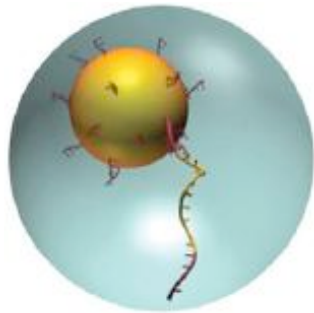


Ultra-deep sequencing of heterogeneous samples

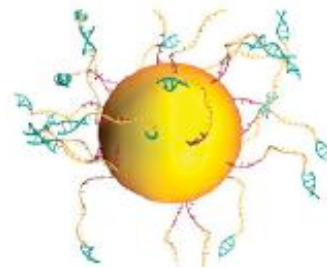
Oswaldo Zagordi
Department of Biosystems Science and
Engineering of ETH Zürich, Basel



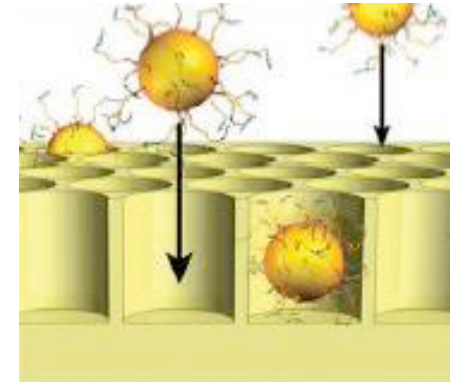
A look at the 454 sequencing technology



DNA fragments are attached to the beads

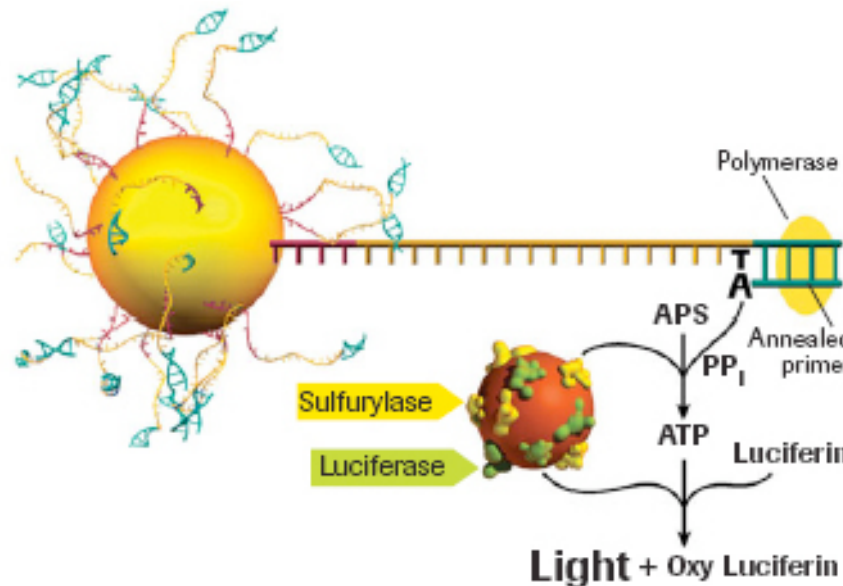


fragments are "amplified"



single beads in ~100.000 wells

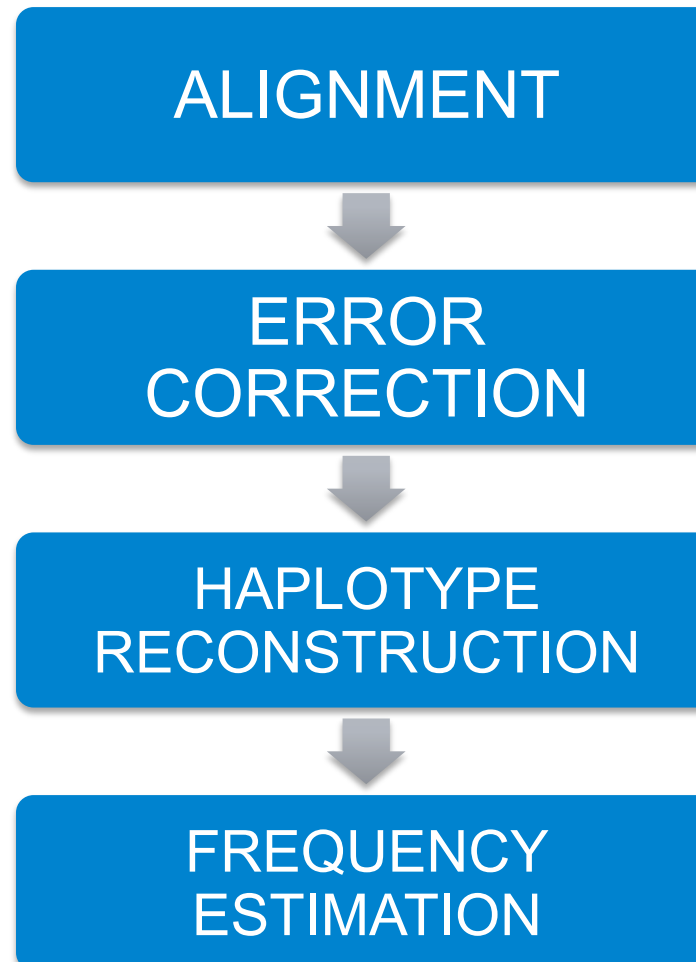
DNA is read by detecting light emission associated to base incorporation



Comparison with Sanger

	Sanger	454/Roche
bps per run	$\sim 10^5$	$\sim 10^8$
read length	700-1000	~ 400
cost per run	~ 1000 \$	~ 15000 \$
cost per Mbp	10K \$	100 \$
accuracy	high	low (in-dels)

From the reads to the haplotypes: four steps

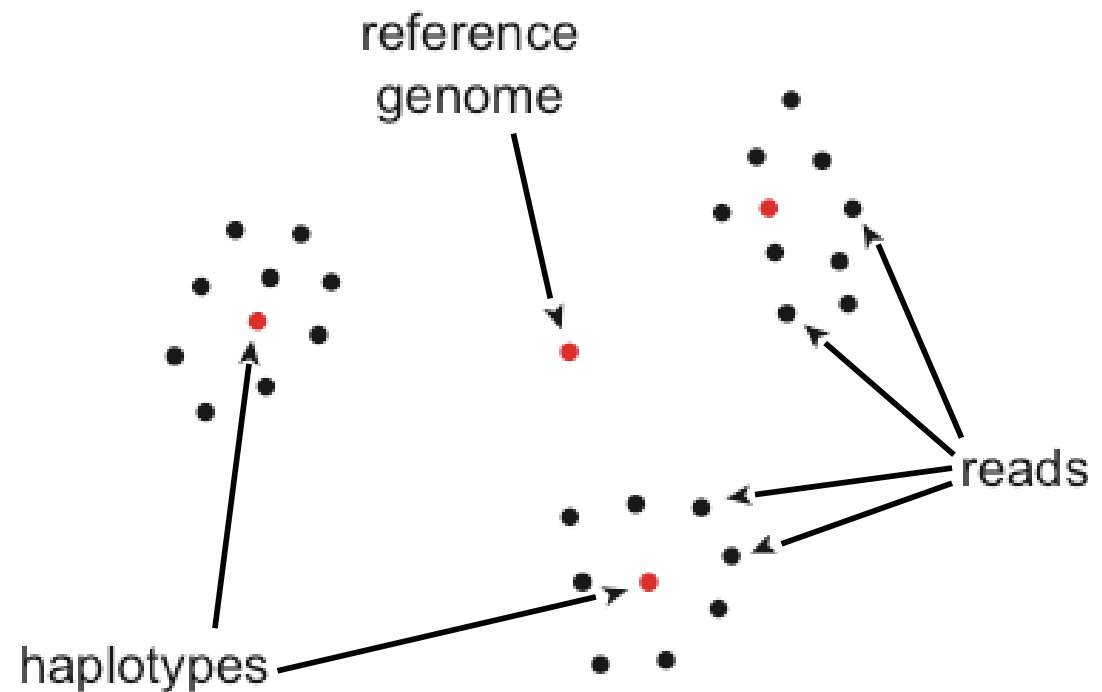


Reads from a heterogeneous sample



The computational approach to error correction

reads tend to cluster
around haplotypes
in sequence space



thousands of reads
are passed to a clustering
algorithm to separate
errors from mutations

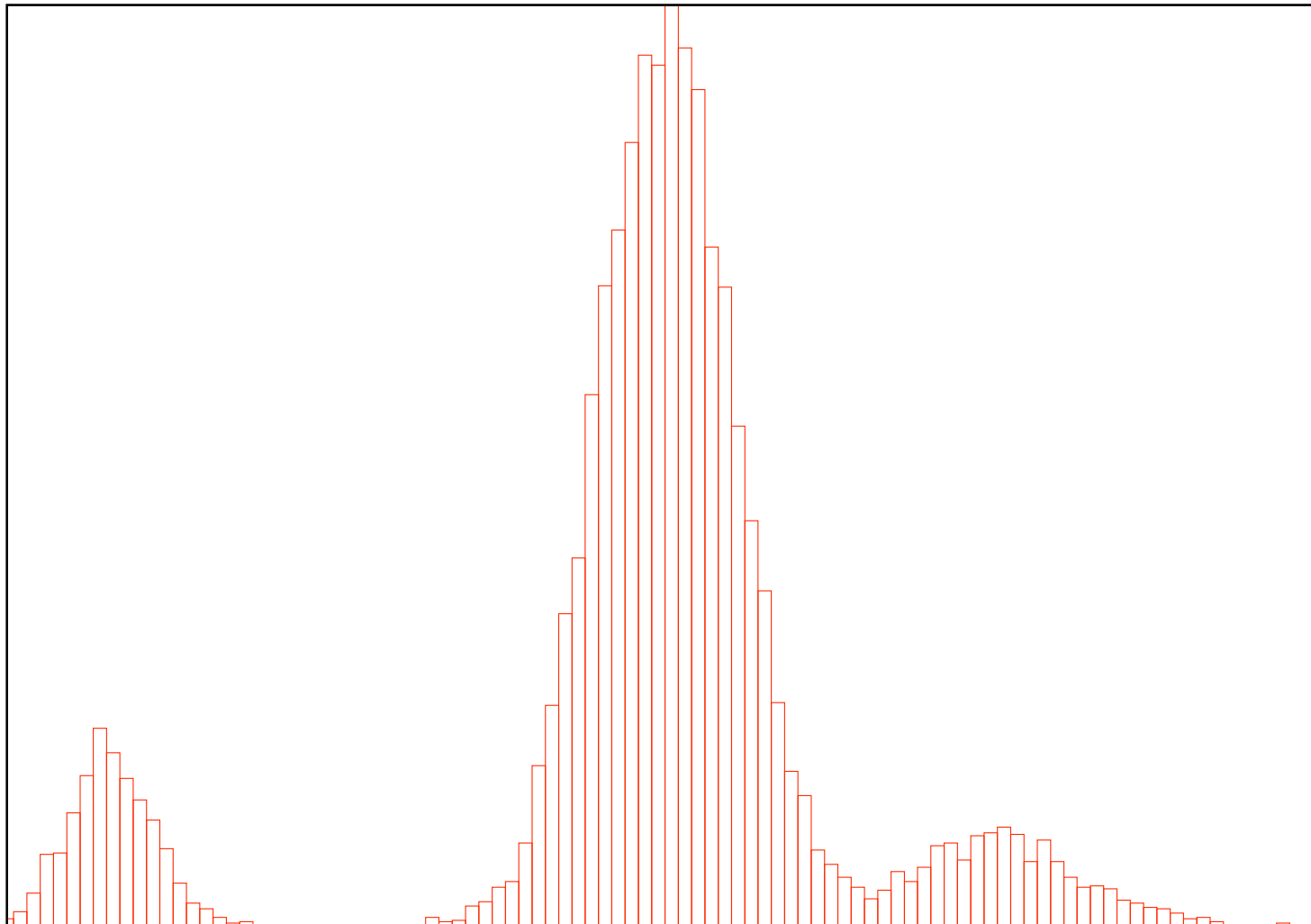
Dirichlet process mixture

- distribution as a mixture of simpler distributions
- a Dirichlet process mixture is a non-parametric prior that can capture uncertainty in the number of components of the mixture

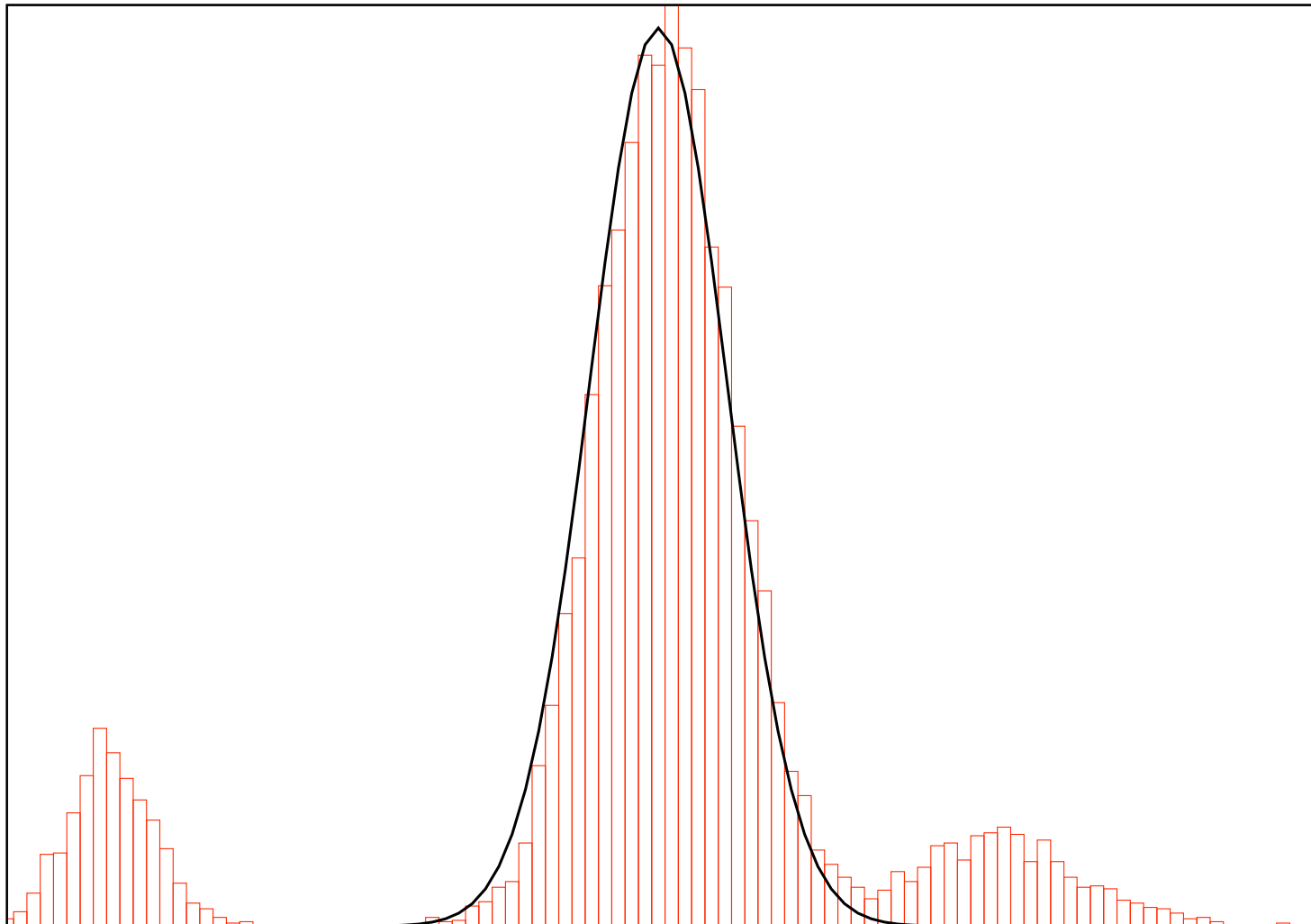
$$P(c_i = c | c_j : j \neq i) = \frac{n_{\setminus i, c}}{n - 1 + \alpha} \quad \text{if class } c \text{ is already populated}$$
$$P(c_i = c | c_j : j \neq i) = \frac{\alpha}{n - 1 + \alpha} \quad \text{if a new class is instantiated}$$

- a prior on mixing proportions that leads to few dominating classes
- just a prior on the proportions that has to be embedded in a complete model
- find the model that better explains the data

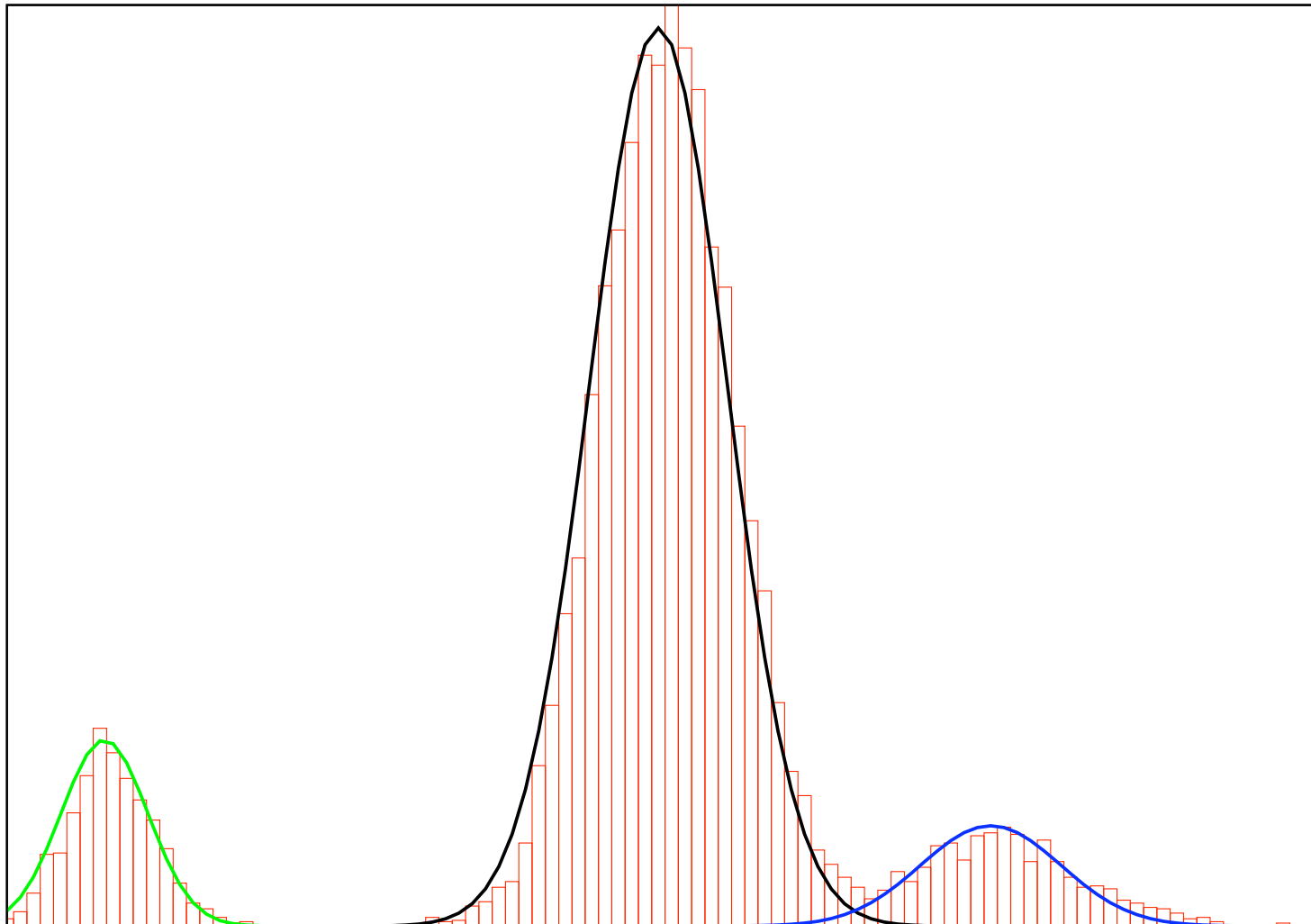
But how?



But how?



But how?



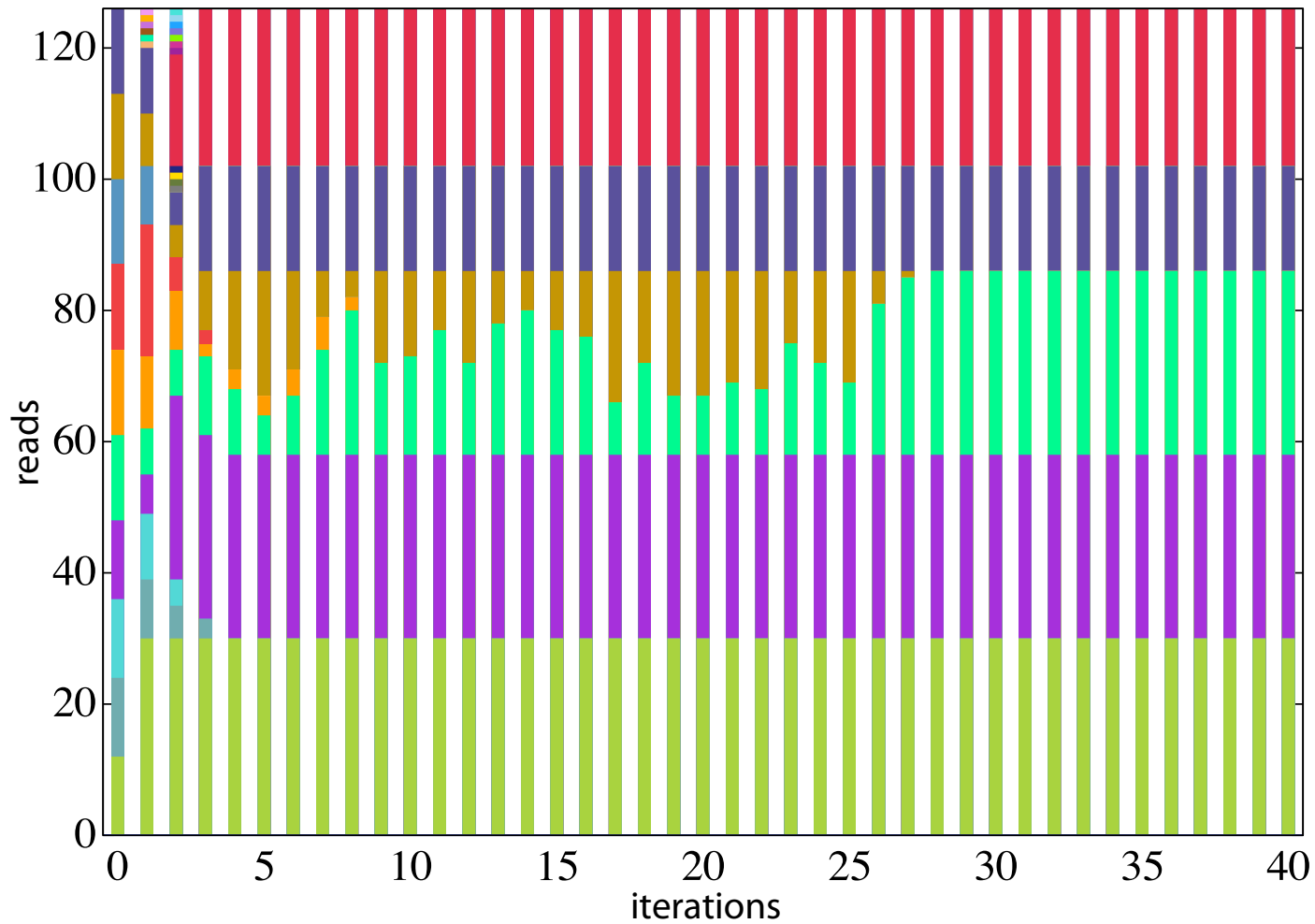
A generative model for the reads from a heterogeneous sample

- given a haplotype, the read shows the same base with probability θ
- base at each position depends only on the corresponding base on the haplotype

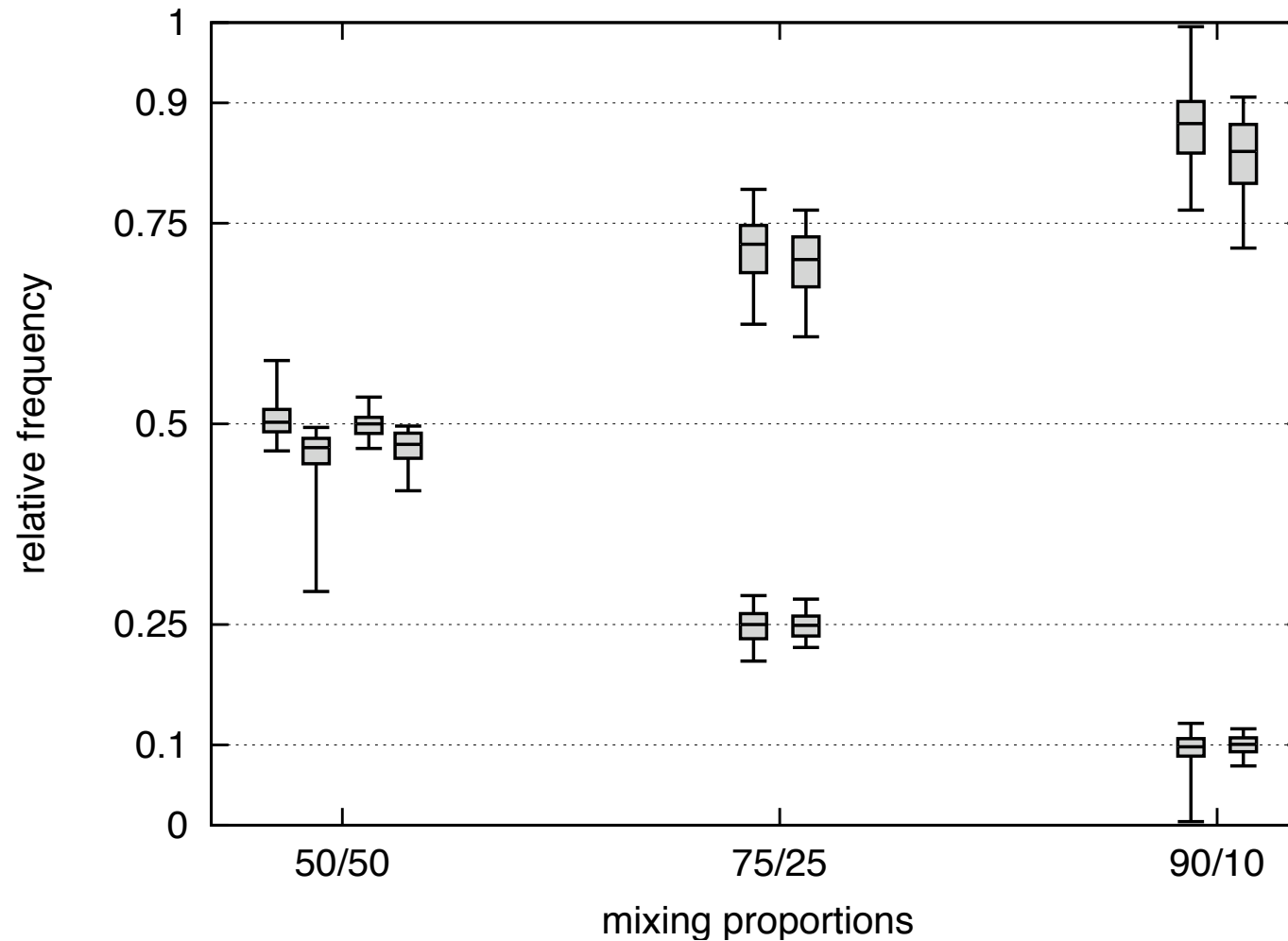
$$p(r_i | c_i = k, h_k, \theta) = \theta^{m_{i,k}} \left(\frac{1 - \theta}{|B| - 1} \right)^{m'_{i,k}}$$

- reads are drawn from haplotypes
- haplotypes are drawn from a reference (not necessarily the wild type)

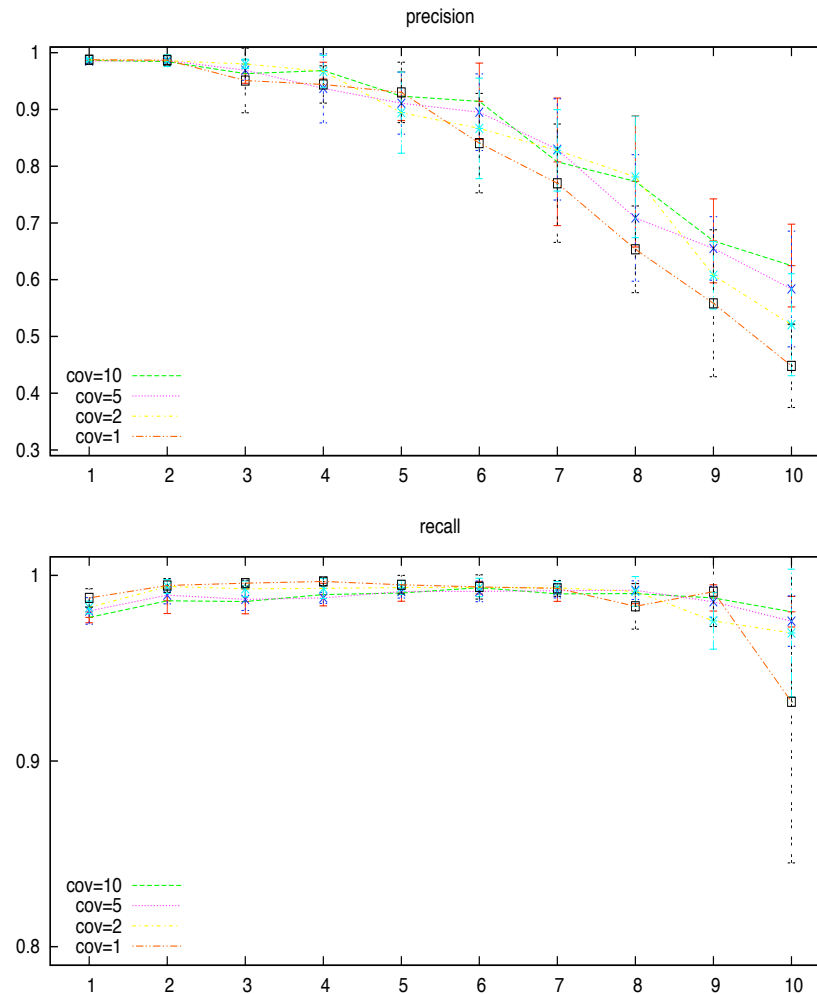
Assignment of the reads in a run



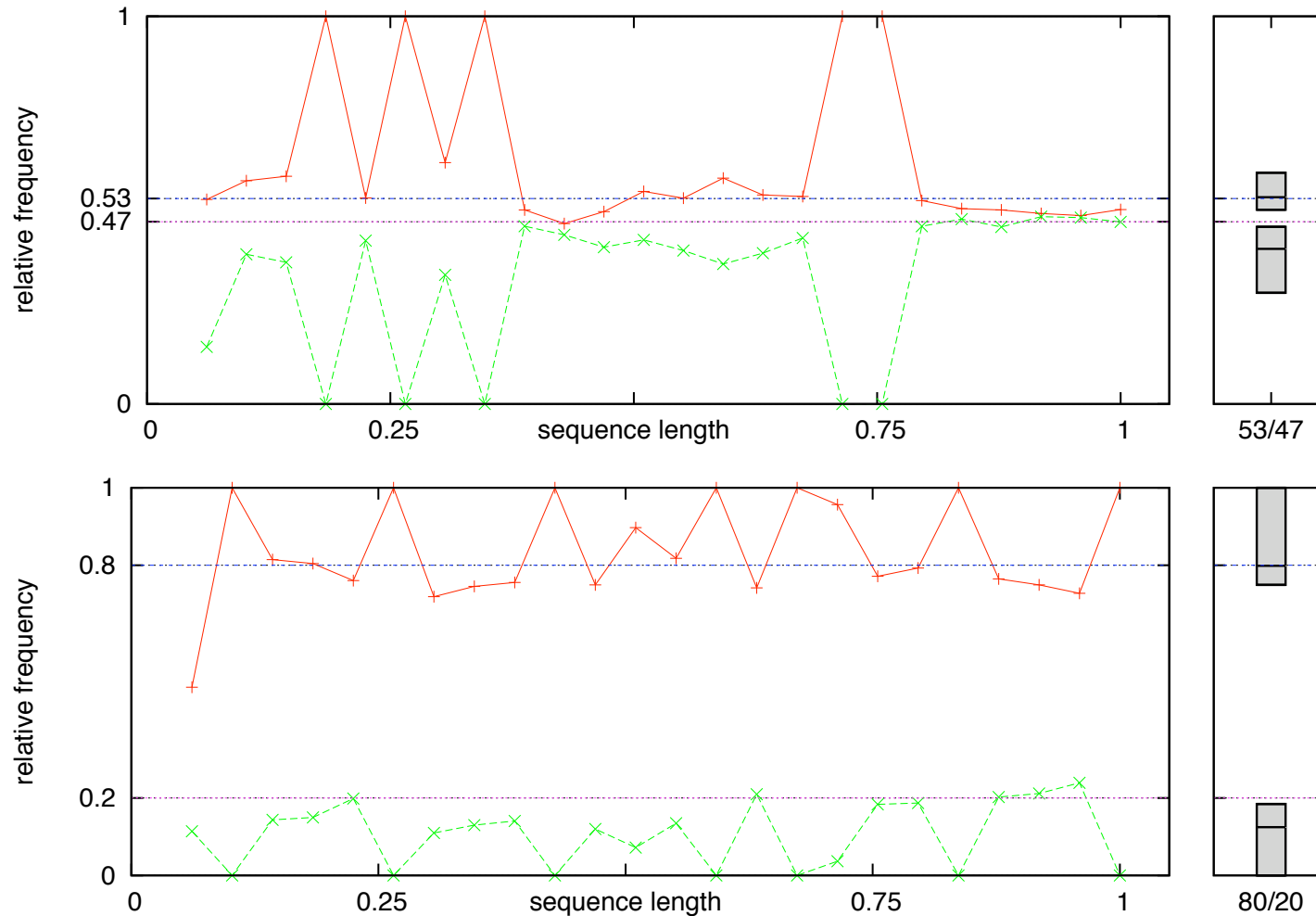
2 haplotypes mixed: 3% and 10 % distance



10 haplotypes at mutual distance 5%, mixed at frequency $\frac{1}{2}$, $\frac{1}{4}$, ...



Mixing reads from 454 runs: two different HIV subtypes



References

- Software implementation ShoRAH
available at www.cbg.bsse.ethz.ch
- Eriksson et al. (without DPM)
Viral population estimation using pyrosequencing.
PLoS Comp Biology (2008)
- Zagordi et al.
Deep sequencing of a genetically heterogeneous sample.
submitted

To do

- applications (HIV drug resistance, cancer)
- comparing 454 and Illumina technologies
- extending the model to paired end reads
- developing a global model

To do

- applications (HIV drug resistance, cancer)
- comparing 454 and Illumina technologies
- extending the model to paired end reads
- developing a global model

To thank

- Niko Beerenwinkel @ BSSE
- Lukas Geyrhofer @ BSSE
- Nicholas Eriksson @ 23andme
- Volker Roth @ UniBasel
- Martin Daeumer @ Immungenetik
- Deep sequencing lab @ BSSE