, but also on a random sampling component, which occurs when some DNA fragments are aliquoted into a PCR reaction. Hence there will be a difference between the observed and expected peak heights. We do not know what this level of variability is for the GNano systems and so include a parameter, *λ*, which we describe below, that models the peak height variability.

For the hierarchical component of the analysis we define:

*  as the mean of the amplification efficiency at locus *l* with prior 
*  as the variance of the amplification efficiency at locus *l* with prior 
*  as the mean of the dye effect for fluorophore *f* with prior 
*  as the variance of the dye effect for fluorophore *f* with prior 

Let, , , , , , ,  and . We seek the posterior distributions of the parameters, given some observed data, **O**, , and can exploit the relationship:



The observed data consists of *C* = 102 profiles, each with observed data **Oc**. For a single profile we have:



And assume that the peaks at each locus are, given **M** and **V**, independent of each other (as per [2]) so that:



The parameters within **M** allow calculation of expected peak heights (as given above) so that:



We model:



where . We note that the model in [2] has two differences to the model we use here. Firstly, they use gamma distribution in a per-profile manner for the term equivalent to *λ*, whereas we use a constant value across all profiles. We do this to simplify the model, but also as a fact that the findings in [2] was that a constant was probably sufficient for modelling data. The second difference is that we model the variance of the observed peak heights as being inversely proportional to the template amount, whereas in [1] the variance is inversely proportional to the expected peak height. We choose our model because the only mass parameter that effects the number of DNA amplicons in the DNA extract (the main driver of peak height variability) is the template. There is no degradation in our model and amplification and dye efficiency terms are PCR effects, not aliquot effects.

We now set to evaluate , using Markov Chain Monte Carlo in the following component-wise manner:

* Component 1 - For each of the *C* profiles, holding the values within **V** constant:
  + Draw values for parameters for profile *c* within **Mc** by random walk
  + Evaluate 
  + Accept or reject proposed parameters by Metropolis-Hasting algorithm
  + Repeat *Y* times
* Component 2 - Holding all values within **M** constant:
  + Draw values for hyper-parameters within **V** by random walk
  + Evaluate 
  + Accept or reject proposed parameters by Metropolis-Hasting algorithm
  + Repeat *Y* times
* Repeat outer loops until converged

We start by running component 1 for 10 000 iterations, then component 2 for 10 000 iterations. We then set *Y* = 1 and run the process for 10 000 loops of the two components.

Random walk standard deviation values were set as:

*T* – 1

*D* – 0.01

*A* – 0.01

 – 0.01

*s* – 0.01

 – 0.01

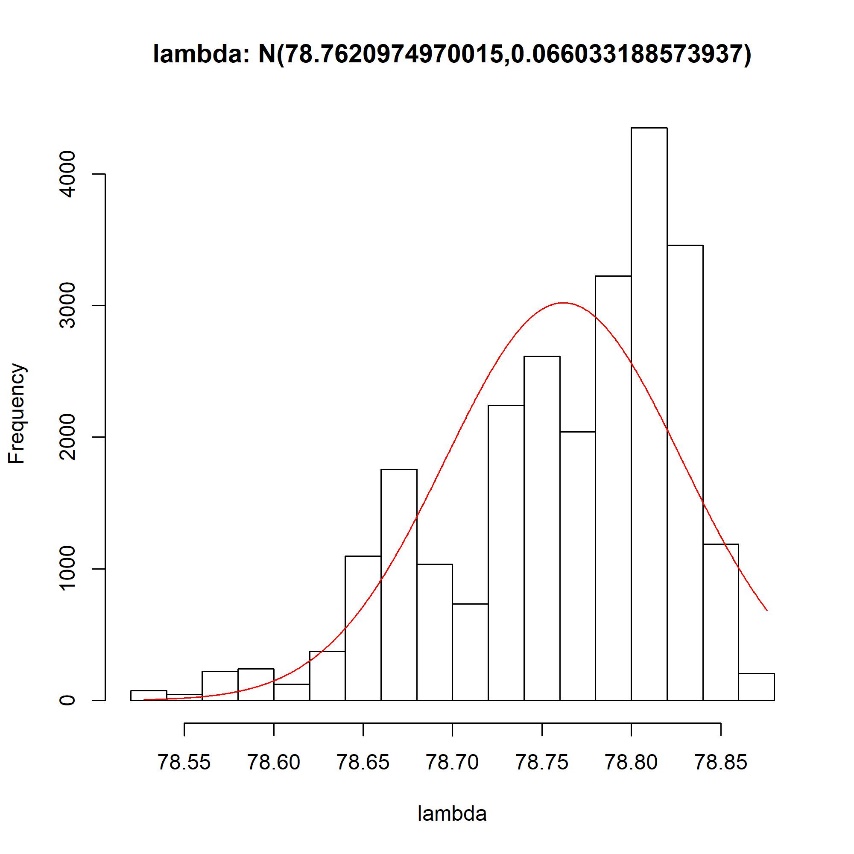
 – 0.01

*λ* – 0.1

We reduce the RWSD step sizes to 1% of their original value over the course of the first half of the MCMC analysis then maintain the reduced level for the remainder of the MCMC.

**RESULTS**

At the conclusion of the MCMC each of the parameters within **V** were modelled with a normal distribution. Figure 2 shows an example of this modelling for the *λ* parameter.



*Figure 2: posterior distribution for λ parameter, with fitted normal distribution*

Posterior distributions for the dye amplification efficiencies are shown in Figure 3. These are produced by randomly drawing values from the fitted posterior normal distributions for  and *s* and then drawing a number of values for *Df*.

|  |  |
| --- | --- |
|  |  |
|  |  |

*Figure 3: Posterior distributions for dye amplification efficiency*

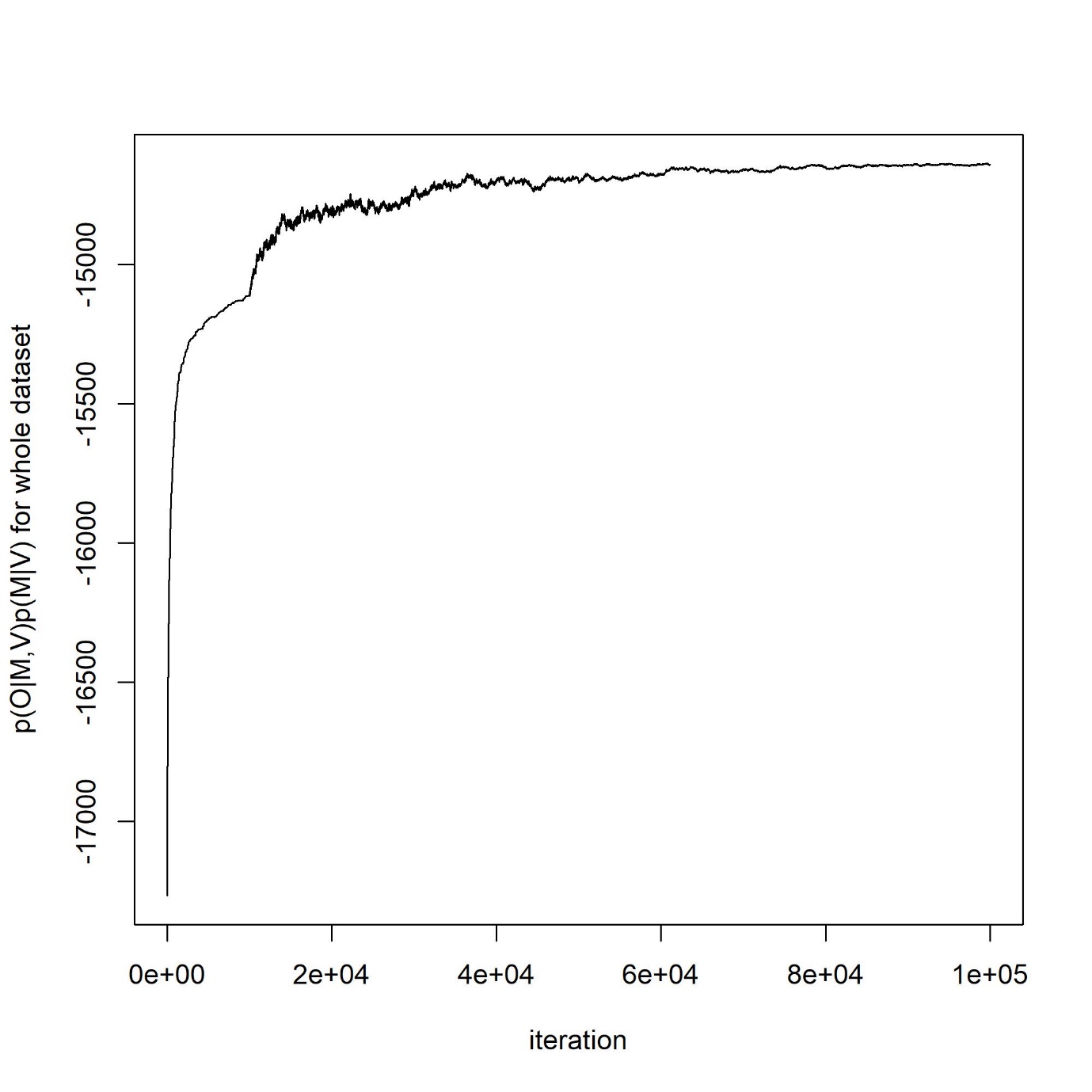
In a similar manner to dye efficiencies, the locus amplification efficiencies are shown in Figure 4.

|  |  |  |
| --- | --- | --- |
|  | |  |
|  | |  |
|  | |  |
|  | |  |
|  | |  |
|  | |  |
|  | |  |
|  | |  |
|  |  | |
|  |  | |
|  |  | |
|  |  | |
|  |  | |
|  |  | |
|  |  | |
|  | *Figure 4: Posterior distributions for locus amplification efficiency. Red line shows the distribution using the mean values for mean and variance.* | |

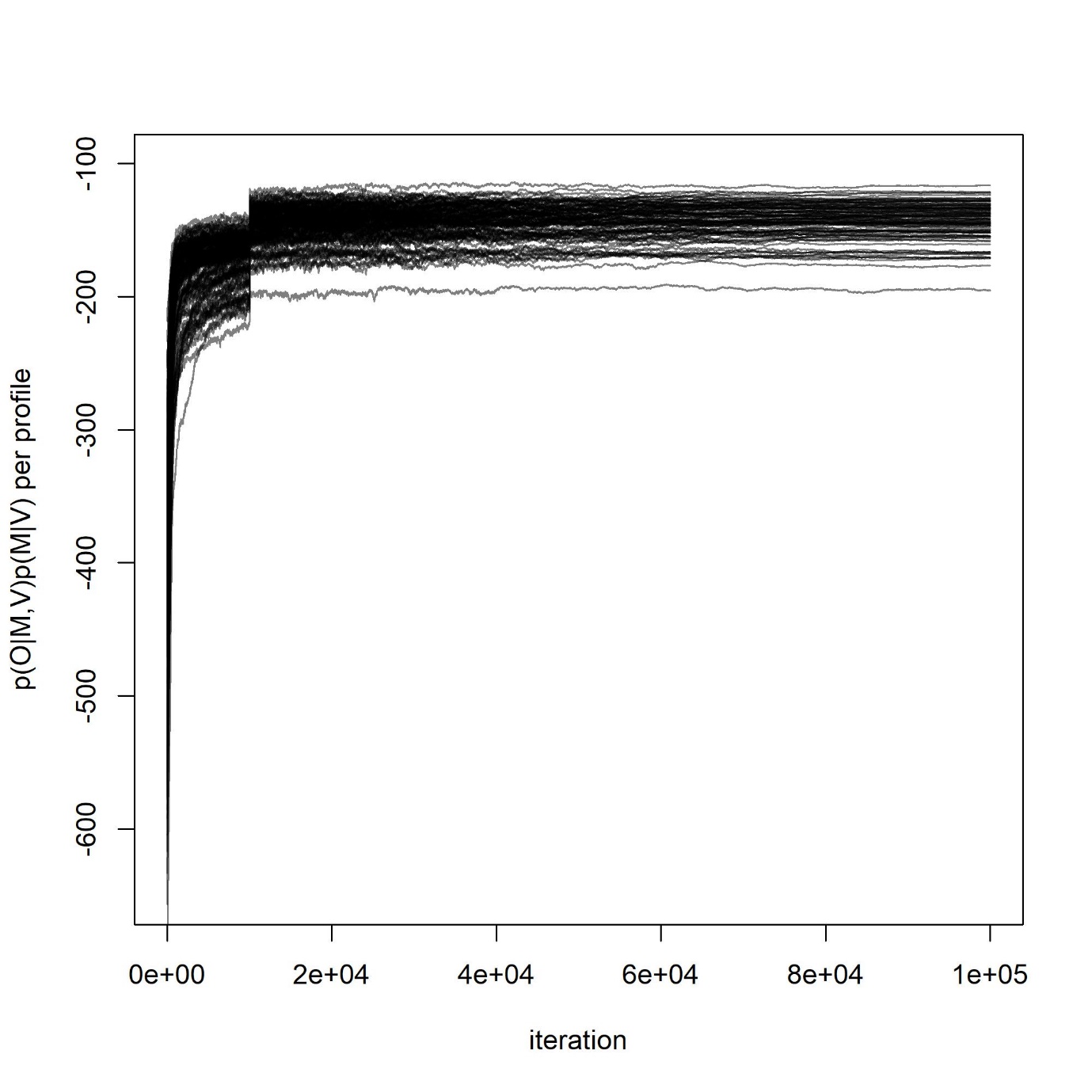
**MODEL CHECKING**

Convergence:

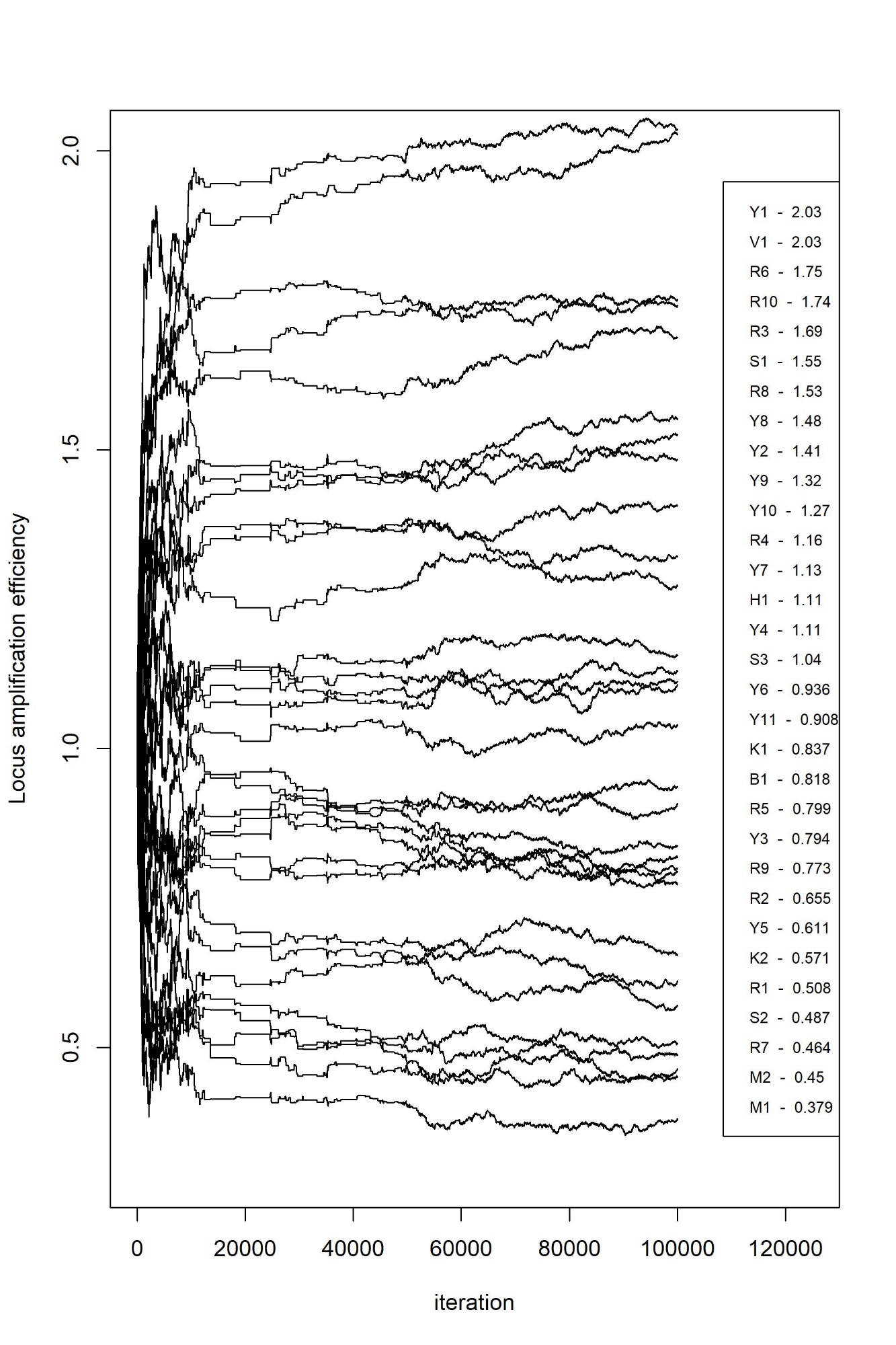
Trace plots for the dataset-wide probability, individual profile probabilities, locus amplification efficiency means, and dye efficiency means are shown in Figures 5, 6, 7 and 8 respectively.



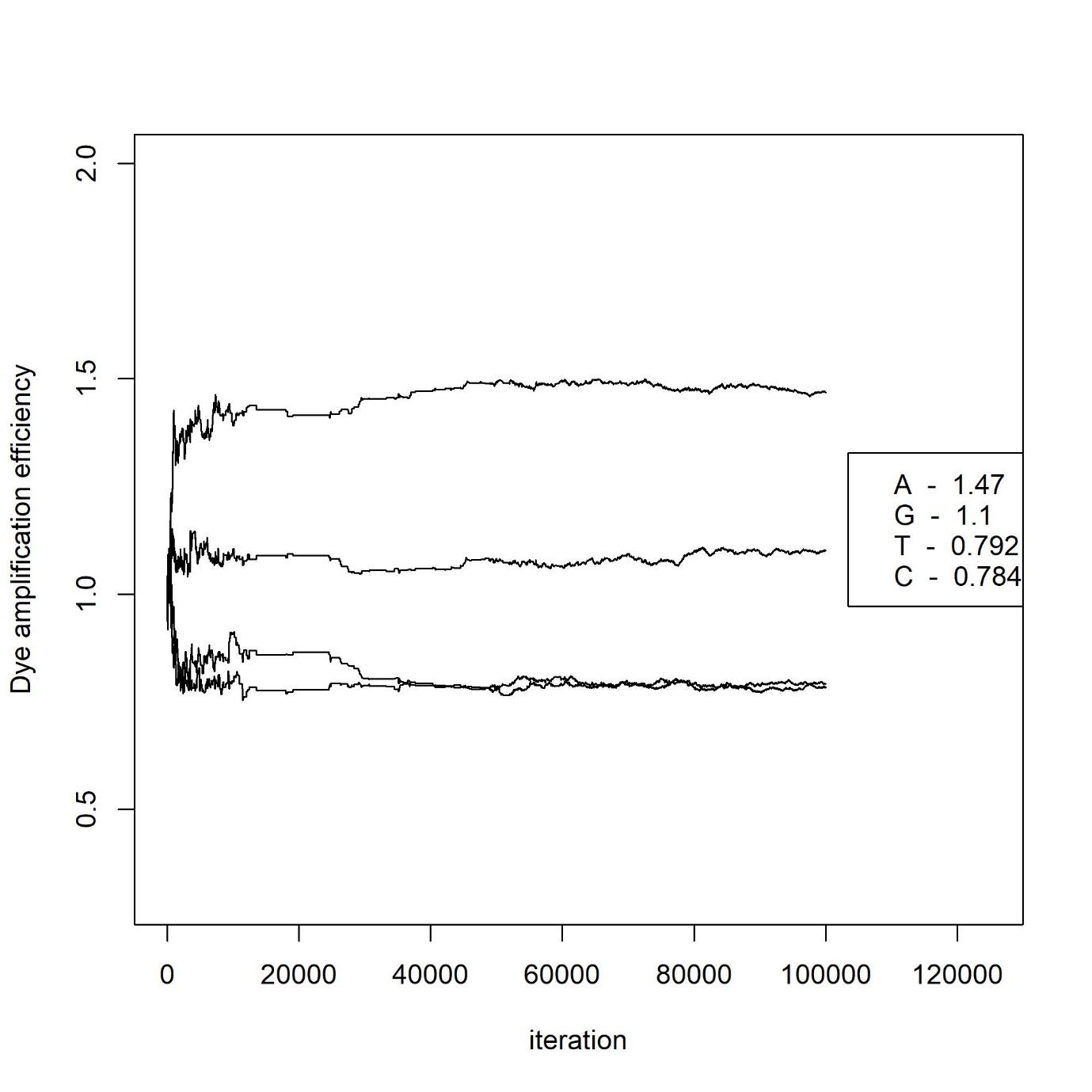
*Figure 5: Dataset wide probability trace plot*



*Figure 6: Profile probability trace plots*

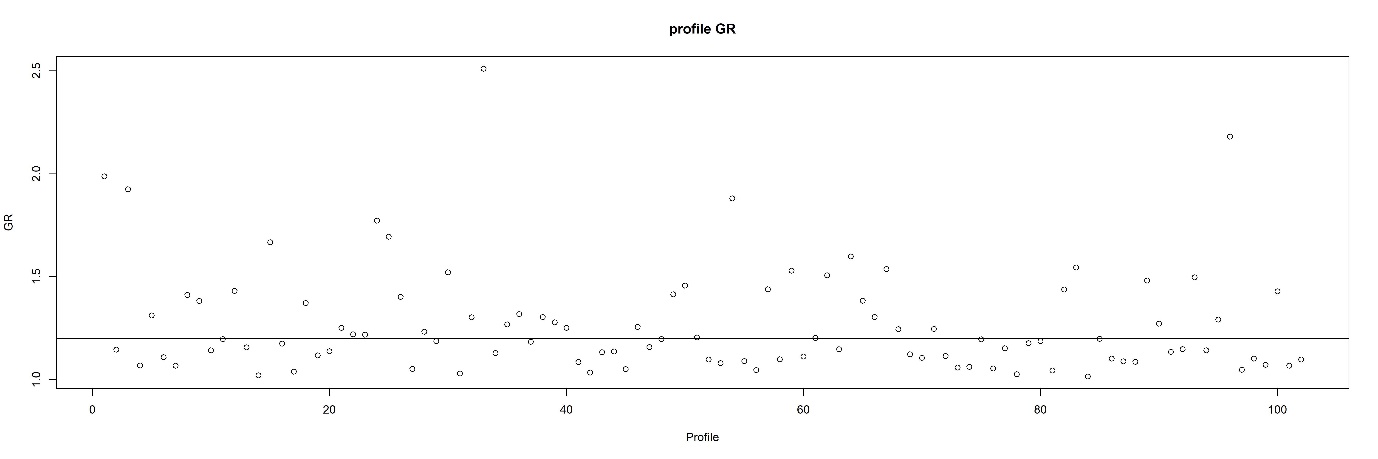


*Figure 7: locus amplification efficiency means trace plots*

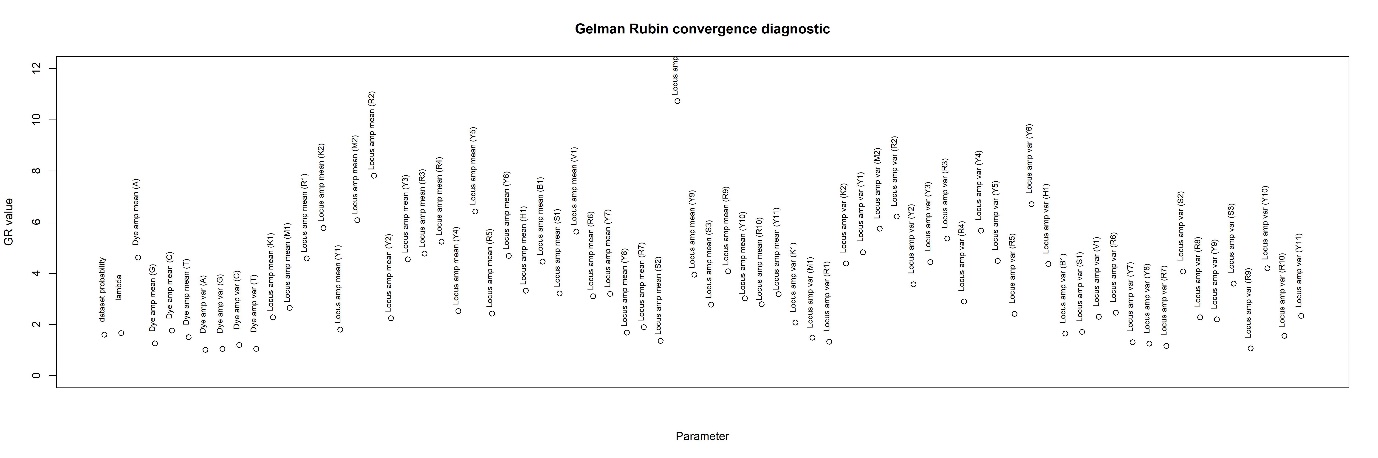


*Figure 8: dye efficiency means trace plots*

We check the models first checking the convergence of the MCMC analysis. Burn-in was considered to be the first 75% of the MCMC after the value of *Y* was set to 1 (as explained above). As the process was run using only 1 chain the convergence was determined by splitting the MCMC values into 4 quarters and calculating the Gelman Rubin convergence diagnostic (GR) on those 4 arrays. Figure 9 shows the GR values for the profile probabilities, individually and Figure 10 shows the GR values for population level parameters:



*Figure 9: GR convergence diagnostics for individual profile probabilities*

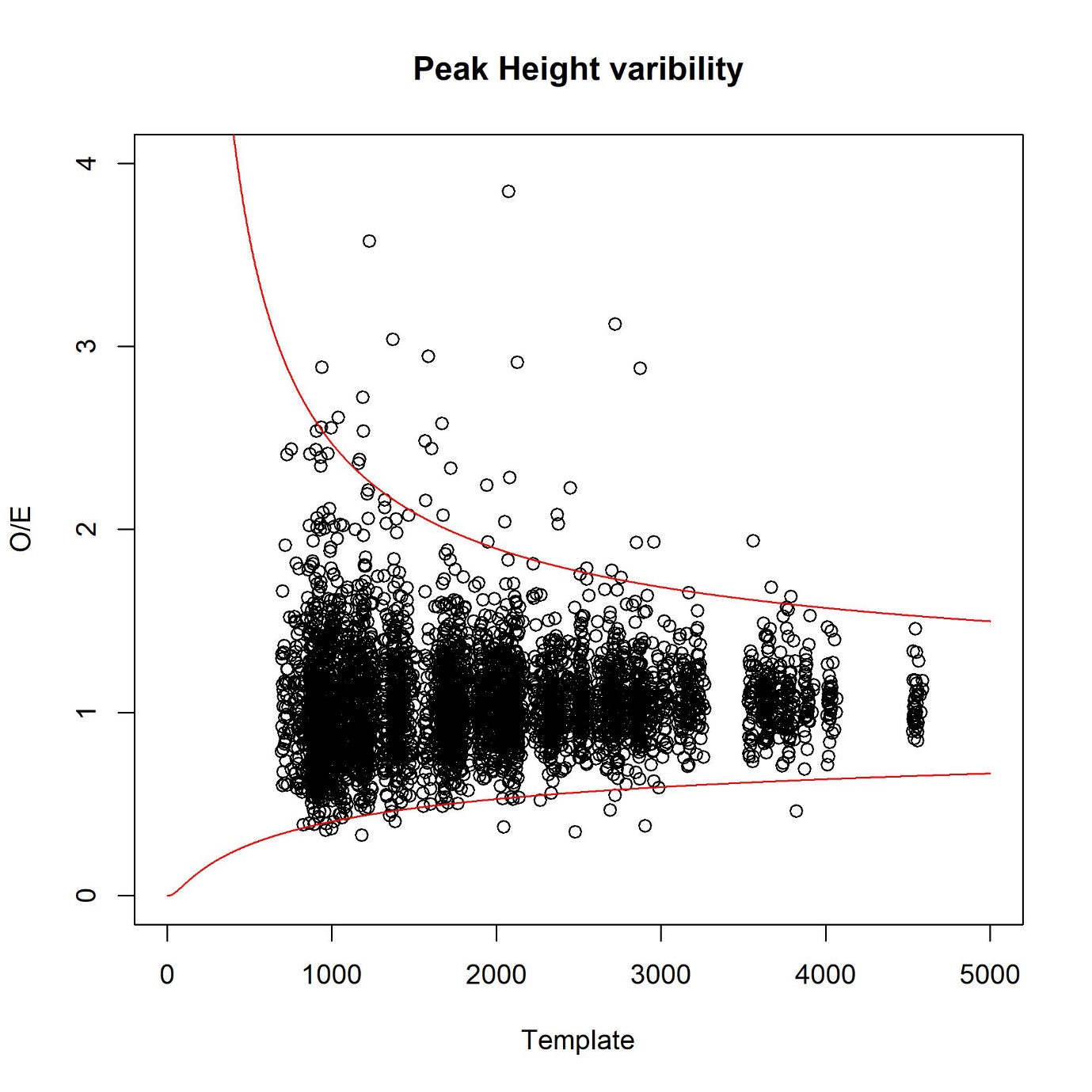


*Figure 10: GR convergence diagnostics for model population level parameters*

Prediction power:

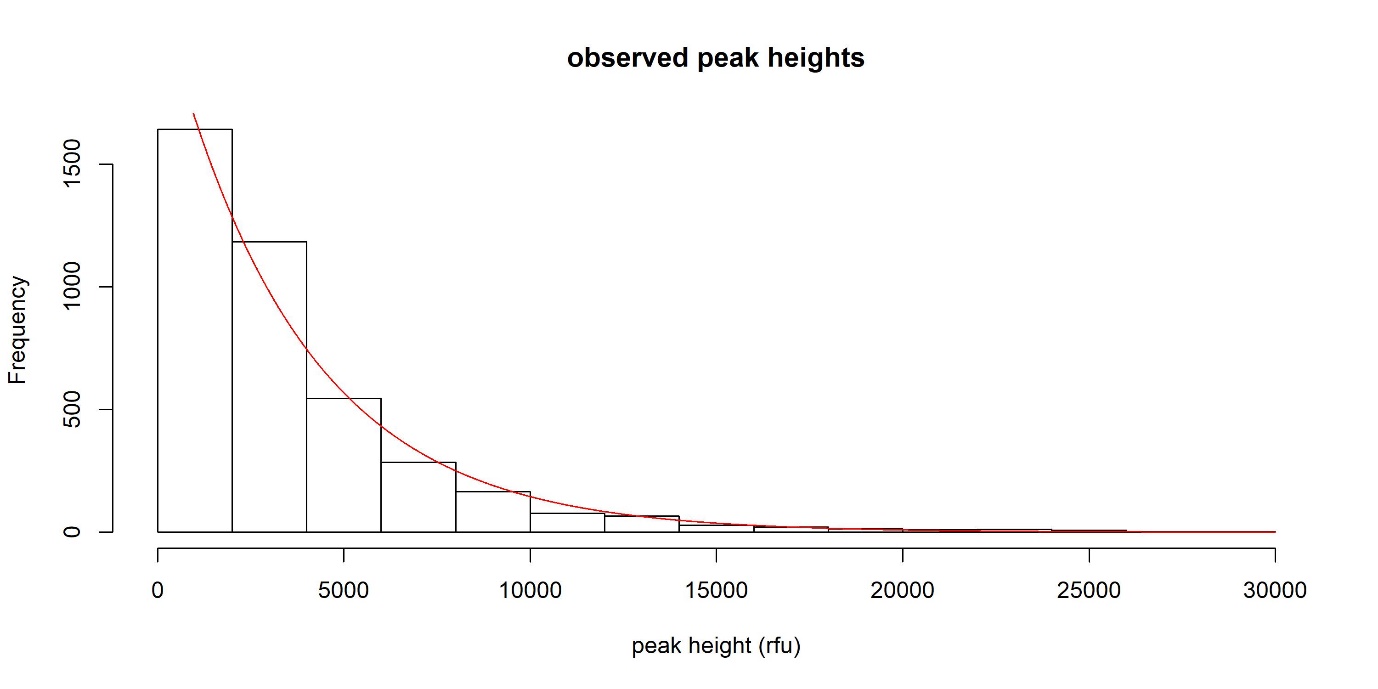
Having a model and posterior distributions for parameters allows the generation of predictions of DNA profile behaviour. These predictions can be compared to the properties of the dataset used in the analysis as a means of assessing the model performance. Ideally, we would test the model performance on a hold-out validation set, however due to the relatively small dataset of 102 profiles, we used all the available data in the modelling.

We test 3 aspects of the model that relate to the three components of the population-wide parameters. The assess the performance of the modelling of *λ* we calculate the ratio of the observed and expected peak heights, for each peak in the dataset, and graph these values over the posterior mean of the template values (from the MCMC analysis). We then consider that 95% of these values should be encompassed by  (as explained in [3]). Figure 11 shows the observed/expected graph and the 95% expected upper and lower bounds on the ratio.



*Figure 11: values of observed / expected peak heights and the 95% bounds expected by our model*

The observed levels of dye efficiencies can be compared to the expected levels by taking the ratio of peak heights for heterozygous peaks in the dataset. By doing this the locus amplification efficiency term does not play a role (as both alleles, although in different dyes, are from the same locus and so presumably affected by the locus amplification efficiency by the same amount). In order to create simulated values we needed to not only consider the distribution for dye efficiencies as given by a ratio of the  terms, but also the peak height variability present in the profile data. Peak height variability also depends on template amount. To align the amount of peak height variability seen in the simulated values with that in the dataset the distribution of peak heights was modelled on the observed heights using an exponential distribution (Figure 12).

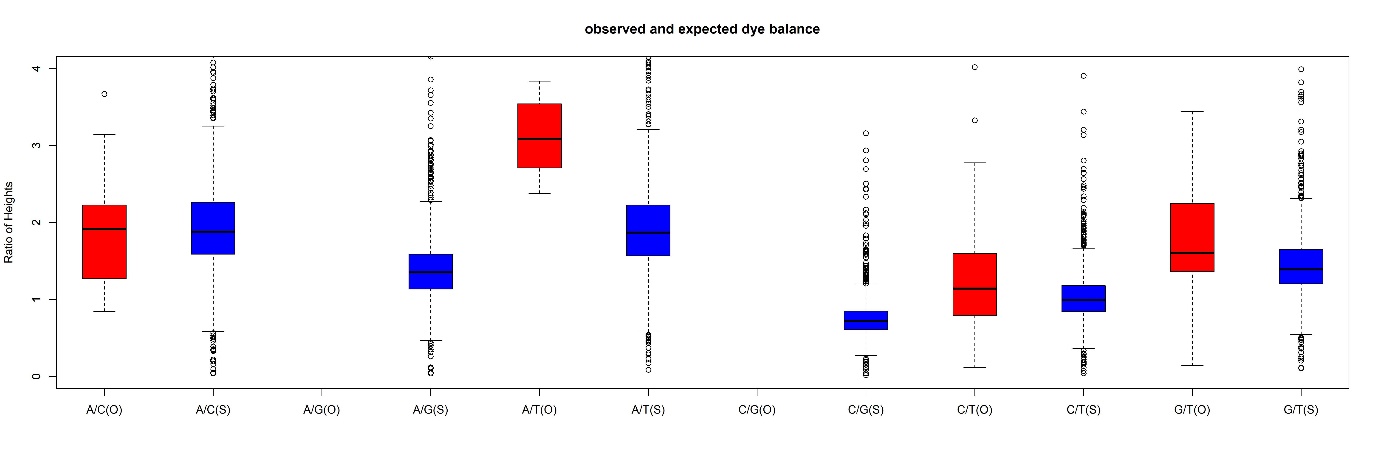
**

*Figure 12: Observed peak height distribution and fitted exponential distribution*

Then, in order to generate a simulated ratio of alleles of different dyes:

1. A template value, *Trandom*, was drawn from the fitted exponential distribution seen in Figure 12
2. Two random observed peak heights, *Orandom1* and *Orandom2*, were drawn from 
3. A dye efficiency value for dye1 (the dye that allele 1 is labelled with) was drawn from  and for dye 2 (the dye that allele 2 is labelled with) was drawn from 
4. The plotted point is then 

The resulting plot is seen in Figure 13 for 1000 randomly generated values (blue) compared to observed values (red).

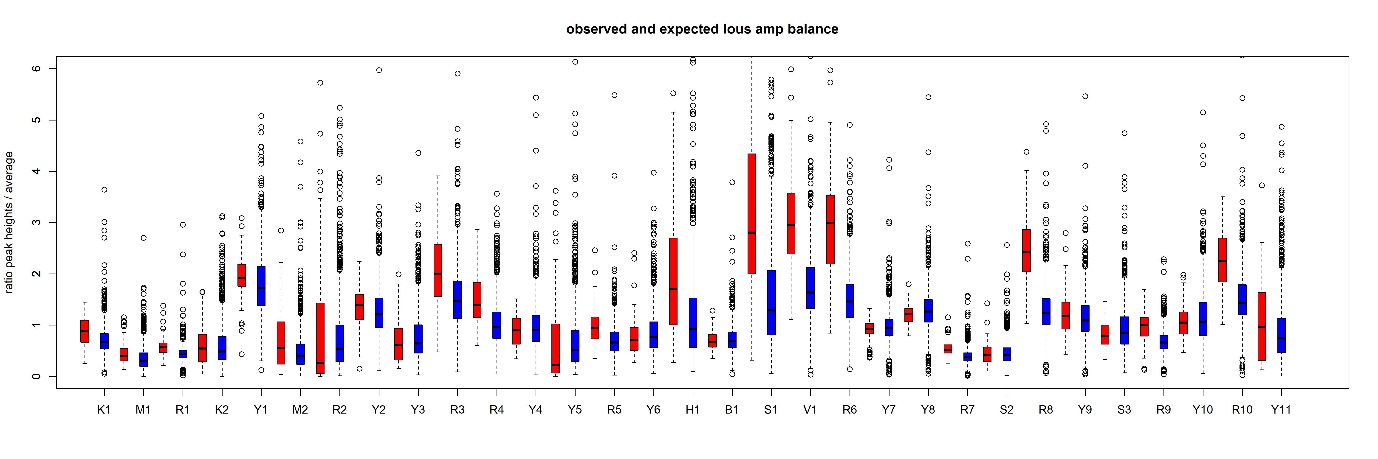


*Figure 13: Observed (O, red) and simulated (S, blue) dye efficiency values*

The locus amplification efficiency model checking can occur in a similar manner to the dye effect check. In order to cancel the effects of template the ratio of the summed peak heights at a locus over the average peak height in the profile was obtained. We note that in this ratio will be a dye effect, which will add noise to the locus signal. We simulate values by:

1. A template value, *Trandom*, was drawn from the fitted exponential distribution seen in Figure 12
2. Random observed peak heights, *Orandom*, were drawn from  for each locus in a simulated profile
3. A locus amplification efficiency value for each locus *l* was drawn from 
4. The plotted point for locus *l* is then 

The resulting plot is shown in Figure 14.



*Figure 14: Observed (red) and simulated (blue) locus amplification efficiency values*

**CONCUSIONS:**

* Most parameters converged adequately
* Model checking gave sensible results
* Slight underestimation in dye effect for A/T and for locus amp effects of S1, V1 and R6 (noting that the ranged of observed and simulated do still overlap), but in general all predictions close to observations.
* Would like to trial simplification of the model where constant variances across dyes or across loci are trialled (TODO, produce graph for variances similar to Figures 7 and 8)
* Trial different priors for hyperparameters other than uninformative over a range (maybe conjugate priors??)
* Despite all this potential further work the HBM used has provided information on posterior distributions for parameters to do with peak height variability as a result of stochastic effects, locus amplification efficiency effects and dye efficiency effects.
* From these results interpretation guidelines could be produced (similar to idea of producing interpretation guideline from probabilistic modelling as was given in [4]), or ultimately a probabilistic analysis system could be produced to provide probabilities of SNP typings in low level or mixed profiles.

**REFERENCES:**

[1] D. Taylor, J.-A. Bright, J. Buckleton, The interpretation of single source and mixed DNA profiles, Forensic Science International: Genetics 7(5) (2013) 516-528.

[2] D. Taylor, J. Buckleton, J.-A. Bright, Factors affecting peak height variability for short tandem repeat data, Forensic Science International: Genetics (2016).

[3] D. Taylor, J.-A. Bright, C. McGoven, C. Hefford, T. Kalafut, J. Buckleton, Validating multiplexes for use in conjunction with modern interpretation strategies, Forensic Science International: Genetics 20 (2016) 6-19.

[4] D. Taylor, J.-A. Bright, J. Buckleton, Using Probabilistic Theory to Develop Interpretation Guidelines for Y-STR Profiles, Forensic Science International: Genetics 21 (2016) 22-34.