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

Elena Stavrovskaya, A.V. Favorov

Moscow State University, Leninskie gory 1-73, Moscow, 119992, Russia , stavrovskaya@gmail.com

Andrey A. Mironov

Institute for Information Transmission Problems , Bolshoy Karetny per. 19, Moscow, 127994, Russia mironov@bioinf.fbb.msu.ru

Vavilov Institute of Genral Genetics RAS , Gubkina str. 3, Moscow, 119333, Russia favorov@gmail.com

State Scientific Center Genetika , 1-st Dorozhniy pr., 1, Moscow, 117545, Russia

State Johns Hopkins University School of Medicine , 550 N Broadway ste 1103 Baltimore, MD 21205 USA

The modern high-throughput sequencing methods provide massive amounts of genome-focused, DNA-positioned data. This data is often represented as a function (e.g. coverage) onf the DNA coordinate. The genome- or chromosome-wide correlations between data from different sources may provide information about functional biological interrelation of the investigated features, e.g., about the trancription 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 a 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. The approach is implemented as a computer program using C++ language. It allows counting of correlation not only for single features, but also for their combinations.

Results and discussion

To test the suggested method, we analyzed the interrelation between the histone modification

marks that are known to be the marks of eu- and hetero-chromatin, and the mRNA-seq signal. All the data were taken from Human Epigenome Atlas (http://www.genboree.org/epigenomeatlas) for Fetal Brain tissue. Fig.1 shows the results. As expected, H3K4me1 is positively correlated with mRNA (*a*), and negatively correlated with H3K27me3 (*b*). The results are in concordance with a similar test for correlation between highly transcribed gene promoters and these marks [1]. Applying the procedure to combinations of euchromatine (correspondingly heterochromatine) modifications yields a much greater effect (*c* and *d*)

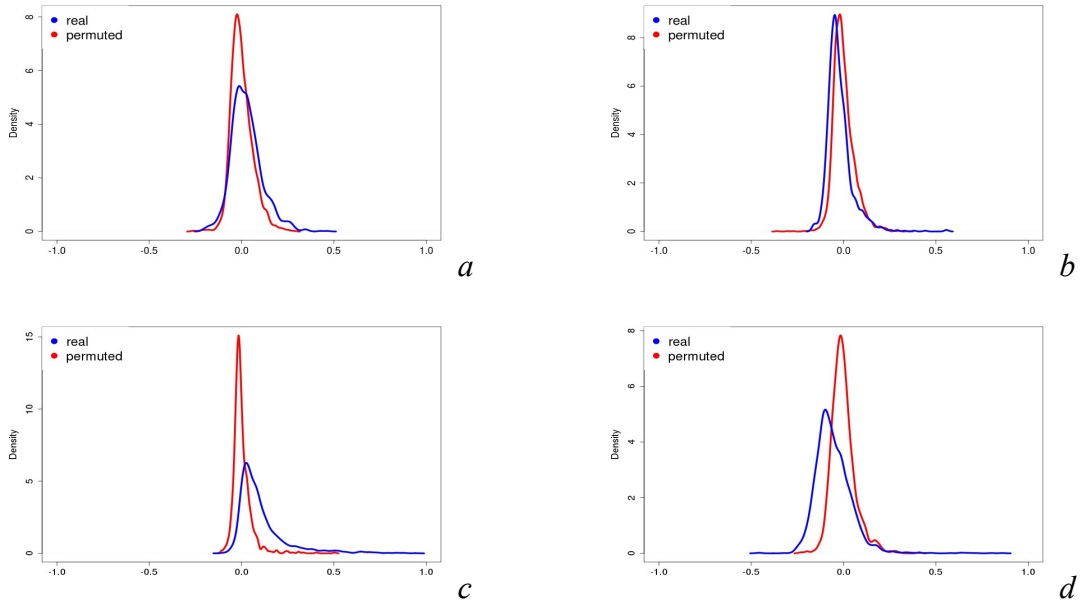

Fig.1 Density for correlation function *a)* for histone modification *H3K4me1* (euchromatine) vs *mRNA-Seq; b)* for histone modification *H3K27me3* (heterochromatine) vs *mRNA-Seq; c)* for combination of histone modifications *H3K4me1, H3K4me1 and H3K9ac* (euchromatine) vs *mRNA-Seq; d)* for combination of histone modifications *H3K27me3* and *H3K9me3* (heterochromatine) vs *mRNA-Seq*. P-value is calculated using Wilcoxon test. Red : background distribution of windows by the in-window correlation; Blue : the observed distribution.

1. A.Favorov et al. (2012) Exploring massive, genome scale datasets with the GenometriCorr package. *PLoS Comput Biol,* **8(5) :**e1002529

2. S.A.Ramsey et al. (2010) Genome-wide histone acetylation data improve prediction of mammalian transcription factor binding sites. *Bioinformatics (Oxford, England),* **26(17):**2071-2075

3. P.J.Bickel et al. (2010) Subsampling methods for genomic inference. *The Annals of Applied Statistics* **4(4):**1660-1697

4. P.J.Bickel et al. (2009) An overview of recent developments in genomics and associated statistical methods. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences* **367(1906):**4313-4337