Topological data analysis of spotting-probes pointclouds

Silvano Garnerone

December 11, 2017

Given a point cloud derived from three-dimensional probes locations we want to analyze the shape of data, for the purpose of studying statistics of chromosomal shapes in a given cell population, comparing different chromosomes and identifying recurrent topological patterns. Topological Data Analysis (TDA) is a set of method allowing us to quantify the shape of the data in a rigorous manner. As Fig.1 shows the idea is to expand spheres centered at each point, and while doing this to keep track of the number of connected components (b_0), holes (b_1) and voids (b_2, not shown in the figure) that emerge as a result of the expanding spheres radious (ϵ). A summary of this information is provided by persistent diagrams (top part of Fig.2) and persistent barcodes (bottom part of Fig.2): One can then compare different topological summaries using the bottleneck and the Wasserstein distances.

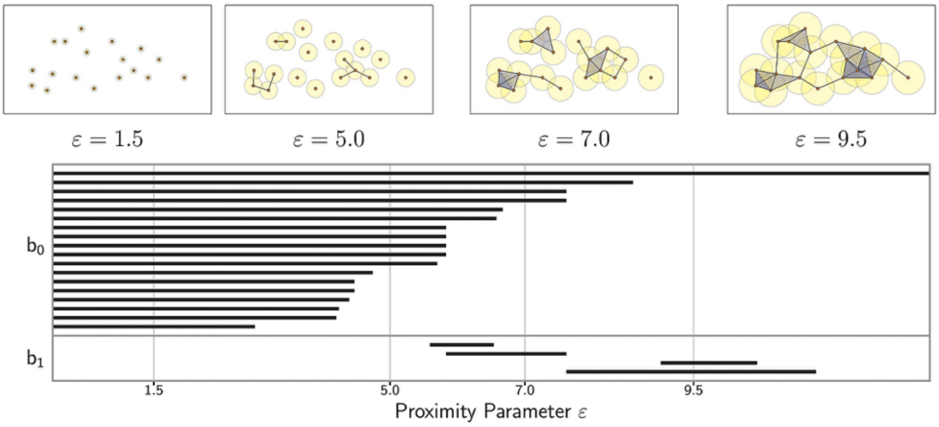

Figure 1: Example of persisten barcode.

1 Example with real data

Let's consider the following dataset (iEG408_003_d0_a594014_006_1.csv): applying TDA we get the following diagrams summarizing the start (birth) and end (death) of clusters (in black), loops (in red) and voids (in blue):

Now let's consider a second dataset (iEG408_003_d0_a594014_006_2.csv), the other homologue in the pair observed in the same nucleus. From this one we have the following plots:

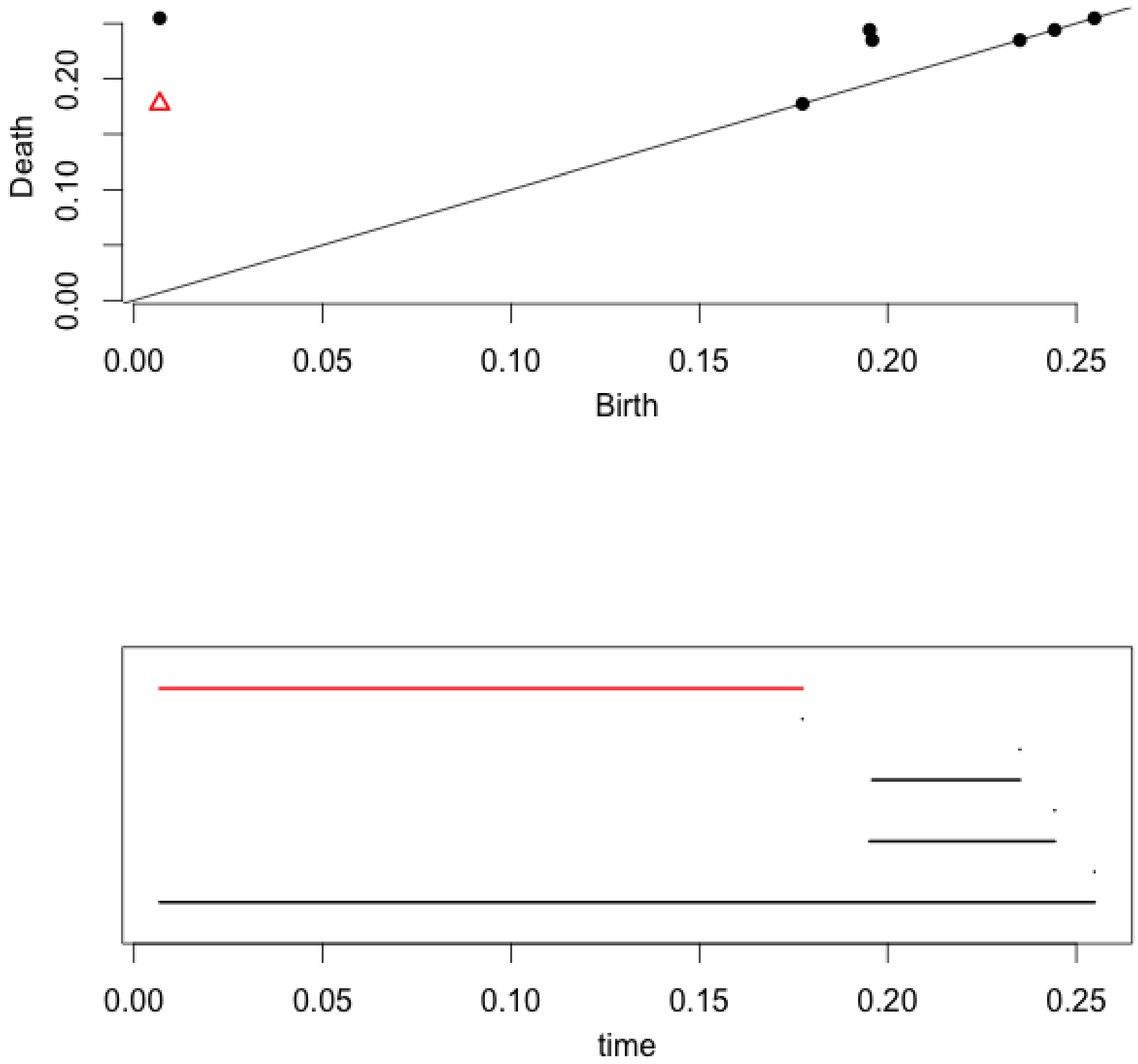

Figure 2: Example of corresponding persistent diagram and persisten barcode, containing the same information.

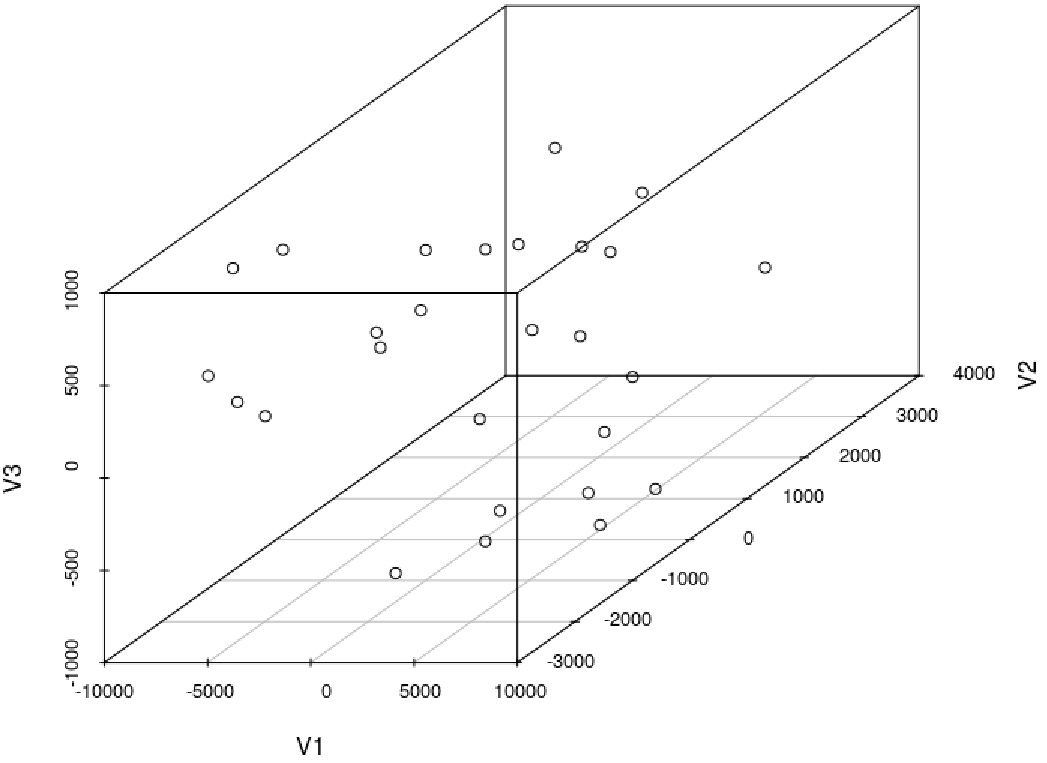

Figure 3: Example of point cloud for chr20, first homologue.

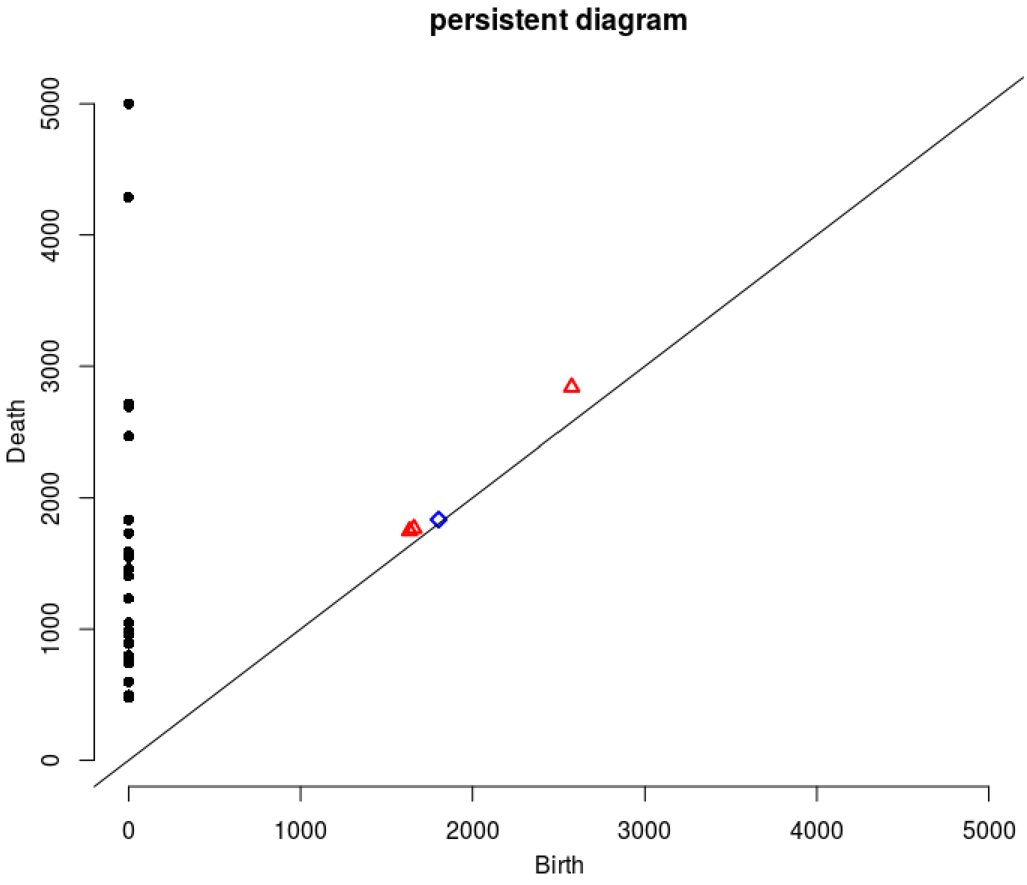

Figure 4: Example of persistent diagram for chr20, first homologue.

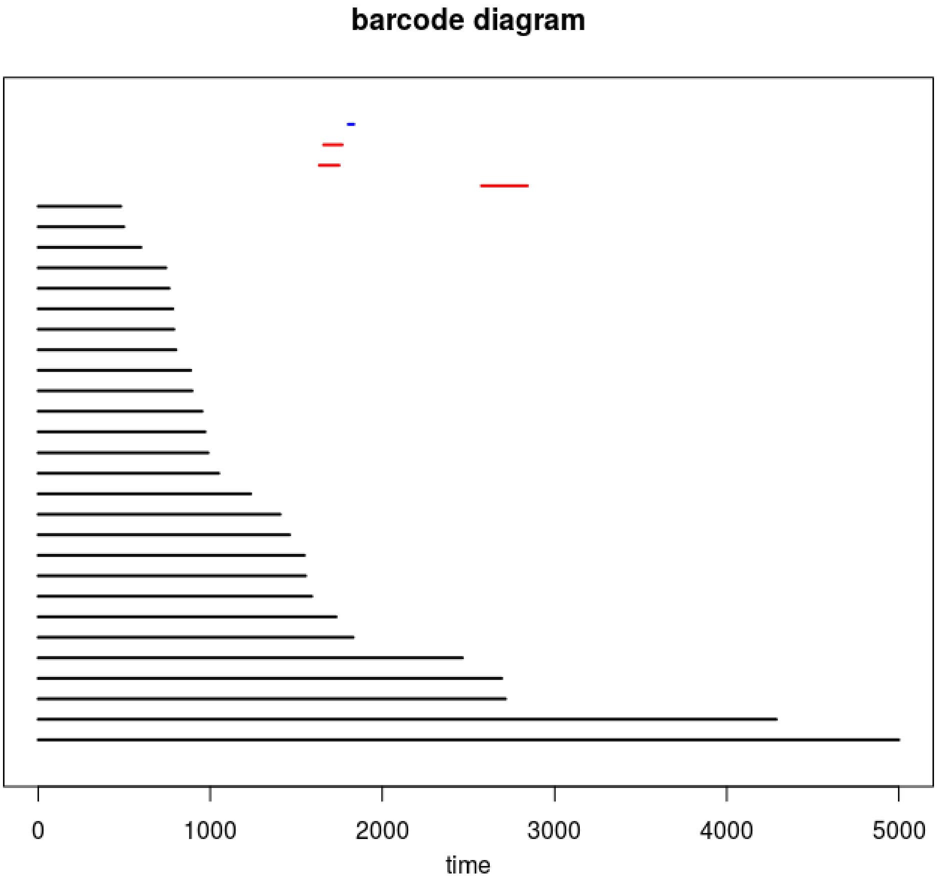

Figure 5: Example of barcode for chr20, first homologue.

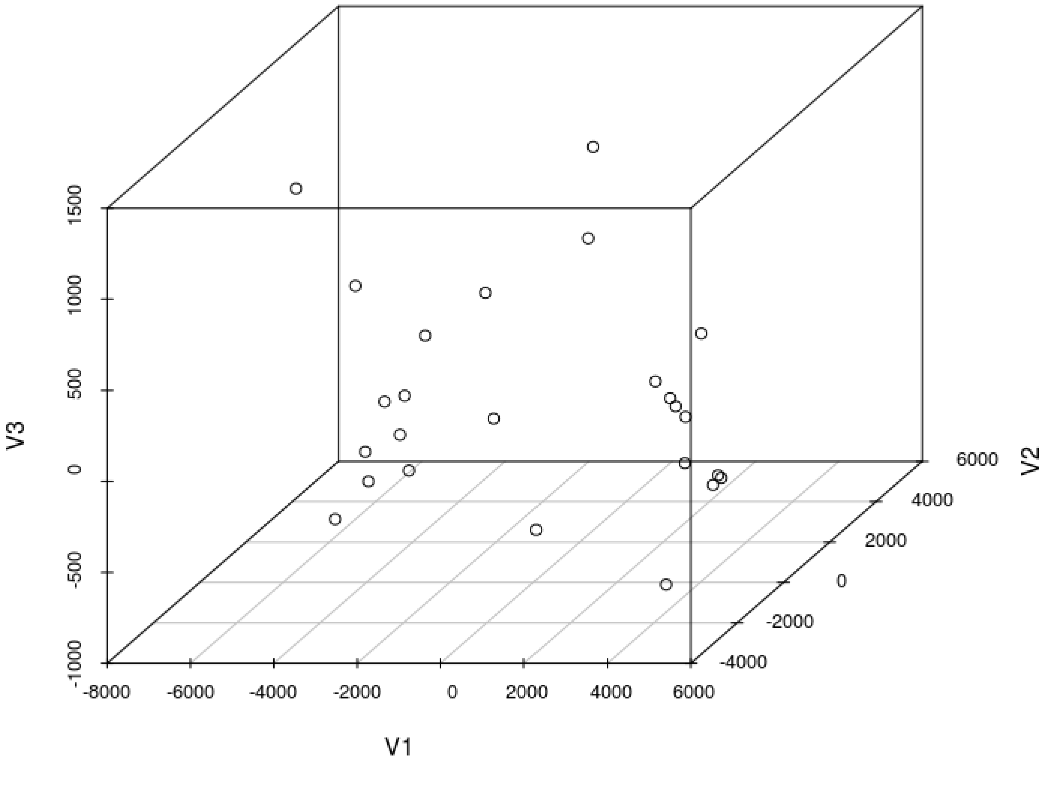

Figure 6: Example of point cloud for chr20, second homologue.

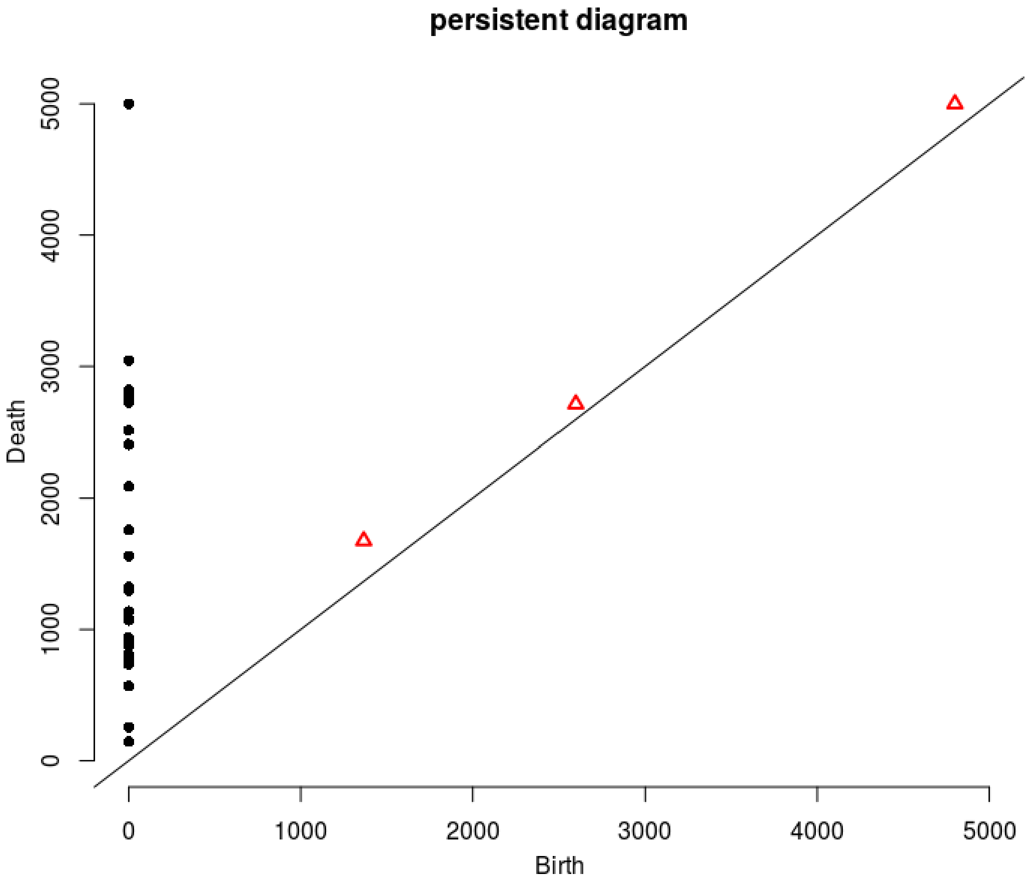

Figure 7: Example of persistent diagram for chr20, second homologue.

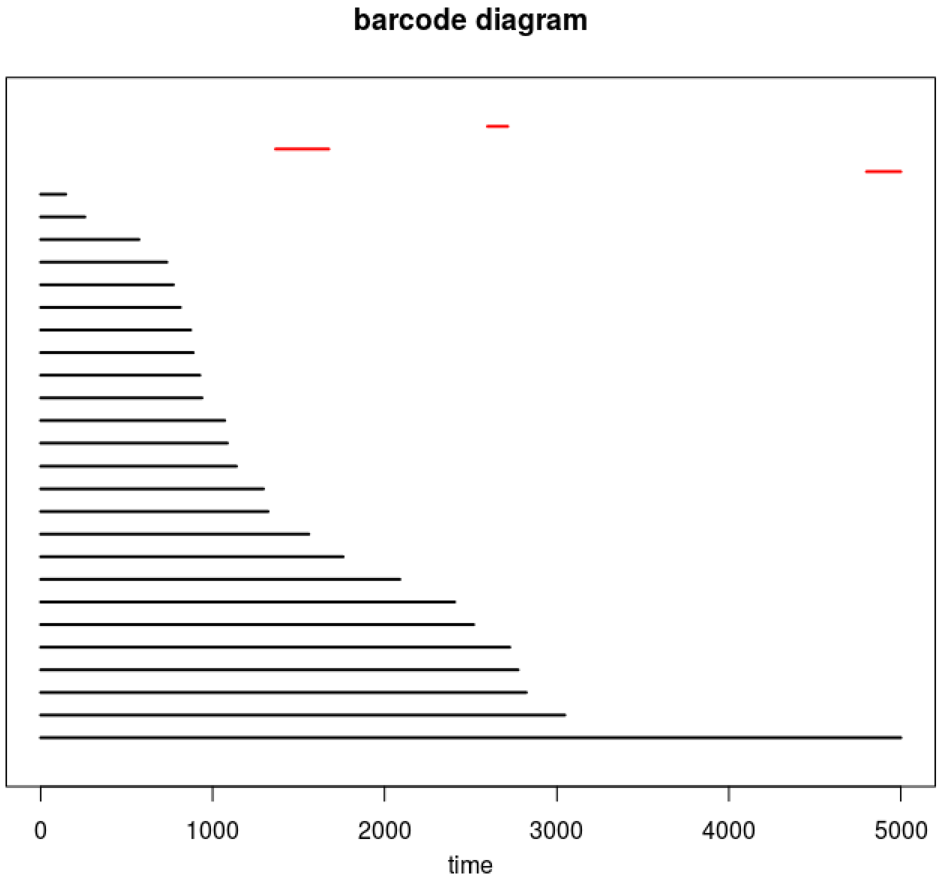

Figure 8: Example of barcode for chr20, second homologue.

Having a pair of point clouds it is also possible to quantify the similarity between their topological features, and quantify the fluctuations present in the topological summary of the point-clouds.

2 Statistics of distances between topological summaries

Given two point clouds, for each we evaluate the persistent diagram and then we take the p-Wassertein distance (with p=2) between the two diagrams, divided by the number of points in the larger cloud of the pair. The distance is a function D depending on chr#1, chr#2, and the dimension of the homological features (i.e. clusters, loops or voids). To start we consider the case where chr#1=chr#2=chromosome 1 (dataset IEG364-004 with 16 clouds with at least 20 points) and we plot the distribution of distances between all point clouds with at least 20 points (see Fig.9,10). We then consider chromosome 20 (dataset iEG408-003-d0-a594 with 28 clouds with at least 20 points, see Fig.11,12), and chromosome 2 (dataset iEG408-003-d0-cy5 with 69 clouds with at least 20 points, see Fig.13,14).

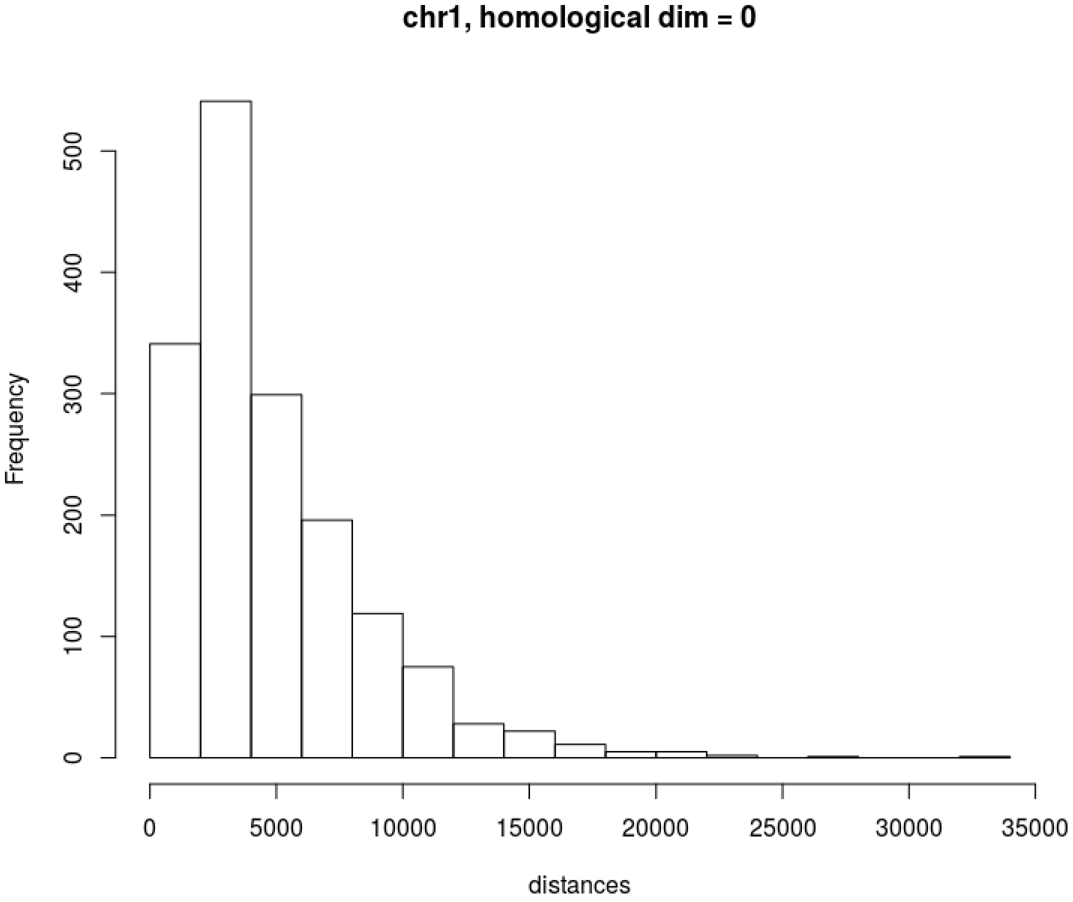

Figure 9: Intra-chromosomal distances for all pair of point clouds, considering only the 0-th homological dimension.

Fig.15 shows a violin plot summarizing the statistics for dim = 0 (clusters), while Fig.16 shows the same for 1-dim features.

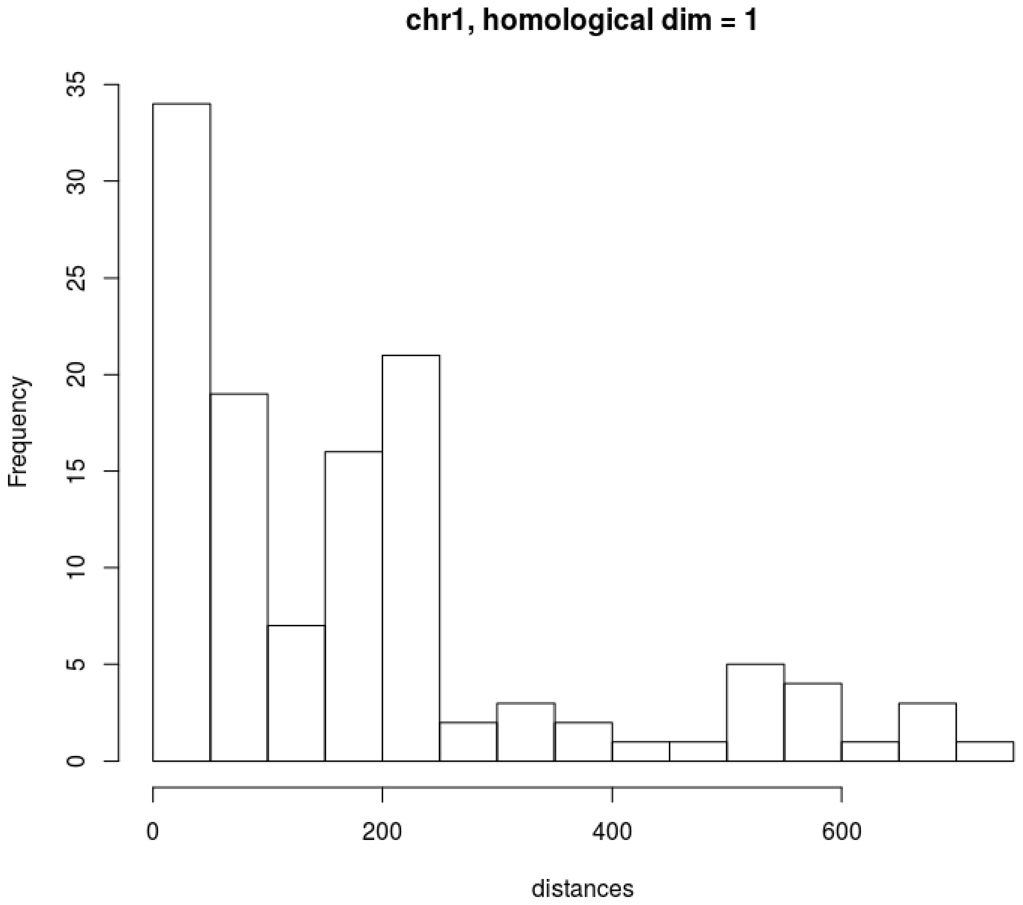

Figure 10: Intra-chromosomal distances for all pair of point clouds, considering only the 1-st homological dimension.

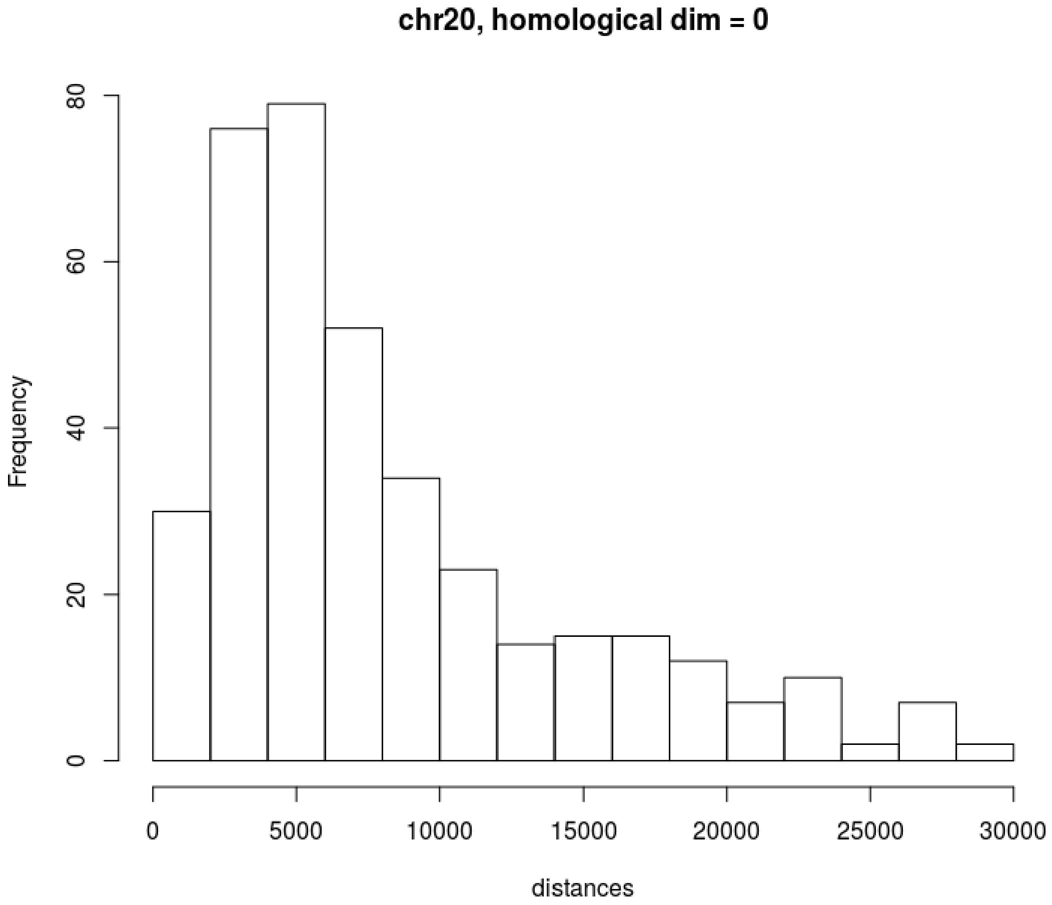

Figure 11: Intra-chomosomal distances in the 0-th homological dimension space.

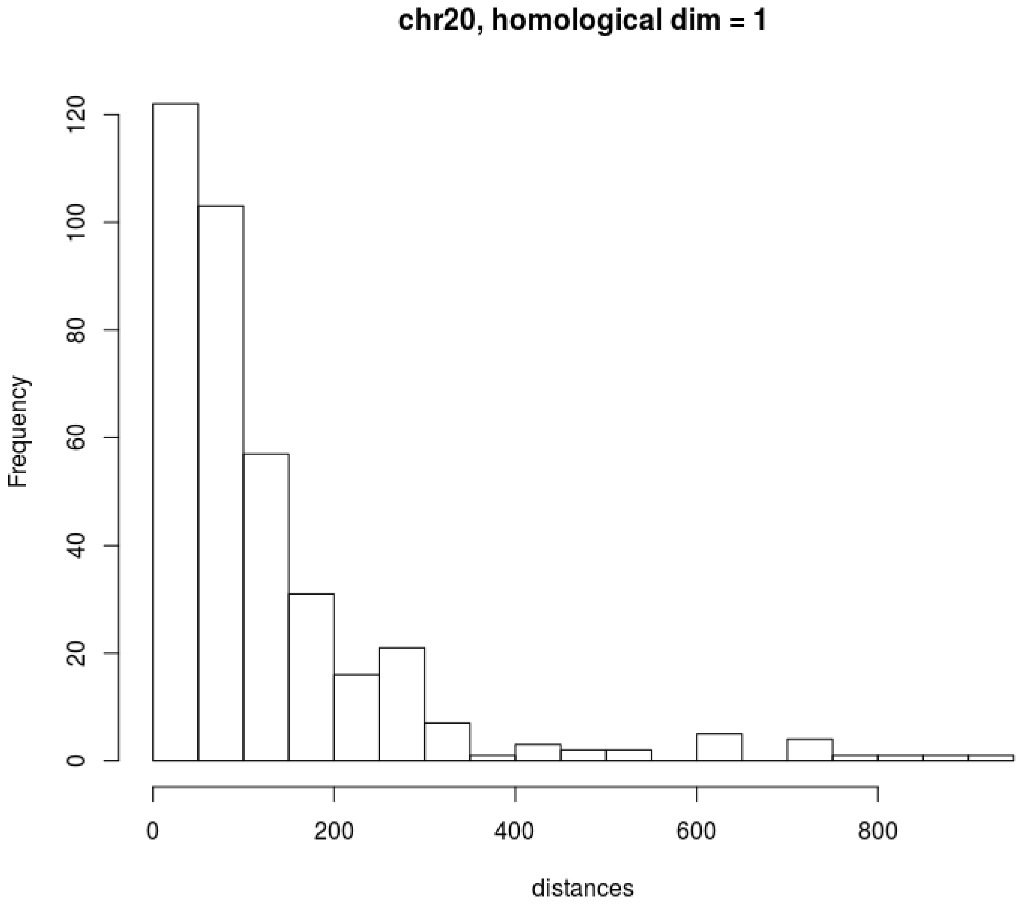

Figure 12: Intra-chomosomal distances in the 1-st homological dimension space.

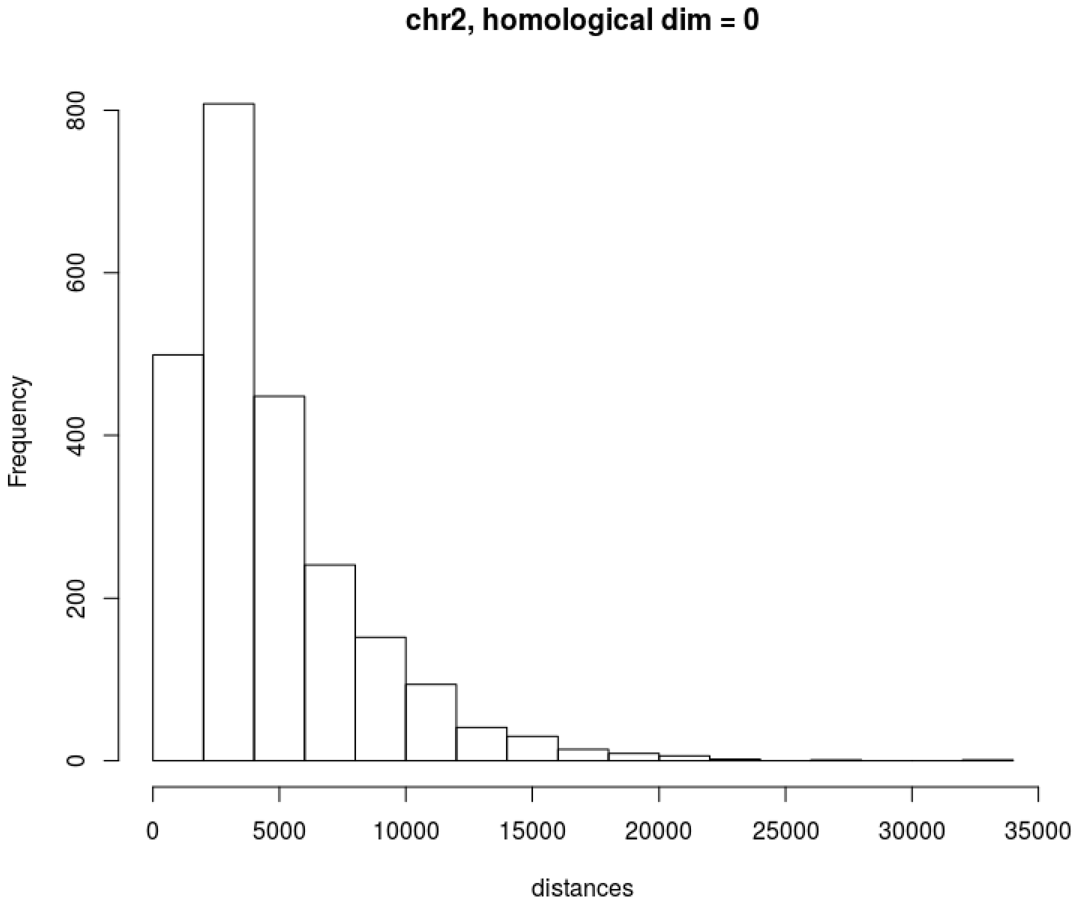

Figure 13: Intra-chomosomal distances in the 0-th homological dimension space.

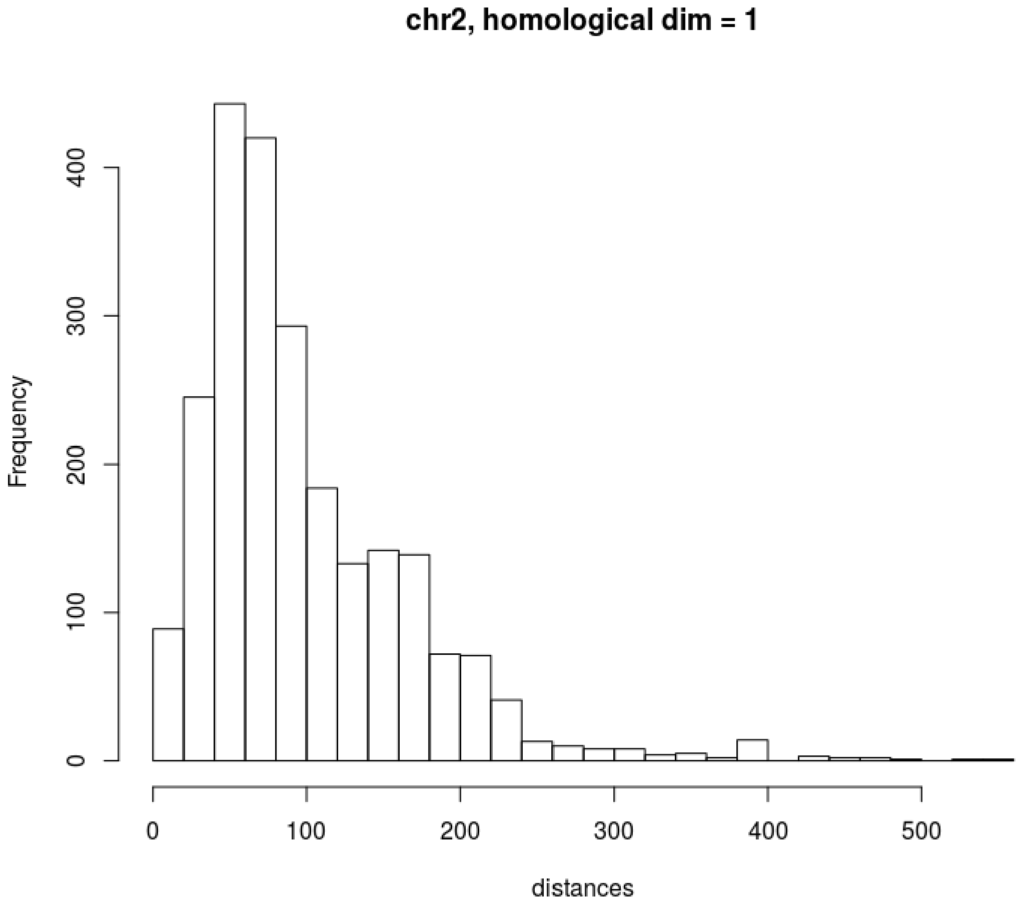

Figure 14: Intra-chomosomal distances in the 1-st homological dimension space.

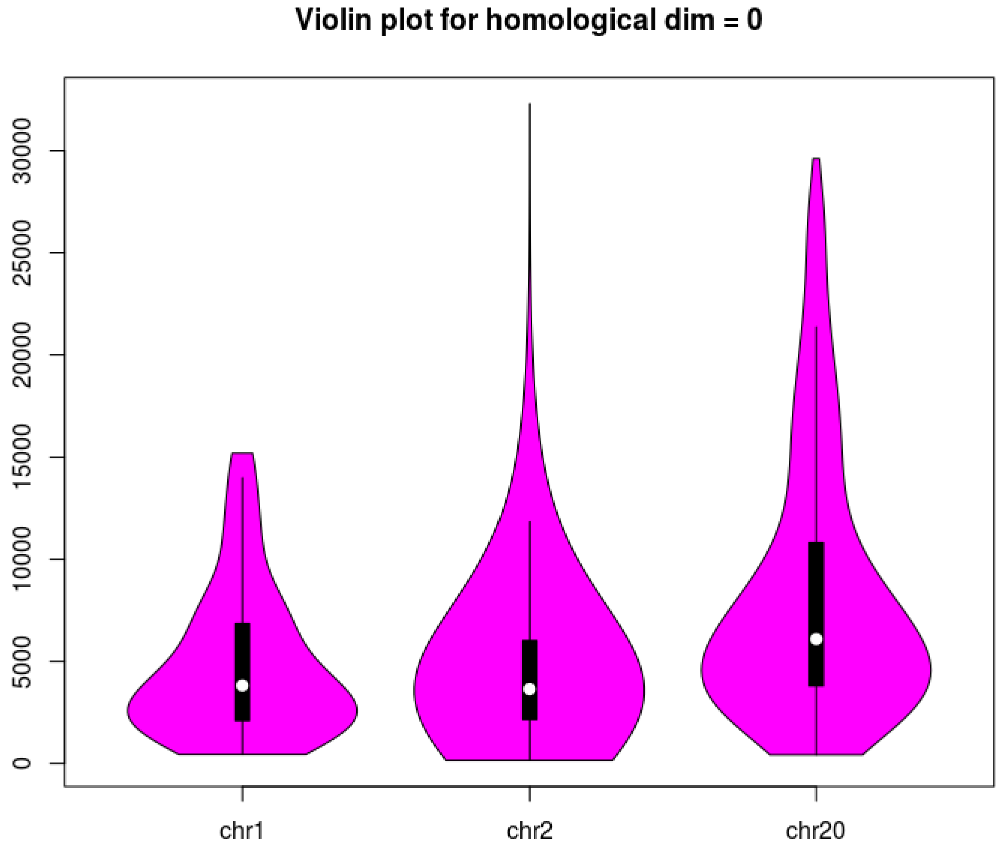

Figure 15: Violin plot of intra-chromosomal 0-dim distances.

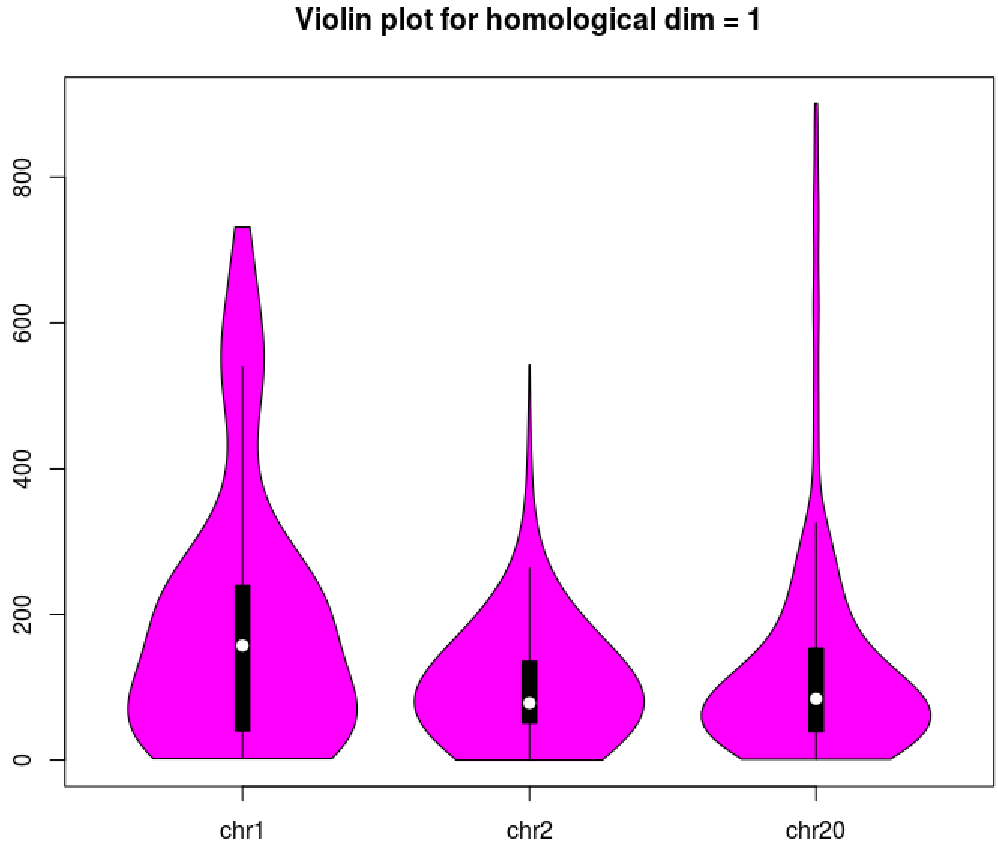

Figure 16: Violin plot of intra-chromosomal 1-dim distances.

Bringing all the different violin plots together we can compare intra- and inter-chromosomal topological fluctuations (see Fig.15,16,17,18).

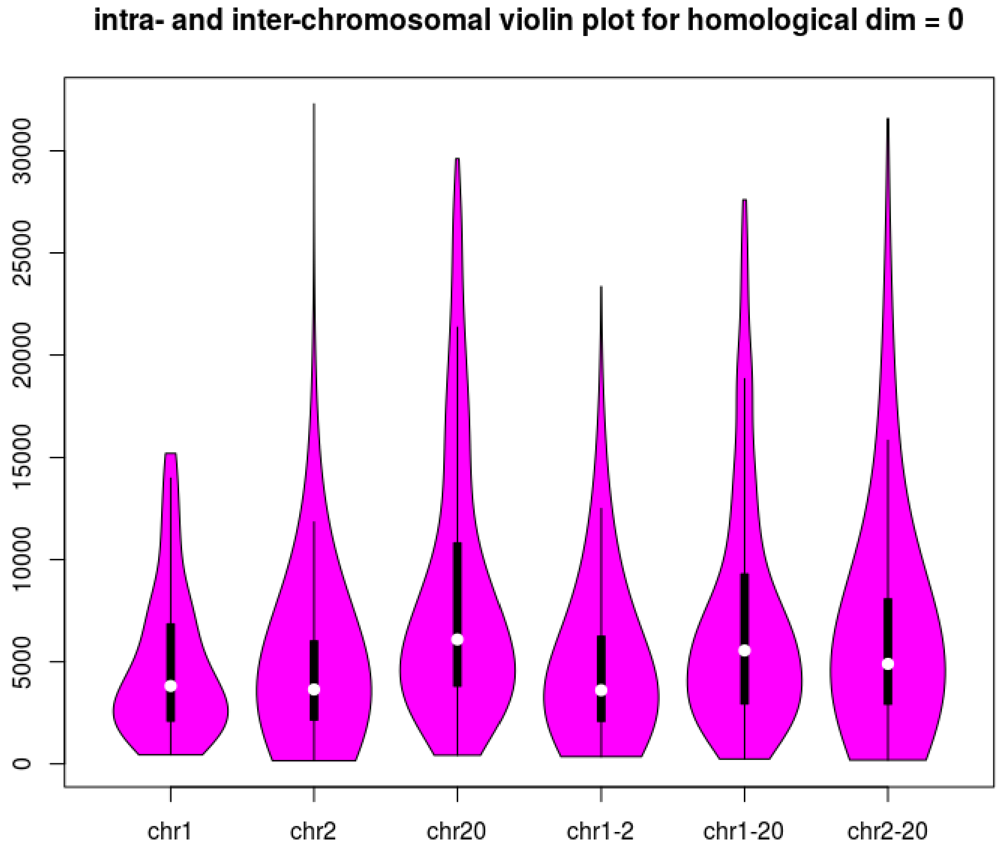

Figure 17: Violin plot of intra and inter-chromosomal homological distances in dim 0.

3 Silhouettes

Persistence silhouettes are real valued functions summarizing the information contained in a persistence diagram [REF]. Consider first the triangle function associated to a persistence diagram with N points, each defined by the pair (b,d) of birth and death coordinates:

$$\Lambda_p(t) = \begin{cases} t-b & t \in [b, \frac{b+d}{2}] \\ d-t & t \in (\frac{b+d}{2}, d] \\ 0 & otherwise \end{cases} , \tag{1}$$

for every $0 < p < \infty$ we define the power-weighted silhouette

$$\phi^{(p)}(t) = \frac{\sum_{j=1}^{N} |d_j - b_j|^p \Lambda_j(t)}{\sum_{j=1}^{N} |d_j - b_j|^p}. \tag{2}$$

For statistical purposes, silhouettes are more convenient than persistent diagrams to deal with, since for example it is possible to unambiguously define the average of a collection of

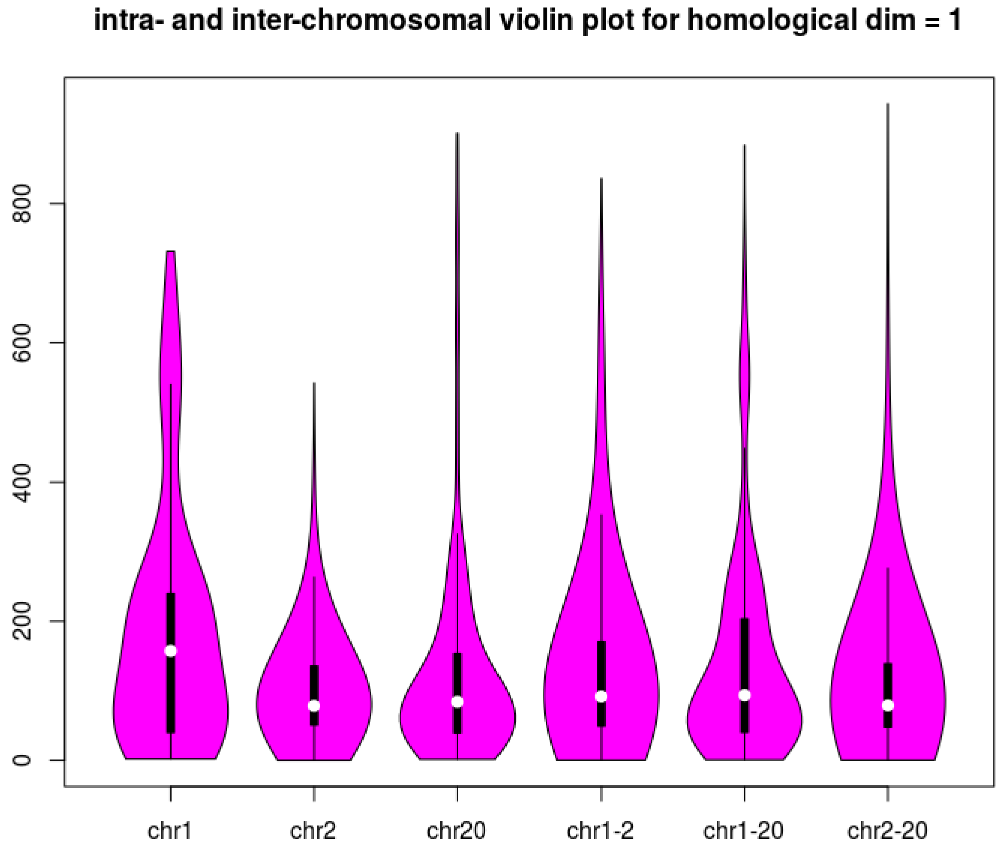

Figure 18: Violin plot of intra and inter-chromosomal homological distances in dim 1.

them. Fig19 to 24 show the average, over different point clouds, and the 95% confidence interval (obtained with a multiplier bootstrap method [REF]) of silhouettes in dim 0 and dim 1; these summaries can be used to classify and compare different chromosomes.

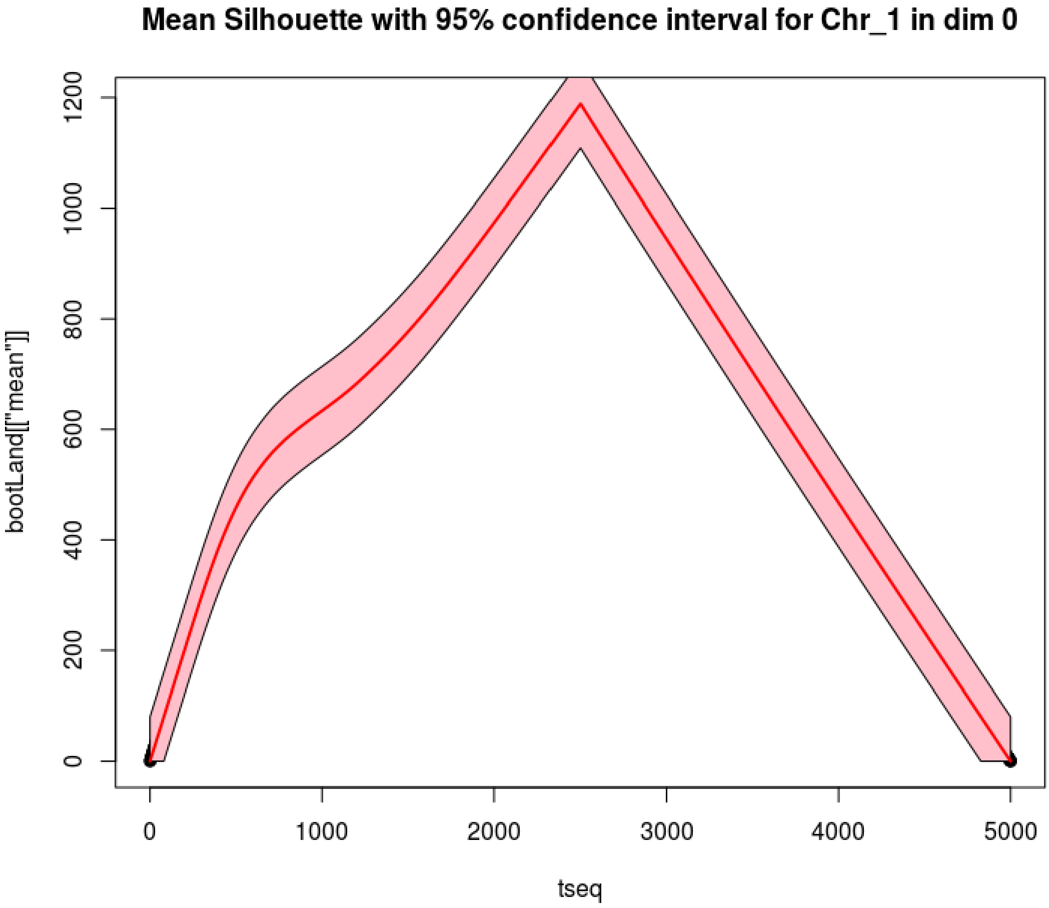

Figure 19: Mean silhouette for chr 1, homological dimension 0.

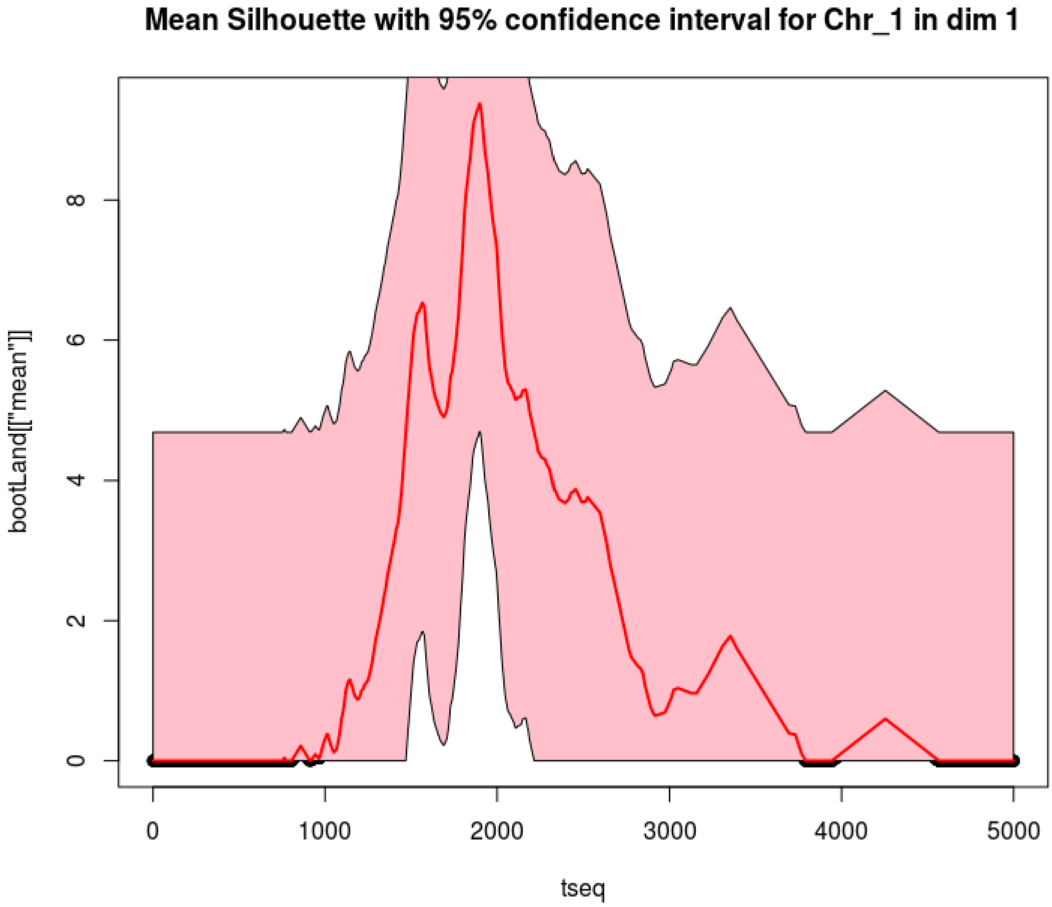

Figure 20: Mean silhouette chr 1, homological dimension 1.

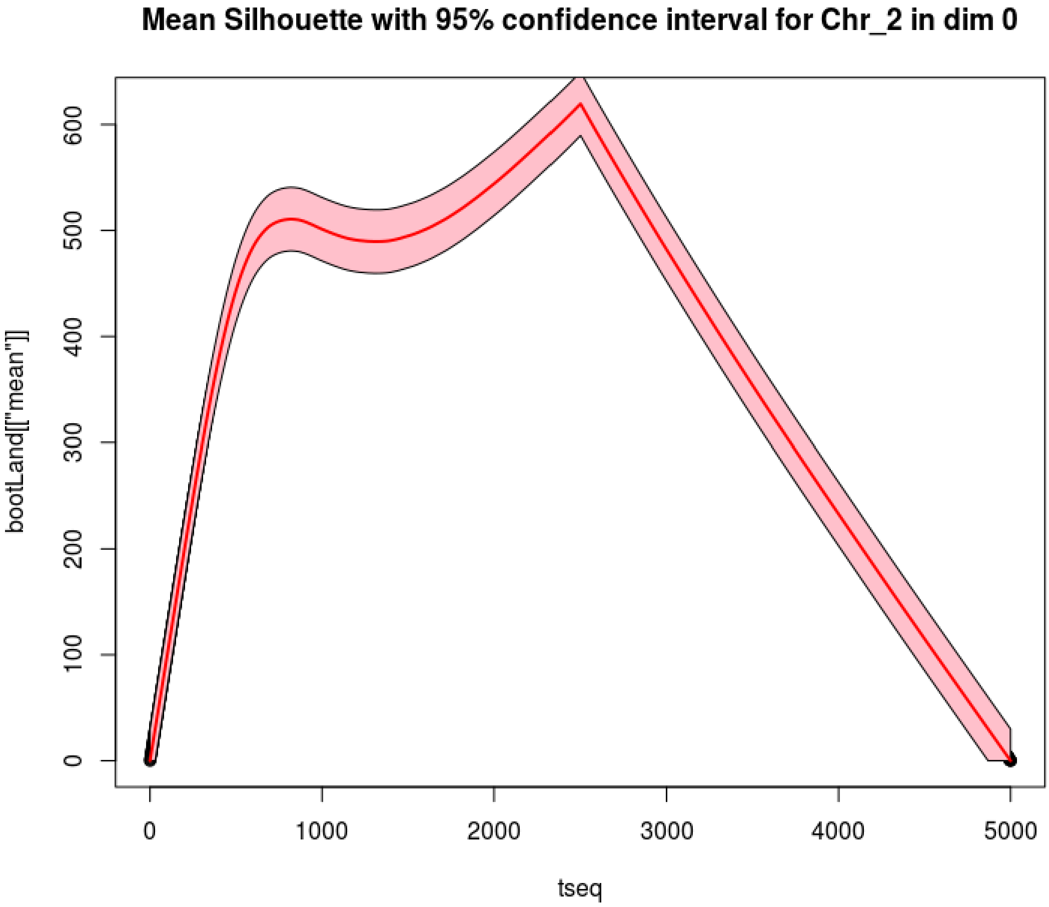

Figure 21: Mean silhouette for chr 2, homological dimension 0.

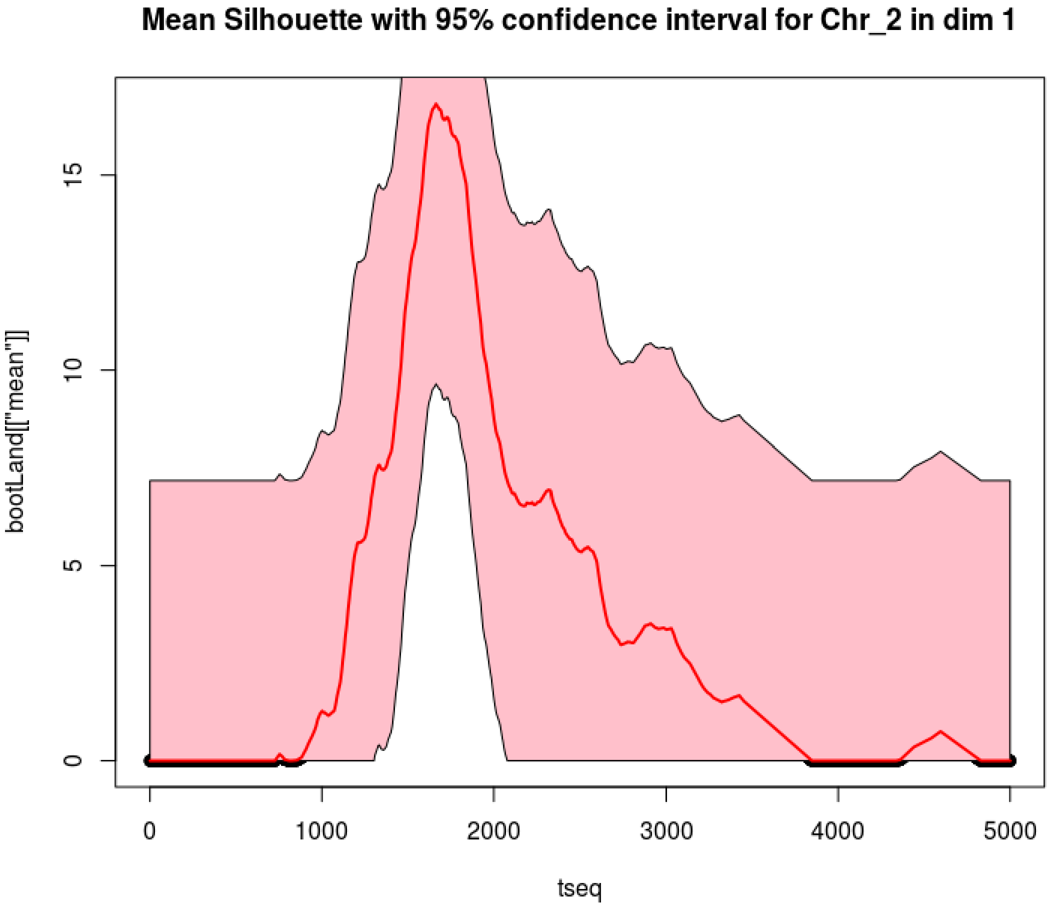

Figure 22: Mean silhouette chr 2, homological dimension 1.

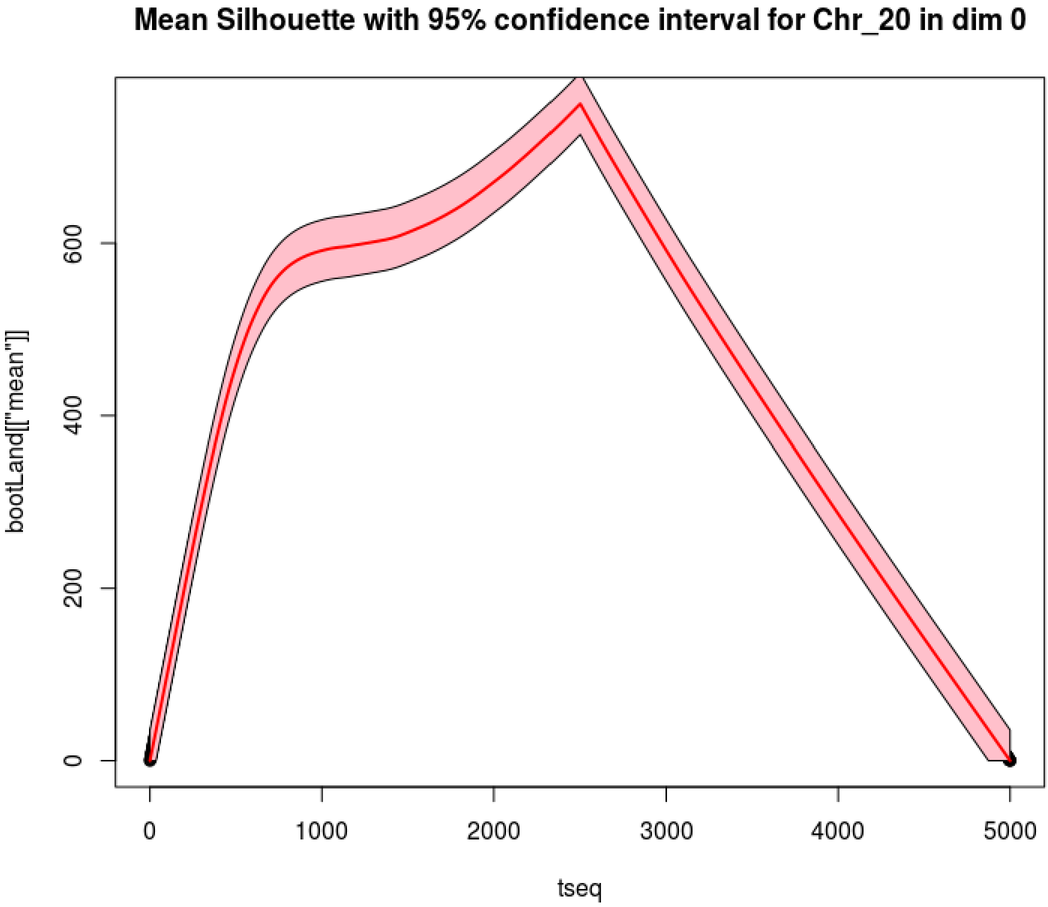

Figure 23: Mean silhouette for chr 20, homological dimension 0.

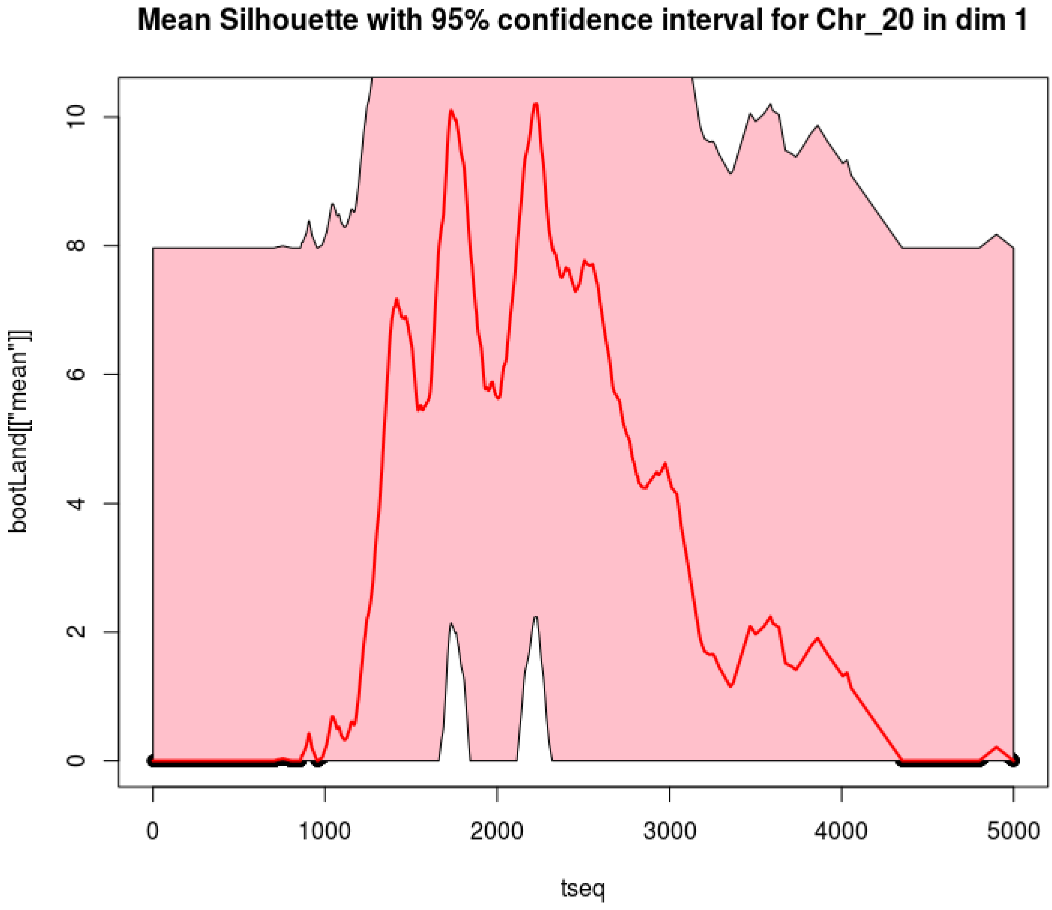

Figure 24: Mean silhouette chr 20, homological dimension 1.