

ETH Zürich  
D-BSSE

# Ultra-deep sequencing of genetically heterogeneous samples

Niko Beerenwinkel<sup>1</sup>, Nicholas Eriksson<sup>2</sup>, Volker Roth<sup>3</sup>, Osvaldo Zagordi<sup>1</sup>

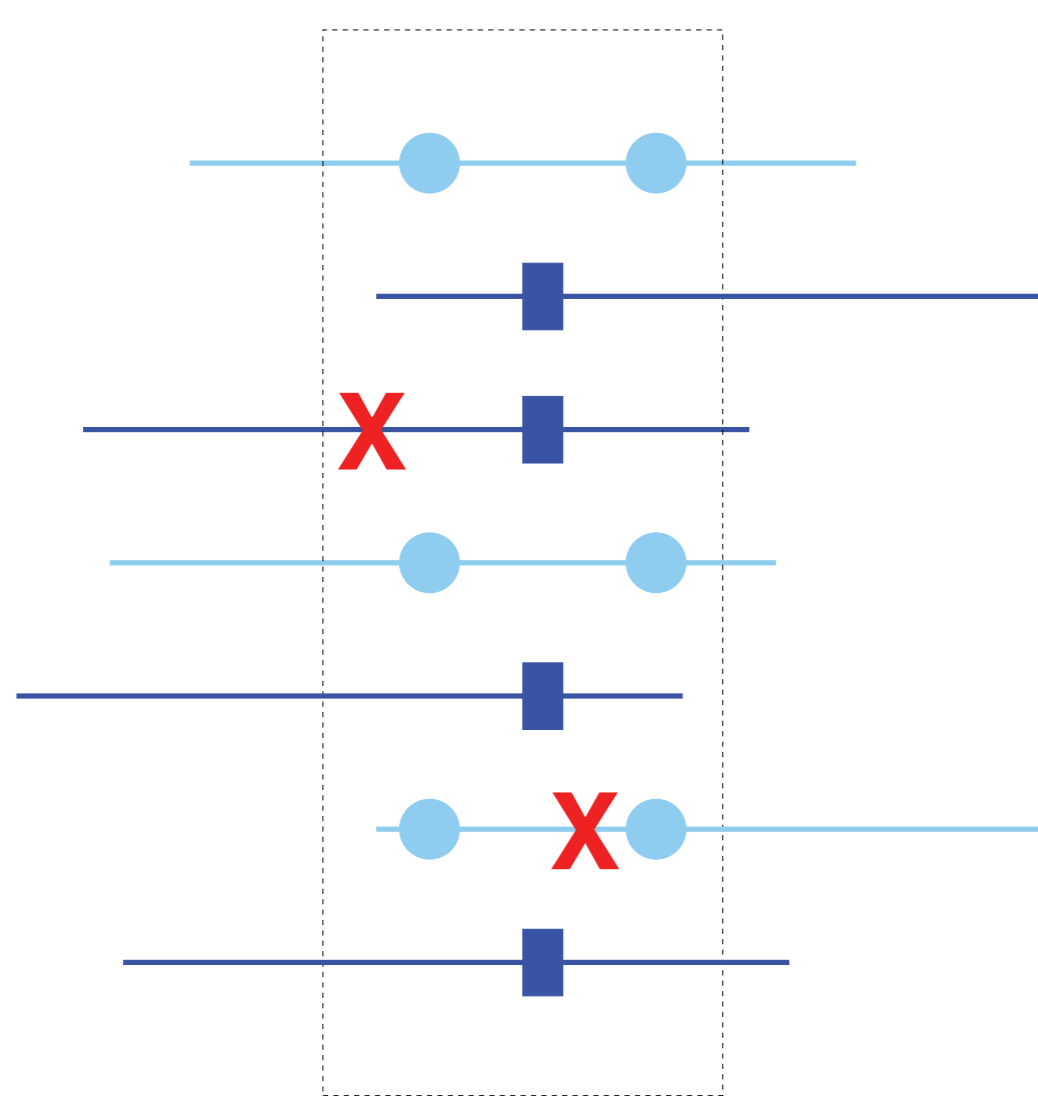
<sup>1</sup> ETH Zürich Department of Biosystems Science and Engineering, Basel, Switzerland

<sup>2</sup> Department of Statistics, University of Chicago, IL, USA

<sup>3</sup> Departement Informatik, University of Basel, Switzerland

## Introduction

Ultra-deep sequencing is a family of new methodologies that, unlike traditional Sanger sequencing, typically give many short error-prone reads. While it is now a commercially available technology for sequencing genetically homogeneous samples, the possibility to use it as a tool to estimate the population variation in a heterogeneous sample is a subject of active research.



Reads from different haplotypes showing both genetic variations (circles and squares) and errors (crosses)

## Computational approach

We propose a methodology to infer the different genomes present in the population (haplotypes) and to estimate their frequencies, which consists in the following four steps:

1. **Alignment:** the existence of a reference genome to which the set of reads can be aligned is assumed;
2. **Error correction:** one can distinguish technical errors from biologically relevant mutations gaining information from multiple reads (the technology is characterized by a high coverage, that compensates for the errors);
3. **Haplotype reconstruction:** reconstructing the smallest pool of haplotypes consistent with the observations;
4. **Haplotype frequency estimation:** finally, one has to infer the population structure, i.e. the probability distribution on the set of haplotypes.

## Error correction via local clustering

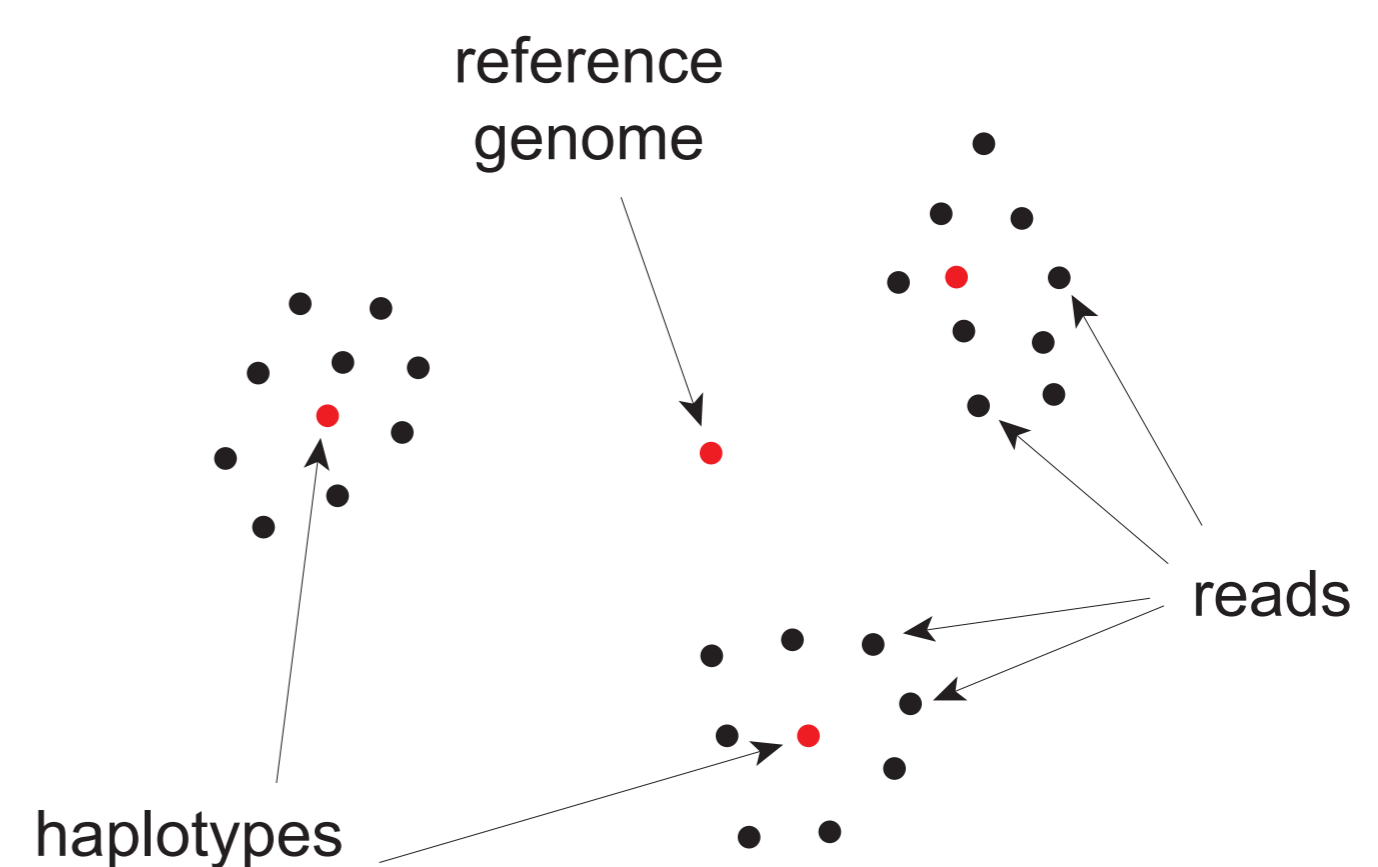
We want to address here the error correction step. The key point is how to consider the differences between the reads and the reference sequence. The procedure adopted in [1] for local error correction consists in a statistical test to discern technical errors from biological variation. If there are more variations than expected under a single haplotype hypothesis, a new one is considered to be present. Once that reads have been clustered (with a general purpose algorithm), the consensus sequence for each cluster (haplotype) is computed.



In this example a set of reads, drawn from three haplotypes with errors, are aligned. Then, in the window, one tries to correct the errors and distinguish the haplotypes.

## Local clustering via Dirichlet process

A different technique pursued by us to achieve error correction consists in clustering reads applying a probabilistic Dirichlet process mixture (DPM). This approach allows us to define a probability distribution starting from the basic features of the system under investigation, and to control the complexity of the model with a single parameter  $\alpha$  that controls the creation of new clusters.



Haplotypes represented by clusters of reads in sequence space. A reference genome from which haplotypes are derived is used to restrain the region of interest in the huge high-dimensional space.

## Application

We are testing the technique described above on both simulated and real data, as for example read sets obtained from sequencing genetically diverse HIV samples derived from several infected patients under antiretroviral therapy.

## References

[1] Nicholas Eriksson, Lior Pachter, Yumi Mitsuya, Soo-Yon Rhee, Chunlin Wang, Baback Gharizadeh, Mostafa Ronaghi, Robert W. Shafer, Niko Beerenwinkel.

Viral population estimation using pyrosequencing.  
arXiv.org:0707.0114; to appear in *Plos Comp Biol*