Stat Biosci (2016) 8:284–309
DOI 10.1007/s12561-016-9145-0

Statistical Challenges in Analyzing Methylation and Long-Range Chromosomal Interaction Data

Zhaohui Qin[1] · Ben Li[1] · Karen N. Conneely[2] · Hao Wu[1] · Ming Hu[3] · Deepak Ayyala[4] · Yongseok Park[5] · Victor X. Jin[6] · Fangyuan Zhang[7] · Han Zhang[4] · Li Li[1] · Shili Lin[4]

Received: 23 November 2015 / Revised: 22 February 2016 / Accepted: 22 February 2016 /
Published online: 7 March 2016

Abstract With the rapid development of high-throughput technologies such as array and next generation sequencing, genome-wide, nucleotide-resolution epigenomic data are increasingly available. In recent years, there has been particular interest in data on DNA methylation and 3-dimensional (3D) chromosomal organization, which are believed to hold keys to understand biological mechanisms, such as transcription regulation, that are closely linked to human health and diseases. However, small sample size, complicated correlation structure, substantial noise, biases, and uncertainties, all present difficulties for performing statistical inference. In this review, we present an overview of the new technologies that are frequently utilized in studying DNA methylation and 3D chromosomal organization. We focus on reviewing recent developments in statistical methodologies designed for better interrogating epigenomic data, pointing out statistical challenges facing the field whenever appropriate.

✉ Shili Lin
shili@stat.osu.edu

1 Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA 30322, USA

2 Department of Human Genetics, Emory University School of Medicine, Atlanta, GA 30322, USA

3 Division of Biostatistics, Department of Population Health, New York University School of Medicine, New York, NY 10016, USA

4 Department of Statistics, The Ohio State University, Columbus, OH 43210, USA

5 Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15261, USA

6 Department of Molecular Medicine, The University of Texas Health Science Center at San Antonio, San Antonio, TX 78229, USA

7 Department of Mathematics & Statistics, Texas Tech University, Lubbock, TX 79409, USA

1 Brief Introduction

Propelled by rapid advances in high-throughput biotechnologies, our understanding of transcriptional regulation, a key mechanism of all living organisms, has improved dramatically over the past decade. It is now clear that DNA sequence alone does not provide full information; the complimentary epigenome carries an entire layer of regulatory information, including nucleosome positioning, DNA methylation, and 3-dimensional (3D) shape of chromatin. Understanding the epigenome sheds light on fundamental cellular processes as well as the molecular basis of human diseases. Despite much progress in genomic analysis, our understanding of the epigenome is lagging behind due to its diversity, complexity, and plasticity. One of the key challenges is the analysis and interpretation of epigenomic data. In this review, we strive to provide an up-to-date overview of the technological advances in this fast-evolving field, key characteristics of the data generated from these technologies, current state-of-the-art statistical methods, and remaining statistical challenges we face when analyzing such data. Our review centers on two important aspects of epigenomics: DNA methylation and spatial (or 3D) chromosomal organization, which we discuss in the following two sections. It is not surprising that many excellent review papers on these topics have already appeared in the literature [1–6]. In this review, we emphasize statistical aspects of epigenomic research, and we strive to present a comprehensive and contemporary view of the fields of DNA methylation and spatial chromosomal organization. Whenever appropriate, we further discuss the latest technologies and open problems to which biostatisticians and bioinformaticians may contribute to help advance epigenetic research.

2 DNA Methylation

DNA methylation is the cornerstone of the field of epigenomics. With rapid advances in sequencing technology, whole genome nucleotide-resolution methylation data are increasingly available, but obtaining such data is still very expensive and out of reach for most laboratories except on a small scale. Technologies for profiling whole genome methylation that provide regional rather than nucleotide resolution are also available and much more economical. Statistical analyses of data from each of these types of technologies present their unique challenges, but common themes exist as well, such as signal biases, small sample sizes, and spatial correlations. These, along with data-type specific issues, are discussed in the following subsections.

2.1 Review of Technologies

Multiple technologies have been developed to profile the methylome. They can be roughly classified into two broad categories: bisulfite conversion-based or capture-based. Both types of technologies have been coupled with microarray and sequencing platforms to produce high-throughput data.

2.1.1 Bisulfite Conversion-Based Technologies

Until the past decade, studies of DNA methylation were conducted on a small scale, but recent development of high-throughput assays has made genome-wide approaches possible. Commercial methylation microarrays produced by Illumina have been widely used due to their accessibility to investigators with a variety of backgrounds and resources. Since 2006, Illumina has produced increasingly dense methylation arrays. The GoldenGate methylation array covered 1,536 CpG sites, selected for their proximity to cancer-relevant genes [7]. The Infinium HumanMethylation27 BeadChip array covered 27,578 CpG sites selected to be in or near the promoter regions and CpG islands associated with 14,495 genes [8]. The Infinium HumanMethylation450 BeadChip array includes 482,686 CpG sites and 3,091 non-CpG loci, covering about 99 % of RefSeq genes and 96 % of CpG islands in the UCSC database [9]. Finally, beginning in 2016, it will be possible to assess >850,000 methylation sites using the Infinium MethylationEPIC BeadChip, including 90 % of sites found on the HumanMethylation450 BeadChip (http://www.illumina.com/techniques/microarrays/methylation-arrays.html).

The Illumina-array-based approaches rely on bisulfite treatment of DNA, which converts unmethylated cytosines to uracils, but leaves 5-methylcytosines unaffected. The converted uracils amplify as thymines during subsequent amplification, so the bisulfite-treated DNA can then be quantitatively "genotyped" to assess the proportion of DNA methylation levels in each sample at single-CpG resolution. All Illumina arrays perform the genotyping via bead-bound probes, though the genotyping assay varies across the three arrays. Respectively, the first three arrays rely on the GoldenGate assay [7], the Infinium I assay [8], and a combination of Infinium I and II assays [9]. Each of these assays allows for the estimation of a methylated (M) and an unmethylated (U) signal intensities; these signals can then be used to estimate the proportion of methylated cells in a sample as a β-value, where β is the ratio of methylated to total signal intensities M/(M+U).

Massively parallel sequencing, also known as next generation sequencing (NGS), has revolutionized genomics and epigenomic research due to its high sensitivity and specificity. Taking advantage of the new technologies, novel and powerful methylome profiling assays have emerged in recent years. Bisulfite sequencing (BS-seq) or MethylC-seq [10,11] also uses bisulfite treatment of DNA to determine the pattern of methylation status. As described above, bisulfite treatment yields specific modifications in the DNA sequence that depend on the methylation status of each individual cytosine. Therefore, BS-seq is able to produce single-base- resolution information about the methylation status of the entire genome, from which one can count the occurrences of methylated and unmethylated reads at a single-nucleotide resolution. However, Whole Genome BS-seq (WGBS) data are still expensive to generate. Reduced Representation BS-seq (RRBS) [12] data are more accessible, but with a much lower genomic coverage (<10 % of CpG sites [13]).

A limitation of bisulfite treatment is that it does not distinguish between 5-methylcytosine-based DNA methylation (5mC) and 5-hydroxymethylcytosine (5hmC), an oxidation product, because 5hmC is also resistant to converting to uracil [14]. Therefore, the measurements of BS-seq actually represent the levels of 5mC and

5hmC combined. This limitation may complicate the interpretation of results from BS-seq experiments, since 5hmC and 5mC have been found to have different functions. 5hmC is related to active gene transcription while 5mC is more likely to be repressive [15–17]. A recently developed approach, Tet-assisted bisulfite sequencing (TAB-seq) [18], provides a way to measure 5hmC specifically. In TAB-seq, 5hmC is first TET enzyme, while leaving the glycosylated 5hmC untouched. After bisulfite treatment, only 5hmC is read as cytosine in the resulting sequencing.

2.1.2 Capture-Based Technologies.

Capture-based technologies rely on the pulldown of methylated DNAs instead of base conversion. The earliest microarray platform was differential methylation hybridization (DMH)[19]. DMH is a high-throughput DNA methylation profiling tool that utilizes methylation-sensitive restriction enzymes to survey methylated fragments by hybridizing them to a CpG island microarray. This array contains probes covering all of the 27,800 CpG islands that were annotated in the UCSC Genome Browser at the time. Quality control and normalization are critical for detecting probes or CpG islands that are differentially methylated under different conditions [20,21]. Other capture-based microarray technologies have also been developed including the popular MeDIP [22]. In recent years, various pulldown technologies have been coupled with NGS, leading to multiple whole genome methylation platforms, including MeDIP-chip [23], MeDIP-seq [24], MethylCap-seq [25], and MBD-seq [26]. For MeDIP-seq, the pulldown is through antibody-based immunoprecipitation [27]. In MethylCap-seq, on the other hand, the pulldown is accomplished by the use of methyl-binding proteins such as modified human MeCP2. Any fragment with at least one methylated CG site will be pulled down [25]. For MBD-seq, fragmented genomic DNA of 50–350 bp in length is subjected to MethylMiner^{TM} methylated DNA kit (Thermo Fisher Scientific, Waltham, MA) enrichment, which uses a recombinant form of the human MBD2 protein, and methylated fractions are eluted with salt [26,28]. In all these methods, these pulldown fragments are then sequenced and aligned to the reference genome.

There are a number of differences between capture-based and bisulfite-converted data. (1) Capture-based data measure the enrichment of a region with a certain length, as such the data are not of "nucleotide resolution," and the pulldown fragments are biased toward dense CG regions. (2) A single-end short read from a pulldown fragment may cover 0, 1, or more CG sites; however, it is unknown as to which CG site(s) is (are) responsible for the fragment being pulled down and sequenced [29]. With a region/window-based approach, there can be "phantom reads," i.e., reads mapped to a window that lacks CG sites, rendering the analysis completely meaningless [29]. This adds one more layer of difficulty in data analysis, as we will discuss below. (3) Capture-based technologies are much cheaper compared to WGBS. Despite the base-resolution accuracy, the cost of WGBS is extremely high. It usually requires one to sequence at least 5× the genome size in order to get complete coverage of the 22–28 million CGs sites in the whole genome. Thus, it is unrealistic to apply it to conduct an experiment with more than tens of patient samples.

Summaries of the various methods, including coverage, resolution, and key preferences, are presented in Table 1.

Table 1 DNA methylation profiling technologies

Methods	Total coverage	Resolution	Reference
Bisulfite converted			
Illumina 450K	482K	Base-pair level	[9]
WGBS	Whole genome	Base-pair level	[10,11]
RRBS	1–4 million	Base-pair level	[12]
TAB-seq	Genome-wide	Base-pair level	[18]
Capture- based			
DMH	Genome-wide	Probe	[19]
MeDIP-chip	Genome-wide	Locus, low.	[23]
MeDIP-seq	Genome-wide	DNA fragment	[24]
MethylCap-seq	Genome-wide	DNA fragment	[25]
MBD-seq	Genome-wide	DNA fragment	[26]

2.1.3 Single-Cell Technology

In addition to the two most popular types of technologies for capturing and quantifying DNA methylation, a single-cell method has been recently proposed. DNA methylation patterns can be extremely heterogeneous even between different cells of the same cell type. The methods discussed above can only capture a summary of the methylation features across many cells because the data generated from those methods are based on thousands or millions of mixed cells. To study the heterogeneity of DNA methylation, single-cell BS-seq (scBS-seq) was developed [30]**.** This technology also depends on bisulfite treatment. However, for scBS-seq, single cells are isolated and lysed during library preparation, prior to bisulfite conversion, PCR amplification, and sequencing. Another distinct difference between the protocol of scBS-seq compared to BS-seq is that tagging to the DNA segment is applied after bisulfite conversion to reduce severe information loss of bisulfite conversion due to DNA degradation. This newly developed post-bisulfite tagging technology minimizes information loss, making scBS-seq possible.

2.2 Biological Problems and Statistical Challenges

2.2.1 Quality Control and Normalization Procedures for Arrays

Processing of data from Illumina methylation arrays via Illumina's GenomeStudio software yields methylated (M) and unmethylated (U) signals intensities for each sample and CpG site, as well as "detection p-values" that indicate whether the total signal is significantly greater than noise, as assessed using negative control probes included on the array. Typical quality control procedures include setting to missing data points with high detection p-values (commonly $>.05$ or $>.01$, though Lehne et al. observed improved reproducibility among technical duplicates when 10^{-16} is used as a cutoff [31]) and removing CpG sites or samples with high proportions (e.g., >5 %) of missing values. These procedures are not available within GenomeStudio, a software

suite developed by Illumina to visualize and analyze data generated on Illumina array platforms (Illumina Inc, Carlsbad, CA), but are performed routinely in most available software or pipelines for analysis of Illumina 450K data (e.g., [31–38]). It is also common to filter out CpG sites that have non-specific probes or include genetic variants in the probe site [39]. Identification of extreme outliers may be performed in GenomeStudio using hierarchical clustering or in downstream analysis using approaches such as principal component analysis (e.g., [40]), multidimensional scaling (e.g., [41]), or identification of outlying values for each CpG site based on the interquartile range (e.g., [31,42]).

Depending on the study design, β-values can be computed from the U and M signals and analyzed without further normalization to adjust for batch differences between CpG sites or between samples. For studies where each CpG site is analyzed separately, this strategy is possible because between-CpG differences will not influence the single-CpG analyses and technical factors that inflate or deflate individual signals will tend to cancel out when the ratio $\beta = M/(M+U)$ is used. However, between-array normalization methods can be used to remove technical differences between samples that may influence the global signal patterns (e.g., [31,37,43]). Similarly, a number of within-array normalization methods have been developed to address the presence of two types of CpG probes on the 450K array (e.g., [37,44–48]). The Infinium I and II probes rely on two different assays, resulting in different sources of bias and different distributions of estimated β-values (described further in [9,49]), and the goal of within-array normalization approaches is to minimize technical differences between β-value distributions across probe types. Approaches for between- and within-array normalization have been reviewed in detail elsewhere [50]. Two studies recently assessed the performance of available methods for between- and within-array normalizations by comparing reproducibility among technical duplicates and performance metrics based on the results of simulated or real analyses. Wu et al. [49] compared four normalization procedures ([45–47,51]) to non-normalized data; they observed high reproducibility in the non-normalized data and noted that while some normalization approaches could slightly increase reproducibility, others led to decreases in reproducibility. Overall results tended to be similar irrespective of how and whether the data were normalized, though single-CpG association analyses yielded the highest proportion of validated results in a split-sample experiment when non-normalized data were analyzed [49]. Lehne et al. compared ten normalization procedures (four variations of quantile normalization as well as the procedures described in [43–47,51]) to non-normalized data and concluded that quantile normalization of the signals (sub-divided by the two probe types as well as probe subtype and color channel) led to the greatest reproducibility, sensitivity, and specificity [31]. Although many normalization procedures have now been proposed, the field has still not reached consensus on the optimal normalization approach for microarrays. Further studies to assess the effects of different normalization procedures on reproducibility of data across duplicate samples may be needed to resolve this lack of consensus. With the release of the denser MethylationEPIC BeadChip in 2016, an emerging challenge will be to characterize relevant features of its design, and assess whether existing pipelines can be adapted for quality control and normalization or whether new approaches are needed.

2.2.2 Quality Control and Sequence Alignment for BS-Seq

Quality control is an important first step for BS-seq data analysis. A common artifact in BS-seq data is the 5′ end bias. It is reported that the bisulfite conversion failure is enriched in the 5′ end, which results in artificially higher methylation level at the start of the read. Also, as with other sequencing data, reads from BS-seq suffer from lower quality toward the 3′ end. An easy remedy to this problem is to trim the reads at both ends. Several tools have been designed to automatically perform such tasks (BisSNP [52], BSeQC [53]).

Alignment of sequence reads from BS-seq is more complicated than in other sequencing experiments. Using aligners designed for general sequencing data (such as Bowtie [54]) is feasible, but will lead to lower alignment efficiency because of the C-T/G-A mismatches caused by bisulfite conversion. Software to align the BS-seq reads has been made available, including Bismark [55], BSMAP[56], RMAPBS [57] and BSmooth-align [58]. The main idea behind these aligners is to modify both the reference genome and the reads *in silico* to mask the C-T/G-A mismatches before alignment. Performance of the different aligners has been compared [59,60], demonstrating various biases affecting the estimation of methylation levels. For example, GC content and number of PCR cycles have been reported to affect the enrichment of highly methylated DNA [61]. During the read mapping step, since bisulfite treatment converts unmethylated cytosines only, sequencing reads with more unmethylated cytosines have more matched bases and are more difficult to align with the reference genome. This may potentially increase the proportion of methylated cytosines and result in inflated methylation levels.

2.2.3 Detection of DML Using Methylation Array Data

Methylation array data are commonly used to perform differentially methylated loci (DML) or epigenome-wide association studies (EWAS), in which each CpG is tested separately for association with a trait, exposure, or biological condition of interest. These single-CpG analyses are often performed as regression-based analyses, with the β-value (or M-value, its logit transform [62]) as the outcome, and the trait/condition of interest and other covariates as independent variables. Because each methylation β-value is a rough approximation of the proportion of methylated DNA among a very large number of DNA strands, the central limit theorem may lead to the validity of the assumption of normally distributed errors if the regression model is correctly specified, with appropriate covariates. Confounding factors need to be taken into consideration in order to remove unwanted signals. As described below, typical confounding factors in cross-sectional studies will likely include age, sex, race/ethnicity of subjects, cell type or tissue heterogeneity between samples, and technical factors such as batch effects.

Both age [63–69] and ancestral population [70,71] have well-documented associations with DNA methylation at many sites across the genome. It is straightforward to include age as a covariate in regression-based analyses. Self-reported measures of ancestral population can be included as covariates or stratum, but often this confounding can be more accurately accounted for by including as covariates principal

components from genome-wide genotype information [71]. If genotypic information is unavailable, the principal components from genome-wide methylation can also be used to adjust for this source of confounding [71].

Cell type heterogeneity is a potential source of confounding in DNA methylation studies of human blood samples, since (1) the component cell types of whole blood have distinct methylation profiles [72] and (2) cell type proportion may vary according to the trait/condition of interest. Houseman et al. [73] proposed a regression calibration method to estimate individual cell type proportions from Illumina array data based on a set of reference samples with known cell type composition (e.g., [72]). This method has been implemented in minfi [32,74] and it has become common to include estimated cell type proportions as covariates in regression-based analysis (e.g., [75–77]) Because this method depends on the availability and quality of reference samples, reference-free approaches to adjust for cell type heterogeneity and other sources of biological confounding have also been proposed [78,79].

Technical sources of confounding include experiment batch effects, chip effects, and positional effects on the chip. While removal of these effects is the goal of the between-sample normalization procedures described above, it is also common to perform adjustment within the regression model by including fixed or random effects for these variables [33], through inclusion of principal components (PC) of the methylation data (e.g., [80]), or through other PC-based approaches [31,81]. Notably, these PC-based approaches have the same goals as the reference-free methods for biological confounding mentioned above, and all of these methods have the potential to adjust for both biological and technical sources of confounding.

Illumina methylation microarrays have the advantage of providing single-CpG resolution at a low cost, thus enabling DML analyses in larger well-powered sample sets. A disadvantage of these arrays is that their coverage is limited (less than 2 % genome-wide) and not equally representative of all CpG sites across the genome (mostly in CpG islands). In genetic association studies, patterns of linkage disequilibrium (i.e., correlation) between genetic variants are long-ranging and static, allowing for the use of genotyped genetic variants as proxies for untyped variants. In epigenetic studies, correlations between CpG sites are dynamic, context-specific, and unpredictable, so are less reliable as proxies for one another or for a region in general. For regional analyses, denser data are likely needed, as described in the sections below.

2.2.4 Detection of DML and DMR Using BS-Seq Data

Differential methylation analysis from BS-seq data can be performed at single-nucleotide or regional levels to detect differentially methylated loci (DML) or regions (DMR). DML analysis is often performed when the data are not from whole-genome scale (e.g., RRBS), or when the methylations are sparse (e.g., hydroxymethylation from TAB-seq). DMR analysis is more typically used when data are from WGBS. Both analyses start by performing individual statistical tests for all CpG sites. Sites with measures of statistical significance (e.g., p values) surpassing a user-specified threshold are deemed DML. To define DMRs, it is required that consecutive CpG sites

are significant. Additional criteria such as minimum region length in base pairs or number of CpG sites are often imposed.

BS-seq data provide single-nucleotide-resolution information of methylation levels, which makes the study of DML possible. However, the study of site-specific methylation levels is heavily dependent on the number of sequencing reads that cover each specific CpG site. In WGBS experiments, most CpG sites are covered by very few sequencing reads, which significantly reduces the detecting power of DML. Therefore, data tiling (methylSig [82], methylKit [83]) or smoothing techniques (BSmooth [58], Biseq [84]) are often applied. If biological events are more likely to be defined by regional rather than single-CpG methylation changes, then identifying differentially methylated regions (DMRs) will provide more stable and biologically meaningful results.

The key component of most DML/DMR detection methods is the use of a statistical test at each CpG site. The observed data can be summarized as counts of total and methylated reads, and the null hypothesis is that the methylation levels are not associated with the biological factors of interest. The count data may be modeled as binomial distribution (when there are no biological replicates) [58] or beta-binomial distribution (when biological replicates are present, to allow over-dispersion to account for biological variance) [85–87]. There are several other important issues to consider in DML/DMR detection. First, with WGBS-seq data, correctly accounting for spatial correlations of methylation levels can greatly improve the power. Currently, several different types of smoothing approaches are available, including BSmooth [58], methylSig [82], and Biseq [84]. However these approaches tend to smooth out the higher frequency signals (such as the sudden drop in methylation typically occurring near CpG island shores), which may hurt the resolution of detected DMRs. Second, estimating the biological variance is vital, especially when the number of replicates is low. To improve the estimation of biological variance, methods such as methylSig [82] and DSS [85] have extended ideas from differential gene expression analysis to borrow information from genome-wide data. These methods employ a beta-binomial model to characterize the count of methylated reads at each CpG site, and derive an empirical Bayes (EB) "shrinkage" estimator for estimating the biological variance (represented by a dispersion parameter). Finally, the sequencing depth of the CpG site needs to be considered in the statistical test. Some methods filter out sites with low depth, but this will result in information loss. Wald- or likelihood ratio-test procedures have been developed and implemented in methylSig [82] and DSS [85] to incorporate sequencing depth information in the test procedure.

Over the last several years, a number of statistical methods and software tools for DML/DMR detection have been developed. Robinson et al. provides an informative review of existing methods [88]. A comprehensive and objective comparison of the methods is still lacking, partly because it is difficult to obtain gold standards. In spite of this issue, there are still ample opportunities for statistical method development in this area. For example, methods for DML/DMR detection under general experimental design (as opposed to simple two-group comparisons) are needed. Currently, the only such methods are RADMeth [89] and BiSeq [84], both based on generalized linear models (GLM). However, running GLM at each CpG site will be very computationally intensive. Moreover, GLM procedures can be numerically unstable, especially when

the methylation levels are close to the boundaries (0 or 1). A more efficient and stable method applicable to general design is still needed. In addition, methods for detecting differential methylation region for 5-hydroxymethyl cytosine (DhMR) from TAB-seq data are also lacking; currently, DMR calling methods are being used with no further adaptation. Since TAB-seq data possess different characteristics (weak spatial correlation, low hydroxymethylation levels, etc.), a customized approach is likely needed. Finally, for analyses of BS- and TAB-seq data from the same sample, it will be of interest to develop methods to jointly detect DMR and DhMR.

2.2.5 Detection of DMR for Capture-Based Data

Data from capture-based methylation experiments (such as MeDIP-seq) have similar characteristics as ChIP-seq data, so DMR detection is often performed using peak calling software designed for ChIP-seq such as MACS [90], HPeak [91] and CisGenome [92]. For comparisons of two or more groups, an easy approach is to call "peaks" separately for each group, and perform overlapping analysis to determine DMRs. However, this approach ignores the quantitative differences of methylation levels and thus could lead to undesirable results. A number of methods have been developed to perform quantitative comparison of ChIP-seq data including, QChIPat [93], DBChIP [94], MAnorm [95], ChIPComp [96], diffReps [97], DIME [98], ChIPnorm [99], and MMDiff [100], all of which could be used for DMR calling from captured data. In particular, MEDIPS [101] is specifically designed for MeDIP-seq data, and implemented as an easy-to-use, well-documented Bioconductor package. Standard statistical tests, such as the Student's t test, have been applied to MBD-seq and other captured data to detect DMR for a predefined region [93,102,103]. However, such "averaging" approaches ignore intrinsic correlations and may wash out non-homogeneous signals.

It is important to note that DMR calling from capture data may be more complicated than quantitative comparison of ChIP-seq data for several reasons. First, since methylation events are much more prevalent than protein binding, the number of peaks detected from MeDIP-seq is likely to be much greater. This results in a larger test space in quantitative comparison, which makes the multiple testing problem more severe. Moreover, the signal-to-noise ratio from MeDIP-seq data is usually much lower than from ChIP-seq, also due to the prevalence of methylation, which further undermines the statistical power of the test. Finally, it is known that the captured methylation data are severely affected by CpG density. Thus, the DMRs called are usually biased toward CpG-dense regions. For these reasons, single-base technology (BS-seq) is generally considered the gold standard for the purpose of DMR detection, but in large-scale population-level studies, the cost of WGBS is often prohibitive. In this case, a solution could be to perform control experiments to provide an estimate of background noise to facilitate unbiased DMR calling.

There are other methods for detecting DMRs using capture data that do not rely on "peak" identification in the first step, including window-based approaches such as MEDIPS [101]. Such methods may model counts as negative binomial after normalization, but they can encounter difficulty in interpretation of the results, including concerns similar to those described for the region-based t test above. A different

method is a two-step approach [29]. The first step is a probabilistic model, PrEMeR-CG, to distribute each read to CGs that may contribute to the pulldown according to the distribution of the fragment-length library. This will create single nucleotide level data, but with relative methylation levels, rather than read counts, for each CpG site. Signals of neighboring CpG sites will be highly correlated [30], and therefore results from a DML analysis would be difficult to interpret. As such, methods for DMR detection have been proposed, but negative binomial modeling can no longer be used in this context. Summing over all signals in a region and performing t-tests is one potential approach, but doing so would ignore the correlation and could lead to higher false positive rates. To take correlation into account, MethMage [29], based on generalized estimating equations (GEE) [104], can be used with an auto-regressive (AR1) spatial structure to construct a working correlation matrix. However, GEE is computationally very expensive. As an attempt to address this issue, a class of procedures based on high-dimensional mean vector tests has been recently proposed as an alternative for the detection of differentially methylated regions. These approaches do not need to assume a specific correlation structure. Moreover, unlike Hotelling's T^2, these approaches can deal with the situation in which the number of CG sites in the region exceeds the number of samples [105].

2.2.6 Other Areas of Inquiry and Novel Statistical Challenges

Assessment of DNA methylation patterns across the genome DNA methylation is heterogeneous even among the same type of tissue within the same individuals. This heterogeneity of DNA methylation patterns may be partially responsible for the heterogeneity of the cell populations. Current NGS technologies provide information from sequencing reads, where each read is from a single cell, thus enabling the study of cell-specific DNA methylation patterns. Xie et. al and Shao et al. used an entropy concept to study genome-wide variation in DNA methylation patterns in individual sequencing reads [106,107]. They model the frequency of distinct methylation patterns observed within a specific genomic region as the probability of an event in a Shannon entropy equation. However, because of the short length of sequencing reads ($\sim$100 bp) and need for each read to have at least several common CpG sites, their approach can only be used to study CpG-dense regions such as CpG islands.

Methods for analysis of single-cell DNA methylation The recently developed technology scBS-seq has opened the door to study cell-specific methylation patterns [30]. Each dataset generated by scBS-seq provides methylation information for a single cell. Although the short length of sequencing reads generated by scBS-seq is not a major concern, the low genomic coverage ($\sim$20 % of CpGs) presents a major statistical challenge in characterizing cell-specific information. Another challenge in the analysis of this type of data in diploid organisms is the presence of allele-specific methylation patterns. These challenges and others must be addressed to facilitate the identification of global and local methylation levels along with spatial methylation patterns within each cell.

3 Three-Dimensional Chromosomal Organization and Long-Range Interaction

The organization of a eukaryotic genome in the three-dimensional (3D) space is not random. The highly structured, hierarchically organized 3D architecture is closely linked to genome functions, cellular processes, and disease mechanisms [108,109]. It has long been observed that, in mammalian genomes, a gene may be regulated by distal enhancers and repressors that are not necessarily on the same chromosome. Such communication between distal elements is achieved through the non-random spatial organization (looping) of the chromosomes, which brings genes and their regulatory elements into close proximity. Due to the complex nature of chromatin interaction data, standard statistical methods are not applicable, and thus methods tailored to such data need to be developed. There are many issues and challenges, which we discuss below.

3.1 Review of Technologies

Since the debut of the chromosome conformation capture (3C) assay in 2002 [110], many variants that are higher throughput, including those coupled with NGS to generate genome-wide chromatin interaction data, have been proposed to great success. The two main technologies to date are Hi-C [111] and ChIA-PET [112], but other newly proposed and high-resolution technologies will be discussed as well.

3.1.1 Hi-C

Traditionally, scientists used microscopy-based techniques to study genome spatial organization [113]. While very successful, these techniques are limited by their low throughput (a few loci, usually less than ten) and low resolution (each locus corresponds to a 40-kb region). More importantly, as a single-cell level assay, it is almost impossible to scale up to measure the structural properties of the entire cell population, which usually consists of millions of cells. To overcome the limitations of microscopic-based techniques, a series of molecular techniques based on the concept of 3C have been developed in recent years. Harnessing the power of next generation sequencing technologies, Lieberman-Aiden et al. devised the revolutionary Hi-C technology, enabling a high-resolution, genome-wide 3D view of chromosomal organization [111]. Hi-C represents a breakthrough in studying chromosomal organization, and the technology was rapidly adopted by scientists and applied to multiple species, resulting in a series of landmark discoveries, which include the demarcation of physical domains [114–117], widespread chromosomal rearrangement during stress [118], and the roles genome organization played in recurrent chromosomal translocations [119]. Multiple variations of the Hi-C technology were also introduced, including tethered conformation capture (TCC) that modifies and enhances the experimental protocol in Hi-C [120]; the in situ Hi-C that produces the finest resolution (10 kb) data to date from the intact nucleus [121], and the capture Hi-C that focuses on a specific set of loci in the genome [123].

3.1.2 4C-Seq

Circular chromosome conformation capture (4C) [122] is an adaptation from 3C that processes a ligated 3C template with another round of DNA digestion and ligation to form small DNA circles. By using primers adjacent to the cutting sites of the viewpoint region of interest, inverse PCR only amplifies sequences with one end coming from the viewpoint region. The library is sequenced and mapped to obtain the genomic positions of the other ends of the ligations, thus obtaining information of loci that are interacting with the viewpoint region [123]. As a one-to-multiple strategy for detecting interactions, 4C-seq focuses on interactions with a single locus of interest, thereby reducing its sequencing cost compared to Hi-C.

3.1.3 ChIA-PET

ChIA-PET is another technology to detect long-range interaction. However, it is more targeted in that it only detects interactions in the genome that are mediated by a particular protein of interest, for example, PolII, AR, or ER. In other words, one can view the set detected by ChIA-PET as just a subset of the total interactions in the genome. The protocol of ChIA-PET is very similar to that of Hi-C, but with an additional pulldown (Immunoprecipitation) step to select only loops involving the particular protein of interest. Figure 3 of Steensel and Dekker [124] provides an excellent summary of the similarity and differences of the Hi-C and ChIA-PET technologies.

3.1.4 Single-Cell Hi-C

While the Hi-C technology was designed for measuring population average genome organization, a modified technology, single-cell Hi-C [125], has been developed to study each cell individually. As a complement of the single-cell level microscopic-based method, single-cell Hi-C shows cell-to-cell variation of chromatin structure. However, interpretation of single-cell Hi-C data is extremely challenging due to the sparsity of such data and the limited sequencing depth. Further optimized experimental protocols and advanced statistical and computational models are necessary to fully process the rich information contained in single-cell Hi-C data, which are currently lacking.

Summaries of the various aspects of the technologies, including sample, resolution, and key references, are presented on Table 2.

3.2 Biological Problems and Statistical Challenges

3.2.1 Hi-C Normalization

Similar to other types of next generation sequencing data such as ChIP-Seq, RNA-Seq, and BS-Seq, Hi-C data contain multiple layers of biases due to complex experimental protocols. Effective and efficient removal of such biases poses great statistical and

Table 2 Methods for characterizing spatial chromosomal interactions

Methods	Sample	Target interactions	References
Hi-C	Population of cells	All interactions	[111]
4C-seq	Population of cells	Interactions originated from a single locus	[122]
TCC	Population of cells	All interactions	[120]
In situ Hi-C	Population of cells	All interactions	[121]
Capture Hi-C	Population of cells	Interactions among a set of loci	[126]
Single-cell Hi-C	Single-cell	All interactions	[125]
ChIA-PET	Population of cells	Specific protein mediated interactions	[112]

bioinformatic challenges. Since the publication of the first Hi-C study [111], two types of computational algorithms have been developed to remove biases in Hi-C data. One type of algorithm focuses on the removal of bias through modeling. For example, Yaffe and Tanay first identified three major bias sources in Hi-C data: restriction enzyme fragment length, GC content, and mappability score [127]. They proposed a highly over-parameterized probabilistic model to remove these three biases. Later on, Hu et al. developed HiCNorm, a Poisson regression model to normalize Hi-C data [128]. Compared to Yaffe and Tanay's method, HiCNorm is much simpler, achieves better bias removal and is more than 1000 times faster. Meanwhile, Cournac et al. proposed the SCN normalization procedure [129], which is designed for removing circulation biases specific to the bacteria circular genome.

The second type of algorithm focuses on the removal of bias through normalization based on matrix balancing theory. As the first method of this kind, Imakaev et al. developed the algorithm ICE, aiming at removing all known and unknown Hi-C biases [130]. A similar algorithm was utilized in a recent ultra-high-resolution Hi-C study [121]. These matrix balancing-based methods assume "equal visibility," i.e., all genomic loci are expected to have equal total number of contact when no bias exists.

Due to the lack of gold standards (such as large-scale microscopic data) for genome spatial organization, a thorough and fair evaluation of the performance of these algorithms is extremely challenging. We expect microscopic data to become increasingly available in the near future, which will motivate the development and evaluation of novel Hi-C normalization algorithms with improved efficiency and enhanced effectiveness.

3.2.2 Identification of Topologically Associated Domains (TADs) and Their Boundaries

Another important bioinformatics problem is the identification of borders of topologically associated domains (TAD). Several TAD border callers are publicly available. One is a hidden Markov model-based TAD border caller, which explicitly models the imbalanced directionality of pair-end reads within TADs and between TADs [115]. Another approach is based on block segmentation by maximizing a likelihood through

dynamic programing [131]. To identify TAD borders from the ultra-high-resolution in situ Hi-C data, an Arrowhead algorithm was proposed and utilized [121].

The hierarchically organized TADs pose a great challenge for developing TAD border callers, making the criterion for calling TAD borders dependent on the specific biological question. Furthermore, the lack of a gold standard poses another challenge for validating TAD border callers. Statistically, a robust TAD border caller should provide highly reproducible results across biological replicates. In addition, a biologically meaningful TAD border should show significant difference in Hi-C contact frequency between intra-TAD interactions and inter-TAD interactions. Biologically, we expect that biologically meaningful TAD borders show significant enrichment of housekeeping genes, key transcription factors such as CTCF and multiple insulators. In the near future, with the accumulation of Hi-C data and gold standard microscopic data, we envision that more statistical and bioinformatic efforts will be devoted toward the development of TAD border callers.

3.2.3 Identification of Interaction Points

Identifying biologically meaningful long-range chromatin interactions is of fundamental biological interest due to their relevance in transcription regulation. Several computational and statistical methods have recently been developed for Hi-C data analysis. Lan et al built a latent class of Poisson regression model to eliminate false positives produced by random ligation [132]. They characterize the proximate ligation events and random ligation events separately using two different Poisson distributions, thus representing the overall ligation events by a latent class model. Jin et al. developed a pipeline to estimate the expected contact frequency accounting for multiple Hi-C biases, and then tested for significant interaction by assuming the observed contact frequency following a negative binomial distribution [133]. Later on, Ay et al. developed Fit-Hi-C, providing more accurate estimates of the contact frequency by fitting non-parametric spline curves across genomic distance [134]. Meanwhile, Rao et al. developed HiCCUPS for analyzing 1-kb resolution in situ Hi-C data [121]. HiCCUPS quantifies the statistical significance of each chromatin interaction from local neighborhood regions. Most recently, Xu et al. proposed a hidden Markov random field-based Bayesian approach to model spatial dependency among adjacent interacting locus pairs, achieving improved robustness and enhanced statistical power [135].

Similar to the afore-mentioned bioinformatics problems in Hi-C data analysis, the key challenge in identifying chromatin interactions is the lack of a gold standard experimental data. More importantly, rigorous statistical approaches are required to explicitly model the null distribution of random chromatin collision. For example, an ANOVA-type statistical approach may provide a promising way to de-convolute biological signals from technical variations in Hi-C experiments, which can lead to valid statistical inferences for biologically meaningful chromatin interactions.

3.2.4 Long-Range Gene Regulation Using ChIA-PET

It has been observed that genes and their regulatory elements can be located far apart from one another [136,137] or even on different chromosomes [138]. ChIA-PET is

designed to detect such long-range gene regulations that are mediated by a specific protein. Two of the goals are (1) detection of true interactions that are not simply random collisions, and (2) detection of changes in interactions and interaction intensities under different conditions. Even though chimeric pairings (that is, pairing that are known to be due to proximity in the 3D space) are excluded, random collisions still exist. Recognizing this problem, Fullwood et al. treated pairs that are connected only once as false pairs [112]. Simple tests such as hypergeometric (HG) and weighted (generalized) HG [139] are used to further filter out false loops. In the weighted HG test, data are "normalized" in the sense that pairs in close proximity in the 1D (linear) genome are treated as more likely to have random collisions. A mixture modeling based approach was also developed to take dependency between pairs into account [140]. This is a "soft-thresholding" method: whether a pair is a true long-range interaction or not is not only dependent on the read counts that connect them, but also on their interactions with other loci as well as genomic annotation information. There is also an interest in detecting changes in gene regulation, for example, when a cell evolves from normal to cancerous, or over a time-course. Two model-based approaches have been proposed for this purpose for comparing long-range regulation under two conditions [141]. Both approaches are mixture modeling based: one is a three-component mixture modeled after the true loops are detected, while the other is a joint approach that considers loop detection and loop intensity variations simultaneously.

Despite the development of various approaches for ChIA-PET data, challenges abound. How to best normalize the data is obviously an important problem, but this issue has not been addressed thoroughly. Integration of other genomic data into loop detection is another important issue, and more work in this area is warranted. How to compare differential looping intensity in more than two groups (e.g., multiple cancer subtypes) is also important but unexplored. Time-course data are being produced but how to analyze them is still unclear. Each of the challenges calls for the development of sophisticated statistical methods.

3.2.5 3D Structure Inference: Optimization Based

Data from Hi-C portraits genome-wide interactions of chromatin and are typically organized into a 2-dimensional square matrix for each experiment, with the (i, j) entry depicting the contact (i.e., interaction) frequency between loci i and j, which are DNA segments in the genome. One of the objectives is to reconstruct the underlying 3D genome structure based on the data contained in this contact matrix. One type of approaches for such a reconstruction is optimization based, with the end result being a "consensus" 3D structure. This type of approach first translates the pairwise interaction frequency into a distance, typically using an inverse relationship. The consensus 3D structure is then obtained by minimizing the total differences between the translated distances and the corresponding ones induced from the 3D architecture to be estimated. Multidimensional scaling (MDS) is a common method for estimating the 3D coordinates of the structure. When the distance measure considered is the Euclidean distance, the MDS method is essentially the principal coordinate analysis, which estimates the 3D coordinates as the eigenvectors of the physical distance matrix [142,143]. Another class of multidimensional scaling algorithms, called non-metric MDS meth-

ods, where the coordinates are estimated by minimizing a cost function, has also been studied [144]. The cost function for such methods penalizes the relative difference between the physical and the induced distances to estimate the 3D coordinates of the loci. These algorithms are generally based on iterative optimization methods such as gradient descent or Newton's method. As the cost functions are typically non-convex, these algorithms may suffer from non-convergence and can be computationally less efficient that one would hope for. To address this problem, a semi-definite programming based approach has been developed [145], which reduces the computational cost significantly. Most recently, a graph theory based approach representing the physical distance as the shortest path and computed using the Floyd–Warshall algorithm has been developed [146].

Despite the availability of numerous optimization-based approaches for constructing 3D structures, there are still challenges that need to be addressed. With studies approaching finer and finer resolution (e.g., 1 kb in Rao et al. [121]), scalability of the methods needs to be studied. To be able to compare graphical representations obtained using different methods, comprehensive statistical measures need to be developed. Work has begun to emerge in this direction. For examples, methods have been developed to compare models of the same resolution [147]. A recent in silico study has also been carried out to evaluate methods using simulated data mimicking various resolutions [148].

3.2.6 3D Structure Inference: Model Based

Model-based statistical approaches for inferring 3D chromatin structure have been proposed in Hi-C data analysis. As the first attempt, Rousseau et al. developed MCMC5C [149], which models each Hi-C contact frequency as a Gaussian random variable. Later on, Hu et al. developed BACH and BACH-MIX [150], two Poisson regression-based algorithms to reconstruct chromatin spatial organizations and characterize chromatin structural dynamics. Although optimization-based methods are more popular for analyzing Hi-C data given their ease of description and relatively better computational efficiency, model-based methods have their own advantages. Covariates, especially those that lead to biases in the measured interaction frequencies, can be incorporated into the model directly. Further, model-based approaches enable one to study the population of potential 3D structures rather than a single "consensus" ones; this is clearly an advantage because most Hi-C studies are designed for a population of cells, and a mixture of 3D structures may be present.

Recognizing the excess of zeroes for higher resolution data such as the in situ Hi-C data [121], the truncated Poisson Architecture Model (tPAM) and the truncated Random effect EXpression model (tREX) were proposed based on a truncated Poisson distribution [148,151]. These methods were shown to be robust when the data contain more zeros than expected under a Poisson model, yet the methods were also shown to be efficient when the data are indeed coming from a Poisson distribution. Their performances have been compared to optimization-based methods for handing data of various resolutions [148].

Optimization-based methods typically run faster and consequently can be more efficient for inferring 3D structures of longer sequences. Model-based methods, on

the other hand, handles uncertainty and noise better as seen in simulation studies [146]. For more definitive performance comparisons, Florescent In Situ Hybridization (FISH) data are needed to serve as the gold standard [111,150]. Only very limited amount of such data are available now, but more is expected to emerge soon.

3.2.7 Integrative Analyses of Hi-C with Other Data Types

Two pioneering works from Lieberman-Aiden et al. and Dixon et al. used Hi-C data to define genomic compartments and domains independently and to link them with certain histone marks [112,116]. A similar but more recent approach developed by Rao et al. using higher sequencing depth Hi-C data provided a more detailed view of six nuclear sub-compartments with distinct patterns of histone marks [121]. On the other hand, work from Lan et al. utilized a clustering algorithm to integrate multiple ENCODE Consortium resources including DNase-seq and ChIP-seq data for 45 transcription factors and nine histone modifications with the Hi-C data [132]. They characterized 12 different sets (clusters) of interacting loci pairs (ILPs) each with different chromatin modification patterns. These sets can be categorized into two types of chromatin linkages (or hubs). Recently Libbrecht et al. [152] developed a graph-based regularization method to exploit chromatin conformation information during genome annotation. They were able to produce a model of chromatin domains in eight human cell types, which revealed five domain types tightly associated with histone marks and gene expression levels. Despite the advances, the above-mentioned integrations are ad hoc rather than grounded on rigorous statistical principles. Thus, challenges remain for statisticians and bioinformaticians to develop new approaches to integrate different data types into Hi-C data analysis.

3.2.8 Visualization

While 3C and 3C-derived technologies have been increasingly used visualization tools for 3C-based data are still under development. Since 2009, a few software programs have been devised to interactively visualize raw data, such as the Hi-C data browsers (http://hic.umassmed.edu/welcome/welcome.php) [111] and (http://yuelab.org/hi-c/index.html) [153]. In addition, WashU EpiGenome browser (http://epigenomegateway.wustl.edu/) is widely used for the joint analysis of Hi-C data and epigenetic data [154]. Most recently, thanks to the progress of the NIH Roadmap Project [155,156], Juicebox (http://www.aidenlab.org/juicebox/) has been developed for visualizing the in situ Hi-C data [121]. In addition, Teng et al. created 4DGenome (http://4dgenome.int-med.uiowa.edu), a comprehensive database to store and share publicly available 3C-based data [157].

With the accumulation of large amounts of Hi-C data [121] and other genetic/epigenetic data, especially those generated from the ENCODE consortium [158] and the Roadmap Epigenome consortium [156], we expect more computationally efficient, user-friendly and interactive Hi-C data visualization tools in the near future, facilitating integrative analysis of Hi-C data and other genomic data.

4 Discussion

Epigenetics has become an area of intense research in the post-genomic era. Rapid advances of high-throughput technologies and experimental techniques bring massive amounts of data, as well as daunting challenges in their analyses. Since epigenetics is a very broad term, epigenetic data are rather diverse, produced from different experiments, experimental platforms, and conditions. As a consequence, a specific analysis strategy is often required for each project, with careful consideration of potential statistical issues. With the expected emergence of further, new experimental techniques, we believe there will be much opportunity for designing and implementing new statistical models and algorithms. A key challenge in analyzing epigenetic data is the interpretation of the results. Because the technologies are new and the high-throughput technologies are prone to many sources of noise and biases, it is often difficult to determine whether the analysis findings are real or mere artifacts. To overcome such challenges, close and effective communication between biological researchers and biostatisticians and bioinformaticians is critical to ensure that the right assumptions can be made when selecting appropriate statistical models. Simulation studies are also critical for evaluating the performance of new methods. To provide meaningful evaluations, realistic assumptions must be used in simulations, which often involves some creative adaptation of real data. Finally, findings need to be scrutinized for possible confounding effects or data quality issues. Ideally, experimental verification is needed to validate findings.

Acknowledgments We thank all members of the Statistical and Applied Mathematical Sciences Institute (SAMSI) Epigenetics Working Group as part of the SAMSI Beyond Bioinformatics Program. We are also grateful for the support of Drs. Sujit Ghosh and Snehalata Huzurbazar at SAMSI. This material was based upon work partially supported by the National Science Foundation (NSF) under Grant DMS-1127914 to SAMSI. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF. This research was also supported in part from NSF Grant DMS-1220772 and NIH Grant 1R01GM114142-01.

References

1. Hu M, Deng K, Qin Z, Liu J (2013) Understanding spatial organizations of chromosomes via statistical analysis of Hi-C data. Quant Biol 1(2):156–174
2. Ay F, Noble WS (2015) Analysis methods for studying the 3D architecture of the genome. Genome Biol 16:183
3. Dekker J, Marti-Renom MA, Mirny LA (2013) Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat Rev Genet 14(6):390–403
4. Fraser J, Williamson I, Bickmore WA, Dostie J (2015) An overview of genome organization and how we got there: from FISH to Hi-C. Microbiol Mol Biol Rev 79(3):347–372
5. Schubeler D (2015) Function and information content of DNA methylation. Nature 517(7534):321–326
6. Bock C (2012) Analysing and interpreting DNA methylation data. Nat Rev Genet 13(10):705–719
7. Bibikova M, Lin Z, Zhou L, Chudin E, Garcia EW, Wu B, Doucet D, Thomas NJ, Wang Y, Vollmer E et al (2006) High-throughput DNA methylation profiling using universal bead arrays. Genome Res 16(3):383–393
8. Bibikova M, Le J, Barnes B, Saedinia-Melnyk S, Zhou L, Shen R, Gunderson KL (2009) Genome-wide DNA methylation profiling using Infinium(R) assay. Epigenomics 1(1):177–200

9. Bibikova M, Barnes B, Tsan C, Ho V, Klotzle B, Le JM, Delano D, Zhang L, Schroth GP, Gunderson KL et al (2011) High density DNA methylation array with single CpG site resolution. Genomics 98(4):288–295
10. Cokus SJ, Feng S, Zhang X, Chen Z, Merriman B, Haudenschild CD, Pradhan S, Nelson SF, Pellegrini M, Jacobsen SE (2008) Shotgun bisulphite sequencing of the Arabidopsis genome reveals DNA methylation patterning. Nature 452(7184):215–219
11. Lister R, O'Malley RC, Tonti-Filippini J, Gregory BD, Berry CC, Millar AH, Ecker JR (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell 133(3):523–536
12. Meissner A, Gnirke A, Bell GW, Ramsahoye B, Lander ES, Jaenisch R (2005) Reduced representation bisulfite sequencing for comparative high-resolution DNA methylation analysis. Nucleic Acids Res 33(18):5868–5877
13. Gu H, Smith ZD, Bock C, Boyle P, Gnirke A, Meissner A (2011) Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. Nat Protoc 6(4):468–481
14. Huang Y, Pastor WA, Shen Y, Tahiliani M, Liu DR, Rao A (2010) The behaviour of 5-hydroxymethylcytosine in bisulfite sequencing. PloS One 5(1):e8888
15. Jin SG, Wu X, Li AX, Pfeifer GP (2011) Genomic mapping of 5-hydroxymethylcytosine in the human brain. Nucleic Acids Res 39(12):5015–5024
16. Rampal R, Alkalin A, Madzo J, Vasanthakumar A, Pronier E, Patel J, Li Y, Ahn J, Abdel-Wahab O, Shih A et al (2014) DNA hydroxymethylation profiling reveals that WT1 mutations result in loss of TET2 function in acute myeloid leukemia. Cell Rep 9(5):1841–1855
17. Wu H, D'Alessio AC, Ito S, Wang Z, Cui K, Zhao K, Sun YE, Zhang Y (2011) Genome-wide analysis of 5-hydroxymethylcytosine distribution reveals its dual function in transcriptional regulation in mouse embryonic stem cells. Genes Dev 25(7):679–684
18. Yu M, Hon GC, Szulwach KE, Song CX, Zhang L, Kim A, Li X, Dai Q, Shen Y, Park B et al (2012) Base-resolution analysis of 5-hydroxymethylcytosine in the mammalian genome. Cell 149(6):1368–1380
19. Huang TH, Perry MR, Laux DE (1999) Methylation profiling of CpG islands in human breast cancer cells. Hum Mol Genet 8(3):459–470
20. Deatherage DE, Potter D, Yan PS, Huang TH, Lin S (2009) Methylation analysis by microarray. Methods Mol Biol 556:117–139
21. Sun S, Chen Z, Yan PS, Huang YW, Huang TH, Lin S (2011) Identifying hypermethylated CpG islands using a quantile regression model. BMC Bioinform 12:54
22. Weber M, Davies JJ, Wittig D, Oakeley EJ, Haase M, Lam WL, Schubeler D (2005) Chromosome-wide and promoter-specific analyses identify sites of differential DNA methylation in normal and transformed human cells. Nat Genet 37(8):853–862
23. Zhang X, Yazaki J, Sundaresan A, Cokus S, Chan SW, Chen H, Henderson IR, Shinn P, Pellegrini M, Jacobsen SE et al (2006) Genome-wide high-resolution mapping and functional analysis of DNA methylation in arabidopsis. Cell 126(6):1189–1201
24. Down TA, Rakyan VK, Turner DJ, Flicek P, Li H, Kulesha E, Graf S, Johnson N, Herrero J, Tomazou EM et al (2008) A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis. Nat Biotechnol 26(7):779–785
25. Brinkman AB, Simmer F, Ma K, Kaan A, Zhu J, Stunnenberg HG (2010) Whole-genome DNA methylation profiling using MethylCap-seq. Methods 52(3):232–236
26. Serre D, Lee BH, Ting AH (2010) MBD-isolated genome sequencing provides a high-throughput and comprehensive survey of DNA methylation in the human genome. Nucleic Acids Res 38(2):391–399
27. Li D, Zhang B, Xing X, Wang T (2015) Combining MeDIP-seq and MRE-seq to investigate genome-wide CpG methylation. Methods 72:29–40
28. Lan X, Adams C, Landers M, Dudas M, Krissinger D, Marnellos G, Bonneville R, Xu M, Wang J, Huang TH et al (2011) High resolution detection and analysis of CpG dinucleotides methylation using MBD-Seq technology. PLoS One 6(7):e22226
29. Frankhouser DE, Murphy M, Blachly JS, Park J, Zoller MW, Ganbat JO, Curfman J, Byrd JC, Lin S, Marcucci G et al (2014) PrEMeR-CG: inferring nucleotide level DNA methylation values from MethylCap-seq data. Bioinformatics 30(24):3567–3574
30. Smallwood SA, Lee HJ, Angermueller C, Krueger F, Saadeh H, Peat J, Andrews SR, Stegle O, Reik W, Kelsey G (2014) Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. Nat Methods 11(8):817–820

31. Lehne B, Drong AW, Loh M, Zhang W, Scott WR, Tan ST, Afzal U, Scott J, Jarvelin MR, Elliott P et al (2015) A coherent approach for analysis of the Illumina HumanMethylation450 BeadChip improves data quality and performance in epigenome-wide association studies. Genome Biol 16:37
32. Aryee MJ, Jaffe AE, Corrada-Bravo H, Ladd-Acosta C, Feinberg AP, Hansen KD, Irizarry RA (2014) Minfi: a flexible and comprehensive bioconductor package for the analysis of infinium DNA methylation microarrays. Bioinformatics 30(10):1363–1369
33. Barfield RT, Kilaru V, Smith AK, Conneely KN (2012) CpGassoc: an R function for analysis of DNA methylation microarray data. Bioinformatics 28(9):1280–1281
34. Du P, Kibbe WA, Lin SM (2008) Lumi: a pipeline for processing Illumina microarray. Bioinformatics 24(13):1547–1548
35. Morris TJ, Butcher LM, Feber A, Teschendorff AE, Chakravarthy AR, Wojdacz TK, Beck S (2014) ChAMP: 450k chip analysis methylation pipeline. Bioinformatics 30(3):428–430
36. Wang D, Yan L, Hu Q, Sucheston LE, Higgins MJ, Ambrosone CB, Johnson CS, Smiraglia DJ, Liu S (2012) IMA: an R package for high-throughput analysis of Illumina's 450K Infinium methylation data. Bioinformatics 28(5):729–730
37. Pidsley R, Y Wong CC, Volta M, Lunnon K, Mill J, Schalkwyk LC (2013) A data-driven approach to preprocessing Illumina 450K methylation array data. BMC Genomics 14:293
38. Davis S, Du P, Bilke S, Triche T Jr, Bootwalla M (2015) methylumi: Handle Illumina methylation data. R package version 2.14.0. https://www.bioconductor.org/packages/3.3/bioc/manuals/methylumi/man/methylumi.pdf
39. Price ME, Cotton AM, Lam LL, Farre P, Emberly E, Brown CJ, Robinson WP, Kobor MS (2013) Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. Epigenetics Chromatin 6(1):4
40. Smith AK, Kilaru V, Kocak M, Almli LM, Mercer KB, Ressler KJ, Tylavsky FA, Conneely KN (2014) Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type. BMC Genomics 15:145
41. Reynolds LM, Taylor JR, Ding J, Lohman K, Johnson C, Siscovick D, Burke G, Post W, Shea S, Jacobs DR Jr et al (2014) Age-related variations in the methylome associated with gene expression in human monocytes and T cells. Nat Commun 5:5366
42. McRae AF, Powell JE, Henders AK, Bowdler L, Hemani G, Shah S, Painter JN, Martin NG, Visscher PM, Montgomery GW (2014) Contribution of genetic variation to transgenerational inheritance of DNA methylation. Genome Biol 15(5):R73
43. Fortin JP, Labbe A, Lemire M, Zanke BW, Hudson TJ, Fertig EJ, Greenwood CM, Hansen KD (2014) Functional normalization of 450k methylation array data improves replication in large cancer studies. Genome Biol 15(12):503
44. Dedeurwaerder S, Defrance M, Calonne E, Denis H, Sotiriou C, Fuks F (2011) Evaluation of the Infinium Methylation 450K technology. Epigenomics 3(6):771–784
45. Maksimovic J, Gordon L, Oshlack A (2012) SWAN: subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. Genome Biol 13(6):R44
46. Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, Beck S (2013) A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. Bioinformatics 29(2):189–196
47. Touleimat N, Tost J (2012) Complete pipeline for Infinium((R)) Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation. Epigenomics 4(3):325–341
48. Triche TJ Jr, Weisenberger DJ, Van Den Berg D, Laird PW, Siegmund KD (2013) Low-level processing of Illumina Infinium DNA methylation beadarrays. Nucleic Acids Res 41(7):e90
49. Wu MC, Joubert BR, Kuan PF, Haberg SE, Nystad W, Peddada SD, London SJ (2014) A systematic assessment of normalization approaches for the Infinium 450K methylation platform. Epigenetics 9(2):318–329
50. Dedeurwaerder S, Defrance M, Bizet M, Calonne E, Bontempi G, Fuks F (2014) A comprehensive overview of Infinium HumanMethylation450 data processing. Brief Bioinform 15(6):929–941
51. Berg BA, Neuhaus T (1991) Multicanonical algorithms for 1st order phase-transitions. Phys Lett B 267(2):249–253
52. Liu Y, Siegmund KD, Laird PW, Berman BP (2012) Bis-SNP: combined DNA methylation and SNP calling for Bisulfite-seq data. Genome Biol 13(7):R61

53. Lin X, Sun D, Rodriguez B, Zhao Q, Sun H, Zhang Y, Li W (2013) BSeQC: quality control of bisulfite sequencing experiments. Bioinformatics 29(24):3227–3229
54. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10(3):R25
55. Krueger F, Andrews SR (2011) Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics 27(11):1571–1572
56. Xi Y, Li W (2009) BSMAP: whole genome bisulfite sequence MAPping program. BMC Bioinform 10:232
57. Smith AD, Chung WY, Hodges E, Kendall J, Hannon G, Hicks J, Xuan Z, Zhang MQ (2009) Updates to the RMAP short-read mapping software. Bioinformatics 25(21):2841–2842
58. Hansen KD, Langmead B, Irizarry RA (2012) BSmooth: from whole genome bisulfite sequencing reads to differentially methylated regions. Genome Biol 13(10):R83
59. Kunde-Ramamoorthy G, Coarfa C, Laritsky E, Kessler NJ, Harris RA, Xu M, Chen R, Shen L, Milosavljevic A, Waterland RA (2014) Comparison and quantitative verification of mapping algorithms for whole-genome bisulfite sequencing. Nucleic Acids Res 42(6):e43
60. Chatterjee A, Stockwell PA, Rodger EJ, Morison IM (2012) Comparison of alignment software for genome-wide bisulphite sequence data. Nucleic Acids Res 40(10):e79
61. Ji L, Sasaki T, Sun X, Ma P, Lewis ZA, Schmitz RJ (2014) Methylated DNA is over-represented in whole-genome bisulfite sequencing data. Front Genet 5:341
62. Du P, Zhang X, Huang CC, Jafari N, Kibbe WA, Hou L, Lin SM (2010) Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinform 11:587
63. Bjornsson HT, Sigurdsson MI, Fallin MD, Irizarry RA, Aspelund T, Cui H, Yu W, Rongione MA, Ekstrom TJ, Harris TB et al (2008) Intra-individual change over time in DNA methylation with familial clustering. JAMA 299(24):2877–2883
64. Bollati V, Schwartz J, Wright R, Litonjua A, Tarantini L, Suh H, Sparrow D, Vokonas P, Baccarelli A (2009) Decline in genomic DNA methylation through aging in a cohort of elderly subjects. Mech Ageing Dev 130(4):234–239
65. Christensen BC, Houseman EA, Marsit CJ, Zheng S, Wrensch MR, Wiemels JL, Nelson HH, Karagas MR, Padbury JF, Bueno R et al (2009) Aging and environmental exposures alter tissue-specific DNA methylation dependent upon CpG island context. PLoS Genet 5(8):e1000602
66. Rakyan VK, Down TA, Maslau S, Andrew T, Yang TP, Beyan H, Whittaker P, McCann OT, Finer S, Valdes AM et al (2010) Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. Genome Res 20(4):434–439
67. Teschendorff AE, Menon U, Gentry-Maharaj A, Ramus SJ, Weisenberger DJ, Shen H, Campan M, Noushmehr H, Bell CG, Maxwell AP et al (2010) Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. Genome Res 20(4):440–446
68. Alisch RS, Barwick BG, Chopra P, Myrick LK, Satten GA, Conneely KN, Warren ST (2012) Age-associated DNA methylation in pediatric populations. Genome Res 22(4):623–632
69. Horvath S (2013) DNA methylation age of human tissues and cell types. Genome Biol 14(10):R115
70. Liu J, Hutchison K, Perrone-Bizzozero N, Morgan M, Sui J, Calhoun V (2010) Identification of genetic and epigenetic marks involved in population structure. PLoS One 5(10):e13209
71. Barfield RT, Almli LM, Kilaru V, Smith AK, Mercer KB, Duncan R, Klengel T, Mehta D, Binder EB, Epstein MP et al (2014) Accounting for population stratification in DNA methylation studies. Genet Epidemiol 38:231–241
72. Reinius LE, Acevedo N, Joerink M, Pershagen G, Dahlen SE, Greco D, Soderhall C, Scheynius A, Kere J (2012) Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. PloS one 7(7):e41361
73. Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, Wiencke JK, Kelsey KT (2012) DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinform 13:86
74. Jaffe AE, Irizarry RA (2014) Accounting for cellular heterogeneity is critical in epigenome-wide association studies. Genome Biol 15(2):R31
75. Liu Y, Aryee MJ, Padyukov L, Fallin MD, Hesselberg E, Runarsson A, Reinius L, Acevedo N, Taub M, Ronninger M et al (2013) Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. Nat Biotechnol 31(2):142–147

76. Cardenas A, Koestler DC, Houseman EA, Jackson BP, Kile ML, Karagas MR, Marsit CJ (2015) Differential DNA methylation in umbilical cord blood of infants exposed to mercury and arsenic in utero. Epigenetics 10(6):508–515
77. Liang L, Willis-Owen SA, Laprise C, Wong KC, Davies GA, Hudson TJ, Binia A, Hopkin JM, Yang IV, Grundberg E et al (2015) An epigenome-wide association study of total serum immunoglobulin E concentration. Nature 520(7549):670–674
78. Houseman EA, Molitor J, Marsit CJ (2014) Reference-free cell mixture adjustments in analysis of DNA methylation data. Bioinformatics 30(10):1431–1439
79. Zou J, Lippert C, Heckerman D, Aryee M, Listgarten J (2014) Epigenome-wide association studies without the need for cell-type composition. Nat Methods 11(3):309–311
80. Bell JT, Tsai PC, Yang TP, Pidsley R, Nisbet J, Glass D, Mangino M, Zhai G, Zhang F, Valdes A et al (2012) Epigenome-wide scans identify differentially methylated regions for age and age-related phenotypes in a healthy ageing population. PLoS Genet 8(4):e1002629
81. Leek JT, Storey JD (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet 3(9):1724–1735
82. Park Y, Figueroa ME, Rozek LS, Sartor MA (2014) MethylSig: a whole genome DNA methylation analysis pipeline. Bioinformatics 30(17):2414–2422
83. Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A, Mason CE (2012) methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biol 13(10):R87
84. Hebestreit K, Dugas M, Klein HU (2013) Detection of significantly differentially methylated regions in targeted bisulfite sequencing data. Bioinformatics 29(13):1647–1653
85. Feng H, Conneely KN, Wu H (2014) A Bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. Nucleic Acids Res 42(8):e69
86. Xu T, Li B, Zhao M, Szulwach KE, Street RC, Lin L, Yao B, Zhang F, Jin P, Wu H et al (2015) Base-resolution methylation patterns accurately predict transcription factor bindings in vivo. Nucleic Acids Res 43(5):2757–2766
87. Wu H, Xu T, Feng H, Chen L, Li B, Yao B, Qin Z, Jin P, Conneely KN (2015) Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. Nucleic Acids Res 43:e141
88. Robinson MD, Kahraman A, Law CW, Lindsay H, Nowicka M, Weber LM, Zhou X (2014) Statistical methods for detecting differentially methylated loci and regions. Front Genet 5:324
89. Dolzhenko E, Smith AD (2014) Using beta-binomial regression for high-precision differential methylation analysis in multifactor whole-genome bisulfite sequencing experiments. BMC Bioinform 15:215
90. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W et al (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biol 9(9):R137
91. Qin ZS, Yu J, Shen J, Maher CA, Hu M, Kalyana-Sundaram S, Yu J, Chinnaiyan AM (2010) HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data. BMC Bioinform 11:369
92. Ji H, Jiang H, Ma W, Johnson DS, Myers RM, Wong WH (2008) An integrated software system for analyzing ChIP-chip and ChIP-seq data. Nat Biotechnol 26(11):1293–1300
93. Liu B, Yi J, Sv A, Lan X, Ma Y, Huang TH, Leone G, Jin VX (2013) QChIPat: a quantitative method to identify distinct binding patterns for two biological ChIP-seq samples in different experimental conditions. BMC Genomics 14(Suppl 8):S3
94. Liang K, Keles S (2012) Detecting differential binding of transcription factors with ChIP-seq. Bioinformatics 28(1):121–122
95. Shao Z, Zhang Y, Yuan GC, Orkin SH, Waxman DJ (2012) MAnorm: a robust model for quantitative comparison of ChIP-Seq data sets. Genome Biol 13(3):R16
96. Chen L, Wang C, Qin ZS, Wu H (2015) A novel statistical method for quantitative comparison of multiple ChIP-seq datasets. Bioinformatics 31:1889–1896
97. Shen L, Shao NY, Liu X, Maze I, Feng J, Nestler EJ (2013) diffReps: detecting differential chromatin modification sites from ChIP-seq data with biological replicates. PLoS One 8(6):e65598
98. Taslim C, Huang T, Lin S (2011) DIME: R-package for identifying differential ChIP-seq based on an ensemble of mixture models. Bioinformatics 27(11):1569–1570
99. Nair NU, Sahu AD, Bucher P, Moret BM (2012) ChIPnorm: a statistical method for normalizing and identifying differential regions in histone modification ChIP-seq libraries. PLoS One 7(8):e39573

100. Schweikert G, Cseke B, Clouaire T, Bird A, Sanguinetti G (2013) MMDiff: quantitative testing for shape changes in ChIP-Seq data sets. BMC Genomics 14:826
101. Lienhard M, Grimm C, Morkel M, Herwig R, Chavez L (2014) MEDIPS: genome-wide differential coverage analysis of sequencing data derived from DNA enrichment experiments. Bioinformatics 30(2):284–286
102. Yan P, Frankhouser D, Murphy M, Tam HH, Rodriguez B, Curfman J, Trimarchi M, Geyer S, Wu YZ, Whitman SP et al (2012) Genome-wide methylation profiling in decitabine-treated patients with acute myeloid leukemia. Blood 120(12):2466–2474
103. Jadhav RR, Ye Z, Huang RL, Liu J, Hsu PY, Huang YW, Rangel LB, Lai HC, Roa JC, Kirma NB et al (2015) Genome-wide DNA methylation analysis reveals estrogen-mediated epigenetic repression of metallothionein-1 gene cluster in breast cancer. Clin Epigenetics 7(1):13
104. Liang K-Y, Zeger SL (1986) Longitudinal data analysis using generalized linear models. Biometrika 73(1):13–22
105. Ayyala DN, Frankhouser DE, Ganbat JO, Marcucci G, Bundschuh R, Yan P, Lin S (2015) Statistical methods for detecting differentially methylated regions based on MethylCap-Seq Data. Brief Bioinform. doi:10.1093/bib/bbv089
106. Xie H, Wang M, de Andrade A, Bonaldo MF, Galat V, Arndt K, Rajaram V, Goldman S, Tomita T, Soares MB (2011) Genome-wide quantitative assessment of variation in DNA methylation patterns. Nucleic Acids Res 39(10):4099–4108
107. Shao X, Zhang C, Sun MA, Lu X, Xie H (2014) Deciphering the heterogeneity in DNA methylation patterns during stem cell differentiation and reprogramming. BMC Genomics 15:978
108. Dekker J (2008) Gene regulation in the third dimension. Science 319(5871):1793–1794
109. Misteli T (2007) Beyond the sequence: cellular organization of genome function. Cell 128(4):787–800
110. Dekker J, Rippe K, Dekker M, Kleckner N (2002) Capturing chromosome conformation. Science 295(5558):1306–1311
111. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO et al (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326(5950):289–293
112. Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, Orlov YL, Velkov S, Ho A, Mei PH et al (2009) An oestrogen-receptor-alpha-bound human chromatin interactome. Nature 462(7269):58–64
113. Speicher MR, Ballard SG, Ward DC (1996) Karyotyping human chromosomes by combinatorial multi-fluor FISH. Nat Genet 12(4):368–375
114. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G (2012) Three-dimensional folding and functional organization principles of the Drosophila genome. Cell 148(3):458–472
115. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485(7398):376–380
116. Hou C, Li L, Qin ZS, Corces VG (2012) Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains. Mol Cell 48(3):471–484
117. Van Bortle K, Nichols MH, Li L, Ong CT, Takenaka N, Qin ZS, Corces VG (2014) Insulator function and topological domain border strength scale with architectural protein occupancy. Genome Biol 15(6):R82
118. Li L, Lyu X, Hou C, Takenaka N, Nguyen HQ, Ong CT, Cubenas-Potts C, Hu M, Lei EP, Bosco G et al (2015) Widespread rearrangement of 3D chromatin organization underlies polycomb-mediated stress-induced silencing. Mol Cell 58(2):216–231
119. Zhang Y, McCord RP, Ho YJ, Lajoie BR, Hildebrand DG, Simon AC, Becker MS, Alt FW, Dekker J (2012) Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. Cell 148(5):908–921
120. Kalhor R, Tjong H, Jayathilaka N, Alber F, Chen L (2012) Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nat Biotechnol 30(1):90–98
121. Rao SS, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES et al (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell 159(7):1665–1680
122. Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, de Wit E, van Steensel B, de Laat W (2006) Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nat Genet 38(11):1348–1354

123. Splinter E, de Wit E, van de Werken HJ, Klous P, de Laat W (2012) Determining long-range chromatin interactions for selected genomic sites using 4C-seq technology: From fixation to computation. Methods 58(3):221–230
124. van Steensel B, Dekker J (2010) Genomics tools for unraveling chromosome architecture. Nat Biotechnol 28(10):1089–1095
125. Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, Laue ED, Tanay A, Fraser P (2013) Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature 502(7469):59–64
126. Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, Wingett SW, Andrews S, Grey W, Ewels PA et al (2015) Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat Genet 47(6):598–606
127. Yaffe E, Tanay A (2011) Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nat Genet 43(11):1059–1065
128. Hu M, Deng K, Selvaraj S, Qin Z, Ren B, Liu JS (2012) HiCNorm: removing biases in Hi-C data via Poisson regression. Bioinformatics 28(23):3131–3133
129. Cournac A, Marie-Nelly H, Marbouty M, Koszul R, Mozziconacci J (2012) Normalization of a chromosomal contact map. BMC Genomics 13:436
130. Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, Dekker J, Mirny LA (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat Methods 9(10):999–1003
131. Levy-Leduc C, Delattre M, Mary-Huard T, Robin S (2014) Two-dimensional segmentation for analyzing Hi-C data. Bioinformatics 30(17):i386–392
132. Lan X, Witt H, Katsumura K, Ye Z, Wang Q, Bresnick EH, Farnham PJ, Jin VX (2012) Integration of Hi-C and ChIP-seq data reveals distinct types of chromatin linkages. Nucleic Acids Res 40(16):7690–7704
133. Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, Yen CA, Schmitt AD, Espinoza CA, Ren B (2013) A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature 503(7475):290–294
134. Ay F, Bailey TL, Noble WS (2014) Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. Genome Res 24(6):999–1011
135. Xu Z, Zhang G, Jin F, Chen M, Furey TS, Sullivan PF, Qin Z, Hu M, Li Y (2015) A hidden Markov random field based Bayesian method for the detection of long-range chromosomal interactions in Hi-C Data. Bioinformatics. doi:10.1093/bioinformatics/btv650
136. Kleinjan DA, van Heyningen V (2005) Long-range control of gene expression: emerging mechanisms and disruption in disease. Am J Hum Genet 76(1):8–32
137. West AG, Fraser P (2005) Remote control of gene transcription. Hum Mol Genet 14(suppl 1):R101–R111
138. Lomvardas S, Barnea G, Pisapia DJ, Mendelsohn M, Kirkland J, Axel R (2006) Interchromosomal interactions and olfactory receptor choice. Cell 126(2):403–413
139. Paulsen J, Rodland EA, Holden L, Holden M, Hovig E (2014) A statistical model of ChIA-PET data for accurate detection of chromatin 3D interactions. Nucleic Acids Res 42(18):e143
140. Niu L, Li G, Lin S (2014) Statistical models for detecting differential chromatin interactions mediated by a protein. PloS One 9(5):e97560
141. Niu L, Lin S (2015) A Bayesian mixture model for chromatin interaction data. Stat Appl Genet Mol Biol 14(1):53–64
142. Duan Z, Andronescu M, Schutz K, McIlwain S, Kim YJ, Lee C, Shendure J, Fields S, Blau CA, Noble WS (2010) A three-dimensional model of the yeast genome. Nature 465(7296):363–367
143. Tanizawa H, Iwasaki O, Tanaka A, Capizzi JR, Wickramasinghe P, Lee M, Fu Z, Noma K (2010) Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. Nucleic Acids Res 38(22):8164–8177
144. Ben-Elazar S, Yakhini Z, Yanai I (2013) Spatial localization of co-regulated genes exceeds genomic gene clustering in the Saccharomyces cerevisiae genome. Nucleic Acids Res 41(4):2191–2201
145. Zhang Z, Li G, Toh KC, Sung WK (2013) 3D Chromosome modeling with semi-definite programming and Hi-C data. J Comput Biol 20(11):831–846
146. Lesne A, Riposo J, Roger P, Cournac A, Mozziconacci J (2014) 3D Genome reconstruction from chromosomal contacts. Nat Methods 11(11):1141–1143
147. Capurso D, Segal MR (2014) Distance-based assessment of the localization of functional annotations in 3D genome reconstructions. BMC Genomics 15:992

148. Park J, Lin S (2016) Impact of data resolution on three-dimensional structure inference methods. BMC Bioinform 17(1):70
149. Rousseau M, Fraser J, Ferraiuolo MA, Dostie J, Blanchette M (2011) Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling. BMC Bioinform 12:414
150. Hu M, Deng K, Qin ZS, Dixon J, Selvaraj S, Feng J, Ren B, Liu JS (2012) Bayesian inference of three-dimensional chromosomal organization. PLoS Comput Biol 9:e1002893
151. Park J, Lin S (2015) Statistical inference on three-dimensional structure of genome by truncated poisson architecture model. In: Choudhary P, Nagaraja C, Ng T (eds) Ordered data analysis, modeling, and health research methods: in honor of H N Nagaraja's 60th birthday. Springer, New York
152. Libbrecht MW, Ay F, Hoffman MM, Gilbert DM, Bilmes JA, Noble WS (2015) Joint annotation of chromatin state and chromatin conformation reveals relationships among domain types and identifies domains of cell-type-specific expression. Genome Res 25:544–557
153. Dixon JR, Jung I, Selvaraj S, Shen Y, Antosiewicz-Bourget JE, Lee AY, Ye Z, Kim A, Rajagopal N, Xie W et al (2015) Chromatin architecture reorganization during stem cell differentiation. Nature 518(7539):331–336
154. Zhou X, Maricque B, Xie M, Li D, Sundaram V, Martin EA, Koebbe BC, Nielsen C, Hirst M, Farnham P et al (2011) The human epigenome browser at Washington University. Nat Methods 8(12):989–990
155. Leung D, Jung I, Rajagopal N, Schmitt A, Selvaraj S, Lee AY, Yen CA, Lin S, Lin Y, Qiu Y et al (2015) Integrative analysis of haplotype-resolved epigenomes across human tissues. Nature 518(7539):350–354
156. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J et al. (2015) Integrative analysis of 111 reference human epigenomes. Nature 518(7539):317–330
157. Teng L, He B, Wang J, Tan K (2015) 4DGenome: a comprehensive database of chromatin interactions. Bioinformatics 31(15):2560–2564
158. Consortium EP, Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET et al (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 447(7146):799–816