Statistical tools for ultra-deep pyrosequencing of fast evolving viruses

David Knowles Susan Holmes

We aim to detect minor variant Hepatitis B viruses (HBV) in 38 pyrosequencing samples from infected individuals. Errors involved in the amplification and ultra deep pyrosequencing (UDPS) of these samples are characterised using HBV plasmid controls. Homopolymeric regions and quality scores are found to be significant covariates in determining insertion and deletion (indel) error rates, but not mismatch rates which depend on the nucleotide transition matrix. This knowledge is used to derive two methods for classifying genuine mutations: a hypothesis testing framework and a mixture model. Using an approximate "ground truth" from a limiting dilution Sanger sequencing run, these methods are shown to outperform the naive percentage threshold approach. The possibility of early stage PCR errors becoming significant is investigated by simulation, which underlines the importance of the initial copy number.

1 Introduction

When an individual becomes infected by a fast evolving virus, such as Human Immunodeficiency Virus (HIV-1) or Hepatitis B (HBV), minor variants rapidly evolve. Although these variants may exist at very low levels, they are hugely important in determining drug resistance. If a minor variant is resistant to the drug that inhibits the primary strain, it will rapidly proliferate under this new selective pressure [1]. The treatment will be ineffective and has helped new drug resistant strains proliferate. As a result, methods to identify minor variants present in an individual are of great interest for directing treatment. With limiting dilution Sanger sequencing variants present at 20% or above are detectable. A relatively new method, which pushes that limit down to around 1%, uses ultra deep pyrosequencing (UDPS) [2]. UDPS allows short reads of viral DNA to be sequenced at enormous coverage (around 5000x) and reasonable cost.

We statistically characterise the errors involved in 454 ultra deep pyrosequencing of HBV using three plasmid controls, and to use this understanding to design appropriate methods to reliably detect genuine minor variants in a dataset of samples from 38 individuals with HBV.

HBV DNA was extracted from the blood plasma of 38 infected individuals and sequenced using 454 pyrosequencing, along with three HBV-1 genomes of known sequence in plasmid vectors. The processes involved are: extraction, limiting dilution Polymerase Chain Reaction (PCR), amplification PCR, dilution, and pyrosequencing. The RNA/DNA is extracted from patient plasma, which might contain around 100,000 copies per ml. After extraction we hope to have an initial copy number of at least 100. Thus by estimating F_0 we can estimate λ. Samples with an initial copy number less than 100 were discarded. Amplification PCR can be performed using several enzymes which give different accuracy-yield trade off. A Taq blend was used for this dataset.

For this dataset four slightly overlapping regions ("amplicons") were amplified using eight custom made primers with known binding sites, making alignment straightforward.

In a previous UDPS study on HIV-1 [3], a Poisson distribution on errors was used in the homopolymeric and non-homopolymeric regions, which was fitted by Expectation Maximi-

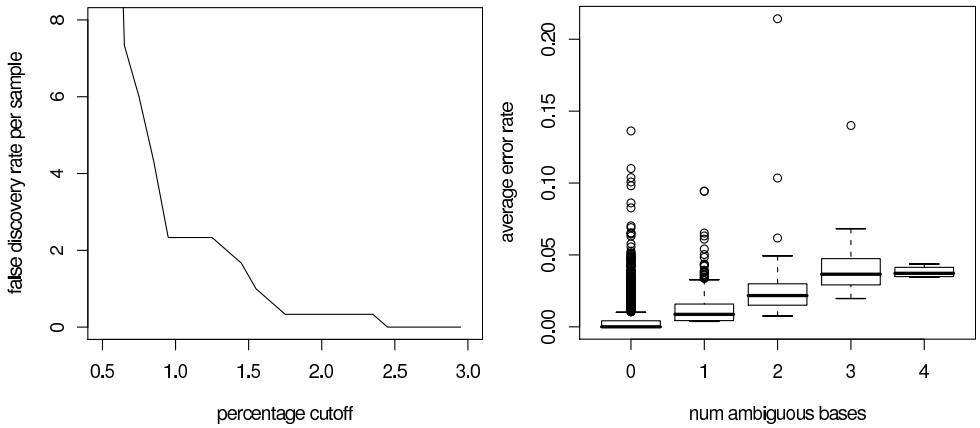
sation (EM). However, the increased error rate for the Taq blend enzyme results in an overdispersed error distribution.

2 Statistical analysis

In order to detect which signals in the data represent genuine variants it is necessary to characterise the amplification and pyrosequencing errors. Three well characterised HBV plasmid vectors were pyro-sequenced using the same experimental method as the patient samples (although the initial copy number was significantly higher for the controls, around 100,000, compared to 100-1000 for the patient samples). Deviations from the consensus sequence represent either PCR or pyrosequencing errors. This control data allows us to fit an error model.

False discovery rate. In previous studies [3], an error was classified as a genuine mutation if it is observed at a given position in more than 1% (for example) of the reads. Thus it is of interest to look at the distribution of these error proportions in the control data to obtain an empirical estimate of the false discovery rate (FDR) for different thresholds.

Figure 1(a) shows the empirical FDR across all three controls for varying percentage cutoff. At seven positions across all three controls mismatches occur in more than 1% of reads, so we estimate the FDR per sample would be $\frac{7}{3} = 2.3$.



(a) Empirical estimate of FDR versus percentage cutoff.

(b) Error rate against number of ambiguous base calls.

Figure 1: Statistical analysis of plasmid control data.

Individual read quality. A previous study [4] found that a small number of poor quality reads contained a disproportionate percentage of the total errors. Similarly, we found the worst 2% of reads account for 20% of errors. In [4] reads with lengths outside the main peaks had increased error rate, which we confirmed. Figure 1(b) shows the strong correlation between the error rate and the number of ambiguous base calls in a read. Note that only 2% of reads contain any ambiguous base calls, so discarding these is recommended.

Errors occur in 454 pyrosequencing because some proportion of the PCR reactions on a bead get out of sync [2]. We would therefore expect a cumulative effect along the length of a read. Error rate against distance from the 5' end of the read is shown in Figure 2. There is significant noise, with systematic peaks appearing across all three controls, due to homopolymeric regions where the indel error rate is increased.

Mismatch rates. We can ignore indels because they are assumed to be PCR/sequencing errors since frameshifts are almost impossible biologically. The mean mismatch error rate is 1.38×10^{-3}, and the maximum at one position is 4.2×10^{-2}. Table 1 shows the mismatch error rates averaged across all three controls, normalised for the relative frequency of each

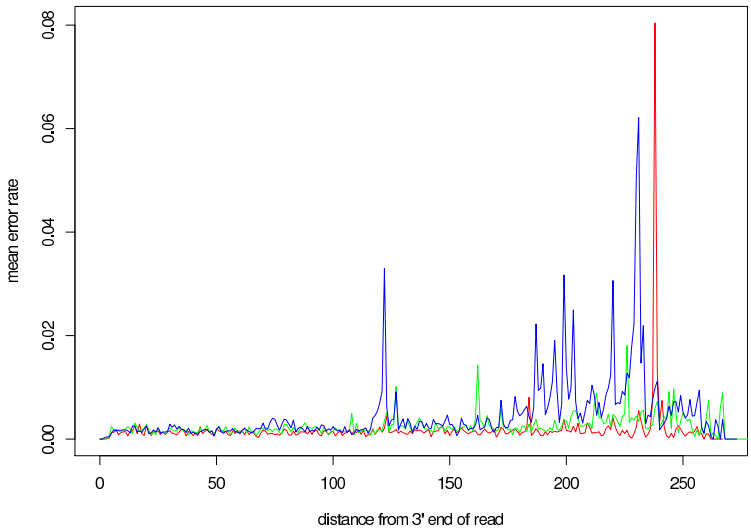


Figure 2: Error rate (mismatch and indel) versus distance from 3' end for each control.

	A	G	T	C
A	9.99e-01	1.39e-03	7.12e-05	3.39e-05
G	4.53e-04	9.99e-01	3.22e-04	1.87e-05
T	1.69e-04	3.73e-05	9.98e-01	1.54e-03
C	4.14e-04	4.74e-05	4.10e-04	9.99e-01

Table 1: Normalised mismatch error rates across controls.

base. As expected, a base is more likely to remain a purine (A or G) or pyrimidine (C or T). For example, $A \rightarrow G$ mismatch errors occur around twenty times more frequently than $A \rightarrow T$ for example.

Homopolymeric regions. 454 pyrosequencing is known to be particularly error prone in homopolymeric regions due to carry forward and incomplete extension (CAFIE) errors [2]. Incomplete extension is when the homopolymer is not completed due to insufficient dNTPs. Carry forward errors occur when a nucleotide from the end of a homopolymer is read a few bases later on due to incomplete dNTP flushing. For example, if the true sequence is AAAATCG, it may be read as AAATCGA. We define a homopolymeric region as three or more identical nucleotides and the immediately flanking nucleotides.

The indel error rate increases from 1.76×10^{-3} to 2.98×10^{-3} in homopolymeric regions. The mismatch error rate is not significantly affected by whether the region is homopolymeric, staying at 1.13×10^{-3}. The increase in the overall error rate in homopolymeric regions is due only to the increase in indel rate.

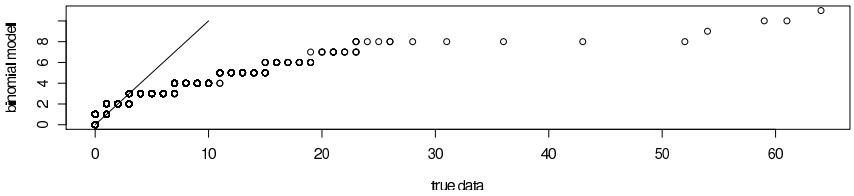
Quality scores. The quality scores from the pyrosequencing software relate to the probability of CAFIE errors, which is somewhat different to Sanger sequencing *phred* scores. There is significant correlation between the mismatch error rate and average quality score, but the effect size is small. Quality scores are not predictive of specific incorrect base calls.

Overdispersion. Even after counting for known covariates the error distribution is overdispersed making the negative Binomial a more appropriate distribution that the Poisson.

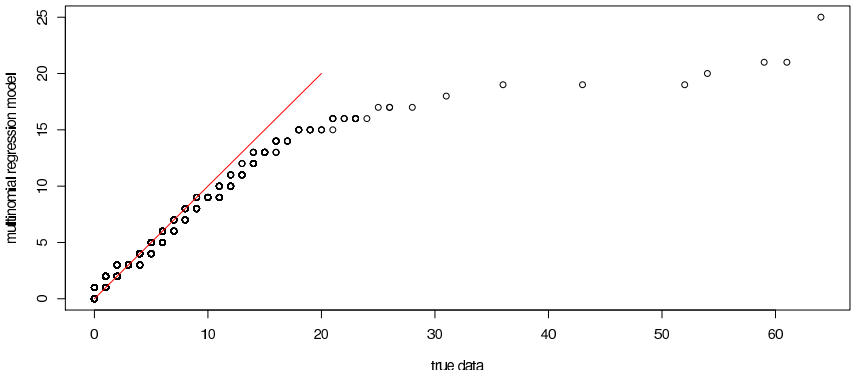
Multinomial regression. Several covariates are available at each sequence position: the consensus sequence, whether the region is homopolymeric and the quality score. Multinomial

regression allows which specific error occurs to be modelled. The function `multinom` from the R package `nnet` was used, which fits the regression using a neural network and allows counts rather than raw data unlike other packages.

Including the consensus, homopolymeric, and quality covariates all reduce the Akaike Information Criterion (AIC), which implies they are all significant. Figure 3(b) shows a qq plot of data simulated from the multinomial regression model versus the true data, for mismatch errors only (a perfect fit would give a straight line). For comparison, Figure 3(a) shows a qq plot of data simulated from a naive binomial model, ignoring all covariates and taking all mismatch errors as equivalent. Clearly the multinomial regression provides a significantly better fit, although the outlying large errors are still not accounted for.



(a) Binomial.



(b) Multinomial regression.

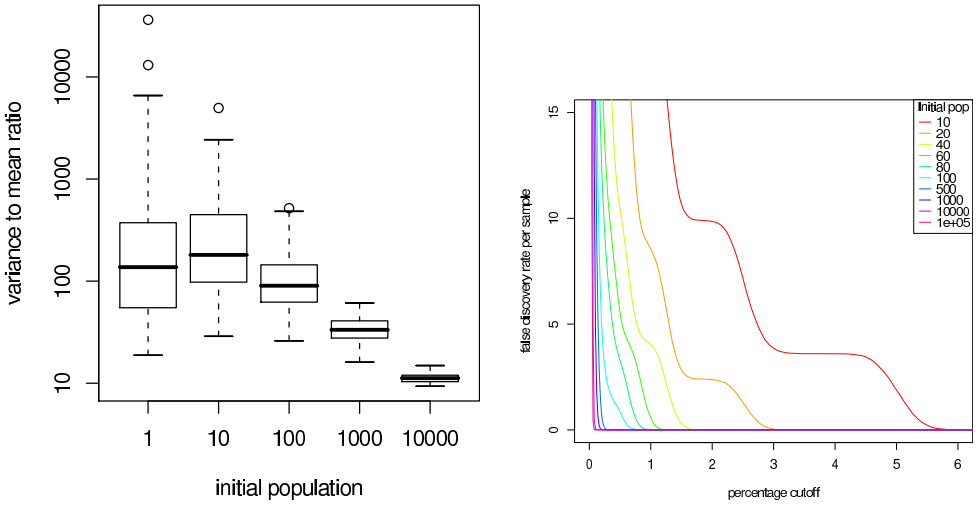
Figure 3: QQ plots of mismatch errors against simulated data.

To test the significance of the parameters in the model a non-parametric bootstrap can be used [5]. The probability model is approximated by its empirical distribution: delta functions of mass $\frac{1}{N}$ at each of the N observations. The sampling distribution of each parameter is estimated by Monte Carlo by sampling with replacement from the original observations. Coefficients with confidence intervals which do not include zero are significant at a 95% confidence level. For mismatch errors the consensus base is significant. The homopolymeric factor is not significant in determining the mismatch error rate. Interestingly the quality score is also not a significant covariate for the mismatch error rates. Deletion and insertion errors rates are the reverse: the consensus base is not significant, but homopolymeric regions are, as are quality scores for insertion errors.

3 PCR Simulation

Early cycle PCR errors could be amplified to a significant proportion of the population and resemble a genuine minor variant, depending on the initial copy number. Since the controls had initial copy numbers on the order of 100,000, compared to just 100 to 1000 for the samples, they cannot answer this question. The PCR amplification was simulated as a stochastic autocatalytic reaction with binary mutations. New DNA molecules inherit mutations from their parent molecule, and gain new mutations at random at a specified rate. Using sparse matrices it is possible to represent final populations of around 10^9 in 1Gb RAM.

The aim of the PCR simulations is to assess the effect of low initial copy number. Figure 4(a) shows boxplots of the variance to mean ratio (a measure of overdispersion) for 100 repeats of the simulation, for varying initial copy number. As expected small initial copy numbers lead to increased overdispersion.



(a) Variance to mean ratios.

(b) FDR based on PCR simulations with varying initial copy number.

Figure 4: PCR simulation results.

Using the final population we can estimate the False Discovery Rate per sample as a function of the percentage cut off, as shown in Figure 4(b), assuming a PCR error rate of 10^{-5}. The multimodal nature of the low copy number error distributions is due to early cycle errors which result in delta functions at exponentially increasing intervals. These are smoothed by stochastic effects in the simulation.

4 Classifying genuine mutations

In this section we develop two methodologies for classifying genuine mutations, using the results of the statistical analysis. Since for mismatch errors rates the quality score and homopolymeric regions are not significant, estimation of the 4 by 4 nucleotide transition matrix, $\mathbf{\Theta}$, will be important for both methods.

Estimating the nucleotide transition matrix. Our data is the count matrix for the controls, $\mathbf{n}$, where element i, j is the number of times nucleotide j was observed when the consensus nucleotide was i. Each row $n_{i:}$ is multinomially distributed with probability vector $\Theta_{i:}$, corresponding to row i of $\mathbf{\Theta}$, the transition probability matrix. We specify a Dirichlet prior on $\mathbf{\Theta}$ with parameter vector $\boldsymbol{\alpha}_{i:}$. Thus:

$$P(\Theta_{i:}|\boldsymbol{\alpha}_{i:}) = \frac{\Gamma(\sum_j \alpha_{ij})}{\prod_j \Gamma(\alpha_{ij})} \prod_j \Theta_{ij}^{\alpha_{ij}-1} \tag{1}$$

We parameterise $\boldsymbol{\alpha}$ as follows:

$$\alpha_{ij} = \begin{cases} a & \text{if } i = j \\ b & \text{if } i \neq j \end{cases} \tag{2}$$

This encodes that no particular mismatch error is more likely, but the probability of no mismatch is different. (An alternative would be to have different prior parameters for transition vs. transversion mismatches). The joint distribution over the data $\mathcal{D}$ and $\mathbf{\Theta}$ can

now be expressed:

$$P(\mathcal{D}, \mathbf{\Theta}|a, b) = P(\mathcal{D}|\mathbf{\Theta}, a, b)P(\mathbf{\Theta}|a, b) \quad (3)$$

$$= \frac{\Gamma(a+3b)^4}{\Gamma(b)^{12}\Gamma(a)^4} \prod_i \Theta_{ii}^{n_{ii}+a-1} \prod_{j\neq i} \Theta_{ij}^{n_{ij}+b-1} \quad (4)$$

Thus the posterior distribution of each row i of $\mathbf{\Theta}$ is Dirichlet$(n_{ij} + \alpha_{ij})$, and its maximum a posterior (MAP) estimate is

$$\Theta_{ij}^{\text{MAP}} = \frac{n_{ij} + \alpha_{ij}}{\sum_k (n_{kj} + \alpha_{kj})} \quad (5)$$

because $n_{ij} + \alpha_{ij}$ is the effective count for nucleotide i going to j. We can fit the hyperparameters a and b using the evidence framework, maximising $P(\mathcal{D}|a, b)$, a Type II maximum likelihood method [6].

$$P(\mathcal{D}|a, b) = \int P(\mathcal{D}, \mathbf{\Theta}|a, b) d\mathbf{\Theta}$$

$$= \frac{\Gamma(a+3b)^4}{\Gamma(b)^{12}\Gamma(a)^4} \prod_i \frac{\Gamma(n_{ii}+a)\prod_{j\neq i}\Gamma(n_{ij}+b)}{\Gamma(n_{ii}+a+\sum_{j\neq i}(n_{ij}+b))}$$

We maximise the log evidence using Newton's method [**?**].

4.1 Hypothesis testing

Once an estimate of the nucleotide transition matrix $\mathbf{\Theta}$ is available the likelihood of a specific error under the model can be calculated. If we have a position where the consensus nucleotide is i but nucleotide j is observed n_e times out of a coverage of n, then n_e is Binomial$(n, \Theta_i j)$ distributed under the null hypothesis of no minor variants. Let $\beta_{ij} = n_{ij} + \alpha_{ij}$ be the parameters of the posterior Dirichlet distribution over $\mathbf{\Theta}$ given the control data $\mathbf{n}$. The marginal distribution of Θ_{ij} is then Beta$(\beta_{ij}, \beta_{i.} - \beta_{ij})$, where $\beta_{i.} = \sum_j \beta_{ij}$. Marginalising Θ we find

$$P(n_{ij} = n_e|\boldsymbol{\beta}) = \binom{n}{n_e} \frac{B(n_e + \beta_{ij}, n - n_e + \beta_{i.} - \beta_{ij})}{B(\beta_{ij}, \beta_{i.} - \beta_{ij})} \quad (6)$$

Since it is possible to perform this integration analytically there is little additional computational cost compared to using a MAP estimate. We can calculate a p-value $P(n_{ij} \geq n_e|\boldsymbol{\beta})$. Since usually $n_e \ll n$ it will be cheaper to calculate the p-value as follows:

$$P(n_{ij} \geq n_e|\boldsymbol{\beta}) = 1 - \sum_{m=0}^{n_e-1} P(n_{ij} = m|\boldsymbol{\beta}) \quad (7)$$

where each term in the sum is evaluated according to Equation 6.

4.2 A mixture model

The observed mismatches are generated by two processes: PCR/sequencing errors and genuine mutations. A two component mixture model can be used to represent this, where the mixture proportions correspond to the probability a particular mismatch is a genuine mutation rather than an error. We use Expectation Maximisation, iterating between fitting the mixing proportions and model parameters, but with the error model from the control data. To model the genuine mutations we use a codon mismatch matrix, since this allows the incorporation three desirable features:

1. Synonymous mutations are more likely than non-synonymous mutations.
2. Non-synonymous mutations which result in a physiochemically different (e.g. different polarity) amino acid, are unlikely because they effect protein function.
3. Mutations which result in a stop codon are very rare because the shortened protein would be non-functional.

The number of parameters in the codon model, $64^2 = 4096$, is very large so Bayesian inference is ideal to prevent overfitting. We model and infer the codon mismatch matrix in the same way as nucleotide mismatch matrix, only now the indices are over codons rather than nucleotides.

Expectation step. We estimate the mixture proportions holding the model parameters fixed. Let m_i be a binary latent variable equal to 1 if codon mismatch i is a genuine mutation, and equal to 0 if it is an error. To calculate the probability that mismatch i is a genuine mutation, π_i, we use Bayes' rule assuming equal priors (i.e. mutation and error are equally likely a prior):

$$\pi_i = P(m_i = 1|\mathcal{D}_i, \Theta_{\text{error}}, \Theta_{\text{mutation}}) = \frac{P(\mathcal{D}_i|m_i = 1, \Theta_{\text{mutation}})}{P(\mathcal{D}_i|m_i = 1, \Theta_{\text{mutation}}) + P(\mathcal{D}_i|m_i = 0, \Theta_{\text{error}})} \tag{8}$$

where $\mathcal{D}_i$ is the data associated with mismatch i (i.e. reference and query codon, how many repeats and coverage), and Θ_{mutation} and Θ_{error} are the current estimates of the codon transition matrix for the mutation and error models respectively.

Maximisation step. This step involves updating the codon mutation matrix Θ_{mutation} using the mixing proportions calculated in the previous step. The counts used now are a weighted sum, with the weights given by the mixing proportions.

$$n_{ij} = \begin{cases} \sum_k n_k \mathbf{1}(r_k = i, q_k = j) & \text{if } i = j \\ \sum_k n_k \pi_k \mathbf{1}(r_k = i, q_k = j) & \text{if } i \neq j \end{cases} \tag{9}$$

where r_k and q_k are the reference and query codons respectively for mutation k, $\mathbf{1}(\cdot)$ is one if the statement is true (zero otherise), and n_k is the number of times this codon mismatch is observed. Note that for counting the number of times mismatches do not *not* occur for a codon, the mixing proportion is effectively one since no mutation (nor error) has occurred.

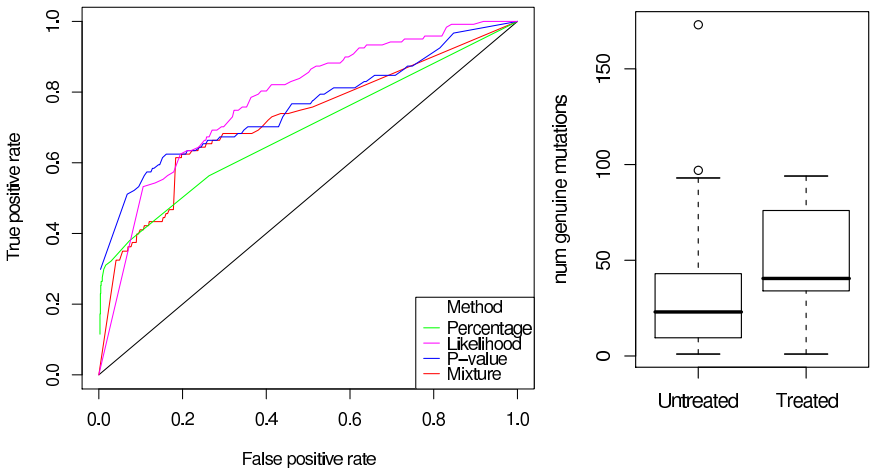
5 Results

No ground truth is available: which mismatches really are genuine mutations? Limiting dilution Sanger sequencing results are available for one of the samples however. This method is not able to detect minor variants at very low levels, so some genuine mutants will not be detected. Never-the-less, this is the closest to ground truth available for comparing the classification methods. The data consist of 95 sequences, which we aligned to the 454 consensus sequence [7]. ROC curves for each method are shown in Figure 5(a). Since some genuine mutations will be missing from the "ground truth" set, the number of false positives will be over-estimated and the number of true positives under-estimated, but the relative performance of the methods can still be assessed.

The more sophisticated methods outperform the percentage cutoff along most of the curve. The mixture model performs worse than the hypothesis testing method at most levels. This may be due to overfitting of the large number of parameters in the mutation model codon mismatch matrix which is currently MAP estimated. Given the large number of parameters in the model integrating over the posterior would be prudent. Figure 5(b) shows a boxplot of the number of genuine mutations classified by the hypothesis testing method for treated versus untreated patients. The number of detectable mutations in the treated patients is significantly higher due to the increased selection pressure. This shows the kind of biologically significant results that these methods can enable.

6 Conclusion

I have statistically characterised the errors inherent in 454 pyrosequencing, and used the results to design methods for detecting genuine variants which outperform the naive threshold method commonly used. I have used computer simulations of the PCR to help understand how initial copy number determines the probability of false positives resulting from early cycle errors.



(a) ROC curve comparing classification method performance.

(b) Boxplots of number of mutations depending on treatment.

Figure 5: Classification results.

As mentioned in Section 5, the somewhat disappointing performance of the mixture model maybe due to overfitting of the large parameter mutation model. To overcome this integration over the posterior of the codon transition matrix should be performed, rather than using a MAP estimate.

A more ambitious aim would be to incorporate more multivariate information into the classification methods. For example, if two mismatches always co-occur, it is highly unlikely they are errors but feasible that they both occur in the same minor variant. A computationally intensive method would be to attempt to infer the hidden phylogeny of the minor variants. Both these methods are complicated by the fact that the sequences only cover some of the region of interest. Each amplicon would have to be considered separately, but the phylogenies would need to be consistent for each.

454 pyrosequencing offers the potential to both answer questions about the evolution of drug resistance in fast evolving viruses and provide an affordable diagnostic alternative to expensive functional assays. I hope the analysis and methods presented here will help achieve these goals.

References

[1] Bndicte Roquebert, Isabelle Malet, Marc Wirden, Roland Tubiana, Marc-Antoine Valantin, Anne Simon, Christine Katlama, Gilles Peytavin, Vincent Calvez, and Anne-Genevive Marcelin. Role of hiv-1 minority populations on resistance mutational pattern evolution and susceptibility to protease inhibitors. *AIDS*, 20(2):287–289, Jan 2006.

[2] Marcel Margulies, Michael Egholm, William E Altman, Said Attiya, Joel S Bader, and Jonathan M Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–380, Sep 2005.

[3] Chunlin Wang, Yumi Mitsuya, Baback Gharizadeh, Mostafa Ronaghi, and Robert W Shafer. Characterization of mutation spectra with ultra-deep pyrosequencing: application to hiv-1 drug resistance. *Genome Res*, 17(8):1195–1201, Aug 2007.

[4] Susan M Huse, Julie A Huber, Hilary G Morrison, Mitchell L Sogin, and David Mark Welch. Accuracy and quality of massively parallel dna pyrosequencing. *Genome Biol*, 8(7):R143, 2007.

[5] B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

[6] D. J. C. Mackay. *Bayesian methods for adaptive models.* PhD thesis, California Institute of Technology, 1991.

[7] M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins. Clustal w and clustal x version 2.0. *Bioinformatics*, 23(21):2947–2948, Nov 2007.