StereoGene: Rapid Estimation of Genomewide Correlation of Continuous or Interval Feature Data

Elena D. Stavrovskaya [1,2], Tejasvi Niranjan [3], Elana J. Fertig [4], Sarah J. Wheelan [3], Alexander Favorov [4,5,6]*, and Andrey Mironov [1,2]

1 Dept. of Bioengineering and Bioinformatics, Moscow State University, Moscow, 119992, Russia
2 Institute for Information Transmission Problems, RAS, Moscow, 127994, Russia
3 Department of Oncology, Sidney Kimmel Comprehensive Cancer Center, The Johns Hopkins University School of Medicine, 21287, Baltimore, MD 21287, USA
4 Department of Oncology, Division of Biostatistics and Bioinformatics, Sidney Kimmel Comprehensive Cancer Center, The Johns Hopkins University, Baltimore, MD 21205, USA
5 Laboratory of Systems Biology and Computational Genetics, Vavilov Institute of General Genetics, RAS, Moscow, 119333, Russia
6 Laboratory of Bioinformatics, Research Institute of Genetics and Selection of Industrial Microorganisms, Moscow, 117545, Russia

*** favorov@sensi.org**

Abstract

Motivation: High throughput sequencing methods produce massive amounts of data. The most common first step in interpretation of these data is to map the data to genomic intervals and then overlap with genome annotations. A major interest in computational genomics is spatial genome-wide correlation among genomic features (e.g. between transcription and histone modification). The key hypothesis here is that features that are similarly distributed along a genome may be functionally related.

Results: Here, we propose a method that rapidly estimates genomewide correlation of genomic annotations; these annotations can be derived from high throughput experiments, databases, or other means. The method goes far beyond the simple overlap and proximity tests that are commonly used, by enabling correlation of continuous data, so that the loss of data that occurs upon reduction to intervals is unnecessary. To include analysis of nonoverlapping but spatially related features, we use kernel correlation. Implementation of this method allows for correlation analysis of two or three profiles across the human genome in a few minutes on a personal computer. Another novel and extraordinarily powerful feature of our approach is the local correlation track output that enables overlap with other correlations (correlation of correlations). We applied our method to the datasets from the Human Epigenome Atlas and FANTOM CAGE. We observed the changes of the correlation between epigenomic features across developmental trajectories of several tissue types, and found unexpected strong spatial correlation of CAGE clusters with splicing donor sites and with poly(A) sites.

Availability: The *StereoGene* C++ source code, program documentation, Galaxy integration scrips and examples are available at the project homepage at `http://stereogene.bioinf.fbb.msu.ru/`

Contact: favorov@sensi.org

Supplementary information: Supplementary data are available online.

Introduction

Modern high throughput genomic methods generate large amounts of data, which can come from experimental designs that compare tissue-specific or developmental stage-specific phenomena for human [7] and model organisms [4]. Single-cell approaches are also rapidly advancing [3]. Such datasets are integrated into several different archive databases [8, 37, 44] and manually curated databases [25].

An important challenge of genome-wide data analysis is to reveal and assess the interactions between biological processes, *e.g.* chromatin profiles and gene expression. A rapidly emerging approach to this challenge is to represent data as functions on genomic positions and to estimate correlations between these functions.

Numerous recent biological publications employ the correlation-based approach. Several research papers [41, 43] focus on relationships between transcription factor binding and chromatin state. These studies also include information on DNA accessibility [1], higher-order chromosomal organization [19], and association of chromatin modifications and alternative splicing [18, 23]. The research field has broadened its focus on analysis of individual and cell/tissue specific variation of epigenomic features and their relationship with diverse traits [31]. An interesting "Comparative epigenomics" paradigm [42] has emerged from an observation that combinations of epigenetic marks are more conserved than the individual marks themselves. This cooperation requires spatial relationships that are difficult to statistically ascertain.

Several bioinformatic methods that estimate the association between genome-wide numerical features have been recently proposed, and powerful aggregation and visualization tools were developed for manual analysis of colocalization of multiple features [12, 36, 37, 40].

Computational assessment of correlations on continuous genomewide data recruits various mathematical and statistical methods. For consistency with existing bioinformatic methods for positional correlation analysis, we use the terms *profile* or *track* for position-defined genomic features. For the colocalization analysis, genomic features are formalized as one of three types: profiles that are represented as a set of intervals on the genome (genes, repeats, CpG islands, etc.); point profiles (binding sites, TSS, splice sites); and continuous profiles, such as coverage data (expression, ChIP etc) resulting from high throughput sequencing experiments.

Many computational approaches have been developed to assess genomic features. An entropy-based approach has been developed for identification of differentially methylated regions [45]. A Bayesian mixture model is used for consistency analysis of different sources of data (ChIP-ChIP and ChIP-seq, [35]. A Hidden Markov Model is used for prediction of generalized chromatin states [10]. A probabilistic approach for the chromatin code landscape is introduced in [48]. A compendium of epigenomic maps is used in [9] to generate genome-wide predictions of epigenomic signal tracks, and a detailed review of machine learning for genome features is given in [21].

Correlations may be direct overlaps, but many of the most interesting relationships are more difficult to discern, as they require a general proximity but not overlap. For example, gene expression (RNA-seq coverage) correlates with transcription factor binding or chromatin state in nearby promoter regions or distant enhancer regions. The distant spatial correlations of interval profiles is addressed in [6, 11].

The interval and point-wise genome-wide correlations are addressed in [6, 10, 11], [16]. A common approach to investigating genomic features is to represent these features as intervals, computed from the original continuous coverage data using a threshold or more sophisticated algorithms [46]. With these methods, the resulting track depends on the algorithm used, and portions of the original data are lost.

[24] work with continuous profiles directly using the Karhunen-Loeve transform. This enables evaluation of both experimental variability and true biological signal (the biological signal tends to be in the higher components). While elegant, this method is slow and precludes the investigational analyses that are so important when analyzing these data.

Here, we propose a fast universal method to assess correlation of genomic profiles. The data can be discrete features (e.g. intervals) or continuous profiles (e.g. coverage data representing the level of histone methylation, protein binding, or expression). The method is based on calculation of the convolution integral with some kernel (kernel correlation, KC), with speedup using Fast Fourier Transform (FFT). The kernel allows calculation of correlation of the profiles that are smoothed over a genomic neighborhood.

The KC measure provides us with an estimate of spatial correlation (overlap, colocalization, or relative distance) of two features. To estimate the statistical significance of the correlation, we split the genome into a set of non-overlapping windows (100kb-1Mb). The foreground signal is computed as the distribution of correlation values for each of the windows. To get the background signal, we shuffle windows and recalculate the correlations. Statistical analysis is based on comparison of foreground and background distributions.

Our implementation is very quick: calculation for a pair of profiles over the human genome takes approximately 1-3 minutes on a standard PC.

StereoGene is presented and source code and some examples are available at the project homepage at `http://stereogene.bioinf.fbb.msu.ru/`.

Materials and Method

Kernel correlation

We consider each genomic feature as a numeric function (profile) on the genomic position x. The standard Pearson correlation of two profiles $f = f(x)$ and $g = g(x)$ is defined as:

$$CC(f,g) = \frac{1}{\sigma_f \sigma_g} \frac{1}{|G|} \int\limits_G \tilde{f}(x)\tilde{g}(x)dx = \frac{Q(\tilde{f},\tilde{g})}{\sqrt{Q(f,f)Q(g,g)}}$$

where $\tilde{f} = (f(x) - \bar{f})$, $\bar{f}$ is the mean value of f; σ_f is the standard deviation of f, $Q(f,g) = \int_G f(x)g(x)dx$; the integration is performed over the genome G. The Pearson correlation relates profile values on exactly the same genomic positions. In biological systems, the relationships of values at proximal but nonoverlapping (in genomic coordinates) positions are also important. These correlations may be mediated by chromatin looping or other interactions. To account for them, we use the following generalization for the covariation integral:

$$Q_\rho(f,g) = \int\limits_G \int\limits_G \tilde{f}(x)\tilde{g}(y)\rho(x-y)dxdy \tag{1}$$

where $\rho(x-y)$ is a kernel function that reflects the expectations of interaction of features at adjoining positions. In the case $\rho(x-y) = \delta(x-y)$, we get the standard covariation integral. Here, we use the Gaussian kernel $\rho(z) = \frac{1}{\sqrt{2\pi\sigma}} exp(-\frac{x^2}{\sigma^2})$, but other non-negative kernel functions can be used.

The two-dimensional integral $Q_\rho(f,g)$ can be rapidly calculated using a Fourier transform:

$$\begin{aligned} Q_\rho(f,g) = & \sum_{i=1} f_k g_k \rho_k; \\ & f(x) = \sum_k f_k \phi_k(x); \\ & g(x) = \sum_k g_k^* \phi_k^*(x); \\ & \rho(y-x) = \sum_k \rho_k \phi_k(y-x) \end{aligned} \tag{2}$$

where $\phi_k(x)$ are the harmonic basis functions $\phi_k(x) = \exp(k \cdot 2\pi i/L)$, and ϕ_k^* means complex conjunction of ϕ_k. The equation takes into consideration that the zero coefficient of a Fourier transform of a function f is the average of the function ($\bar{f} = f_0$). Thus, the kernel correlation KC is defined as:

$$KC(f,q) = \frac{1}{\sigma_f^\rho \cdot \sigma_g^\rho} \sum_{i=1} f_k \cdot g_k^* \cdot \rho_k \tag{3}$$

where $\sigma_f^\rho = \sqrt{(Q_\rho(f,f))} = \sqrt{\sum f_k f_k^* \rho_k}$, $\sigma_g^\rho = \sqrt{\sum g_k g_k^* \rho_k}$. The value $KC(f,g)$ satisfies the inequality: $-1 \le KC\,(f,g) \le 1$. The Fourier transform can be calculated by the discrete Fast Fourier Transform (FFT) algorithm [22] that have the computational complexity $O(|G| \cdot \log |G|)$, where $|G|$ is the genome length. Complexity of correlation coefficient calculation consists of the complexity of Fourier transform and the complexity of summation $O(|G|)$. Hence the calculation time can be evaluated as $O(|G| \cdot \log |G|)$.

Cross-correlation (Distance correlation)

For two given profiles, $f(x)$ and $g(x)$ the cross-correlation function can be calculated:

$$c(x) = \frac{1}{\sigma_f \sigma_g} \frac{1}{|G|} \int_G \tilde{f}(t) \tilde{g}(t-x) dt \tag{4}$$

The cross-correlation function reflects a distance dependence of the profiles.

This function can also be calculated using Fourier transform:

$$c(x) = \frac{1}{\sigma_f \sigma_g} \frac{1}{|G|} \int_G \tilde{f}(t) \tilde{g}(t-x) dt = \frac{1}{\sigma_f \sigma_g} \frac{1}{|G|} FT^{-1}(f_k \cdot g_k^*) \tag{5}$$

where FT^{-1} means the reverse Fourier transform that can also be calculated using FFT algorithm.

Local correlation profile generation

Along with the integration of the correlation measure along the genome, *StereoGene* can generate a new profile that describes the kerneled local correlation of two profiles.

$$LC(x) = \frac{g(x) \int_G \rho(x-t) f(t) dt + f(x) \int_G \rho(x-t) g(t) dt}{2\sigma_f \sigma_g} \tag{6}$$

The integrals in this equation can be represented via Fourier transform, and the correlation profile is expressed as

$$LC(x) = \frac{1}{2\sigma_f \sigma_g} \left(g(x) \cdot FT^{-1}(\rho_k f_k) + f(x) \cdot FT^{-1}(\rho_k g_k) \right) \tag{7}$$

This profile is necessary to investigate relationships that are non-uniform along the genome, revealing more or less correlated segments. In particular, it can be used for a gene set enrichment analysis or correlated with a third genomic profile, and thus it can be involved in a 3-way correlation analysis that is analogous to liquid correlation [20]. This is a powerful and unique approach to dissecting complex relationships among genomewide datasets. Note that the value of LC is not restricted by ± 1 boundaries and can take any values.

Partial correlation

Nonrandom correlation of the two profiles may occur due to their correlation with a third profile (confounder) that systematically biases both signals (e.g. level of mapability). To computationally exclude such an influence, *StereoGene* can correlate projections of the two profiles orthogonal to the confounder profile a subspace:

$$\hat{f}(x) = f(x) - a(x)\frac{< af >}{< aa >} \tag{8}$$

A typical example of a confounder could be a common input track for two ChIP-seqs from the same cell type.

Statistical significance

The KC value provides useful information about the relative genomewide correlation of features, but it does not carry any information on statistical significance. To obtain the latter, KC is calculated in a set of adjacent large windows that cover the genome. Then, a shuffling procedure is used that randomly matches windows of one profile to another, and KC calculation is repeated in all the window pairs. Thus, two distributions, a foreground distribution of the real KC values and a background of permuted values, are obtained. The statistical significance is provided by a Mann-Whitney test of these two sets of values.

Program implementation

As input, *StereoGene* accepts two or more input files in one of the standard Genome Browser formats: BED, WIG, BedGraph, and BroadPeak. In the first step, *StereoGene* converts input profiles to an internal binary format and saves the binary tracks for future runs. If a project refers to the saved profile and the parameters have not changed, *StereoGene* reuses the saved tracks. *StereoGene* also requires chromosome length information provided in any standard UCSC form.

Output depends on parameters and will provide the following files: *.bkg — array of correlations for shuffled windows; *.fg — correlations in coherent windows; *.dist — distance distribution (correlation function) for background, foreground, and chromosomes; *.wig — a wig file for local kernel correlations; *.chrom — statistics by chromosomes; 'statistics' — a file that stores statistics for all runs and provides a summary, including total correlation, Z-score for Mann-Whitney statistics, and p-value.

For a quick and intuitive depiction of results, the *StereoGene* optionally generates an R script that graphs the output in two plots (Fig.1). The first plot displays foreground and background (permuted) distributions of genomic windows of the kernel correlation. A right shift of the foreground distribution relative to the background distribution represents positive correlation, and vice versa. The plot also displays more complicated features, such as multimodality, which show that the correlation is not uniform over the genome, or that multiple classes of features with different correlation profiles exist. The second plot, the (not kernel) cross-correlation function on possible

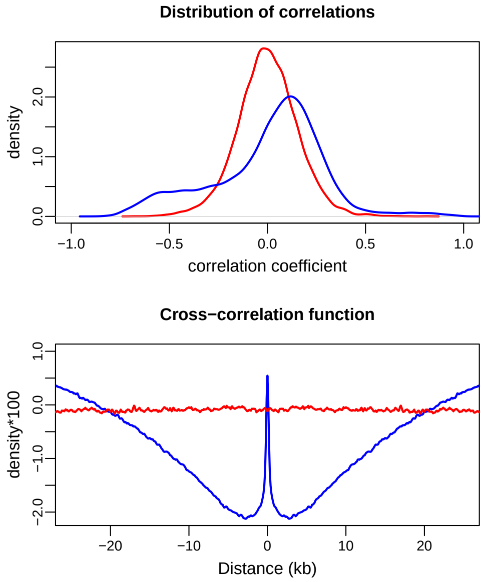


Figure 1. Example of output plots: H3K27me3 vs H3K36me3 in Fetal brain cells. The upper plot shows the distribution of the Kerneled Correlations over the windows; the lower plot shows the cross-correlation function. Red lines reflects the background distribution; blue line – the coherent windows

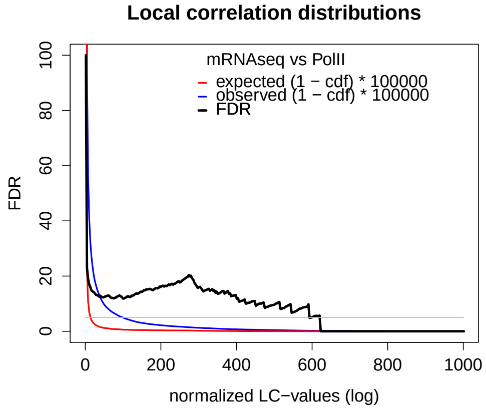


Figure 2. Distributions of the local correlations for mRNA Seq vs PolII in IMR90 cell line. Red line – background distribution $(1 - cdf)$; Blue line – Observed distribution $(1 - cdf)$; Black line FDR.

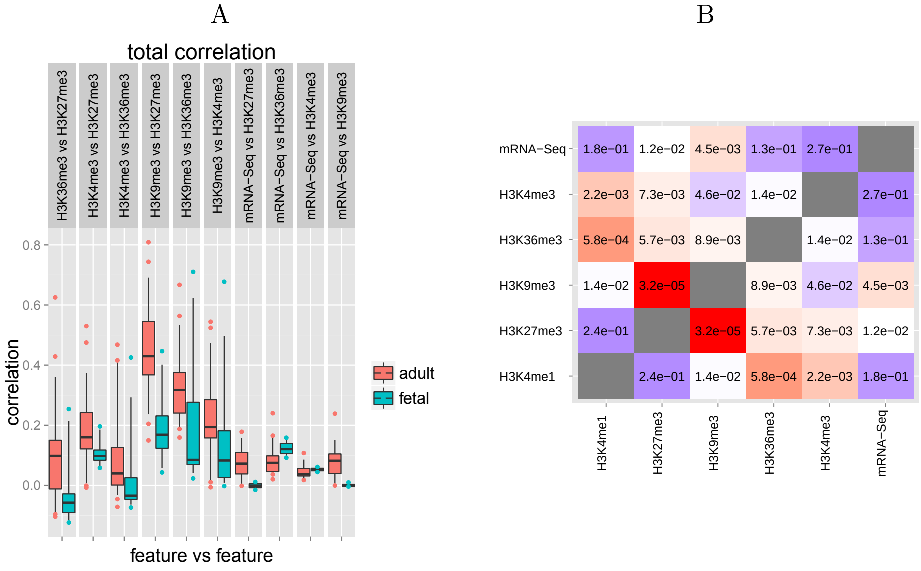
feature-to-feature shifts, represents local relationships between them. The program can provide mass analysis using track lists. The local correlation track is supported with data table for the observed and the expected local correlation distributions and the FDR q-values calculated for each local correlation values (Fig. 2). More detailed information about the the *StereoGene* software is presented in the supplementary file ProgDescr.pdf and in the program documentation

StereoGene is implemented in C++. The time required for the binary file preparation depends on file size. On a standard computer, the preparation takes from a few seconds to 1-2 minutes. The calculation of correlation with shuffling requires roughly one minute. A complete description of the keys and the parameters of *StereoGene* and output files formats are presented on *StereoGene* homepage.

Data source

Data by Roadmap Epigenomics Project [2] was obtained via the Human Epigenome Atlas (`http://www.genboree.org/`). Data for FANTOM4 CAGE clusters [30] was obtained from the UCSC website (RIKEN CAGE Loc tracks, GEO accession IDs were GSM849326 for nucleus GSM849356 for cytosole in H1 Human Embryonic Stem Cell Line, RRID:CVCL_9771). The datasets with the tracks are listed in Supplementary file 1.

1 Results



	H3K4me1	H3K27me3	H3K9me3	H3K36me3	H3K4me3	mRNA−Seq
mRNA−Seq	1.8e−01	1.2e−02	4.5e−03	1.3e−01	2.7e−01	
H3K4me3	2.2e−03	7.3e−03	4.6e−02	1.4e−02		2.7e−01
H3K36me3	5.8e−04	5.7e−03	8.9e−03		1.4e−02	1.3e−01
H3K9me3	1.4e−02	3.2e−05		8.9e−03	4.6e−02	4.5e−03
H3K27me3	2.4e−01		3.2e−05	5.7e−03	7.3e−03	1.2e−02
H3K4me1		2.4e−01	1.4e−02	5.8e−04	2.2e−03	1.8e−01

Figure 3. A. Genome-wide correlations for different types of cells. 12 fetal and 39 adult cell types data were used B.p-values for difference of correlation distributions between fetal and adult tissues (Wilcoxon test).

Human Epigenome Atlas Pairwise Correlation Anthology

As a straightforward test of our method, we prepared an anthology of pairwise correlations of the profiles from the Human Epigenome Atlas [2]. We built a pipeline that analyzes colocalization at all pairs of different profiles from the same tissue (or cell line) and all the pairs of the same profiles from different tissues. The results are displayed here `http://stereogene.bioinf.fbb.msu.ru/epiatlas.html` Interestingly, the majority of the comparisons of Epigenomics Roadmap profiles show a significant positive correlation, while negative correlations appear rarely. In many cases the cross-correlation function has a narrow peak centered at zero. For example, Fig.1 shows the distributions of the KC and the cross-correlation function for tracks H3K27me3 vs H3K36me3 in fetal brain cells. Though further evidence is needed, such behavior could reflect very precise nucleosome positioning; indeed, recent results from reChIP [15] suggests that the same nucleosome can carry different modifications.

To prepare an overview of this complex and multifaceted dataset, we split the Human Epigenome Atlas data into two collections: mature tissues and fetal tissues (refer to Supplementary file 1 for the track URL's). We focused on correlations of the most frequently studied epigenetic marks (*i.e.*), H3K4me1, H3K4me3, H3K9me3, H3K27me3, and H3K36me3, as well as RNA-seq. Fig 3A showed distributions of genome-wide correlations for the pairs of profiles. The highest difference of feature-to-feature correlation between the collections was observed for the H3K9me3 and H3K27me3 pair: they were significantly more correlated in adult tissues than in fetal ones. A comparison of correlation between H3K9me3 and H3K27me3 in the same tissue for fetal and adult gave a $p-value = 3.2 \cdot 10^{-5}$ (Wilcoxon test). This result is consistent with the prior observation that at early stages, different genomic regions are separately regulated by H3K9me3 and H3K27me3, but during tissue maturation, these heterochromatin marks became more synchronized [5]. One possible explanation is that H3K27me3 initiates chromatin compaction by recruitment of H3K9me3. The colocalization of H3K27me3 vs H3K36me3 relates to monoallelic gene expression [26]. Figure 3 shows significant increase of correlation of these marks in adult tissues in comparison with fetal tissues. The observation is consistent with the recent studies [27]. Other pairs of epigenomic marks behaved similarly, but with more moderate effect (Fig.3B).

In some cases, a bimodal shape is observed among the distribution of correlations in a feature-to-feature comparison; this may indicate that subsets of a feature fall into multiple classifications, each with different correlation properties. The correlation of H3K4me3 and H3K27me3 in adult lung tissue provides a good example of such bimodal behavior (fig.4A). These marks are widely assumed to have opposite effects: H3K4me3 is associated with active genes, while H3K27me3 is associated with closed chromatin. Simultaneously, the trimethylation of H3K4 and H3K27 presumably delineates bivalent domains in which developmental genes are poised for expression as the cell differentiates, and in general they are to be repressed in adult tissue [33]. To provide an example of *StereoGene* application, we analyzed genes associated with regions that carry high correlation of H3K4me3 vs H3K27me3 in the adult lung. To do this, we took the local correlation track (*.wig *StereoGene* output file) and selected 3000 of the highest peaks using MACS-1.4.2 [47]. Then we selected the genes with TSS, which were located in the interval $\pm 5k$ around these peaks. The resulting list of genes was mined for biological enrichment using DAVID 6.7 software [14]. The most interesting terms that were found under $FDR < 5\%$ threshold were alternatively spliced mRNAs, cell motion regulation and apoptosis (Supplementary file 2).

Chromosome-specific correlation of promoter and polycomb marks

We compared the relationship between two well-investigated histone marks: the promoter-related H3K4me3, and the heterochromatin polycomb-related H3K27me3, in the adult lung, chromosome by chromosome (see fig. 4B). The genome-wide correlation (e.g. with all the chromosomes pooled) distribution for these marks is bimodal with a rather high peak on positive correlations. At the same time, the correlation distribution on chromosome 19 has a significantly different shape and is significantly less. This result could be explained by the fact that chromosome 19 has very high gene density and contains a lot of housekeeping genes. The correlation distribution on chromosome X also slightly differs from the genome-wide distribution and has a peak on very high correlations.

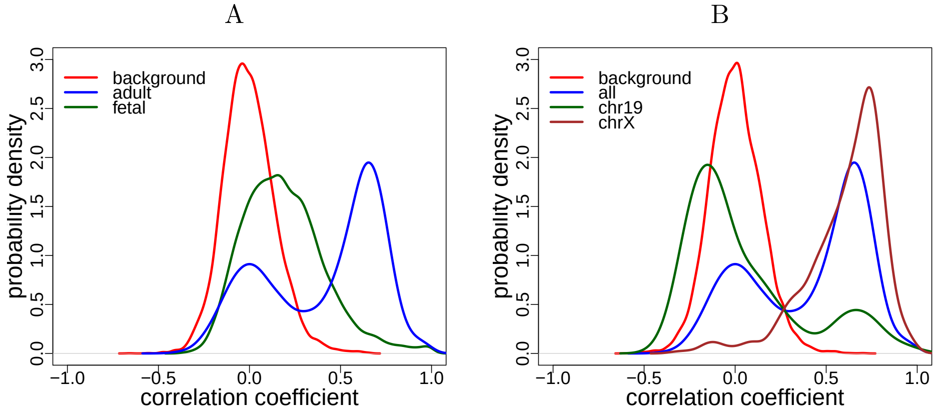


Figure 4. Distributions of correlations. A. H3K27me3 vs H3K4me3 in lung tissues. The background (red) coincides for adult lung (blue) and for fetal (green). B. Correlation distribution for H3K27me3 vs H3K4me3 in female adult lung. Red – background distribution; blue – correlation distribution over genome; green - correlation distribution for chr19, brown - correlations for X-chromosome.

Example of partial correlations

The H3K4me3 is an 'active promoter' mark and is expected to be positively correlated with RNA-seq. Indeed, fig. 5A shows some weak positive but statistically significant correlation. Interestingly, using a projection mode to remove H3K27me3 binding from the correlation of H3K4me3 with RNA-seq profile (fig 5B) produces a much stronger, and bimodal, correlation. This suggests that the relationship H3K4me3 to gene expression is modulated by H3K27me3 in some way. This observation is consistent with "poised promoters," in which the activating and repressive histone marks are both bound; these promoters are a subset of all genes and this unusual behavior runs contrary to what is seen in the majority of promoters. Here, we have uncovered multiple promoter states in addition to the multiple modes of interaction between H3K4me3 and H3K27me3.

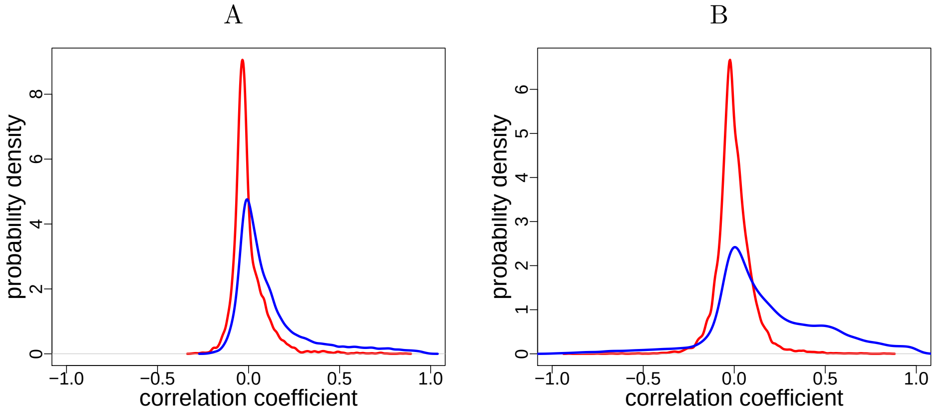


Figure 5. Distribution of correlations over the windows: A. Correlation of H3K4me3 vs mRNA-seq in Brain Hippocampus Middle; B. The same correlation with H3K27me3 removed as a confounder by partial correlation. Black line – foreground distribution; Gray line – background correlation distribution

Chromatin marks vs gene features

We separated genes on three fractions by expression level: genes with high expression level (top 25% of mRNA-seq level), genes with low expression (bottom 25% of mRNA-seq level) and moderately expressed genes (other genes) for certain cell type (Brain Cingulate gyrus) and plot cross-correlation function of histone marks vs gene features – start/end, and intron beg/end (Fig.6). Generally, we can see:

- Specific distribution of H3K4meX and H3K9ac near TSS – two high peak left and right to TSS and a gap at TSS position. This behavior is in an agreement with other research [9].
- Some specificity of the histone modification near intron boundaries. The sharp break of the H3K4meX and H3K9ac level of at the ends of the introns may be related to splicing definition. The H3K36me3 mark is usually related to PolII activity and the observed cross-correlation function can be explained by possible regulation of the intron and polyadenilation sites definition.
- The H3K27me3 mark usually is related to bivalent promoters. We can see that for active genes there exist rather strong peak right to TSS while for low expressed and silent genes this mark has wide peak.

Cohesin and histone modifications

We calculated the positional correlations of cohesin protein Rad21 with CTCF and different histone modifications in H1 stem cells (RRID:CVCL_9771) and in the K562 (RRID:CVCL_5145) cell line (table 1). We observed very strong positional correlation of the CTCF binding with cohesin protein Rad21. Promoter and enhancer regions (H3K4meX) were co-localized with cohesin while active transcribed regions and repressed regions were not related to cohesin. These observations are consistent with [38].

CAGE vs gene annotation

We analyzed the positional relationship of CAGE (FANTOM4 [30]) data, a genome-wide map of capped mRNA, for the nucleus and for cytosol of H1-hESC cells and the RefSeq [28] gene annotations.

The correlation functions are presented in fig.7. CAGE clusters are highly correlated with transcription start sites (fig. 7A), as expected. In addition, we observed two unexpected phenomena: strong positional correlation of CAGE clusters (panel B) with intron start sites and strong positional correlation of CAGE clusters with transcription termination sites (panel C). Both observations were relevant only when the CAGE clusters and genes were on the same strand, further supporting a meaningful biological relationship. More detailed analysis showed very precise localization of CAGE clusters at donor sites and at polyadenylation sites (fig. 7D). To check statistical significance of this observation, we selected equivalent random positions at 500 bp downstream from the donor splice sites or polyadenylation sites, as a control set. The resulting contingency tables are here 2. The Exact Fisher test for these contingency tables gave p-values less than $2.2 \cdot 10^{-16}$ in both cases.

CAGE association with intron starts may be explained by the activity of debranching enzymes [32]. After lariat debranching, the freed $5'$ end of the intron may become available for capping, and this cap would be detected by CAGE. Taft et al. [39] observed short (18-30 nucleotides) RNAs associated with donor splice sites. The authors

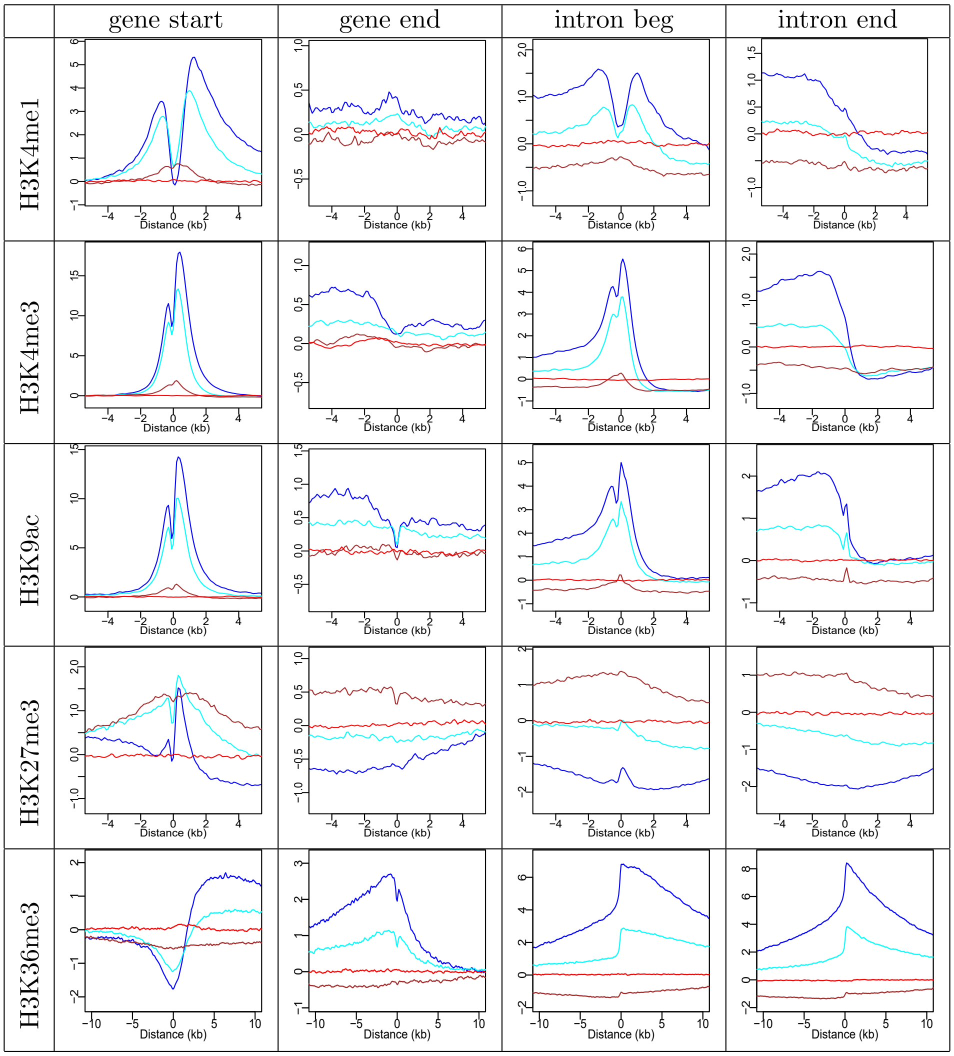


Figure 6. Cross-correlation function for different epigenomic marks and gene features. Blue line – active genes (top 25%); Cyan line – middle expressed genes Brown line – passive genes Red line – background cross-correlation function

suggested a model where RNA polymerase produced such transcripts on donor splice sites during mRNA transcription. The transcriptional stop site correlation is less evident, though suggests that occasional capping of the free 5′ end after cleavage by the polyadenylation complex is possible.

Discussion

We present a new method with unprecedented speed for estimation of genomewide positional correlations. As seen on public datasets, the approach yields biologically plausible results. The correlation distribution graphs depict multiple varieties of genomewide relationships. Local correlation tracks can be used for traditional gene enrichment analysis or to describe the relationship between genomic features. Currently, we use a permutation procedure to estimate statistical significance, still we can compare

Table 1. Correlations of cohesine Rad21 track vs histone modifications

Cohesine	Histone	avCorr	p-value
H1 stem cells			
Rad21	H3K9me3	0.01	9.18E-007
Rad21	H3k36me3	0.02	2.73E-023
Rad21	H3k79me2	0.04	1.70E-050
Rad21	H3K27me3	0.1	0.00E+000
Rad21	H4k20me1	0.12	0.00E+000
Rad21	H3k27ac	0.13	0.00E+000
Rad21	H3k4me3	0.14	0.00E+000
Rad21	H3k9ac	0.15	0.00E+000
Rad21	H3k4me2	0.19	0.00E+000
Rad21	H3k4me1	0.21	0.00E+000
Rad21	H2AZ	0.27	0.00E+000
Rad21	CTCF	0.9	0.00E+000
K562 cell line			
Rad21	H3k36me3	0.01	9.52E-001
Rad21	H3K27me3	0.02	1.19E-018
Rad21	H3K9me3	0.08	6.04E-283
Rad21	H3k79me2	0.11	0.00E+000
Rad21	H3k27ac	0.13	0.00E+000
Rad21	H3k9ac	0.15	0.00E+000
Rad21	H3K9me1	0.15	0.00E+000
Rad21	H3k4me3	0.17	0.00E+000
Rad21	H4k20me1	0.18	0.00E+000
Rad21	H3k4me2	0.18	0.00E+000
Rad21	H3k4me1	0.19	0.00E+000
Rad21	H2AZ	0.25	0.00E+000
Rad21	CTCF	0.78	0.00E+000

Table 2. Contingency tables for numbers of CAGE clusters starting at specific positions in comparison with +500bp control position.

Donor splice site

	CAGE	no CAGE
intron start	66181	320252
intron start+500	50	386383
p-value	$< 2.2 \cdot 10^{-16}$	

Poly-A sites

	CAGE	no CAGE
gene end	2393	42003
gene end+500	2	44394
p-value	$< 2.2 \cdot 10^{-16}$	

KC distributions for different comparisons. Thus, for coverage tracks that are supplied with the input track for control, we can use the values of KC between a feature and other feature's input or even between two inputs as a background instead of the

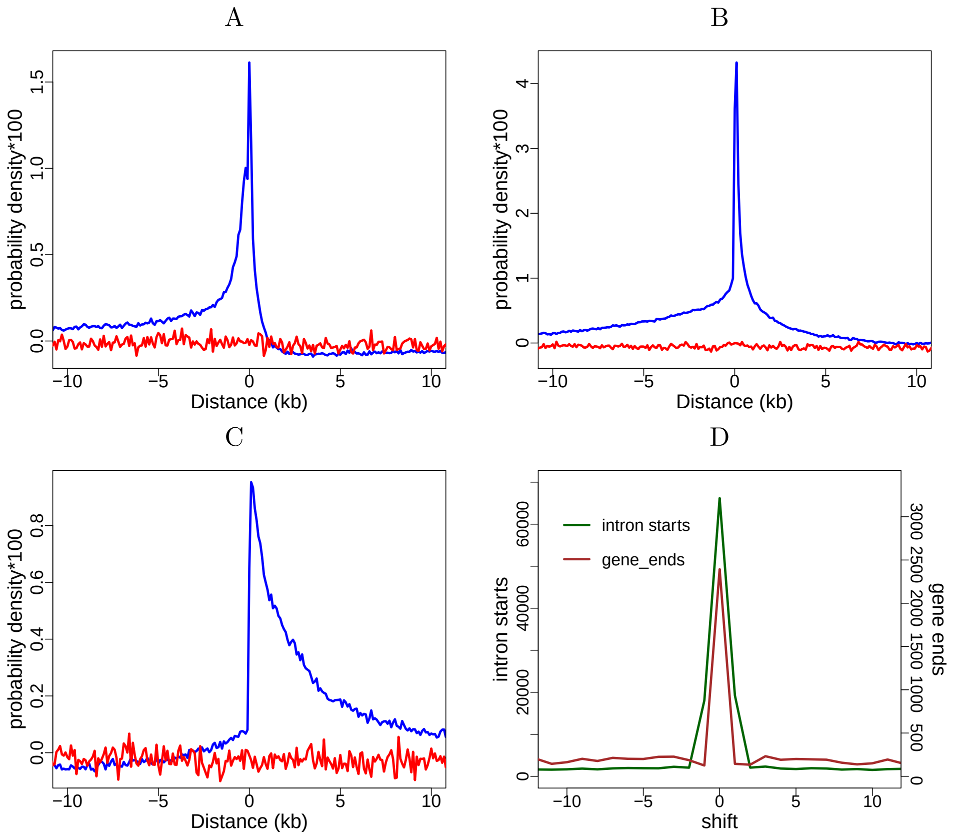


Figure 7. The cross-correlation function for CAGE vs gene annotation. Red lines show background (shuffled windows), blue lines show foreground (coherent windows). A. CAGE vs gene starts; B. CAGE vs intron starts; C. CAGE vs gene ends; D. CAGE vs intron starts (green) and gene ends (brown) at single nucleotide resolution.

permutation results.

StereoGene is also available as a Galaxy plugin, and we provide two examples, one using 2-way correlation and the other using partial correlation, to illustrate usage. In both cases, the user can save the correlation track and use these data for more complex queries.

We compare (Table 3) *StereoGene* with commonly used tools. Notably, very few programs can compute on continuous data (bedGraph, wig, etc) and require establishment of often arbitrary thresholds in order to create intervals for analysis. KLTepigenome [24] is able to work with continuous profiles, but is limited to sparse data and is quite slow even when compared to *StereoGene* doing the same computation on the full profile. We test the difference of the binarized profiles with different thresholds and continuous profiles. Usage of high threshold lead to overestimated correlations (Fig.8). *StereoGene* has additional, unique functions such as partial correlation analysis and the ability to compute over a linear combination of different profiles.

We applied *StereoGene* to continuous, interval, and pointwise genomic data, including experimental results and annotation tracks. In all cases, *StereoGene* produced reliable and sometimes nonobvious, yet intuitive, results that stimulate further investigation. *StereoGene* is thus a powerful and promising method for identifying genome-level biological patterns. The potential for guided 3-way (liquid) correlation is particularly novel and enables elucidation of the phenomena underlying complex

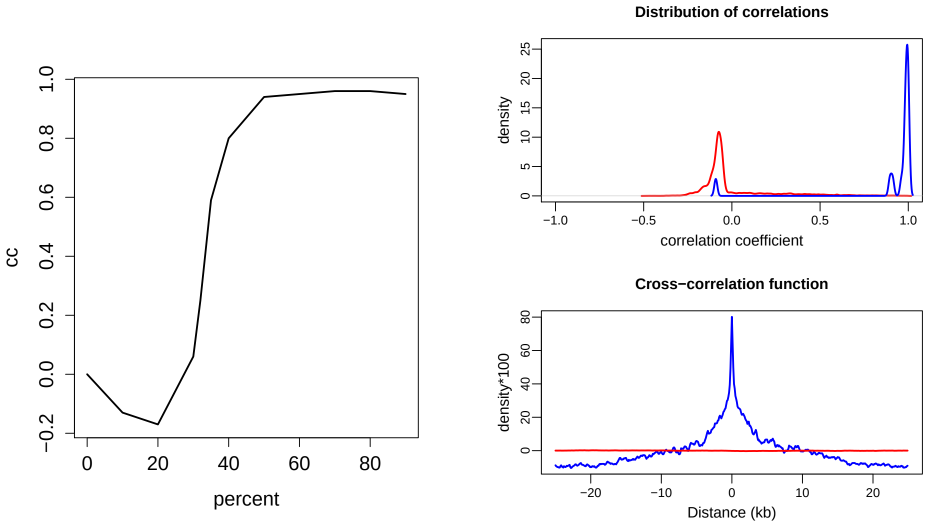


Figure 8. Left – dependence of the KC on binarization threshold. X-axis – threshold in percent of the maximum observed value in the profile; Y-axis – the KC value. Right – distribution of the KC over the windows (upper plot) and cross-correlation (lower plot) for binarized profiles for the threshold value = 50% of max value. The data is the same as on Fig.1

Table 3. Comparison of functionality for correlation analysis programs.

Program features	IntervalStats [6]	BEDTools [29]	GenomicRanges [17]	GenometriCorr [11]	Genome Track Analyzer [16]	KLTepigenome [24]	Genomic Hyper-Browser [34]	GAT [13]	Stereogene
correlate non-local features	+	+	+	+	+	+	−	+	+
interval profiles	+	+	+	+	+	+	+	+	+
work with continuous data	−	−	−	−	−	+	−	−	+
statistical evaluation	+	−	−	+	+	+	+	+	+
partial correlation	−	−	−	−	−	−	−	−	+
stratification by annotation	+	−	−	−	−	−	−	−	−
liquid correlation	−	−	−	−	−	−	−	−	+
produce correlation profile	−	−	−	−	−	−	−	−	+
cross correlation function	−	−	−	−	−	−	−	−	+

relationships.

Funding

This work was supported by Russian Scientific Foundation (grant 14-24-00155) and by National Institutes of Health (grant P30 CA006973). A.M. and A.F. were supported by

Russian Foundation for Basic Research (grants 14-04-01872 and 14-04-00576) S.J.W and T.N. were supported by Allegheny Health Network-Johns Hopkins Cancer Research Fund and JHU IDIES/Moore Foundation

Acknowledgments

We are grateful to Roman Kudrin, Ekaterina Khrameeva and Alexandra Golytsyna for testing the program. Thanks to Renat Arufilov, Artur Zalevsky and to Dmitriy Vinogradov for technical solutions and for support. Thanks to Aleksey Stupnikov for his ideas for the future. Thanks to Leslie Cope for his advice. Thanks to Patricia Palmer for her help with the text of the manuscript.

Supplementary files

prog_descr.pdf Description of the program

supplement1.pdf List of references to data sources.

supplement2.xls The file is a result of David [14] analysis of gene enrichment of top-3000 genes with high correlation of marks H3K4me3 vs H3K27me3 near TSS.

References

1. A. Arvey, P. Agius, W. S. Noble, and C. Leslie. Sequence and chromatin determinants of cell-type-specific transcription factor binding. *Genome research*, 22(9):1723–1734, 2012.

2. B. E. Bernstein, J. A. Stamatoyannopoulos, J. F. Costello, B. Ren, A. Milosavljevic, A. Meissner, M. Kellis, M. A. Marra, A. L. Beaudet, J. R. Ecker, P. J. Farnham, M. Hirst, E. S. Lander, T. S. Mikkelsen, and J. A. Thomson. The NIH roadmap epigenomics mapping consortium. *Nature Biotechnology*, 28(10):1045–1048, 2010.

3. L. Bintu, J. Yong, Y. Antebi, K. McCue, Y. Kazuki, N. Uno, M. Oshimura, and M. Elowitz. Dynamics of epigenetic regulation at the single-cell level. *Science*, 351(6274):720–724, 2016.

4. A. Boyle, C. Araya, C. Brdlik, P. Cayting, C. Cheng, Y. Cheng, K. Gardner, L. Hillier, J. Janette, L. Jiang, D. Kasper, T. Kawli, P. Kheradpour, A. Kundaje, J. Li, L. Ma, W. Niu, E. Rehm, J. Rozowsky, M. Slattery, R. Spokony, R. Terrell, D. Vafeados, D. Wang, P. Weisdepp, Y. Wu, D. Xie, K. Yan, E. Feingold, P. Good, M. Pazin, H. Huang, P. Bickel, S. Brenner, V. Reinke, R. Waterston, M. Gerstein, K. White, M. Kellis, and M. Snyder. Comparative analysis of regulatory information and circuits across distant species. *Nature*, 512(7515):453–456, 2014.

5. T. Chen and S. Dent. Chromatin modifiers and remodellers: regulators of cellular differentiation. *Nat Rev Genet.*, 15(2):93–106, 2014.

6. M. Chikina and O. Troyanskaya. An effective statistical evaluation of chipseq dataset similarity. *Bioinformatics*, 28(5):607–613, 2012.

7. E. P. Consortium. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, 2012.

8. N. R. Coordinators. Database resources of the national center for biotechnology information. *Nucleic Acids Res*, 44(D1):D7–D19, 2016.

9. J. Ernst and M. Kellis. Large-scale imputation of epigenomic datasets for systematic annotation of diverse human tissues. *Nat Biotechnol*, 33(4):364–376, 2015.

10. J. Ernst, P. Kheradpour, T. Mikkelsen, N. Shoresh, L. Ward, C. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, M. Ku, T. Durham, M. Kellis, and B. Bernstein. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–49, 2011.

11. A. Favorov, L. Mularoni, L. Cope, Y. Medvedeva, A. Mironov, V. Makeev, and S. Wheelan. Exploring massive, genome scale datasets with the genometricorr package. *PLoS Comput Biol*, 8(5):e1002529–e1002529, 2012.

12. K. Halachev, H. Bast, F. Albrecht, T. Lengauer, and C. Bock. Epiexplorer: live exploration and global analysis of large epigenomic datasets. *Genome Biol*, 13(10):R96–R96, 2012.

13. A. Heger, C. Webber, M. Goodson, C. P. Ponting, and G. Lunter. Gat: a simulation framework for testing the association of genomic intervals. bioinformatics. *Bioinformatics*, 29(16):2046–2048, 2013.

14. W. Huang da, B. Sherman, and L. R.A. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nat Protoc*, 4(1):44–57, 2009.

15. S. Kinkley, J. Helmuth, J. Polansky, I. Dunkel, G. Gasparoni, S. Fröhler, W. Chen, J. Walter, A. Hamann, and H. Chung. rechip-seq reveals widespread bivalency of h3k4me3 and h3k27me3 in cd4(+) memory t cells. *Nat Commun*, 7:12514–12514, 2016.

16. Y. Kravatsky, V. Chechetkin, N. Tchurikov, and G. Kravatskaya. Genome-wide study of correlations between genomic features and their relationship with the regulation of gene expression. *DNA Res*, 22(1):109–119, 2015.

17. M. Lawrence, W. Huber, H. Pagès, P. Aboyoun, M. Carlson, R. Gentleman, M. Morgan, and V. Carey. Software for computing and annotating genomic ranges. *PLoS Computational Biology*, 9, 2013.

18. G. Lev Maor, A. Yearim, and G. Ast. The alternative role of dna methylation in splicing regulation. *Trends Genet*, 31(5):274–280, 2015.

19. G. Li, X. Ruan, R. K. Auerbach, K. S. Sandhu, M. Zheng, P. Wang, H. M. Poh, Y. Goh, J. Lim, J. Zhang, H. S. Sim, S. Q. Peh, F. H. Mulawadi, C. T. Ong, Y. L. Orlov, S. Hong, Z. Zhang, S. Landt, D. Raha, G. Euskirchen, C.-L. Wei, W. Ge, H. Wang, C. Davis, K. I. Fisher-Aylor, A. Mortazavi, M. Gerstein, T. Gingeras, B. Wold, Y. Sun, M. J. Fullwood, E. Cheung, E. Liu, W.-K. Sung, M. Snyder, and Y. Ruan. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, 148(1-2):84–98, 2012.

20. K. Li, C. Liu, W. Sun, S. Yuan, and T. Yu. A system for enhancing genome-wide coexpression dynamics study. *Proc Natl Acad Sci U S A*, 101(44):15561–15566, 2004.

21. M. Libbrecht and W. Noble. Machine learning applications in genetics and genomics. *Nat Rev Genet*, 16(6):321–332, 2015.

22. C. V. Loan. *Computational Frameworks for the Fast Fourier Transform.* SIAM, 1992.

23. R. Luco, M. Allò, I. Schor, A. Kornblihtt, and T. Misteli. Epigenetics in alternative pre-mrna splicing. *Cell*, 144(1):16–26, 2011.

24. P. Madrigal and P. Krajewski. Uncovering correlated variability in epigenomic datasets using the karhunen-loeve transform. *BioData Min*, 8(14):20–20, 2015.

25. Y. Medvedeva, A. Lennartsson, R. Ehsani, I. Kulakovskiy, I. Vorontsov, P. Panahandeh, G. Khimulya, T. Kasukawa, FANTOM Consortium, and F. Drabløs. Epifactors: a comprehensive database of human epigenetic factors and complexes. *Database (Oxford)*, 2015:bav067–bav067, 2015.

26. A. Nag, V. Savova, H. Fung, A. Miron, G. Yuan, K. Zhang, and A. Gimelbrant. Chromatin signature of widespread monoallelic expression. *Elife*, 31(2):e01256, 2013.

27. A. Nag, S. Vigneau, V. Savova, L. Zwemer, and A. Gimelbrant. Chromatin signature identifies monoallelic gene expression across mammalian cell types. *G3(Bethesda)*, 5(8):1713–1720, 2015.

28. K. Pruitt, T. Tatusova, W. Klimke, and D. Maglott. Ncbi reference sequences: current status, policy and new initiatives. *Nucleic Acids Res*, 37((Database issue)):D32–6, 2009.

29. A. R. Quinlan and I. M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

30. T. Ravasi, H. Suzuki, C. V. Cannistraci, S. Katayama, V. B. Bajic, K. Tan, A. Akalin, S. Schmeier, M. Kanamori-Katayama, N. Bertin, P. Carninci, C. O. Daub, A. R. R. Forrest, J. Gough, S. Grimmond, J.-H. Han, T. Hashimoto, W. Hide, O. Hofmann, A. Kamburov, M. Kaur, H. Kawaji, A. Kubosaki, T. Lassmann, E. v. Nimwegen, C. R. MacPherson, C. Ogawa, A. Radovanovic, A. Schwartz, R. D. Teasdale, J. Tegnér, B. Lenhard, S. A. Teichmann, T. Arakawa, N. Ninomiya, K. Murakami, M. Tagami, S. Fukuda, K. Imamura, C. Kai, R. Ishihara, Y. Kitazume, J. Kawai, D. A. Hume, T. Ideker, and Y. Hayashizaki. An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, 140(5):744–752, 2010.

31. Roadmap Epigenomics Consortium, A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. Ziller, V. Amin, J. Whitaker, M. Schultz, L. Ward, A. Sarkar, G. Quon, R. Sandstrom, M. Eaton, Y. Wu, A. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R. Harris, N. Shoresh, C. Epstein, E. Gjoneska, D. Leung, W. Xie, R. Hawkins, R. Lister, C. Hong, P. Gascard, A. Mungall, R. Moore, E. Chuah, A. Tam, T. Canfield, R. Hansen, R. Kaul, P. Sabo, M. Bansal, A. Carles, J. Dixon, K. Farh, S. Feizi, R. Karlic, A. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. Mercer, S. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. Sallari, K. Siebenthall, N. Sinnott-Armstrong, M. Stevens, R. Thurman, J. Wu, B. Zhang, X. Zhou, A. Beaudet, L. Boyer, P. De Jager, P. Farnham, S. Fisher, D. Haussler, S. Jones, W. Li, M. Marra, M. McManus, S. Sunyaev, J. Thomson, T. Tlsty, L. Tsai, W. Wang, R. Waterland, M. Zhang, L. Chadwick, B. Bernstein, J. Costello, J. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. Stamatoyannopoulos, T. Wang, and M. Kellis. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, 2015.

32. B. Ruskin and M. Green. An rna processing activity that debranches rna lariats. *Science*, 229(4709):135–140, 1985.

33. M. Sachs, C. Onodera, K. Blaschke, K. Ebata, J. Song, and R.-S. M. Bivalent chromatin marks developmental regulatory genes in the mouse embryonic germline in vivo. *Cell Rep.*, 3(6):1777–1784, 2013.

34. G. K. Sandve, S. Gundersen, H. Rydbeck, I. K. Glad, L. Holden, M. Holden, K. Liestøl, T. Clancy, E. Ferkingstad, M. Johansen, V. Nygaard, E. Tøstesen, A. Frigessi, and E. Hovig. The genomic hyperbrowser: inferential genomics at the sequence level. *Genome Biology*, 11(12):1–12, 2010.

35. M. Schäfer, O. Lkhagvasuren, H. Klein, C. Elling, T. Wüstefeld, C. Müller-Tidow, L. Zender, S. Koschmieder, M. Dugas, and K. Ickstadt. Integrative analyses for omicsdata: a bayesian mixture model to assess the concordance of chip-chip and chip-seq measurements. *J Toxicol Environ Health A*, 75(8-10):461–470, 2012.

36. J. Severin, M. Lizio, J. Harshbarger, H. Kawaji, C. Daub, Y. Hayashizaki, FANTOM Consortium, N. Bertin, and A. Forrest. Interactive visualization and analysis of large-scale sequencing datasets using zenbu. *Nat Biotechnol*, 32(3):217–219, 2014.

37. M. Speir, A. Zweig, K. Rosenbloom, B. Raney, B. Paten, P. Nejad, B. Lee, K. Learned, D. Karolchik, A. Hinrichs, S. Heitner, R. Harte, M. Haeussler, L. Guruvadoo, P. Fujita, C. Eisenhart, M. Diekhans, H. Clawson, J. Casper, G. Barber, D. Haussler, R. Kuhn, and W. Kent. The ucsc genome browser database: 2016 update. *Nucleic Acids Res*, 44(D):D717–D25, 2016.

38. L. Steiner, V. Schulz, Y. Makismova, K. Lezon-Geyda, and P. Gallagher. Ctcf and cohesinsa-1 mark active promoters and boundaries of repressive chromatin domains in primary human erythroid cells. *PLoS One*, 11(5):e0155378, 2016.

39. R. Taft, C. Simons, S. Nahkuri, H. Oey, D. Korbie, T. Mercer, J. Holst, W. Ritchie, J. Wong, J. Rasko, D. Rokhsar, B. Degnan, and J. Mattick. Nuclear-localized tiny rnas are associated with transcription initiation and splice sites in metazoans. *Nat Struct Mol Biol*, 17(8):1030–1034, 2010.

40. H. Thorvaldsdòttir, J. Robinson, and J. Mesirov. Integrative genomics viewer (igv): high-performance genomics data visualization and exploration. *Brief Bioinform*, 14(2):178–192, 2013.

41. Z. Wang, C. Zang, J. A. Rosenfeld, D. E. Schones, A. Barski, S. Cuddapah, K. Cui, T.-Y. Roh, W. Peng, M. Q. Zhang, and K. Zhao. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature genetics*, 40(7):897–903, 2008.

42. S. Xiao, D. Xie, X. Cao, P. Yu, X. Xing, C. Chen, M. Musselman, M. Xie, F. West, H. Lewin, T. Wang, and S. Zhong. Comparative epigenomic annotation of regulatory dna. *Cell*, 149(6):1381–1392, 2012.

43. Y. Xu, M. Zhang, W. Li, X. Zhu, X. Bao, B. Qin, A. Hutchins, and M. Esteban. Transcriptional control of somatic cell reprogramming. *Trends Cell Biol*, 26(4):272–88, 2016.

44. D. Zerbino, N. Johnson, T. Juetteman, D. Sheppard, S. Wilder, I. Lavidas, M. Nuhn, E. Perry, Q. Raffaillac-Desfosses, D. Sobral, D. Keefe, S. Gräf, I. Ahmed, R. Kinsella, B. Pritchard, S. Brent, R. Amode, A. Parker,

S. Trevanion, E. Birney, I. Dunham, and P. Flicek. Ensembl regulation resources. *Database (Oxford)*, 2016:–, 2016.

45. Y. Zhang, H. Liu, J. Lv, X. Xiao, J. Zhu, X. Liu, J. Su, i. X. L, Q. Wu, F. Wang, and C. Y. Qdmr: a quantitative method for identification of differentially methylated regions by entropy. *Nucleic Acids Res*, 39(9):e58, 2011.

46. Y. Zhang, T. Liu, C. Meyer, J. Eeckhoute, D. Johnson, B. Bernstein, R. Nusbaum C., and Myers, M. Brown, W. Li, and X. Liu. Model-based analysis of chip-seq (macs). *Genome Biol*, 9(9):R137, 2008.

47. Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu. Model-based analysis of ChIP-seq (MACS). *Genome Biology*, 9(9):R137, 2008.

48. J. Zhou and O. Troyanskaya. Probabilistic modelling of chromatin code landscape reveals functional diversity of enhancer-like chromatin states. *Nat Commun*, 7:10528–10528, 2016.