The method for homologous recombination detection within bacterial species

Anastasia S. Kalinina[1], Alexandra L. Suvorikova[1,2,4],
Vladimir G. Spokoiny[1,3,4], Mikhail S. Gelfand[1,5]

[1]A.A. Kharkevich Institute for Information Transmission Problems, RAS, Moscow, Russia; [2]IRTG 1792; [3]WIAS; [4]PreMoLab; [5]M.V. Lomonosov Moscow State University, Department of Bioengineering and Bioinformatics, Moscow, Russia, as.kalinina@gmail.com

In the absence of sexual reproduction in bacteria, transfer of DNA fragments from cell to cell and introducing these fragments into a genome by homologous recombination plays a major role in spreading beneficial mutations through the population. During the last decade a large number of completely sequenced bacterial genomes became available, yielding a need in methods for detection of homologous recombination which could deal with tens and hundreds of genomes. Sophisticated Bayesian approaches, such as ClonalFrame [1,2], produce accurate and reliable results, but are very time-consuming. Gubbins [3] and BratNextGen [4] packages are also accurate, but much faster on large datasets. The former uses spatial scan statistics to detect loci with high local density of nucleotide substitutions, which are potential recombination sites. The former is based on the Bayesian change-point clustering model with predefined clusters of genomes. Unlike Gubbins, BratNextGen also predicts sources of recombination, but is reported to be conservative [4, 5].

We represent here a rapid and transparent pipeline for detection of homologous recombination events in closely related bacterial strains. As in Gubbins, loci with high local density of nucleotide substitutions are supposed to be recombined. Instead of spatial scan statistics, we use Adaptive Weight Smoothing (AWS) [6] for segmentation. Since probability of homologous recombination of DNA fragments from donor and recipient cells decreases with increasing sequence diversity [7], most recombination events occur within bacterial species. So, we focused on detection recombination events, where donor and recipient genomes belong to same species, but to different phylotypes.

The first step of the algorithm is to convert a multiple alignment to a Bernoulli sequence of “0” and “1” for each strain. We used here an empirical rule where a position in alignment is significant and turns to “1”, the considered strain differs if in this position from more than

half of members of its phylotype and coincides with more than half of members of another phylotype. A high local density of "1" in a sequence indicates that a haplotype, which is characteristic for the other phylotype, was introduced to a considered genome by homologous recombination. Then, AWS is applied to the resulting Bernoulli sequence to estimate the local density of "1" in each position. The threshold on the density, above which a segment is considered recombined, is tuned manually, since it depends on the diversity of considered strains. Description of this procedure for *Escheriachia coli* as an example is provided below.

To check the reasonableness of our method, we applied it to the dataset with simulated recombination events and obtained precision 94%, recall 87% and F1-measure 91%. Then, we compare AWS with with BratNextGen on another artificial dataset and demonstrated that the precision and recall are higher for AWS.

AWS was applied to 20 *E. coli* genomes from the phylogroups A, B1, B2 and E. It has been shown in [8] that the shape of the distribution of the number of differences in non-overlapping fixed-size windows in a pairwise comparison of *E. coli* genomes depends only on the genetic distance between the considered strains. Two regimes with clear change-point at 3-5 differences per 1000 bp are observed in the histograms: for vertically inherited segments with a low number of differences and for recombined segments with a high number of differences. The average genetic distances between phylogroup members are 6-8 differences per 1000 bp. Since these values are the average of recombined and vertically inherited segments, the threshold on the differences density should be lower than 0.006. Hence we set the threshold to 0.005 to minimize the number of false positive recombined segments.

The estimates of recombination fluxes between phylogroups and parameters of recombination are similar to those in [9]. To obtain estimates for the fraction of recombined segments with an unknown source, we applied AWS to each pair of genomes within phylogroup (position was considered significant if the strains differ in this position). Approximately 30% of potentially recombined segments in pairwise comparisons in phylogroups A and B1 could not be explained by the homologous recombination with a source in another phylogroup. This

estimate is robust and does not depend on genetic distances between strains in considered pairs.

1. X.Didelot et al. (2007) Inference of bacterial microevolution using multilocus sequence data, *Genetics,* **175**:1251-66.

2. X.Didelot et al. (2010) Inference of homologous recombination in bacteria using whole-genome sequences, *Genetics,* **186**:1435-49.

3. N.J.Croucher et al. (2015) Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins, *Nucleic Acids Res,* **43**:e15.

4. P.Marttinen et al. (2012) Detection of recombination events in bacterial genomes from large population samples, *Nucleic Acids Res,* **40**:e6.

5. M.de Been et al. (2013) Recent recombination events in the core genome are associated with adaptive evolution in Enterococcus faecium, *Genome Biol Evol,* **5**:1524-35.

6. J.Polzehl et al. (2000) Adaptive Weights Smoothing with applications to image restoration, *J R Stat Soc Ser B Stat Methodol,* **62**:335-354.

7. J.Majewski et al. (1999) DNA sequence similarity requirements for interspecific recombination in Bacillus, *Genetics,* **153**:1525-33.

8. P.D.Dixit et al. (2014) Quantifying evolutionary dynamics of the basic genome of E. coli. *arXiv*:1405.2548 [q-bio.PE].

9. X.Didelot et al. (2012) Impact of homologous and non-homologous recombination in the genomic evolution of Escherichia coli, *BMC Genomics,* **13**:256.