Statistical tools for ultra-deep pyrosequencing of fast evolving viruses

David Knowles
Cambridge University
Engineering Department

Susan Holmes
Statistics Department
Stanford University

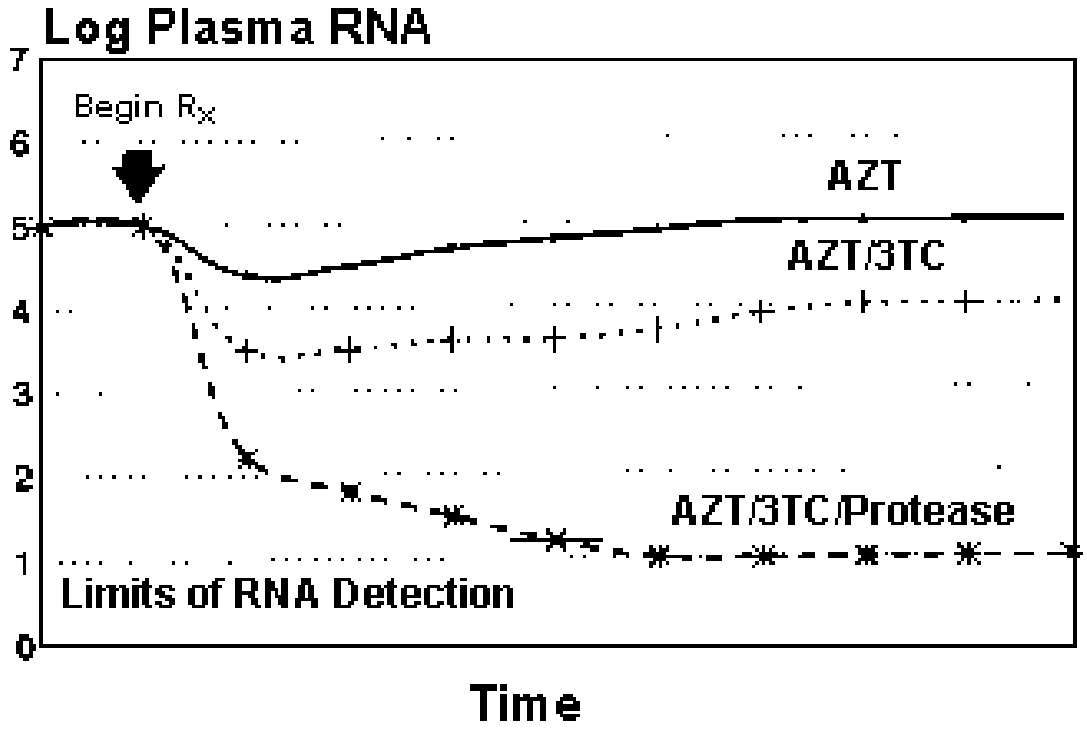
UNIVERSITY OF CAMBRIDGE

STANFORD UNIVERSITY

Motivation

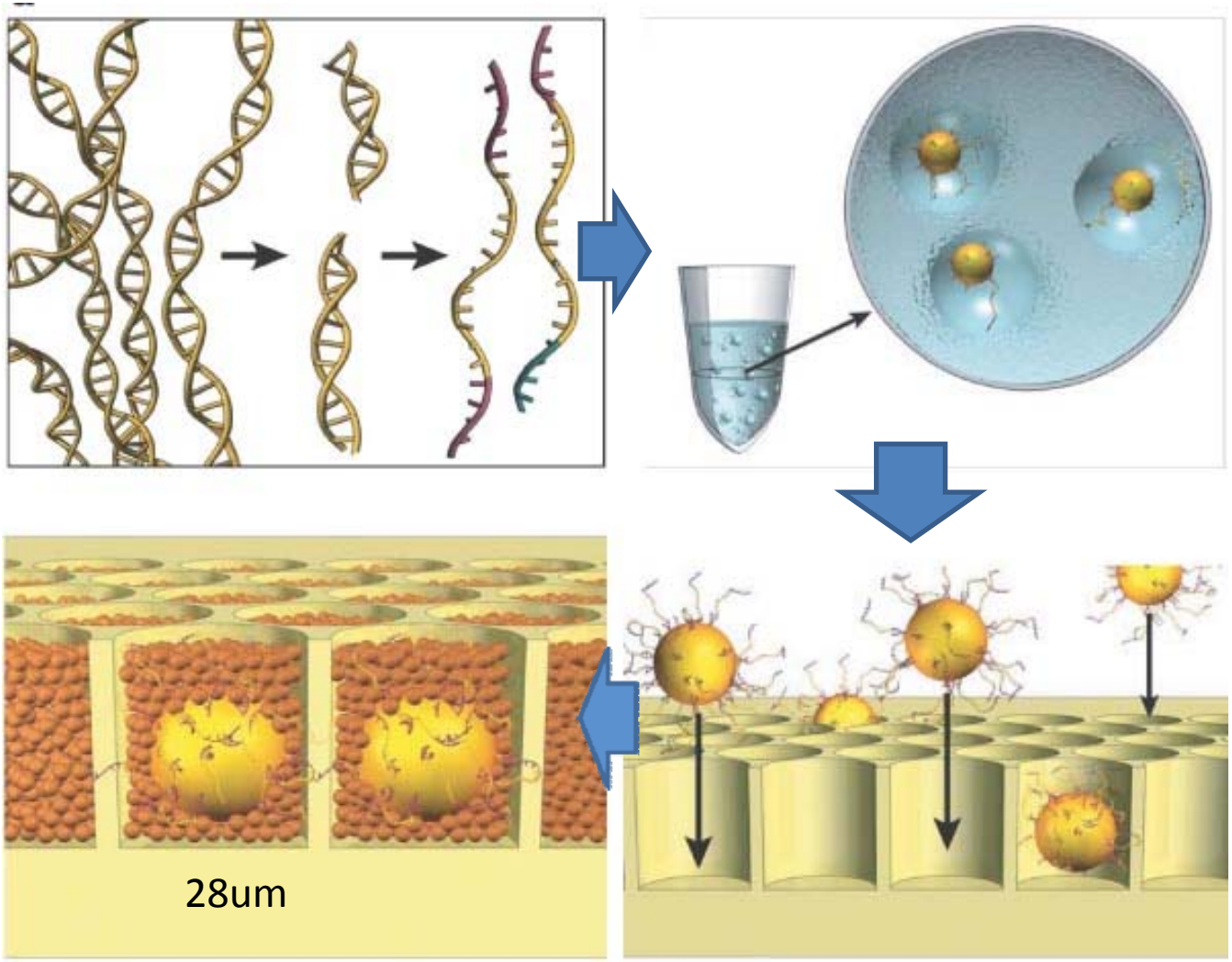
- Multiple strains of a virus are often present in an individual
- A minor strain may be resistant to the prescribed drug
- Treatment is ineffective, and increases the prevalence of a resistant strain (not to mention expense)

Motivation

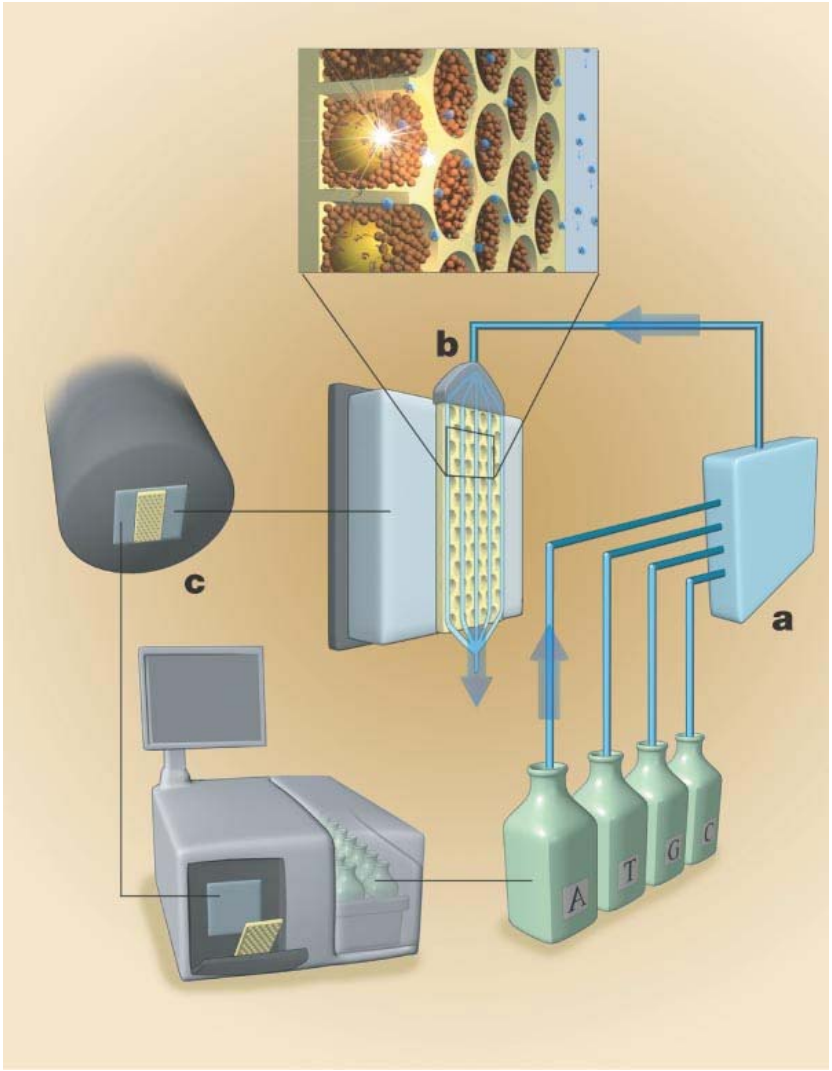


[John Hopkins AIDS service]

454 Pyrosequencing



454 Pyrosequencing



Detecting minor Hep-B variants

- Limiting dilution Sanger sequencing can only find variants present at >20%
- Ultra deep pyrosequencing gives 5000x coverage at reasonable cost
- So how rare a variant can we detect with this?

Sources of error

10^5 copies/ml 10^2 copies/ml

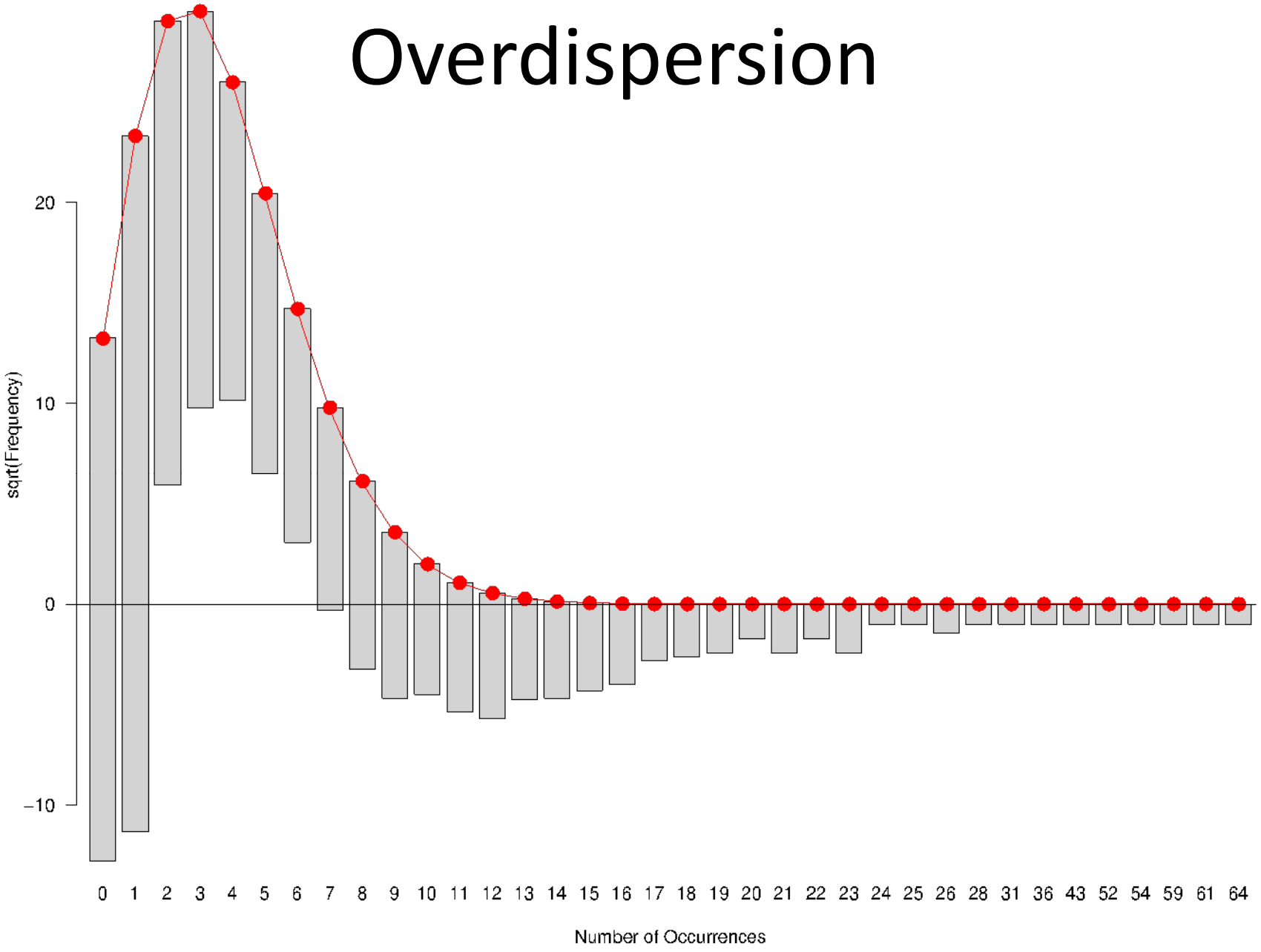
Jargon

- **Homopolymeric:** repeated bases, usually three or more in a row
- **Mismatch:** a different nucleotide is called than the true nucleotide. Can be a...
- **Transition** (purine to purine or pyridimine to pyrimidine) or a...
- **Transversion** (purine to pyrimidine or visa versa).
- **Indel**: an insertion (an extra nucleotide is read) or deletion (a nucleotide is skipped)
- **Ambiguous base call**: the sequencer couldn't determine the nucleotide at a particular position
- **Initial copy number**: the number of DNA molecules after extraction but before PCR amplification

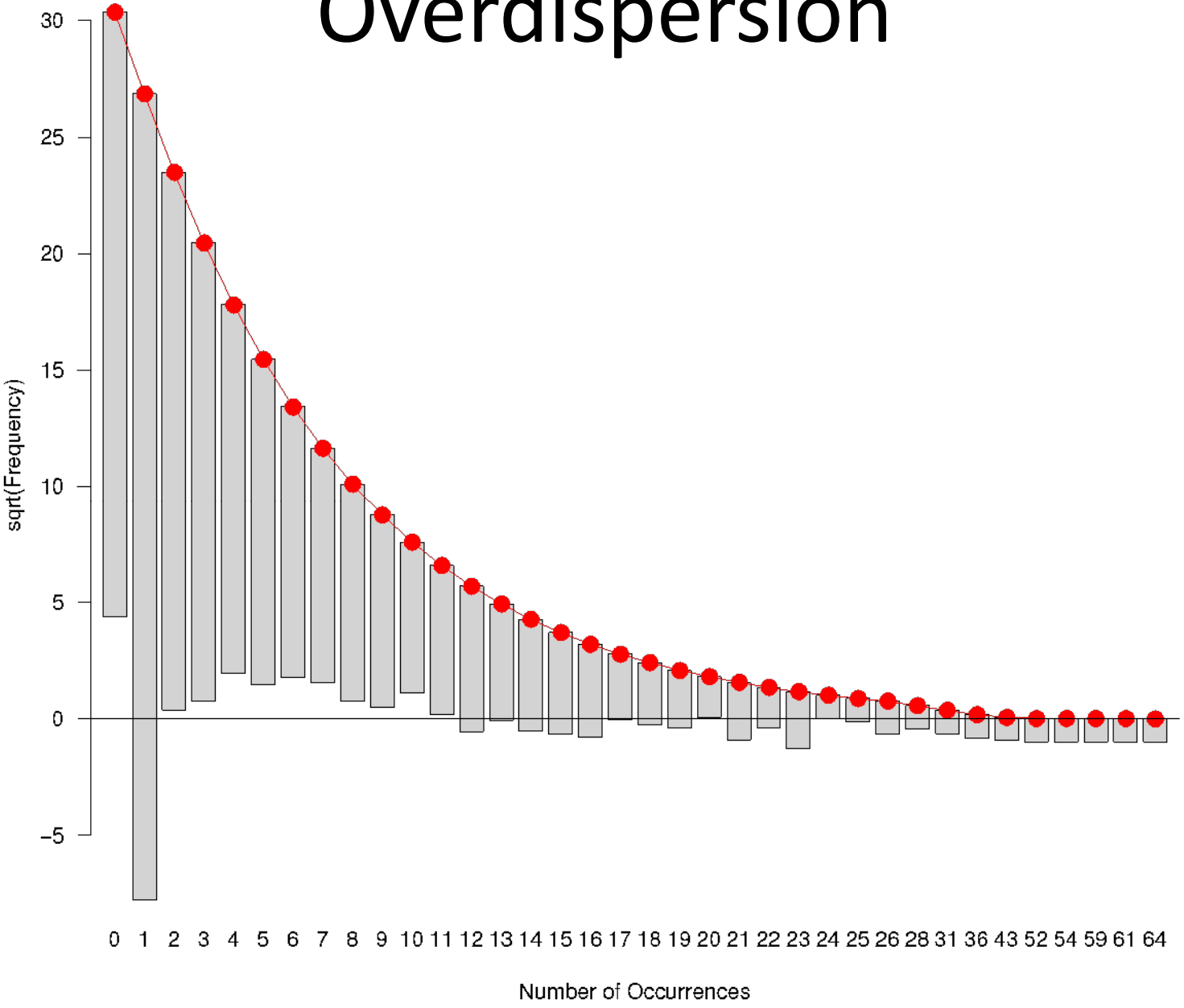
Fitting the “noise”

- Three plasmid controls with known sequence
- Deviations can therefore be assumed to be pyrosequencing errors

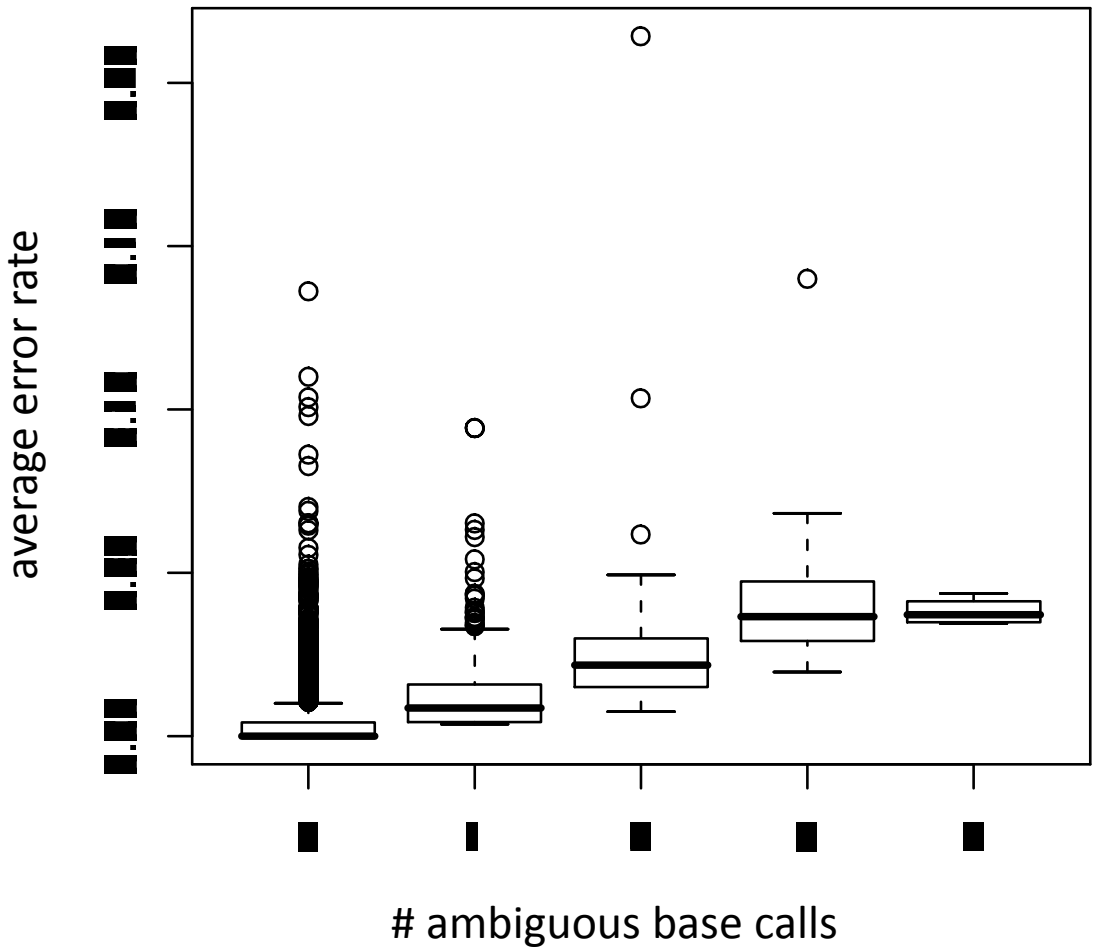
Overdispersion



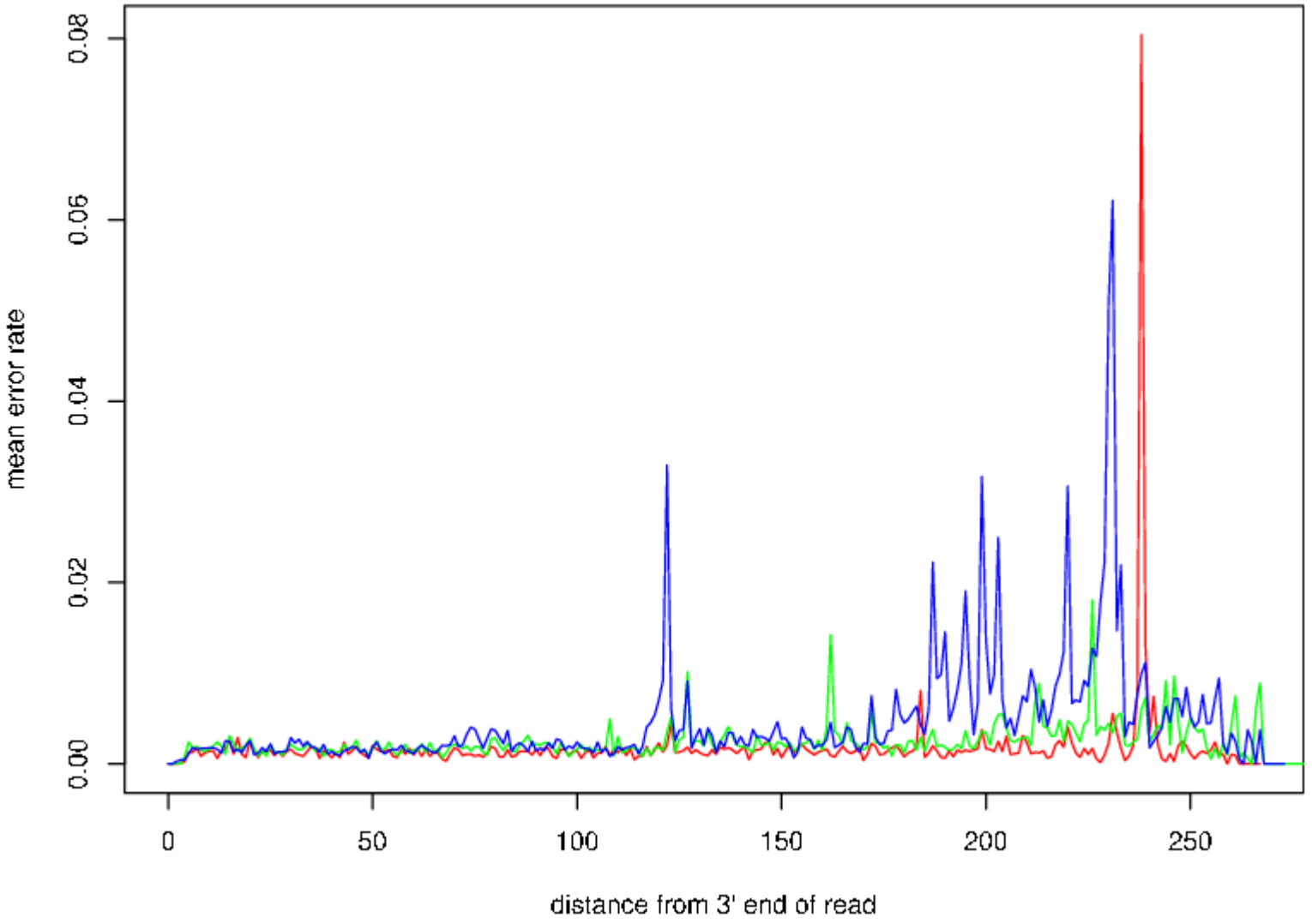
Overdispersion



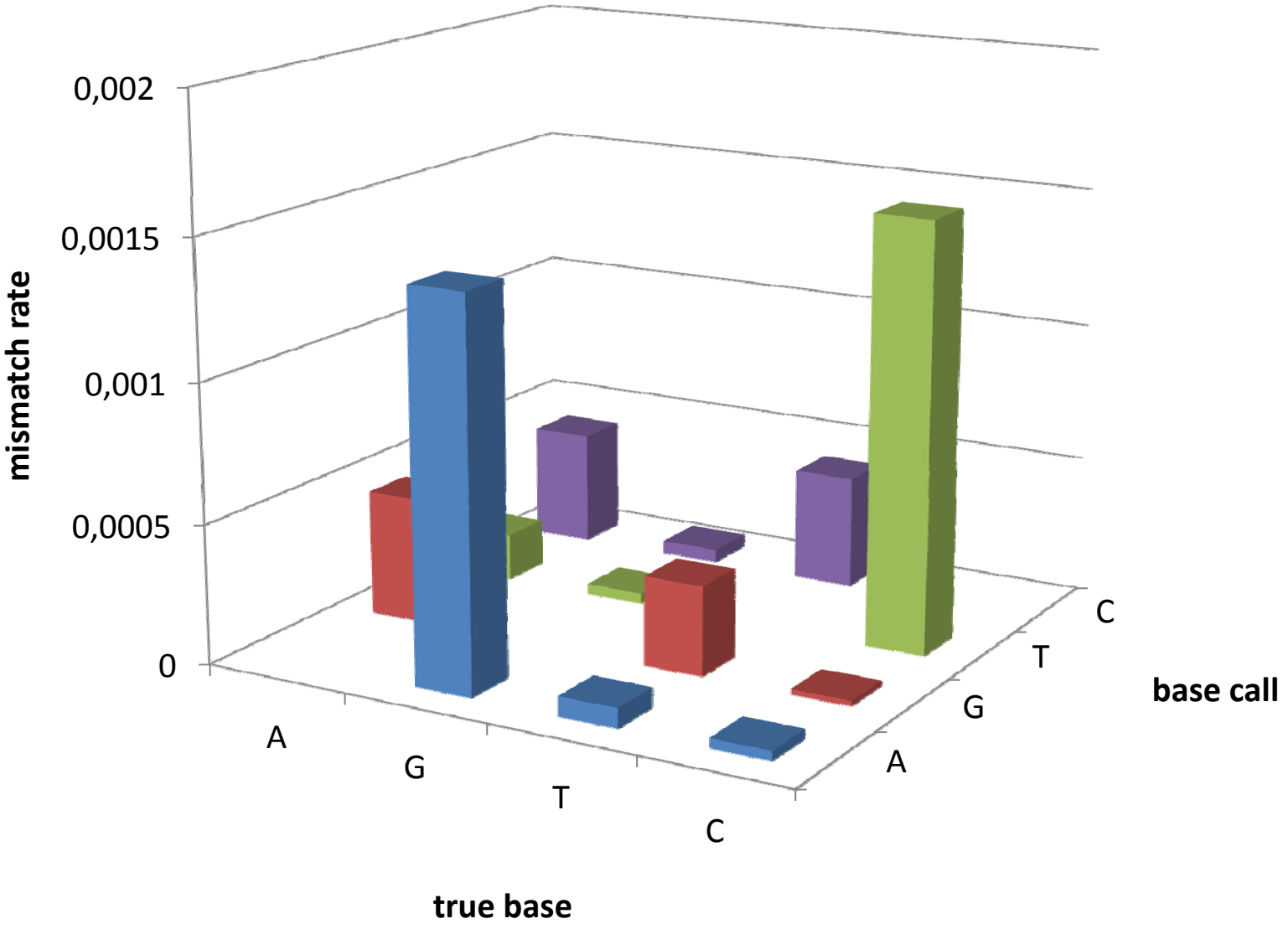
Some reads are terrible



Indels accumulate along a read



Some mismatches are common



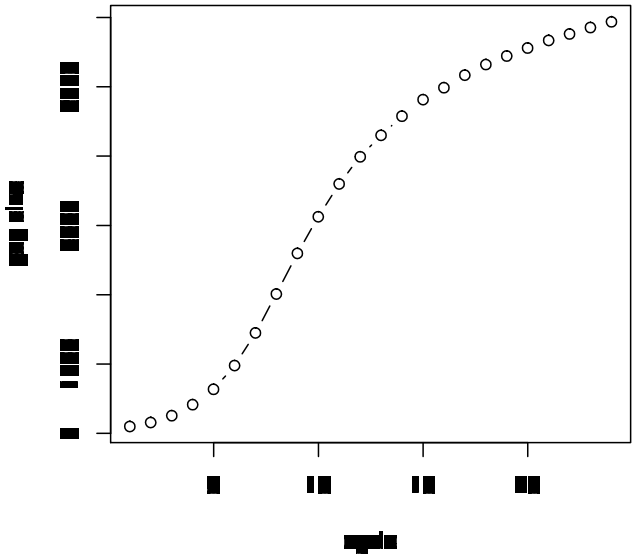
PCR simulation

- Could early stage PCR errors lead to significant "variants" in the final population?
- Initial copy number = n, amplicon length = l, error rate per base per duplication = e
- Assuming all molecules duplicated each round:

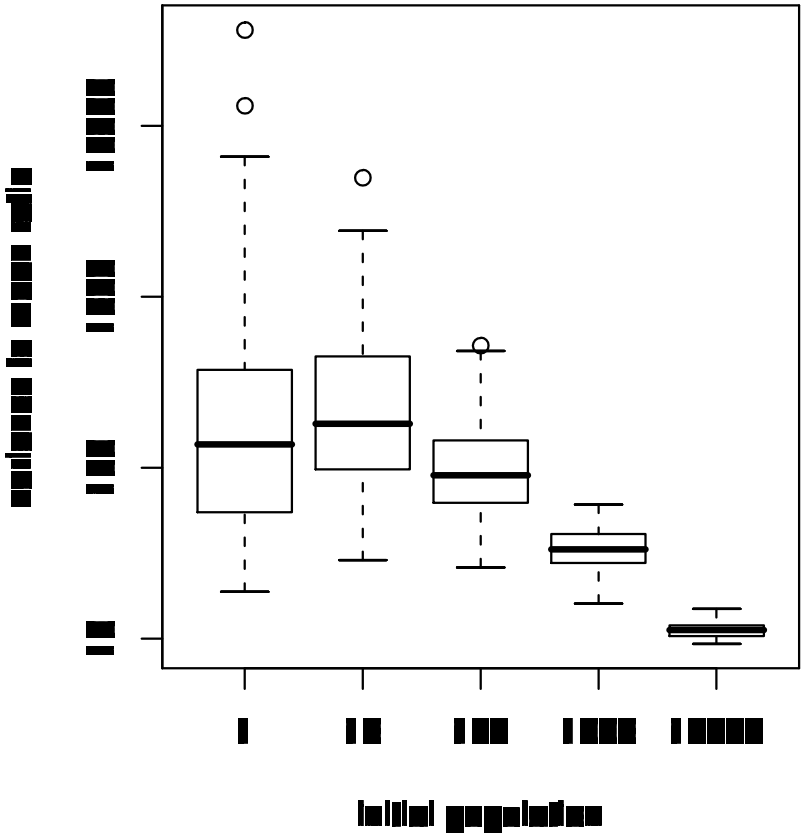
	Probability	Final proportion
Round 1	2nle	1/(2n)
Round 2	4nle	1/(4n)

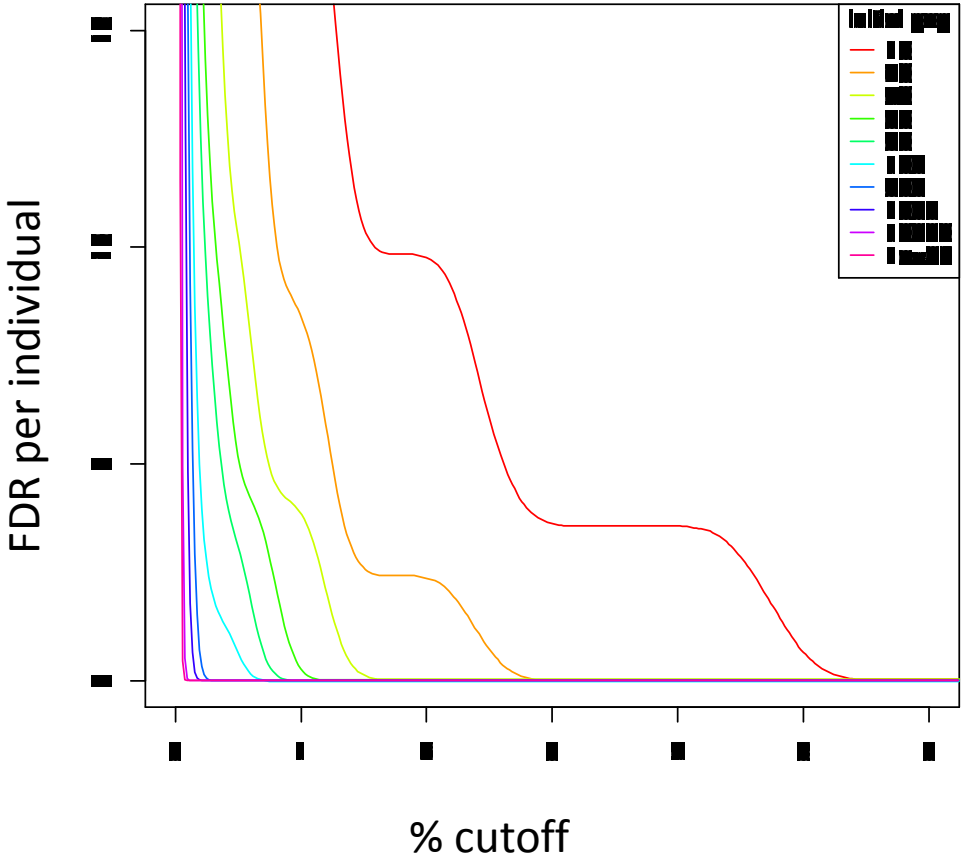
PCR simulation

- Modelled as a stochastic autocatalytic reaction with parameters from the literature
- Per base per duplication error rate fitted to give correct global error rate for particular enzyme



PCR simulation





Estimating the mismatch matrix

Dirichlet prior on rows of the mismatch error matrix Θ:

$$P(\Theta_{i:}|\alpha_{i:}) = \mathrm{Dir}(\Theta_{i:};\alpha_{i:}) \quad (1)$$

where

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

After observing counts n the posterior over Θ is

$$P(\Theta_{i:}|n_{i:}, a, b) = \mathrm{Dir}(\Theta_{i:};\alpha_{i:} + n_{i:}) \quad (3)$$

Estimating the mismatch matrix

The joint distribution over the counts n and Θ can now be expressed:

$$P(n, \mathbf{\Theta}|a, b) = P(n|\mathbf{\Theta}, a, b)P(\mathbf{\Theta}|a, b) \quad (4)$$

$$= \frac{\Gamma(a+3b)^4}{\Gamma(b)^{12}\Gamma(a)^4} \prod_i \Theta_{ii}^{n_{ii}+a-1} \prod_{j,j\neq i} \Theta_{ij}^{n_{ij}+b-1} \quad (5)$$

Marginalise over Θ

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

We fit a and b by maximising the log of this expression using a Newton scheme.

Hypothesis testing

If the reference is i, what is the probability of sequencing j at particular position m times if the total coverage is c?

$$P(n_{ij} = m|\Theta_{ij}) = \binom{c}{m}\Theta_{ij}^{m}(1-\Theta_{ij})^{c-m} \tag{1}$$

Marginally,

$$P(\Theta_{ij}|n,\alpha) = \text{Beta}(\beta_{ij}, \beta_{i.} - \beta_{ij}) \tag{2}$$

where $\beta_{ij} = n_{ij} + \alpha_{ij}$ and $\beta_{i.} = \sum_j \beta_{ij}$. Integrating over Θ:

$$P(n_{ij} = m|n,a,b) = \int P(n_{ij} = m|\Theta_{ij})P(\Theta_{ij}|n,a,b)d\Theta_{ij} \tag{3}$$

$$= \binom{c}{m}\frac{B(m+\beta_{ij}, n-m+\beta_{i0}-\beta_{ij})}{B(\beta_{ij}, \beta_{i0}-\beta_{ij})} \tag{4}$$

Hypothesis testing

Calculate a p-value for each mismatch: the probability of this event, *or any more extreme event*, under the null hypothesis; in this case $P(n_{ij} \geq m|n, a, b)$. Since we usually have $m \ll n$ it will be cheaper to calculate the p-value as follows:

$$P(n_{ij} \geq m|n, a, b) = 1 - P(n_{ij} < m|n, a, b) \quad (5)$$

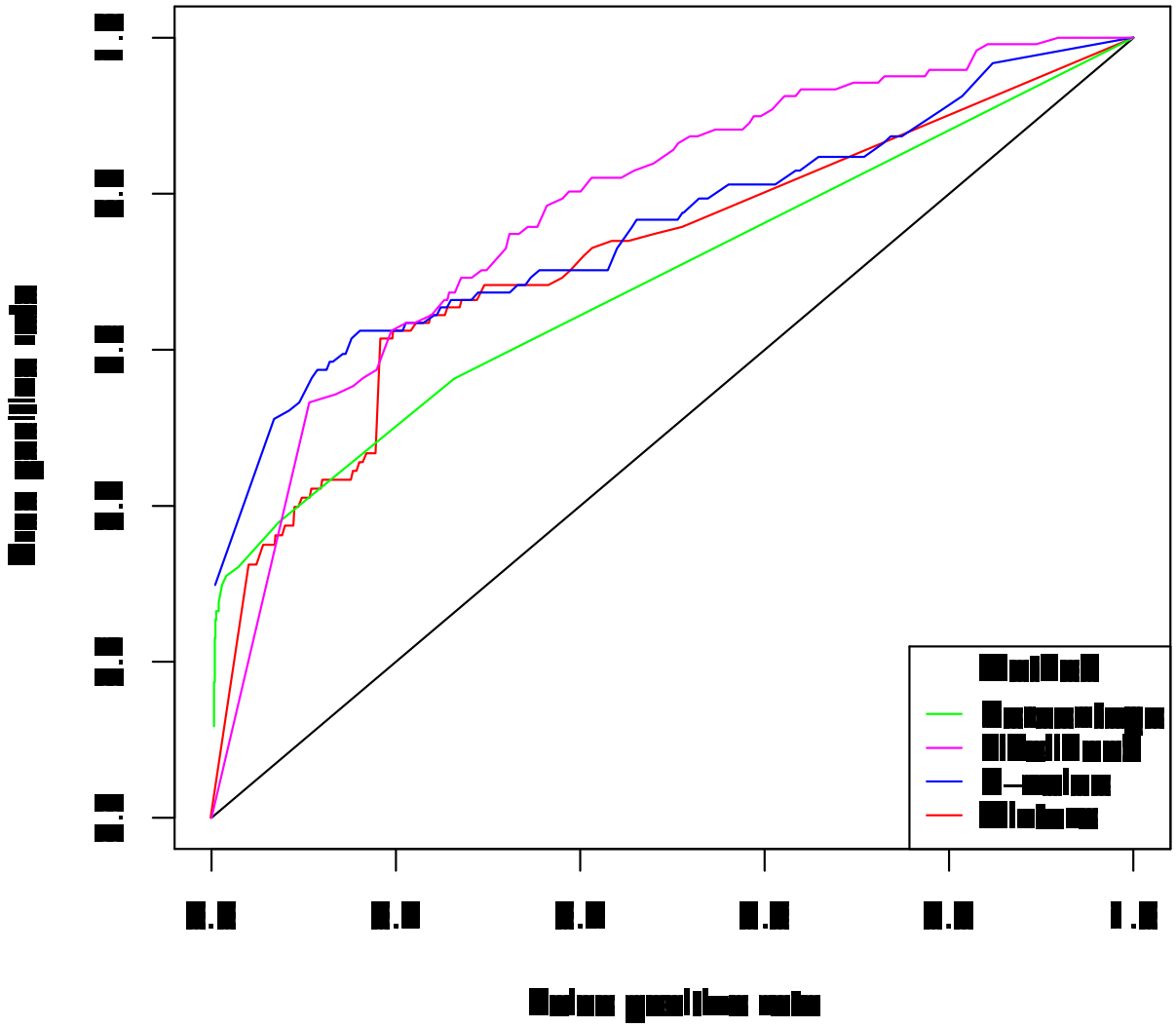
$$= 1 - \sum_{m=0}^{m-1} P(n_{ij} = m|n, a, b) \quad (6)$$

Mixture model

1. Component one – sequencing errors (fitted on plasmids)
2. Component two – mutations (codon mismatch matrix). Fitted by EM on full dataset.

Mixing components give probability of being a genuine variant.

Compare to 95 Sanger limiting dilution sequences



Conclusion

- Statistically characterised the errors involved in 454 pyrosequencing
- PCR simulations used to understand how initial copy number effects FDR
- Modelling the nucleotide mismatch matrix improves classification performance

Thanks

Professor Robert Shafer
Infectious Diseases
Stanford University

Professor Susan Holmes
Department of Statistics
Stanford University