StereoGene: a tool for fast correlation assessment and its application to the analysis of high throughput data

Elena Stavrovskaya[1,2,*], A.V. Favorov[3,4,5], Tejas Niranjan[5], Sarah Wheelan[5], Andrey A. Mironov[1,2]

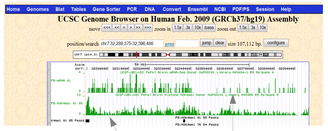
1 Moscow State University, Leninskie gory 1-73, Moscow, 119992, Russia ,
2 Institute for Information Transmission Problems , Bolshoy Karetny per. 19, Moscow, 127994, Russia
3 Vavilov Institute of Genral Genetics RAS , Gubkina str. 3, Moscow, 119333, Russia
4 State Scientific Center Genetika , 1-st Dorozhniy pr., 1, Moscow, 117545, Russia
5 Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, 550 N Broadway ste 1103 Baltimore, MD 21205 USA
*corresponding author: stavrovskaya@gmail.com

Introduction

The modern high-throughput sequencing methods provide massive amounts of genome-focused, DNA-positioned data. This data is often represented as a function of the DNA coordinate (e.g. coverage). The genome- or chromosome-wide correlations between data from different sources may provide information about functional biological interrelation of the investigated features, e.g., transcription and histone modification. The key idea of the correlation studies is that two features that are similarly distributed along a chromosome may be functionally related. The correlation could also be treated as a function on genomic coordinate, and so we can not only assess the interrelations, but also to investigate their localization inside the genome.

Previously, methods of correlation analysis were applied for numerical annotations and some biological results were obtained. But these methods do not allow to analyze positional correlations. The task to compute the spatial correlation was successfully solved only for interval annotations.

Here we present StereoGene that is a fast and powerful tool for estimation of correlations.

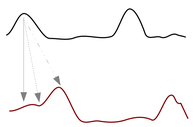


Are these two features correlated ?

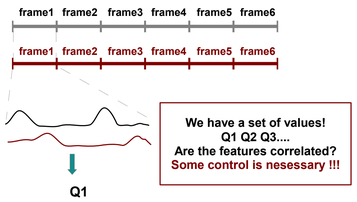
Some measure is necessary!!!

The idea of Kernel

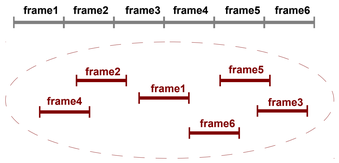
Allows to account for correlation of neighbour positions rather than the same only!



Problem: genome is too long!!!



Permutations



Method

Correlation:

$$Q = \frac{\int_0^L f(x)g(x)dx}{\sqrt{\int_0^L f^2(x)dx \int_0^L g^2(x)dx}}$$

An equivalent transformation of the previous integral:

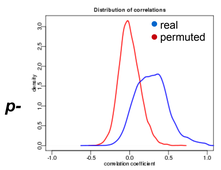
$$Q = \frac{\int_0^L \int_0^L f(x)g(y)\delta(y-x))dxdy}{\sqrt{\int_0^L f^2(x)dx \int_0^L g^2(x)dx}}$$

We can use another kernel function instead of delta-function :

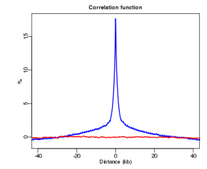
$$Q = \frac{\int_0^L \int_0^L f(x)g(y)\rho(y-x))dxdy}{\sqrt{\int_0^L f^2(x)dx \int_0^L g^2(x)dx}}$$

For the whole genome: **Mann–Whitney U test *p-value***

For each real frame: permutation ***p-value*** and **FDR *q-value.***



Cross-correlation function on coordinate shift shows typical picture of mutual positioning of two tracks

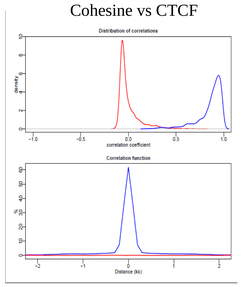


Tool

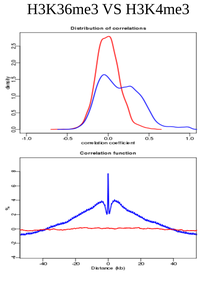
- Very fast (3 min per genome)
- works with quantitative and qualitative data
- The kernel-based approach allows complex geometry (shifts, smoothing, etc)
- Along with predefined kernel, calculates the results for set of standard shifts.
- Produce correlation track that can be used as input for further correlation (liquid association)
- Allows also to scale and sum profiles and compare profile combinations

Correlation

Cohesine vs CTCF

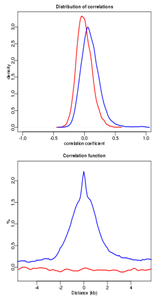


H3K36me3 VS H3K4me3

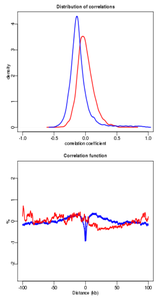


Bivalent promoters

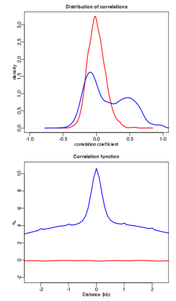
H3K9me3 vs H3K4me3 ***Stem cell H1***



H3K9me3 vs H3K4me3 ***Fetal Brain***



H3K9me3 vs H3K4me3 ***Brain Angular Gyrus***



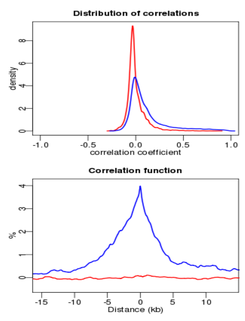
Thee-way correlation: projection.

Projection correlation is intended for analysis correlations of two profiles f,g with exclusion of correlation of these profiles with third one (confounder)

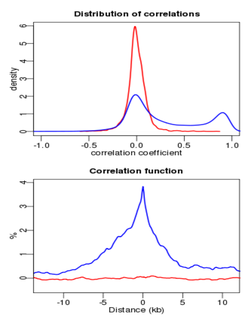
$$\hat{f}(x) = f(x) - a(x)\frac{< af >}{< aa >}$$

Correlation of H3K4me3 with mRNA-Seq, both projected on H3K4me1:

No projection



Projection

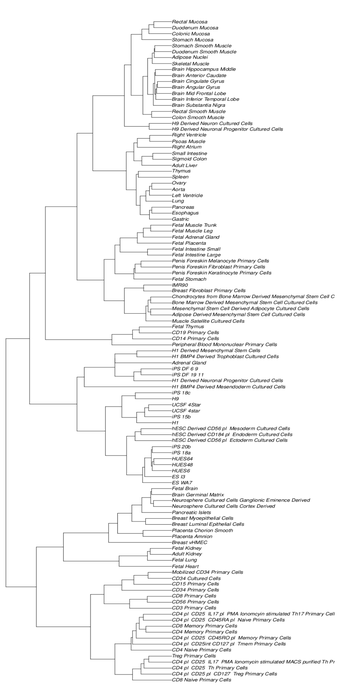


Tissue clustering (the tree)

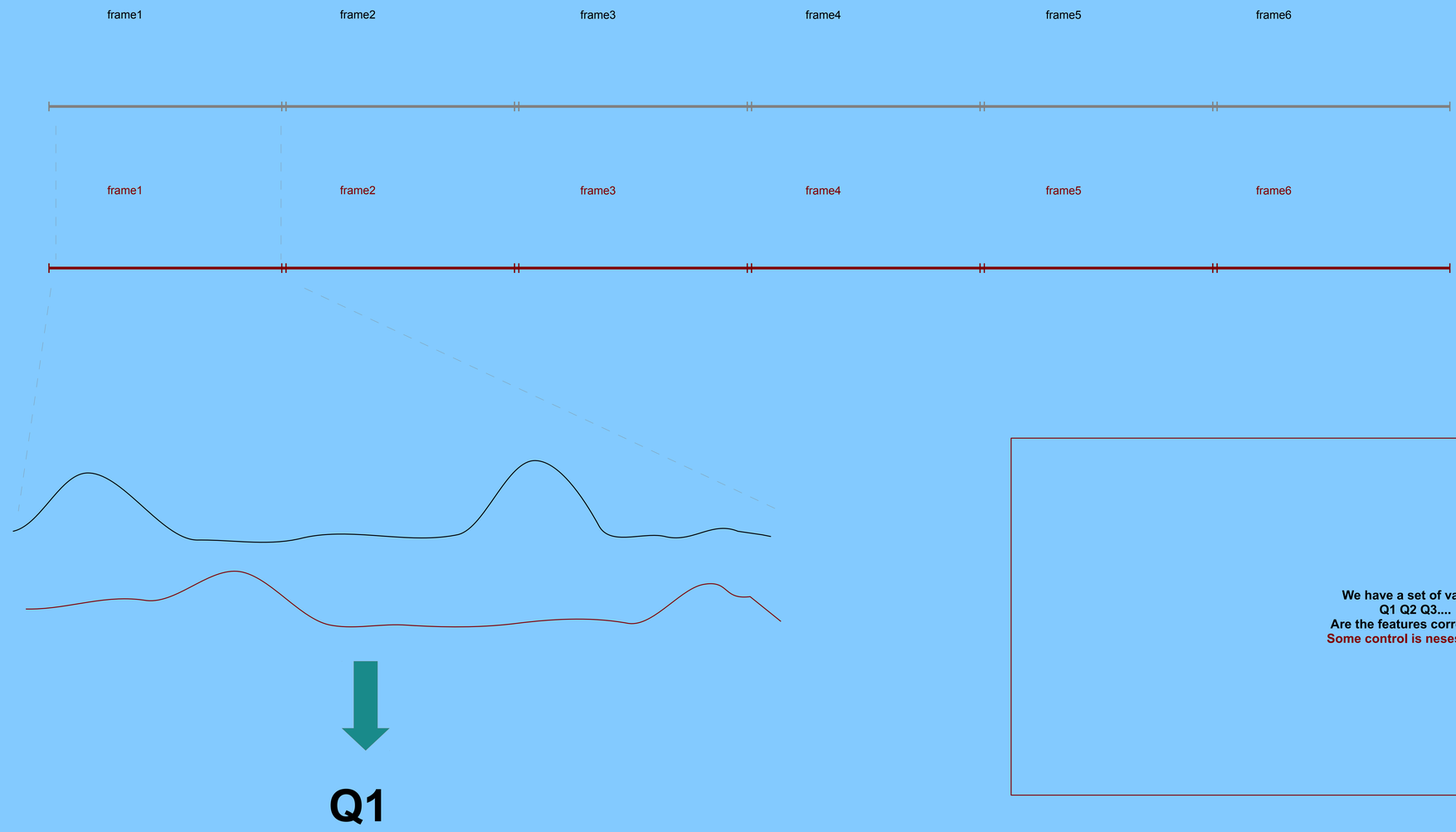
- 9 marks, 111 tissues
- For each mark we build a distance matrix, based on pairwise correlation between all tissues.
- Based on the matrix, we build hierarchical cluster tree
- For each pair of tissues, we count the maximal level of common subtree containing them both (divergence level), or the minimal path length inside the tree
- To count the mean divergence level for the pair of tissues, we average the DL in all trees that contain the pair
- Then we build the new distance matrix 1/mean(trees) level and run the hierarchical clustering again

Used marks:

H3K4me1, H3K4me3, H3K9me3, H3K9ac, H3K27me3, H3K27ac, H3K36me3



Problem: genome is too long!!!



We have a set of values!
Q1 Q2 Q3....
Are the features correlated?
Some control is nesessary !!!