BIOINFORMATICS REVIEW Vol. 24 no. 1 2008, pages 1–10
doi:10.1093/bioinformatics/btm546

Genome analysis

Computational epigenetics

Christoph Bock* and Thomas Lengauer
Max-Planck-Institut für Informatik, Saarbrücken, Germany

Received on August 25, 2007; revised and accepted on October 28, 2007
Advance Access publication November 17, 2007
Associate Editor: Jonathan Wren

ABSTRACT

Epigenetic research aims to understand heritable gene regulation that is not directly encoded in the DNA sequence. Epigenetic mechanisms such as DNA methylation and histone modifications modulate the packaging of the DNA in the nucleus and thereby influence gene expression. Patterns of epigenetic information are faithfully propagated over multiple cell divisions, which makes epigenetic regulation a key mechanism for cellular differentiation and cell fate decisions. In addition, incomplete erasure of epigenetic information can lead to complex patterns of non-Mendelian inheritance. Stochastic and environment-induced epigenetic defects are known to play a major role in cancer and ageing, and they may also contribute to mental disorders and autoimmune diseases. Recent technical advances such as ChIP-on-chip and ChIP-seq have started to convert epigenetic research into a high-throughput endeavor, to which bioinformatics is expected to make significant contributions. Here, we review pioneering computational studies that have contributed to epigenetic research. In addition, we give a brief introduction into epigenetics—targeted at bioinformaticians who are new to the field—and we outline future challenges in computational epigenetics.

Contact: cbock@mpi-inf.mpg.de

1 INTRODUCTION

Epigenetics is commonly defined as the 'study of mitotically and/or meiotically heritable changes in gene function that cannot be explained by changes in DNA sequence' (Russo *et al*., 1996). Its fundamental objective is to elucidate how genetic information encoded in the DNA sequence and non-genetic aspects such as the way the DNA is packaged inside the nucleus jointly control gene expression. This touches upon two central problems of biology: How do cells specialize when a complex multi-cellular organism develops from a single fertilized egg (Reik, 2007)? And which molecular mechanisms contribute to phenotypic inheritance (Richards, 2006)?

The field of epigenetics has recently received a boost of attention and is currently among the fastest moving areas in molecular biology. Unprecedented technological advances enable genome-scale analysis of epigenetic mechanisms and render comprehensive epigenome projects feasible (Bernstein *et al*., 2007). Epigenetic analysis of embryonic stem (ES) cells has started to unveil the basic circuitry of mammalian development (Surani *et al*., 2007). And in cancer research, epigenetics opens up novel approaches for early diagnosis and treatment (Jones and Baylin, 2007).

Various bioinformatic challenges arise from the analysis of epigenetic data, and computational methods have already played a role in solving important epigenetic problems. In this review, we introduce the basic concepts of epigenetics and we summarize relevant computational and bioinformatic work performed in this area. Furthermore we outline future directions, arguing that upcoming epigenome projects will constitute a major challenge for the emerging field of computational epigenetics.

2 TWO FACETS OF EPIGENETIC INHERITANCE

Epigenetic mechanisms influence phenotype through heritable regulation of gene expression. The constitutive property of epigenetic inheritance is that it is encoded in covalent modifications of the DNA and the chromatin proteins attached to it, rather than in the DNA sequence itself (as is the case for genetic inheritance). Because such modifications are more readily altered than the DNA sequence, epigenetic information can be reprogrammed dynamically during cellular differentiation, but is also propagated with substantially lower fidelity than genetic information. An error rate of 10^{-3} has been estimated per site and cell division for DNA methylation (Ushijima *et al*., 2003), in contrast to values in the order of 10^{-8} per basepair and cell division for genetic mutations (Drake *et al*., 1998). Epigenetic inheritance occurs both between generations of cells (mitotic inheritance) and between generations of a species (meiotic inheritance).

Epigenetic mitotic inheritance is critically involved in cellular differentiation and cell fate decisions. Recent research has provided a mechanistic understanding of the key phases of epigenetic regulation during development (Reik, 2007). To start with, germ cells carry highly specialized and parent-specific epigenetic information. Shortly after fertilization, a fundamental reprogramming step resets most epigenetic information to a default state, which might be derived from properties of the DNA sequence. This reprogrammed epigenetic state seems to be crucial for the pluripotency of ES cells (i.e. for their ability to differentiate into diverse tissue types). During cellular differentiation, ES cells reprogram their epigenetic state once again when tissue-specific transcription factors are activated and pluripotency-specific genes become silenced. In terminally differentiated cells, epigenetic information is faithfully propagated during cell division. However, cellular ageing leads to

*To whom correspondence should be addressed.

increasing heterogeneity within a cell population and can also contribute to tumor development (Fraga and Esteller, 2007). Finally, the specialized cells of the germline reprogram epigenetic information in a parent-specific way, before it is passed on to the offspring as sperm or egg.

Epigenetic meiotic inheritance is caused by incomplete reprogramming in the early embryo, which results in the propagation of epigenetic information from parent to offspring. This phenomenon gives rise to patterns of phenotypic inheritance that are inconsistent with Mendelian rules. First, imprinted genes are inherited and expressed in a parent-specific way, i.e. only the maternal allele is transcribed while the paternal allele is epigenetically silenced or vice versa. Imprinted genes play a central role in the development of placenta and brain, and thay have been linked to several rare neurogenetic disorders as well as to cancer (Solter, 2006). Second, acquired traits can be epigenetically transmitted over multiple generations. While this type of inheritance is relatively rare in mammals (Peaston and Whitelaw, 2006), for plants it seems to be a common way of adapting gene regulation to a changing environment (Grant-Downton and Dickinson, 2006).

3 MECHANISMS OF EPIGENETIC REGULATION

Epigenetic regulation exploits the fact that the packaging of DNA inside the nucleus directly influences gene expression (Dillon, 2006). In general, the tighter a gene's DNA is wrapped up, the more likely it is switched off. Conversely, the more accessible it is to the transcription machinery, the more likely it is actively transcribed. Physically, the genome of eukaryotic cells is stored in a highly regulated protein–DNA complex called chromatin, which controls DNA accessibility for cellular processes such as transcription, replication and DNA repair (Woodcock, 2006). Epigenetic mechanisms can be both activating (i.e. fostering open chromatin structure, called euchromatin) or repressive (i.e. fostering condensed chromatin structure, called heterochromatin), and different epigenetic mechanisms frequently act synergistically. Three biochemical mechanisms are commonly referred to as epigenetic: (i) DNA methylation, (ii) histone modifications and (iii) binding of non-histone proteins such as Polycomb and trithorax group complexes.

DNA methylation (Bird, 2002; Weber and Schübeler, 2007) is the only epigenetic modification that directly affects the DNA. Biochemically, a hydrogen atom of the cytosine base is replaced by a methyl group (Fig. 1, left). This does not alter the way in which the cytosine is transcribed into mRNA, but it fosters a locally more compact chromatin structure and affects transcription factor binding. In mammals, DNA methylation is largely confined to cytosines in a CpG context ('CpG' stands for **c**ytidine and **g**uanosine, separated by a **p**hosphate atom), which has two important implications. First, any genomic position that can be methylated is symmetric, i.e. there is a—methylated or unmethylated—cytosine on the forward strand as well as on the reverse strand. Therefore, after DNA replication a specific enzyme can read the DNA methylation pattern of the parent strand and faithfully copy it to the newly synthesized strand, thereby maintaining heritable DNA methylation patterns. Second, in mammalian genomes CpG dinucleotides occur in clusters, and the genomic regions with highest CpG density—termed CpG islands—exhibit the lowest levels of DNA methylation. This phenomenon is most likely caused by the fact that mutation rates are substantially higher for methylated CpGs than for unmethylated CpGs, hence absence of DNA methylation at least in the germline seems to be constitutive for long-term maintenance of most CpG islands.

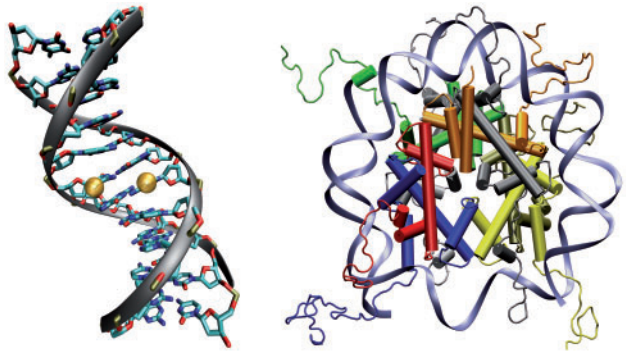

Fig. 1. Carriers of epigenetic information: DNA and nucleosome. The left panel shows a DNA double helix that is methylated symmetrically on both strands (orange spheres) at its center CpG (PDB structure: 329d). DNA methylation is the only epigenetic mechanism that directly targets the DNA. The right panel shows a nucleosome spindle consisting of eight histone proteins (center), around which two loops of DNA are wound (PDB structure: 1KX5). The nucleosome is subject to covalent modifications of its histones and to the binding of non-histone proteins.

Histone modifications (Kouzarides, 2007) are post-translational modifications of the core histone proteins that constitute the nucleosome (Fig. 1, right). The long and unstructured N-terminal tails by which histone proteins interact with neighboring nucleosomes are subject to various types of covalent modifications, including lysine and arginine methylation, lysine acetylation and serine phosphorylation. Histone modifications influence the nucleosome's assembly into higher-order packaging structures by moderating its DNA-binding affinity and by recruiting further chromatin remodeling complexes. The concept of the histone code (Turner, 2007) suggests that histone modifications are used combinatorially to program genes for activation during subsequent steps of cellular differentiation. Although the generality of this concept is controversial and remains difficult to test experimentally, it provides a plausible model for a number of recent observations, such as the programmed activation of tissue-specific transcription factors during differentiation of ES cells (Bernstein *et al*., 2006).

Non-histone proteins influence chromatin structure by interacting with histones and DNA in a number of ways. ATP-dependent chromatin remodeling complexes act like molecular machines and can directly move or displace nucleosomes along the DNA (Gangaraju and Bartholomew, 2007). A second group of proteins, which includes HP1 as well as the Polycomb and trithorax group complexes can be thought of as the readers and writers of the epigenome. They bind to the DNA or to specifically modified histones and catalyze other histone modifications or DNA methylation. The Polycomb group complex 2 for example catalyzes repressive histone methylations and recruits DNA methylation through its

Table 1. Methods for genome-wide mapping of epigenetic information

ChIP-on-chip combines **ch**romatin **i**mmuno**p**recipitation (ChIP) for enriching specific DNA fragments with the power of tiling microarrays for detecting differences between sample and control DNA (Buck and Lieb, 2004). Initially, cells are treated with formaldehyde to cross-link any DNA-bound proteins to the DNA. Next, the chromatin is extracted and sheared into small fragments, which are typically around 500 basepairs in length (this step limits the method's resolution). Using an antibody against a histone modification or a chromatin protein the corresponding fragments are enriched. The DNA is then released from these fragments and hybridized to a tiling microarray. Regions that are significantly overrepresented in the immunoprecipitated DNA relative to control DNA are regarded as epigenetically modified or protein-bound, depending on the antibody used. In a variant of ChIP-on-chip that is called **me**thyl-**D**NA **i**mmuno**p**recipitation (MeDIP), purified DNA is immunoprecipitated with an antibody against methylated cytosines, giving rise to genomic maps of DNA methylation. While these methods are used in a large number of laboratories worldwide, antibody quality remains a matter of concern and must be monitored carefully. Furthermore, background noise introduced by cross-hybridization and varying oligomer affinities should be accounted for during data processing. To foster data quality and standardization, the Microarray and Gene Expression Data Society has released a checklist of minimum required information about ChIP-on-chip experiments that are to be reported for any dataset (Microarray and Gene Expression Data Society, 2005). While ChIP-on-chip is increasingly replaced by ChIP-seq (see below) for genome-wide studies, it continues to be important for studies of localized genomic regions such as all promoter regions or a specific chromosome.
ChIP-seq is a variant of ChIP-on-chip that uses high-throughput DNA sequencing rather than tiling arrays for detecting differences between sample and control DNA (Barski et al., 2007; Mikkelsen et al., 2007). This method has two advantages over ChIP-on-chip: (i) data normalization is less of an issue because sequencing results in absolute read counts rather than relative hybridization scores and (ii) recent progress in sequencing-by-synthesis methods (e.g. by Roche/454 and Illumina/Solexa) makes ChIP-seq highly cost-efficient. ChIP-seq shares ChIP-on-chip's dependence on high-quality antibodies and has the additional drawback that extra steps are required to restrict ChIP-seq to specific subregions of the genome. Nevertheless, its unparalleled throughput makes ChIP-seq the prime candidate for a comprehensive human epigenome project.
Bisulfite sequencing exploits the ability of bisulfite to convert the DNA methylation state of cytosines into a methylation-dependent SNP (Hajkova *et al.*, 2002). The application of bisulfite sequencing is restricted to DNA methylation, for which it continues to be the gold standard due to its single-basepair resolution. DNA methylation patterns that are specific to a single cell can be obtained by combining bisulfite treatment with vector cloning and sequencing of a number of clones. However, for cost reasons bisulfite-treated DNA is often subjected to direct sequencing, which destroys any information about co-methylation in a particular cell but is sufficient for deriving profiles of average methylation.

interaction with a DNA methyltransferase (Schuettengruber *et al.*, 2007). Transcription factors can also affect chromatin structure, e.g. through recruitment of histone acetylases. Interestingly, there is evidence that sometimes transcription factor binding is maintained during cell division and would therefore qualify as mitotically heritable (Zhou *et al.*, 2005). Nevertheless, by convention rather than by definition transcription factor binding is not usually regarded as epigenetic.

In summary, a variety of epigenetic mechanisms jointly control the packaging of the DNA, thereby regulating which genes are accessible for transcription. Epigenetic mechanisms are highly interwoven and regulate their target genes (and each other) in a complex network of synergistic and antagonistic interactions. Disentangling this network both biochemically for a small number of representative genes and statistically from a whole-genome perspective, and relating the results to development and disease are important goals of epigenetic research. In the remainder of this review, we discuss arising bioinformatic challenges, and we show how computational methods have contributed and will continue to contribute to answering important epigenetic questions.

4 GENERATION, LOW-LEVEL PROCESSING AND QUALITY CONTROL OF EPIGENETIC DATA

Various experimental techniques have been developed for genome-wide mapping of epigenetic information (Table 1). These techniques follow a basic three-stage design. First, the epigenetic information is biochemically converted into genetic information, e.g. by enriching genomic regions that carry a particular histone modification in a DNA library. Second, standard DNA techniques such as tiling microarrays or sequencing are applied. Third, computational algorithms are used to infer the epigenetic information from the tiling array data or sequencing output. All experimental methods for epigenome mapping generate large amounts of data and require efficient ways of low-level data processing and quality control.

For ChIP-on-chip (Table 1), the key bioinformatic challenge is to derive a ranked list of overrepresented genomic regions from raw probe intensities. Although there are some similarities to the analysis of tiling array data for transcriptome mapping (see Royce *et al.*, 2005 for a review), most available algorithms are specifically targeted to peak finding in ChIP-on-chip data. The initial and still widely used solution employs a three-step process (Cawley *et al.*, 2004). First, the microarrays are quantile-normalized and standardized to a common median intensity. Second, a Wilcoxon rank sum test is applied locally on a sliding window to test for differential hybridization and to derive a Z-score for each probe. Third, significant probes are merged into regions of overrepresentation if sufficiently close to each other, and these regions are ranked by their combined Z-score. More recently, hidden Markov models were introduced to improve the detection accuracy (first implemented in HMMTiling, Li *et al.*, 2005), linear models were applied to control for differences in probe sensitivity [implemented in MAT (Johnson *et al.*, 2006) for Affymetrix one-color arrays and in MA2C (Song *et al.*, 2007) for NimbleGen two-color

arrays] and probabilistic binding models were used to improve spatial resolution (implemented in the JBD algorithm, Qi *et al*., 2006). Furthermore, several peak finding toolkits have been developed to facilitate routine processing of ChIP-on-chip datasets. TileMap is an easy-to-use peak finder for Affymetrix tiling array data, which has been applied in a number of independent studies (Ji and Wong, 2005); Ringo is a Bioconductor package for the analysis of ChIP-on-chip data from the widely used NimbleGen platform (Toedling *et al*., 2007); ChIPOTle is a basic peak finding macro for Excel, which does not take platform-specific information into account (Buck *et al*., 2005); and Tilescope is a fully integrated analysis pipeline that is applicable to data from both the Affymetrix and the NimbleGen platform (Zhang *et al*., 2007). In spite of the abundance of algorithms published recently, the peak finding problem for ChIP-on-chip data cannot be regarded as solved. In particular, current peak finders have problems with histone modifications that cover extended genomic regions and they seem to miss a substantial number of weak binding sites. In order to select a biologically meaningful cutoff that distinguishes between significant peaks and random fluctuations, experimental validation of a moderate number of detected peaks continues to be crucial. To guide this process, a framework has been proposed that can help identify most informative regions for validation (Du *et al*., 2006).

The key bioinformatic step of ChIP-seq (Table 1) is the fast and accurate mapping of short sequence reads to the reference genome. In principle, any seed-based alignment program such as blastn (http://www.ncbi.nlm.nih.gov/BLAST) or BLAT (Kent, 2002) is applicable to this task. Nevertheless, seed alignment strategies that are specifically optimized for reads from a particular sequencing platform have been reported to yield substantial increases in speed and coverage (Synamatix Sdn. Bhd., 2007). Two commercial solutions for short ChIP-seq reads are currently available, namely, the ELAND tool included in the Solexa analysis pipeline (http://www.solexa.com/) and the SXOligosearch software (http://www.synamatix.com/). In addition, a customized alignment protocol has been developed at the Broad Institute (Mikkelsen *et al*., 2007). Unlike relative probe intensities in ChIP-on-chip, each sequence read in a ChIP-seq experiment directly corresponds to a single chromatin fragment that was bound by the antibody during immunoprecipitation. For this reason, it is commonly assumed that ChIP-seq requires almost no normalization and that data analysis can be based directly on sequence read counts (Barski *et al*., 2007) or sliding window read counts (Mikkelsen *et al*., 2007). However, an important caveat is that the process of mapping tags to the reference genome can bias the analysis toward genomic regions with unique and complex sequence patterns. This is because short sequencing reads that (partially) overlap with low-complexity regions or with interspersed repeats stand a higher chance of being discarded for lack of unique genomic alignment.

Bisulfite sequencing (Table 1) requires customized analysis software that accounts for the 'fifth base', 5-methyl-cytosine. When bisulfite-treated DNA is sequenced directly (i.e. without vector cloning), the average methylation levels can be estimated using the ESME software (Lewin *et al*., 2004, freely available from http://www.epigenome.org/index.php?page=download). This software corrects for systematic bias induced by different molecular weights at methylation-specific SNPs and facilitates quality control. When subclones of bisulfite-treated DNA are sequenced, which is regarded as the gold standard for DNA methylation analysis, methylation patterns are inferred by aligning the clonal sequences to the genomic sequence. The BiQ Analyzer software (Bock *et al*., 2005) has been developed to simplify this analysis, to perform stringent quality control and to visualize the results. In addition, specialized primer design programs exist, of which Methyl Primer Express (freely available from http://www.appliedbiosystems.com/) is probably the most widely used. However, manual refinement is often necessary, suggesting that further improvements of primer design programs are needed.

5 EPIGENOME DATA ANALYSIS

Rapid progress of experimental technologies has given rise to several epigenome mapping initiatives (Table 2). These projects have been breaking ground not only in terms of applying and improving large-scale experimental methods, but also in terms of developing bioinformatic methods for analyzing their data.

This is particularly true for the ENCODE project, which has been designed from the onset as a close cooperation between experimental and computational biologists. Although the ENCODE project aims to map functional elements in the human genome rather than to resolve epigenetic questions, the methods and tools that emerged from this project contribute to epigenome data analysis in a number of ways. First, a method for unsupervised segmentation of chromatin data was developed based on wavelet smoothing and hidden Markov models (Thurman *et al*., 2007). When applied to selected ChIP-on-chip datasets from the ENCODE pilot phase, the algorithm neatly recovered the two main chromatin states: open and transcriptionally competent euchromatin as well as inaccessible and transcriptionally silent heterochromatin. Second, the joint statistical analysis of all 105 ChIP-on-chip datasets from the ENCODE pilot phase (Zhang *et al*., 2007) provides an example of exploratory data analysis on a large and heterogeneous dataset that includes substantial amounts of epigenetic information. Third, several alternative prediction methods for annotating functional promoters were developed and evaluated (Trinklein *et al*., 2007), indicating that epigenetic data can substantially improve the accuracy of promoter annotation. Fourth, a rigorous statistical test was developed that assesses the significance of overlap between two sets of genomic features, e.g. between CpG islands and unmethylated genomic regions (ENCODE Project Consortium, 2007). The authors show that—under relatively weak assumptions—their Genome Structure Correction method yields realistic *P*-values while other randomization-based methods tend to over-estimate significance. Fifth, the ENCODE project was accompanied by the systematic incorporation of epigenome datasets into the UCSC Genome Browser (Thomas *et al*., 2007), which now provides integrated visualization and standardized retrieval of various genome and epigenome datasets. Finally, the successful collaboration of experimental and bioinformatic researchers in the ENCODE project

Table 2. Large-scale epigenome mapping projects as of October 2007

Initiator	Summary	Current state	References
AHEAD Task Force (international)	The goal of the 'Alliance for Human Epigenomics and Disease' (AHEAD) is to initiate and coordinate a comprehensive human epigenome mapping project. Initially, focus is set on developing a suitable bioinformatic infrastructure and on performing epigenome mapping in a selection of normal tissues, which may provide the reference for subsequent mapping in abnormal cells	In the May 2007 roadmap update, the NIH selected epigenetics as one of two roadmap initiatives to be started immediately. This decision was partially based on a proposal submitted by the AHEAD Task Force	Alliance for Human Epigenomics and Disease (2007) Jones and Martienssen (2005)
ENCODE Project Consortium (international)	The NIH-funded 'Encyclopedia of DNA Elements' (ENCODE) project aims to map all functional elements in the human genome sequence. Although epigenome mapping is not its main goal, the project includes large-scale mapping of DNA methylation, histone modifications and other epigenetic information	The pilot project comprehensively analyzed 1% of the genome with results published in June 2007. In the production phase, selected analyses are performed on the entire human genome	ENCODE Project Consortium (2007) ENCODE Project Consortium (2004)
HEP Project Consortium (UK/D/F)	The partially EU-funded 'Human Epigenome Project' (HEP) analyzed DNA methylation in 43 unrelated individuals at single basepair resolution. Although the analysis was confined to selected regions on three chromosomes, it is the largest high-resolution, multi-individual epigenome dataset published to date	The results of the pilot phase dataset were published in 2004 and the results of the main phase were published in 2006	Eckhardt *et al.* (2006) Rakyan *et al.* (2004)
HEROIC Project Consortium (EU)	The 'High-throughput Epigenetic Regulatory Organisation In Chromatin' (HEROIC) project is a multi-center EU project that applies ChIP-on-chip, chromosome interaction analysis and whole-genome nuclear localization assays to understanding human genome regulation	This EU 'Integrated Project' is funded from 2005 to 2010 and does not involve synchronized pilot or production phases	HEROIC Project Consortium (2005)
Broad Institute of MIT and Harvard (US)	In a large single-center study, ChIP-seq was used to derive genome-wide maps of chromatin state for mouse ES cells, neural progenitor cells and embryonic fibroblasts	Initial results were published in July 2007	Mikkelsen *et al.* (2007)
National Heart, Lung, and Blood Institute of the NIH (US)	In a large single-center study, ChIP-seq was used to derive genome-wide maps of chromatin state for human T-cells	Initial results were published in June 2007	Barski *et al.* (2007)

has raised the awareness of synergies between wet lab and computational research. The AHEAD task force for example acknowledges the critical importance of bioinformatic methods and infrastructure in their proposal for a human epigenome project (Alliance for Human Epigenomics and Disease, 2007).

Although the bioinformatic focus of the other large-scale epigenome projects (Table 2) was less pronounced than in the ENCODE project, important bioinformatic progress arose from them as well. The HEROIC project played a catalyzing role for the development of epigenome data storage, visualization and analysis infrastructure in Europe. In fact, in its regulatory builds the Ensembl genome browser (Hubbard *et al.*, 2007) will increasingly incorporate epigenetic information such as genome-wide maps of DNA methylation and histone modifications (P.Flicek, personal communication). The HEP project for the first time explored the challenges and opportunities of high-resolution epigenome analysis in multiple unrelated individuals. And the two large-scale ChIP-seq projects that have been completed recently underline the relevance of analyzing various epigenetic mechanisms simultaneously in a single cell type (Barski *et al.*, 2007) and at multiple stages during cellular differentiation (Mikkelsen *et al.*, 2007). While the general picture emerging from these studies is consistent with mammalian epigenomes being segmented into alternating regions of open and condensed chromatin, many more sophisticated concepts become visible only at high resolution and when analyzing various epigenetic mechanisms simultaneously. For example, it has been shown recently that computational integration of several histone modification maps can be used to predict the locations of enhancers in the human genome, even where these are invisible to phylogenetic methods (Heintzman *et al.*, 2007; Roh *et al.*, 2007).

However, these pioneering epigenome mapping projects also highlight two major impediments to epigenome data analysis: the unsolved problem of public data storage and the lack of experimental standardization. Public data storage in databases such as GenBank and ArrayExpress has played an important role for bioinformatic research, by making primary data available for meta-analysis and benchmarking

studies. However, with the advent of ChIP-seq, the central collection of primary data is hitting technical limitations. A typical three-day run on a Solexa sequencer gives rise to hundreds of gigabytes of primary image data and several gigabases of sequence reads, and in less than a year a single Solexa sequencer could generate the equivalent of all sequence data stored in GenBank until 2005. In addition to developing more efficient methods for data processing and storage, it will therefore be necessary to work out policies that regulate how primary data should be archived and how the benefits of publicly available primary data can be maintained when central storage is no longer an option. The second problem, lack of experimental standardization, hampers the computational integration of epigenetic datasets from different studies. Because epigenetic information is tissue-specific and because methods such as ChIP-on-chip are highly sensitive to variation in the experimental protocol, most epigenome datasets that have been published to date are—strictly speaking—incomparable. Nevertheless, several meta-analyses of ChIP-on-chip data have been published and significant correlations have been observed for epigenetic modifications that are associated with an open chromatin structure (Bock *et al*., 2007; Parisi *et al*., 2007; Zhang *et al*., 2007), while an initial comparison for repressive histone modifications indicated substantially less correlation between different datasets (C.Bock, unpublished data). Although complete standardization is neither realistic nor desirable, it seems advisable to focus different epigenome mapping projects on the same set of cell lines, as is done in the ENCODE project.

6 EPIGENOME PREDICTION: INFERRING EPIGENETIC STATES FROM THE DNA SEQUENCE

A substantial amount of bioinformatic research has been devoted to the prediction of epigenetic information from characteristics of the genomic sequence. Such predictions serve a dual purpose. First, accurate epigenome predictions can substitute for experimental data, to some degree, which is particularly relevant for newly discovered epigenetic mechanisms and for species other than human and mouse. Second, prediction algorithms build statistical models of epigenetic information from training data and can therefore act as a first step toward quantitative modeling of an epigenetic mechanism.

Promoter prediction—an important topic in bioinformatics since the early 1990s—can be regarded as the first attempt to predict epigenetic states from the DNA sequence. This is because active promoters are characterized by an open and transcriptionally permissive chromatin structure and exhibit specific epigenetic properties such as absence of DNA methylation and enrichment for histone acetylation. A large number of promoter prediction methods have been developed during the last two decades, most of which use DNA sequence characteristics combined with a machine-learning algorithm to identify candidate promoters (see Bajic *et al*., 2004 for a comprehensive overview and benchmarking analysis). In the highly annotated human genome, promoter prediction has lost some of its relevance and researchers are increasingly focusing on advanced questions of transcription control, such as inferring tissue-specific signals (Smith *et al*., 2007) and reconstructing transcriptional networks (Bulcke *et al*., 2006).

CpG island prediction has some overlap with promoter prediction because the majority of promoters in mammalian genomes co-localize with CpG islands (Antequera, 2003). However, CpG islands play a more general role as mediators of open chromatin structure, and they frequently overlap with enhancers and other regulatory elements. CpG islands were originally discovered by a striking absence of DNA methylation (Cooper *et al*., 1983), which is regarded as a constitutive feature of CpG islands. The absence of DNA methylation in the germline reduces CpG-to-TpG mutation rates inside CpG islands, leading to overrepresentation of CpGs relative to the genomic average. CpG islands are often predicted solely based on their GC and CpG frequencies, and multiple variants of the original definition (Gardiner-Garden and Frommer, 1987) are in use. However, a recent study showed that these definitions yield high false positive rates, and a refined concept of bona fide CpG islands based on large-scale epigenome prediction was proposed (Bock *et al*., 2007).

DNA methylation prediction is conceptually easier than the prediction of more volatile epigenetic mechanisms because DNA methylation patterns exhibit relatively low tissue specificity compared to other epigenetic information. Therefore, it is not surprising that similar approaches applied to DNA methylation data for blood (Bock *et al*., 2006) and brain tissue (Das *et al*., 2006; Fang *et al*., 2006) yielded comparable results. In all three cases, machine-learning methods were used to derive a classifier for presence or absence of DNA methylation in a given region. Prediction accuracies were high, and the most predictive attributes included CpG-rich sequence patterns (Bock *et al*., 2006; Das *et al*., 2006; Fang *et al*., 2006), specific DNA structure properties and repetitive DNA elements (Bock *et al*., 2006) as well as certain transcription factor binding sites (Fang *et al*., 2006). Interestingly, a similar method could also predict which genomic regions are prone to becoming methylated in a cell line overexpressing the DNA methyltransferase DNMT1 (Feltus *et al*., 2003).

Prediction of nucleosome positioning is based on the observation that the sequence composition of DNA molecules strongly affects their nucleosome affinity, i.e. how easily they can be wound around a nucleosome (Satchwell *et al*., 1986). Several recent papers showed that this *in vitro* effect has significant impact on the genomic positioning of nucleosomes *in vivo* (Ioshikhes *et al*., 2006; Peckham *et al*., 2007; Segal *et al*., 2006). Although all three papers focus their analysis on yeast, the highly conserved nature of the nucleosome suggests a general applicability of these results. Indeed, Segal *et al*. observe that the predictions change little when training is performed on nucleosome positioning data from chicken instead of yeast, and Ioshikhes *et al*. find that an alignment of multiple yeast species can increase prediction accuracy.

Successful prediction has also been reported for several other epigenetic phenomena: DNase I hypersensitive sites could be distinguished from a random control set using support vector machines with *k*-mer sequence motifs as prediction attributes (Noble *et al*., 2005). Polycomb/trithorax response elements in Drosophila were identified by sequence criteria

(Ringrose *et al*., 2003), a finding that may not easily translate to humans because mammalian Polycomb/trithorax response elements exhibit less identifiable sequence patterns. Imprinted genes were predicted using a wide range of genomic features (sequence motifs, CpG islands, repeats, predicted transcription factor binding sites) and a commercial support vector machine-based data mining suite (Luedi *et al*., 2005). Finally, genes that escape X-chromosome inactivation were predicted by a support vector machine and found to be enriched in Alu repeats and CpG-rich sequence motifs (Wang *et al*., 2006). However, a conclusive assessment of prediction methods for imprinted genes and for genes that escape inactivation seems problematic due to the small number of affected genes, their clustering in small genomic regions and the difficulty of independent experimental validation.

In summary, a large number of genomic regions exhibit clearly detectable epigenetic footprints in their DNA sequence. This has practical applications for genome annotation and also challenges the notion of *genome* and *epigenome* as two largely independent systems of inheritance working at different time scales. Rather, the genome seems to encode not only genes and *cis*-regulatory elements, but also a default epigenetic state that becomes active in the absence of other regulatory influences such as the binding of transcription factors or the activity of chromatin remodeling complexes. This interpretation is consistent with the emerging concept of multitasking genomes, which simultaneously (and on top of each other) encode genes and their regulation (Kapranov *et al*., 2007). Furthermore, this model provides an explanation for the fact that only a small subset of suitable consensus binding motifs are actually used by transcription factors *in vivo*. A new generation of *in silico* methods for detecting transcription factor binding has already started to benefit from epigenome prediction in order to distinguish functional from non-functional sites (Narlikar *et al*., 2007).

7 CANCER EPIGENETICS: TOWARD IMPROVED DIAGNOSIS AND THERAPY

It has been known for a long time that mutations and chromosomal deletions can irreversibly destroy tumor suppressor genes and are pivotal events in cancer progression. In contrast, the importance of epigenetic mechanisms for tumor development has been appreciated more recently (see Feinberg and Tycko, 2004 for a historical account of cancer epigenetics). It is now clear that a substantial proportion of silenced tumor suppressor genes are lost due to epigenetic deactivation rather than sequence damages (Esteller, 2007; Jones and Baylin, 2007). Furthermore, a comparison between the epigenetic characteristics of cancer cells and stem cells suggests that epigenetic deregulation may program cells for cancer-like behavior long before they are visually identifiable as tumor cells (Feinberg *et al*., 2006).

The important role of epigenetic defects for cancer opens up new opportunities for improved diagnosis and therapy. Early diagnosis profits from the fact that epigenetic aberrations occur early during tumorigenesis and are frequently detectable in peripheral blood when destroyed tumor cells leak DNA into the bloodstream (Laird, 2003). Epigenetic cancer therapy exploits the fact that—in contrast to genomic damage—epigenetic aberrations are pharmacologically reversible (Yoo and Jones, 2006). These active areas of research give rise to two questions that are particularly amenable to bioinformatic analysis. First, given a list of genomic regions exhibiting epigenetic differences between tumor cells and controls (or between different disease subtypes), can we detect common patterns or find evidence of a functional relationship of these regions to cancer? Second, can we use bioinformatic methods in order to improve diagnosis and therapy by detecting and classifying important disease subtypes?

Keshet *et al*. faced a typical instance of the first question, after MeDIP analysis in two cancer cell lines and in a set of primary tumors had detected hundreds of genes whose promoters were selectively methylated in cancer (Keshet *et al*., 2006). They applied several bioinformatic methods in order to identify common patterns among these genes, including overrepresentation analysis of Gene Ontology terms, sequence motif discovery, genomic clustering analysis and comparison with public gene expression data. Based on these computational analyses, they concluded that only a small percentage of epigenetically silenced genes in cancer cells are tumor suppressor genes. In contrast, many of the genes that are unlikely to be tumor suppressor genes exhibit certain sequence patterns, which may predispose them for epigenetic silencing—as a side effect rather than cause of tumor development. A recent study elaborated on this finding by applying a more advanced motif discovery pipeline and could identify additional sequence motifs on the same dataset (Eden *et al*., 2007). The observation that epigenetically silenced genes often share certain sequence motifs in their promoters has also been used in order to detect new candidates for cancer-specific hypermethylation (Goh *et al*., 2007). To address the substantial class bias—only a small percentage of genes become hypermethylated in a particular cancer—and the lack of an experimental control set, Goh *et al*. devised an algorithm that iteratively combines unsupervised clustering and supervised prediction. Furthermore, the recent discovery of a link between DNA hypermethylation in cancer and Polycomb binding in ES cells using a combination of bioinformatic comparisons and experimental validation (Ohm *et al*., 2007; Schlesinger *et al*., 2007; Widschwendter *et al*., 2007) highlights the synergistic power of computational and experimental methods in cancer epigenetics. Future studies toward understanding the epigenetic characteristics of cancer cells will benefit from the recently launched PubMeth database, which aggregates literature data about which genes have been reported hypermethylated for which cancer (Ongenaert *et al*., 2007).

The second question is aimed at the discovery and validation of biomarkers for cancer diagnosis, prognosis and therapy optimization (Laird, 2003). In an early study on DNA methylation patterns in leukemia, support vector machines applied to DNA methylation microarray data could accurately distinguish between two important disease subtypes, acute lymphoblastic leukemia and acute myeloid leukemia (Model *et al*., 2001). In a series of papers, Siegmund and coworkers developed (Marjoram *et al*., 2006; Siegmund *et al*., 2004) and applied (Weisenberger *et al*., 2006) clustering methods for unsupervised

discovery of epigenetically distinct cancer subtypes. They could show that a well-defined subset of colon cancer patients exhibit substantially elevated levels of DNA hypermethylation, and they developed a biomarker for diagnosing this disease subtype. Epigenetic biomarkers also play an increasing role for therapy optimization. For example, clinical trials showed that cancer-specific DNA methylation of the *MGMT* gene promoter can make glioblastomas (brain tumors) more susceptible to chemotherapy with alkylating agents (Hegi *et al*., 2005). A combination of bioinformatic methods and experiments was recently used to optimize DNA methylation analysis of *MGMT* and to develop it into a routine clinical biomarker for personalized cancer therapy (Mikeska *et al*., 2007). However, in spite of the fast progress in epigenetic cancer diagnosis, few epigenetic cancer biomarkers have yet been validated in large patient cohorts and substantial work remains to be done before epigenetic cancer diagnosis will start having a measurable positive effect on disease burden in the population.

8 FUTURE DIRECTIONS

The first wave of research in the field of computational epigenetics was driven by rapid progress of experimental methods for data generation, which required adequate computational methods for low-level data processing and quality control, prompted epigenome prediction studies as a means of understanding the genomic distribution of epigenetic information, and provided the foundation for initial projects on cancer epigenetics. While these topics will continue to be highly relevant areas of research and the mere quantity of epigenetic data arising from epigenome projects poses a major bioinformatic challenge, we also expect that the focus of computational studies will significantly broaden and deepen. First, epigenome data analysis will increasingly take the proteins into account that read and write epigenetic information, as well as their interaction partners and regulatory networks. Such reverse engineering of epigenetic regulation could lead to a quantitative model and, ultimately, rational manipulation of the core circuitry that controls cell fate and pluripotency (Boyer *et al*., 2005). Second, the decreasing cost of epigenome mapping will enable quantitative analysis of epigenetic variation in human populations. Recent twin studies suggest that both environmental influences (Fraga *et al*., 2005) and genetic variation (Heijmans *et al*., 2007) influence epigenetic variation. It will be a daunting bioinformatic task to distill putative functional connections from the integration of epigenome data with gene expression profiles and haplotype maps for a large sample from a heterogeneous population. Third, epigenome mapping in multiple species will add an evolutionary perspective to computational epigenetics. Initial results suggest that orthologous regions in different mammals carry similar epigenetic information (Bernstein *et al*., 2005; Enard *et al*., 2004), which is expected because the DNA encodes parts of its epigenetic state (Bock *et al*., 2007; Segal *et al*., 2006). It will be interesting to see whether comparative epigenomics can significantly improve our ability to identify functionally important sites in the human genome, as is the case for comparative genomics. Fourth, theoretical modeling will provide a way to fathom our mechanistic and quantitative understanding of epigenetic mechanisms. For example, two recent studies could show that co-operativity among the proteins that write epigenetic information is required for stably maintaining the state of an epigenetic switch in the presence of highly dynamic fluctuations at the molecular level (Dodd *et al*., 2007; Sontag *et al*., 2006). Modeling studies can thus help explain how the high-level phenomena that we observe for epigenetic regulation emerge from the dynamic interplay of various epigenetic mechanisms. Fifth, the development of powerful and easy-to-use 'statistical genome browsers' will enable biologists to perform complex epigenome data analysis without requiring strong statistical or programming skills. The Galaxy web service (Blankenberg *et al*., 2007)—which lets users design and execute genome analyses through an intuitive web front-end—is a first step in this direction and further tools that are more specifically targeted to epigenetic data will follow. Sixth, epigenetic mechanisms could turn out to play a role in diseases other than cancer, as there is strong circumstantial evidence for epigenetic regulation being involved in mental disorders, autoimmune diseases and other complex diseases (Bjornsson *et al*., 2004; Feinberg, 2007). Bioinformatic methods such as text mining and exploratory data mining may play a role in identifying and prioritizing concrete hypotheses for experimental validation.

In conclusion, exciting times are ahead for research in computational epigenetics!

ACKNOWLEDGEMENTS

We would like to thank Jörn Walter and Paul Flicek for critically reading the manuscript. Funding by the Max Planck Society is greatly acknowledged.

Conflict of Interest: none declared.

REFERENCES

Alliance for Human Epigenomics and Disease (2007) Proposal for an International AHEAD Pilot Project. Available: http://www.aacr.org/Uploads/DocumentRepository/TaskForces/ahead_pilot_project_proposal_may2007.pdf (28 October 2007, date last accessed).

Antequera,F. (2003) Structure, function and evolution of CpG island promoters. *Cell. Mol. Life Sci*., **60**, 1647–1658.

Bajic,V.B. *et al*. (2004) Promoter prediction analysis on the whole human genome. *Nat. Biotechnol*., **22**, 1467–1473.

Barski,A. *et al*. (2007) High-resolution profiling of histone methylations in the human genome. *Cell*, **129**, 823–837.

Bernstein,B.E. *et al*. (2005) Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell*, **120**, 169–181.

Bernstein,B.E. *et al*. (2006) A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, **125**, 315–326.

Bernstein,B.E. *et al*. (2007) The mammalian epigenome. *Cell*, **128**, 669–681.

Bird,A. (2002) DNA methylation patterns and epigenetic memory. *Genes Dev*., **16**, 6–21.

Bjornsson,H.T. *et al*. (2004) An integrated epigenetic and genetic approach to common human disease. *Trends Genet*., **20**, 350–358.

Blankenberg,D. *et al*. (2007) A framework for collaborative analysis of ENCODE data: making large-scale analyses biologist-friendly. *Genome Res*., **17**, 960–964.

Bock,C. *et al*. (2005) BiQ Analyzer: visualization and quality control for DNA methylation data from bisulfite sequencing. *Bioinformatics*, **21**, 4067–4068.

Bock,C. *et al*. (2006) CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats and predicted DNA structure. *PLoS Genet*., **2**, e26.
Bock,C. *et al*. (2007) CpG island mapping by epigenome prediction. *PLoS Comput*. *Biol*., **3**, e110.
Boyer,L.A. *et al*. (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, **122**, 947–956.
Buck,M.J. and Lieb,J.D. (2004) ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics*, **83**, 349–360.
Buck,M.J. *et al*. (2005) ChIPOTle: a user-friendly tool for the analysis of ChIP-chip data. *Genome Biol*., **6**, R97.
Bulcke,D. *et al*. (2006) Inferring transcriptional networks by mining 'omics' data. *Curr*. *Bioinformatics*, **1**, 313.
Cawley,S. *et al*. (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell*, **116**, 499–509.
Cooper,D.N. *et al*. (1983) Unmethylated domains in vertebrate DNA. *Nucleic Acids Res*., **11**, 647–658.
Das,R. *et al*. (2006) Computational prediction of methylation status in human genomic sequences. *Proc*. *Natl Acad*. *Sci*. *USA*, **103**, 10713–10716.
Dillon,N. (2006) Gene regulation and large-scale chromatin organization in the nucleus. *Chromosome Res*., **14**, 117–126.
Dodd,I.B. *et al*. (2007) Theoretical analysis of epigenetic cell memory by nucleosome modification. *Cell*, **129**, 813–822.
Drake,J.W. *et al*. (1998) Rates of spontaneous mutation. *Genetics*, **148**, 1667–1686.
Du,J. *et al*. (2006) A supervised hidden markov model framework for efficiently segmenting tiling array data in transcriptional and chIP-chip experiments: systematically incorporating validated biological knowledge. *Bioinformatics*, **22**, 3016–3024.
Eckhardt,F. *et al*. (2006) DNA methylation profiling of human chromosomes 6, 20 and 22. *Nat*. *Genet*., **38**, 1378–1385.
Eden,E. *et al*. (2007) Discovering motifs in ranked lists of DNA sequences. *PLoS Comput*. *Biol*., **3**, e39.
Enard,W. *et al*. (2004) Differences in DNA methylation patterns between humans and chimpanzees. *Curr*. *Biol*., **14**, R148–R149.
ENCODE Project Consortium (2004) The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science*, **306**, 636–640.
ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
Esteller,M. (2007) Cancer epigenomics: DNA methylomes and histone-modification maps. *Nat*. *Rev*. *Genet*., **8**, 286–298.
Fang,F. *et al*. (2006) Predicting methylation status of CpG islands in the human brain. *Bioinformatics*, **22**, 2204–2209.
Feinberg,A.P. (2007) Phenotypic plasticity and the epigenetics of human disease. *Nature*, **447**, 433–440.
Feinberg,A.P. and Tycko,B. (2004) The history of cancer epigenetics. *Nat*. *Rev*. *Cancer*, **4**, 143–153.
Feinberg,A.P. *et al*. (2006) The epigenetic progenitor origin of human cancer. *Nat*. *Rev*. *Genet*., **7**, 21–33.
Feltus,F.A. *et al*. (2003) Predicting aberrant CpG island methylation. *Proc*. *Natl Acad*. *Sci*. *USA*, **100**, 12253–12258.
Fraga,M.F. and Esteller,M. (2007) Epigenetics and aging: the targets and the marks. *Trends Genet*., **23**, 413–418.
Fraga,M.F. *et al*. (2005) Epigenetic differences arise during the lifetime of monozygotic twins. *Proc*. *Natl Acad*. *Sci*. *USA*, **102**, 10604–10609.
Gangaraju,V.K. and Bartholomew,B. (2007) Mechanisms of ATP dependent chromatin remodeling. *Mutat*. *Res*., **618**, 3–17.
Gardiner-Garden,M. and Frommer,M. (1987) CpG islands in vertebrate genomes. *J*. *Mol*. *Biol*., **196**, 261–282.
Goh,L. *et al*. (2007) Genomic sweeping for hypermethylated genes. *Bioinformatics*, **23**, 281–288.
Grant-Downton,R.T. and Dickinson,H.G. (2006) Epigenetics and its implications for plant biology 2. The 'epigenetic epiphany': epigenetics, evolution and beyond. *Ann*. *Bot*. *(Lond.)*, **97**, 11–27.
Hajkova,P. *et al*. (2002) DNA-methylation analysis by the bisulfite-assisted genomic sequencing method. *Methods Mol*. *Biol*., **200**, 143–154.
Hegi,M.E. *et al*. (2005) MGMT gene silencing and benefit from temozolomide in glioblastoma. *N*. *Engl*. *J*. *Med*., **352**, 997–1003.
Heijmans,B.T. *et al*. (2007) Heritable rather than age-related environmental and stochastic factors dominate variation in DNA methylation of the human IGF2/H19 locus. *Hum*. *Mol*. *Genet*., **16**, 547–554.
Heintzman,N.D. *et al*. (2007) Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat*. *Genet*., **39**, 311–318.
HEROIC Project Consortium (2005) High-throughput Epigenetic Regulatory Organisation In Chromatin - Project Fact Sheet. Available: http://cordis.europa.eu/fetch?CALLER=FP6_PROJ&ACTION=D&DOC=1&CAT=PROJ&QUERY=1183993108794&RCN=78439 (28 October 2007, date last accessed).
Hubbard,T.J. *et al*. (2007) Ensembl 2007. *Nucleic Acids Res*., **35**, D610–D617.
Ioshikhes,I.P. *et al*. (2006) Nucleosome positions predicted through comparative genomics. *Nat*. *Genet*., **38**, 1210–1215.
Ji,H. and Wong,W.H. (2005) TileMap: create chromosomal map of tiling array hybridizations. *Bioinformatics*, **21**, 3629–3636.
Johnson,W.E. *et al*. (2006) Model-based analysis of tiling-arrays for ChIP-chip. *Proc*. *Natl Acad*. *Sci*. *USA*, **103**, 12457–12462.
Jones,P.A. and Baylin,S.B. (2007) The epigenomics of cancer. *Cell*, **128**, 683–692.
Jones,P.A. and Martienssen,R. (2005) A blueprint for a Human Epigenome Project: the AACR Human Epigenome Workshop. *Cancer Res*., **65**, 11241–11246.
Kapranov,P. *et al*. (2007) Genome-wide transcription and the implications for genomic organization. *Nat*. *Rev*. *Genet*., **8**, 413–423.
Kent,W.J. (2002) BLAT–the BLAST-like alignment tool. *Genome Res*., **12**, 656–664.
Keshet,I. *et al*. (2006) Evidence for an instructive mechanism of de novo methylation in cancer cells. *Nat*. *Genet*., **38**, 149–153.
Kouzarides,T. (2007) Chromatin modifications and their function. *Cell*, **128**, 693–705.
Laird,P.W. (2003) The power and the promise of DNA methylation markers. *Nat*. *Rev*. *Cancer*, **3**, 253–266.
Lewin,J. *et al*. (2004) Quantitative DNA methylation analysis based on four-dye trace data from direct sequencing of PCR amplificates. *Bioinformatics*, **20**, 3005–3012.
Li,W. *et al*. (2005) A hidden Markov model for analyzing ChIP-chip experiments on genome tiling arrays and its application to p53 binding sequences. *Bioinformatics*, **21** (**Suppl. 1**), i274–i282.
Luedi,P.P. *et al*. (2005) Genome-wide prediction of imprinted murine genes. *Genome Res*., **15**, 875–884.
Marjoram,P. *et al*. (2006) Cluster analysis for DNA methylation profiles having a detection threshold. *BMC Bioinformatics*, **7**, 361.
Microarray and Gene Expression Data Society (2005) The MIAME Checklist – update January 2005. Available: http://www.mged.org/Workgroups/MIAME/MIAMEchecklist_chipchip.pdf (28 October 2007, date last accessed).
Mikeska,T. *et al*. (2007) Optimization of Quantitative MGMT Promoter Methylation Analysis Using Pyrosequencing and Combined Bisulfite Restriction Analysis. *J*. *Mol*. *Diagn*., **9**, 368–381.
Mikkelsen,T.S. *et al*. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, **448**, 553–560.
Model,F. *et al*. (2001) Feature selection for DNA methylation based cancer classification. *Bioinformatics*, **17** (**Suppl. 1**), S157–S164.
Narlikar,L. *et al*. (2007) Nucleosome occupancy information improves de novo motif discovery. In: Speed,T.P. and Huang,H. (eds.) *Research in Computational Molecular Biology, 11th Annual International Conference, RECOMB 2007, Oakland, CA, USA, April 21–25, 2007, Proceedings*. Springer-Verlag, New York.
Noble,W.S. *et al*. (2005) Predicting the in vivo signature of human gene regulatory sequences. *Bioinformatics*, **21** (**Suppl. 1**), i338–i343.
Ohm,J.E. *et al*. (2007) A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. *Nat*. *Genet*., **39**, 237–242.
Ongenaert,M. *et al*. (2007) PubMeth: a cancer methylation database combining text-mining and expert annotation. *Nucleic Acids Res*.
Parisi,F. *et al*. (2007) Identifying synergistic regulation involving c-Myc and sp1 in human tissues. *Nucleic Acids Res*., **35**, 1098–1107.
Peaston,A.E. and Whitelaw,E. (2006) Epigenetics and phenotypic variation in mammals. *Mamm*. *Genome*, **17**, 365–374.
Peckham,H.E. *et al*. (2007) Nucleosome positioning signals in genomic DNA. *Genome Res*., **17**, 1170–1177.
Qi,Y. *et al*. (2006) High-resolution computational models of genome binding events. *Nat*. *Biotechnol*., **24**, 963–970.

Rakyan,V.K. *et al*. (2004) DNA methylation profiling of the human major histocompatibility complex: a pilot study for the human epigenome project. *PLoS Biol*., **2**, e405.
Reik,W. (2007) Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature*, **447**, 425–432.
Richards,E.J. (2006) Inherited epigenetic variation – revisiting soft inheritance. *Nat. Rev. Genet*., **7**, 395–401.
Ringrose,L. *et al*. (2003) Genome-wide prediction of Polycomb/Trithorax response elements in Drosophila melanogaster. *Dev. Cell*, **5**, 759–771.
Roh,T.Y. *et al*. (2007) Genome-wide prediction of conserved and nonconserved enhancers by histone acetylation patterns. *Genome Res*., **17**, 74–81.
Royce,T.E. *et al*. (2005) Issues in the analysis of oligonucleotide tiling microarrays for transcript mapping. *Trends Genet*., **21**, 466–475.
Russo,V.E.A. *et al*. (1996) *Epigenetic Mechanisms of Gene Regulation*. Cold Spring Harbor Laboratory Press, Plainview, N.Y.
Satchwell,S.C. *et al*. (1986) Sequence periodicities in chicken nucleosome core DNA. *J. Mol. Biol*., **191**, 659–675.
Schlesinger,Y. *et al*. (2007) Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. *Nat. Genet*., **39**, 232–236.
Schuettengruber,B. *et al*. (2007) Genome regulation by polycomb and trithorax proteins. *Cell*, **128**, 735–745.
Segal,E. *et al*. (2006) A genomic code for nucleosome positioning. *Nature*, **442**, 772–778.
Siegmund,K.D. *et al*. (2004) A comparison of cluster analysis methods using DNA methylation data. *Bioinformatics*, **20**, 1896–1904.
Smith,A.D. *et al*. (2007) Tissue-specific regulatory elements in mammalian promoters. *Mol. Syst. Biol*., **3**, 73.
Solter,D. (2006) Imprinting today: end of the beginning or beginning of the end? *Cytogenet. Genome Res*., **113**, 12–16.
Song,J.S. *et al*. (2007) Model-based analysis of two-color arrays (MA2C). *Genome Biol*., **8**, R178.
Sontag,L.B. *et al*. (2006) Dynamics, stability and inheritance of somatic DNA methylation imprints. *J. Theor. Biol*., **242**, 890–899.
Surani,M.A. *et al*. (2007) Genetic and epigenetic regulators of pluripotency. *Cell*, **128**, 747–762.
Synamatix Sdn. Bhd. (2007) SXOligoSearch Supporting Document. Available: http://synasite.mgrc.com.my:8080/sxog/files/SXOligoSearch_benchmark.pdf (28 October 2007, date last accessed).
Thomas,D.J. *et al*. (2007) The ENCODE Project at UC Santa Cruz. *Nucleic Acids Res*., **35**, D663–D667.
Thurman,R.E. *et al*. (2007) Identification of higher-order functional domains in the human ENCODE regions. *Genome Res*., **17**, 917–927.
Toedling,J. *et al*. (2007) Ringo – an R/Bioconductor package for analyzing ChIP-chip readouts. *BMC Bioinformatics*, **8**, 221.
Trinklein,N.D. *et al*. (2007) Integrated analysis of experimental data sets reveals many novel promoters in 1% of the human genome. *Genome Res*., **17**, 720–731.
Turner,B.M. (2007) Defining an epigenetic code. *Nat. Cell Biol*., **9**, 2–6.
Ushijima,T. *et al*. (2003) Fidelity of the methylation pattern and its variation in the genome. *Genome Res*., **13**, 868–874.
Wang,Z. *et al*. (2006) Evidence of influence of genomic DNA sequence on human X chromosome inactivation. *PLoS Comput. Biol*., **2**, e113.
Weber,M. and Schübeler,D. (2007) Genomic patterns of DNA methylation: targets and function of an epigenetic mark. *Curr. Opin. Cell Biol*., **19**, 273–280.
Weisenberger,D.J. *et al*. (2006) CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. *Nat. Genet*., **38**, 787–793.
Widschwendter,M. *et al*. (2007) Epigenetic stem cell signature in cancer. *Nat. Genet*., **39**, 157–158.
Woodcock,C.L. (2006) Chromatin architecture. *Curr. Opin. Struct. Biol*., **16**, 213–220.
Yoo,C.B. and Jones,P.A. (2006) Epigenetic therapy of cancer: past, present and future. *Nat. Rev. Drug Discov*., **5**, 37–50.
Zhang,Z.D. *et al*. (2007a) Statistical analysis of the genomic distribution and correlation of regulatory elements in the ENCODE regions. *Genome Res*., **17**, 787–797.
Zhang,Z.D. *et al*. (2007b) Tilescope: online analysis pipeline for high-density tiling microarray data. *Genome Biol*., **8**, R81.
Zhou,G.L. *et al*. (2005) Memory mechanisms of active transcription during cell division. *Bioessays*, **27**, 1239–1245.