Fast assessment of the correlation between different coverage-like genomic features

E.D. Stavrovskaya[1,2], A.V. Favorov[3,4], A.A. Mironov[1,2]

1. Faculty of Bioengineering and Bioinformatics, Lomonosov Moscow State University, Moscow, Russia
2. Institute for Information Transmission Problems, Moscow, Russia
3. State Scientific Center GosNIIGenetika, Moscow, Russia
4. Johns Hopkins University School of Medicine, Baltimore, MD

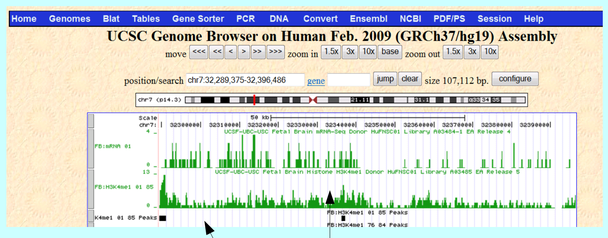
Abstract

1

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 task to compute the correlation was already successfully solved for interval annotations ([1]) as well as for coverage (functional) data ([2], [3], [4]). The key idea of the correlation studies is that two features that are similarly distributed along a chromosome may be functionally related. The point we are addressing here is that peaks of dependent functional features can be located in a similar, although somewhat different, way. To account for these similarities, we propose here a fast method for calculation of kerneled correlation between two numeric annotations of the genome. The kernel represents the mutual position of related features; e.g., a Gaussian shape corresponds to 'somewhere around', etc.

Problem statement

2



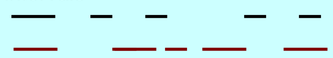
Are these two features correlated ?

Some measure is necessary!!!

3

Consider each feature as a curve and calculate some distance between these curves

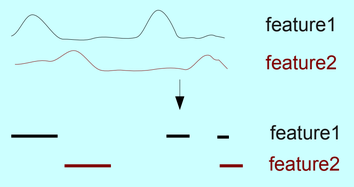
You can select some threshold and compare these two sets of intervals



We can use measure similar to Jaccard similarity to compute correlation of the interval sets:

$$Q = \frac{\int_0^L f(x)g(x)dx}{\sqrt{\int_0^L f^2(x)dx \int_0^L g^2(x)dx}} \qquad f(x) = \begin{cases} 1 \\ 0 \end{cases} \quad g(x) = \begin{cases} 1 \\ 0 \end{cases}$$

Distant correlations:



One feature can influence other feature, but genome position may differ

It is worth searching for correlation not only in current position, but also in its neighborhood

Method

4

An equivalent transformation of the previous integral (box 3):

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

And substitute delta-function with another kernel function:

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

Applying Fourier transform. Consider standard Fourier basis. The numerator of our correlation function takes on the following form:

$$\phi_k(x) = e^{ikx \cdot 2\pi/L}$$

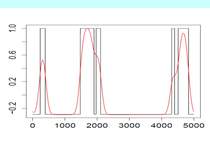
$$f(x) = \sum_k f_k e^{\frac{ikx2\pi}{L}}$$

$$g(y) = \sum_m g_m e^{\frac{imy2\pi}{L}} = \sum_m g_m^* e^{\frac{-imy2\pi}{L}}$$

$$\rho(y-x) = \sum_l \rho_l e^{\frac{il(y-x)2\pi}{L}}$$

$$Q = Re(\int_0^L \int_0^L f(x)g(y)\rho(y-x)dxdy)$$
$$= Re(\int_0^L \int_0^L \sum_k f_k e^{\frac{ikx2\pi}{L}} \sum_m g_m^* e^{\frac{-imy2\pi}{L}} \sum_l \rho_l e^{\frac{-ilx2\pi}{L}} e^{\frac{ily2\pi}{L}} dxdy)$$
$$= Re(\sum_{k,m,l} f_k g_m^* \rho_l \int_0^L e^{\frac{ikx2\pi}{L}} e^{\frac{-ilx2\pi}{L}} dx \int_0^L e^{\frac{ily2\pi}{L}} e^{\frac{-imy2\pi}{L}} dy)$$

$$Q = Re(\sum_k f_k g_k^* \rho_k)$$



g(x)(black) and inverse Fourier transform for $g_k \rho_k$ (red)

Idea: what if we use this measure not only for interval functions, but also for any pair of functions?

Pipeline

5

Problem: Genome sequences are very long
Solution: Cut sequence into shorter frames.

It also allows us to select genome areas with strongly correlated features and analyze them more precisely.

Count correlation for each frame

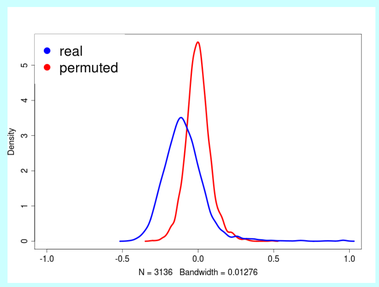
↓

Distribution of correlation for the complete genome

↓

We can use p-value to estimate correlation for the complete genome

p-value and *q-value*



We calculate the same correlation function ***Q*** for the permuted frame pairs and thus we estimate background distribution

We use **Mann–Whitney U test *p-value*** of the comparison of the real and the permuted distributions of **Q**'s of frames as a correlation measure for the whole genome.

For the value of ***Q*** for each real frame, we calculate the permutation ***p-value*** and **FDR *q-value*.** We can use these values to select genome areas, where the features are strongly correlated.

Performance

6

Program implementation: C++

PC: AMD Athlon(tm) 64 Processor 3000+, cpu 1000 mHz

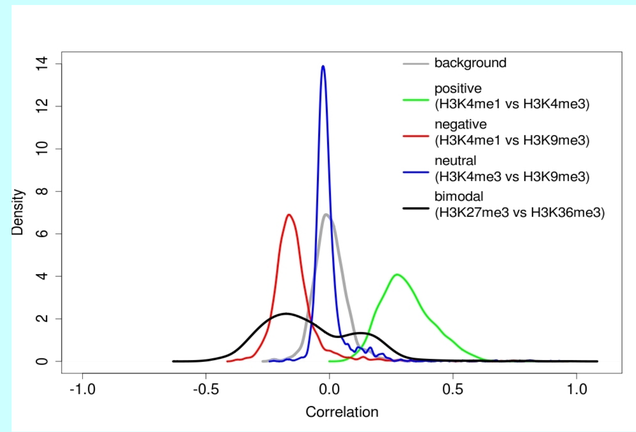
Input: two features (tracks) for a complete human genome or for a chromosome

Running time for full genome: < 3 min

Results

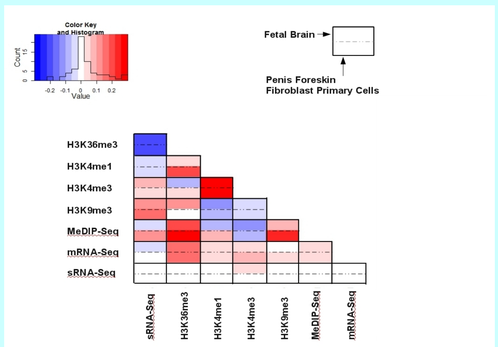
- Source: Human Epigenome Atlas (http://www.genboree.org/epigenomeatlas)
- 129 Tissues
- 34 Features
- about 150 000 Comparisons

Observed distribution types

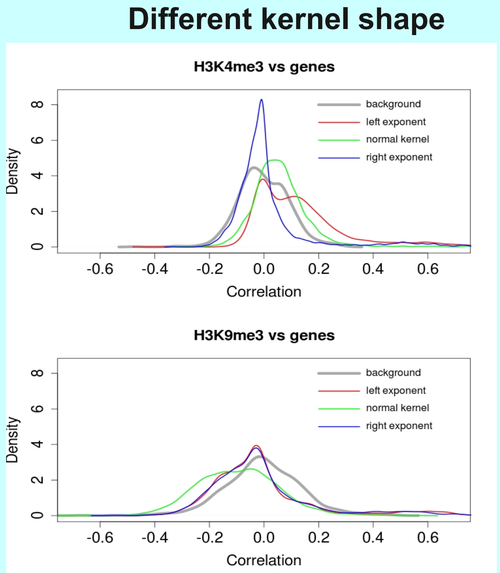


Features correlation in different tissues

7

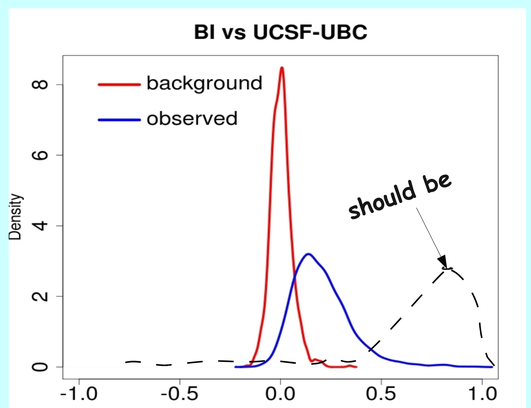


Different kernel shape

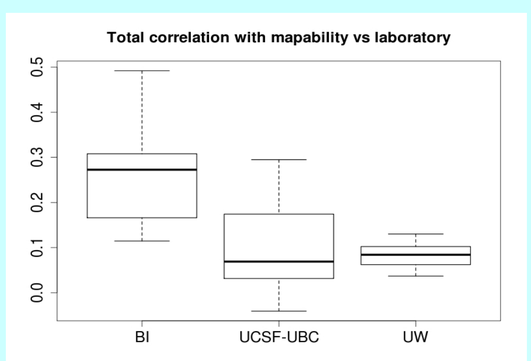


Same feature/tissue – different labs

8



We can check data accuracy!!!



Discussion

9

Here we present a FFT-based StereoGene algorithm for fast assessing a chromosome- or genomewide correlation between a pair of genomic features and obtain both statistical measures and the outlier regions. The features are represented as numeric functions on the chromosome coordinate. Different kernel shapes allow to take the feature-specific assumptions into account (shift, anisotropy)

Full genome in 3 minutes!!!

Plans: Encode, mapping, profile combinations...

References

1. Favorov, A., Mularoni, L., Cope, L.M., Medvedeva, Y., Mironov, A.A., Makeev,V.J., Wheelan, S.J.: Exploring massive, genome scale datasets with the GenometriCorr package. PLoS Comput Biol 8(5) (May 2012) e1002529
2. Ramsey, S.A., Knijnenburg, T.A., Kennedy, K.A., Zak, D.E., Gilchrist, M., Gold, E.S., Johnson, C.D., Lampano, A.E., Litvak, V., Navarro, G., Stolyar, T., Aderem, A., Shmulevich, I.: Genome-wide histone acetylation data improve prediction of mammalian transcription factor binding sites. Bioinformatics (Oxford, England) 26(17) (September 2010) 2071{2075 PMID: 20663846.
3. Bickel, P.J., Boley, N., Brown, J.B., Huang, H., Zhang, N.R.: Subsampling methods for genomic inference. The Annals of Applied Statistics 4(4) (December 2010) 1660{1697 Zentralblatt MATH identier: 05910045; Mathematical Reviews number (MathSciNet): MR2829932.
4. Bickel, P.J., Brown, J.B., Huang, H., Li, Q.: An overview of recent developments in genomics and associated statistical methods. Philosophical transactions. Series A, Mathematical, physical, and engineering sciences 367(1906) (November 2009) 4313{4337 PMID: 19805447.

Acknowledgements

Russian foundation of basic reserch (11-04-0216_a, 12-04-91333)
Federal Agency on Education (№8283)
Russian Academy of Sciences, the program "Molecular and Cellular Biology"
Johns Hopkins University Framework for the Future, and the Commonwealth Foundation