EPIGENOMICS: BEYOND CpG ISLANDS

Melissa J. Fazzari and John M. Greally†*

Epigenomic studies aim to define the location and nature of the genomic sequences that are epigenetically modified. Much progress has been made towards whole-genome epigenetic profiling using molecular techniques, but the analysis of such large and complex data sets is far from trivial given the correlated nature of sequence and functional characteristics within the genome. We describe the statistical solutions that help to overcome the problems with data-set complexity, in anticipation of the imminent wealth of data that will be generated by new genome-wide epigenetic profiling and DNA sequence analysis techniques. So far, epigenomic studies have succeeded in identifying CpG islands, but recent evidence points towards a role for transposable elements in epigenetic regulation, causing the fields of study of epigenetics and transposable element biology to converge.

Epigenetic inheritance involves the transmission of information not encoded in DNA sequences from cell to daughter cell or from generation to generation. Covalent modifications of the DNA or its packaging histones are responsible for transmitting epigenetic information. Epigenomics can be defined as a genome-wide approach to studying epigenetics. This term encompasses whole-genome studies of epigenetic processes and the identification of the DNA sequences that specify where the epigenetic processes are targeted. The former aspect of epigenomics has been the subject of previous reviews[1–6], but the mining of genomic sequence annotations has added an interesting facet to our understanding of epigenomics and is the main focus of this review. The central goal of epigenomics is to define the DNA sequence features that direct epigenetic processes. That such features exist is evident from the example of CpG islands, originally defined as representing the hypomethylated fraction of the genome[7] and subsequently characterized in terms of DNA base composition[8]. We look critically at CpG island biology in the context of the more detailed genome sequence that has been generated since these original studies. The hope for identifying additional DNA sequence features that direct epigenetic processes is now founded on a two-pronged approach that involves genome-wide epigenetic assays and sequence data mining. The results of such complex, large-scale studies are complicated by issues of correlation and causality — for example, the DNA sequence feature might be the effect of the epigenetic process rather than mechanistically involved in directing it. As new techniques to characterize epigenetic processes throughout the genome are being applied, we have the potential to generate large amounts of data to facilitate epigenomic studies. It is a good time now to consider these issues so that we can design our analytical approaches appropriately.

*Departments of *Epidemiology and Social Medicine, †Medicine (Hematology) and Molecular Genetics, Albert Einstein College of Medicine, Bronx, New York 10461, USA. Correspondence to M.J.F. or J.M.G. e-mails: mfazzari@aecom.yu.edu; jgreally@aecom.yu.edu*
doi:10.1038/nrg1349

Mediators of epigenetic regulation

Since it was realized that CpG dinucleotides in mammals represent the target for the covalent modification of DNA[9], it has been apparent that DNA sequence characteristics can influence the targeting of epigenetic processes. This methyl group protrudes from the cytosine nucleotide into the major groove of the DNA and has two main effects: it displaces transcription factors that normally bind to the DNA[10,11]; and it attracts methyl-binding proteins, which in turn are associated with gene silencing and chromatin compaction[12] (probably through interactions with complexes that modify the tails of histone proteins). Histone proteins form octamers around which DNA loops to form the nucleosome, the individual packaging unit of genomic DNA. The histone tails that extrude from the nucleosomes can be modified by methylation[13], acetylation[14], phosphorylation[15] or ubiquitylation[16] at different sites, creating

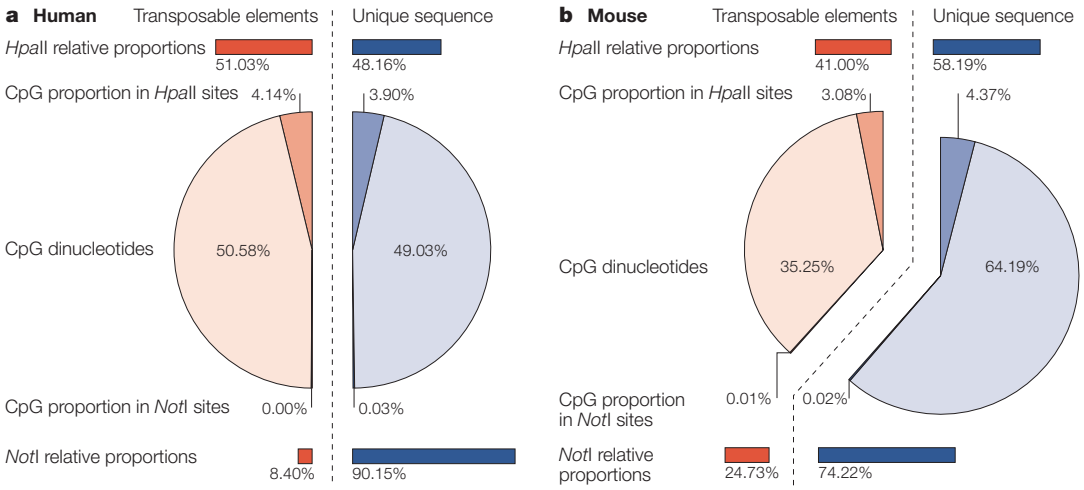


Figure 1 | **Genomic distribution of CpG dinucleotides.** Sequence data from the UCSC Genome Browser for the human (hg16, July 2003; panel **a**) and mouse (mm4, October 2003; panel **b**) genomes were analysed using a Perl program (see online supplementary information S1 (box)) to count CG, CCGG and GCGGCCGC motifs in unique and repetitive sequences. Those occurring in simple or low-complexity repeats were excluded from the repetitive DNA sample to allow a direct comparison with a previous study[27].The proportion of CpG dinucleotides in repetitive and unique sequences are illustrated by the pie charts, with the proportions of total CpG content contained in *Hpa*II (CCGG) and *Not*I (GCGGCCGC) also represented (see online supplementary information S1 (box)). The data show that only a small proportion of CpGs are located within *Hpa*II sites in unique sequence, demonstrating the limitation of our current 'whole-genome' methylation approaches that depend on methylation-sensitive restriction enzymes and hybridization. In addition, only 35.25% of *Hpa*II sites lie within transposable-element-derived DNA in the mouse genome, indicating that a substantial proportion of *Hpa*II sites in unique sequence is, in fact, methylated, contrary to the conclusions drawn from more limited bioinformatics studies[27].

potential combinations that have been referred to as a 'histone code'[17] in which gene regulatory information is encrypted. Cell-type-specific cytosine methylation and histone-tail modifications could contribute to the differences in gene expression patterns between cell types. This possibility has prompted the search for global epigenetic patterns that distinguish or are variable between cell types. These patterns have been named the 'epigenome'[4] or the 'methylome'[18]. The Human Epigenome Project (see online links box), which is discussed below, aims to define the sites at which the cytosine methylation component of epigenomic regulation differs between cell types. Between this project and the use of genomic microarrays to define the sites of cytosine methylation or binding of transcription factors by CHROMATIN IMMUNOPRECIPITATION (ChIP), whole-genome annotation of these epigenetic patterns is now underway.

CHROMATIN IMMUNOPRECIPITATION
Intact nuclei are gently fixed to maintain the physical relationship of DNA-binding molecules to genomic DNA. The chromatin (DNA plus bound molecules) is sheared to small fragments and exposed to an antibody that immunoprecipitates one of the bound molecules selectively. The sites of binding of the molecule (usually protein) of interest are apparent from their enrichment in the immunoprecipitated fraction of the genome.

Cytosine methylation and CpG islands

The identification of CpG islands, almost 20 years ago, remains a great model for today's epigenomics studies. These genomic features, which direct an epigenetic process, were identified as a result of whole-genome molecular studies. Initially, CpG islands were identified on the basis of the strikingly discordant patterns of digestion of genomic DNA by restriction enzyme isoschizomers that differed only by their sensitivity to cytosine methylation[19]. The methylation-insensitive *Msp*I (5′-CCGG-3′) enzyme digests the genome to completion, whereas most of the DNA exposed to its methylation-sensitive isoschizomer *Hpa*II remains high-molecular weight, as 55–70% of these sites are methylated in animal genomes[20–22]. The hypomethylated minority of the genome digested by *Hpa*II became known as *Hpa*II tiny fragments (HTFs)[7]. When these were cloned and sequenced, it became apparent that these were strikingly (G+C) and CpG rich[7]. Empirical genomic criteria to define these CpG islands were established[8] based on relatively few sequence descriptors, a small number of HTF sequences and the 1985 GenBank database, resulting in criteria that are still in use today. To be recognized as a CpG island, a sequence must satisfy the following critera: (G+C) content of 0.50 or greater; an observed to expected CpG dinucleotide ratio of 0.60 or greater; and both occurring within a sequence window of 200 bp or greater.

CpGs are vastly underrepresented genome-wide compared to what would be expected by chance (0.23 in the human genome and 0.19 in the mouse genome, respectively) (see online supplementary information S1 (box)). This is because deamination of cytosine gives rise to uracil, which is easily recognized as foreign within the DNA strand and replaced, whereas deamination of methylcytosine gives rise to thymine, which is less readily recognized as foreign and therefore prone to mutation and depletion in the genome[23].

CpG island definition based on sequence composition identifies these elements at the promoter sites of approximately half of the genes in the human genome[24], most of which are expressed in most or all tissues, hence their designation as 'housekeeping' genes[25]. But, based on this definition, the CpG-rich promoters of some endogeneous retro-elements are defined as CpG islands. Takai and Jones[26] revisited the above criteria in order to refine them, focusing on the sequences of chromosomes

21 and 22. They found that increasing the size threshold to 500 bp and the (G+C) content threshold to 0.55 biased the definition away from repeated sequences, towards unique sequence.

CpG islands have been proposed to be invariantly unmethylated[27]. Although the basis for this influential proposal was largely indirect, few data have been available to challenge it. Based on the observation that the proportion of *Hpa*II restriction sites located within transposable elements in the human genome (in fact, in a sample of 606 kb, ~1/500th of the genome) was approximately equivalent to the proportion of *Hpa*II sites that are methylated in the mouse genome (both ~60%), the authors suggested that most CpGs at *Hpa*II sites that are not located within transposable elements remain free from methylation[27]. Because CpG islands seemed to be the main remaining location for *Hpa*II sites, it was proposed, by exclusion, that CpG islands were always free from methylation. Exceptions to this rule include those CpG islands at loci that undergo GENOMIC IMPRINTING[28] and those that are subject to X-chromosome inactivation[29] (see below). The observation indicates that there is an invariant pattern of cytosine methylation — always targeted to transposable elements and never to CpG islands (apart from imprinted or X-inactivated loci). The existence of this pattern implies that cytosine methylation can have no role in establishing differences in epigenetic regulation between cell types.

When the *Hpa*II analysis is extended to the entire human and mouse genomes (using data from the UCSC Genome Browser[30] (see online links box)), 50.58% of *Hpa*II sequences are located within transposable elements in the human genome and 35.25% in the mouse genome (FIG. 1). Both values are below the 55–70% *Hpa*II sites that are methylated in animal genomes[20–22], which indicates that a variable and substantial proportion of *Hpa*II sites within unique sequences are methylated in any given cell type. Only a few *Hpa*II sites in the mouse (14%) and the human (22%) genomes are located within CpG islands; therefore, these data do not address whether cytosine methylation at CpG islands is responsible for this variation. However, most *Not*I sites (90%) in both species are located within CpG islands, and the restriction landmark genomic scanning (RLGS) technique (see below), which studies methylation at these restriction sites, has repeatedly shown tissue-dependent methylation[31,32]. It is crucial to address whether cytosine methylation varies among tissues, as the entire rationale for performing whole-genome analyses to study the redistribution of cytosine methylation assumes that such a redistribution of methylation occurs. Our updated bioinformatic analyses (FIG. 1) now support the likelihood that cytosine methylation is physiologically variable.

The dynamic nature of cytosine methylation becomes especially evident during tumorigenesis — methylation is decreased genome-wide, whereas the CpG islands at promoters of tumour-suppressor genes acquire methylation[33], which leads to their silencing and subsequent tumour progression. Hypomethylation is also linked to chromosomal instability, a common phenomenon in human tumours[34], which has been observed in mice with hypomethylated genomes due to engineered methyltransferase deficiencies[35]. A practical current focus of epigenomics research is the genome-wide characterization of cytosine methylation changes in cancer[5], which promises to reveal additional loci that contribute to the neoplastic process through epigenetic dysregulation rather than mutation.

Genome-wide epigenetic assays

One of the first techniques that successfully analysed epigenetic patterns genome-wide was RLGS[36]. It identifies methylation differences at *Not*I sites (5′-GCGGCGCC-3′) between DNA samples, using radiolabelling and 2D gel electrophoresis. What this technique lacks in ease of use is made up for with a proven track record of sensitivity — for example, differentially methylated sites between ANDROGENETIC and PARTHENOGENETIC embryos were identified using this technique to allow the identification of several new imprinted genes[36,37].

More recently, techniques that use methylation-sensitive restriction enzymes and genomic DNA microarrays have been developed to isolate methylated sequences throughout the genome. The techniques have been named differential methylation hybridization (DMH)[38], amplification of inter-methylated sites (AIMS)[39] and methylation target array (MTA)[40]. In each case, the methylated fraction of the genome is enriched, in a manner that depends on restriction digestion of unmethylated sequences using a methylation-sensitive enzyme, followed by the failure to PCR amplify the digested fragments. For DMH, two restriction enzymes at a time are used: *Mse*I (5′-TTAA-3′) to reduce the average size of the DNA while preserving CpG-rich sequences, followed by a 5′ methylcytosine-sensitive restriction enzyme (for example, *BstU*I (5′-CGCG-3′) or *Hpa*II). AIMS uses the methylation-sensitive *Sma*I restriction enzyme (5′-CCCGGG-3′), whereas MTA uses a similar approach to DMH, cutting initially with an enzyme that spares CpG-rich sequences followed by the use of a methylation-sensitive enzyme such as *BstU*I or *Hpa*II. The analysis of the methylation patterns for DMH and MTA requires a subsequent hybridization to genomic microarrays, whereas the limited number of *Sma*I sites in the genome means that a fingerprinting approach using electrophoresis is sufficient to identify differentially methylated sites.

The use of restriction enzymes limits the proportion of CpGs in the genome that can be tested using these techniques. The proportion of CpGs that are located within *Hpa*II sites in the human genome is 4.14% in transposable elements + 3.90% in unique sequence (8.04% in total), and 7.45% in the mouse genome (FIG. 1). Only those that reside in unique sequence can be tested using hybridization-based techniques, reducing the totals to 3.90% and 4.37% for human and mouse, respectively. The use of *Not*I sites in RLGS is even more limited in terms of representation, but most *Not*I sites in human and mouse (75% and 63%, respectively) are located in unique sequence, of which three-quarters are

GENOMIC IMPRINTING
The epigenetic marking of a locus on the basis of parental origin, which results in monoallelic gene expression.

ANDROGENETIC
A diploid offspring that is produced from two sets of haploid paternal gametes and no maternal contribution.

PARTHENOGENETIC
A diploid offspring that is produced from two sets of haploid maternal gametes and no paternal contribution.

located in canonical CpG islands for both species (see online supplementary information S1 (box)). The RLGS technique samples fewer sites than the other techniques described here, but it is more CpG-island-specific (see online supplementary information S1 (box)).

All of these techniques are especially useful for identifying the dynamic nature of CpG methylation in normal cell differentiation and disease. DMH has been most widely applied so far, with several reports describing the methylation profiles in different tumour cells or in response to pharmacological treatment[38,41–43]. The limitations imposed by the use of restriction enzymes have been bypassed in the Human Epigenome Project, a multigroup collaborative project that is using a combination of BISULPHITE SEQUENCING and MALDI MASS SPECTROMETRY to perform large-scale analysis of cytosine methylation in human cells. A combination of cell types is used as the source of DNA, so that CpGs at which methylation varies (methylation-variable positions, MVPs) can be identified from the mixed methylated/unmethylated pattern observed[4]. The Human Epigenome Project promises to provide insights into CpG methylation whether in restriction sites or not, with the goal of mapping MVPs in 30,000 human genes using 200 cell types[44].

The analysis of cytosine methylation is complemented by experimental approaches that use ChIP to determine the composition of chromatin. The modification of ChIP to allow the analysis of chromatin composition genome-wide using microarrays (ChIP on chip) represents the other main, genome-wide approach to epigenetic organization. This technique has been primarily developed in *Saccharomyces cerevisiae*, but studies have also been successfully performed in mammalian cells. The most striking studies that demonstrate the power of this technique have been based on whole-chromosome representations by oligonucleotide microarrays to map transcription-factor binding sites[45,46]. Other genomic resources used for hybridization have included CpG island[47,48] and promoter[49] microarrays. Challenges in these ChIP-on-chip experiments involve the creation of suitable genomic microarrays and the amplification of sparse starting material for hybridization (reviewed by Buck and Lieb[50]). Again, only unique sequences can be tested using these hybridization-based approaches, but the main limitation at present is in common with that of the whole-genome cytosine methylation approaches — the need for widespread availability of suitable genomic microarrays. The whole-chromosome oligonucleotide microarrays indicate a very promising avenue for epigenomics exploration.

Monoallelic expression and flanking sequences

Genome-wide cytosine methylation and chromatin compositional studies are imminent sources of large amounts of data that will provide insights into the nature of the DNA sequences that are targeted by epigenetic processes. In the absence of such data so far, a handful of studies have looked at loci that are known to undergo specific types of epigenetic modification. These studies, which mine and analyse annotated DNA sequence features, provide the technical foundation for the analyses of data that will be generated by whole-genome epigenetic studies.

The first example of such a study was prompted by a molecular cytogenetic finding that L1 LINES are strongly overrepresented on most of the X chromosome in humans[51] and mice[52]. Random X-chromosome inactivation in females depends on a single region within the X chromosome[53], which spreads an inactivation signal along its ~160 Mb length. How the signal is propagated over such immense distances has been puzzling, leading to the proposal that booster elements or waystations might exist on the X chromosome specifically for this purpose. Given the increased L1 LINE density on the X chromosome, these transposable element sequences were proposed as candidates for mediating the spreading of the X-chromosome inactivation[54].

Analysis of the draft human genome sequence confirmed the L1 LINE enrichment on the X chromosome; moreover, it turns out that L1 LINE content is lower in the regions that contain increased numbers of genes that escape X-chromosome inactivation[55]. A subsequent study took a gene-centred approach, comparing the sequence features that flank the genes that escape or are subject to X-chromosome inactivation[56]. The results showed that MIR (mammalian-wide interspersed repeat) SINES are less common near genes that escape inactivation, whereas CpG islands are more common at those that are subject to it. The study did not, however, find L1 LINEs to be associated with inactivating genes, a finding which was confirmed in our laboratory's unpublished analysis of these same gene samples. Considering the scale of the initial study, which looked at very broad genomic landscapes for L1 LINE density, it is possible that the 100-kb window of flanking sequence that was used in the subsequent studies was insufficient to detect differences in L1 LINE accumulation, indicating that a functional role for these transposable elements would be exerted over even greater distances. It is also possible that L1 LINEs are not involved in propagating epigenetic signals or that they accumulate on the X chromosome for reasons that are causally unrelated to X-chromosome inactivation (see below).

X-chromosome inactivation is an example of monoallelic gene expression, with effects extending throughout the chromosome. A second example is genomic imprinting. It occurs when the epigenetic state of a genomic region differs between homologous chromosomes, resulting in gene activation that depends on its gamete of origin. As with X-chromosome inactivation, the silenced allele at an imprinted locus is characterized by cytosine methylation[28], histone-tail modifications[57,58] and chromatin conformation differences[59]. Intrigued by previous observations that suggested that imprinted genes might have unusual sequence characteristics[60], we studied the influence that flanking sequences at the site of integration can have on transgene expression[61,62] — so-called position effects[63]. To this end, we described the regions that flank imprinted loci in terms of available sequence annotations, expanding on previous studies in terms of the size of the imprinted gene sample and the

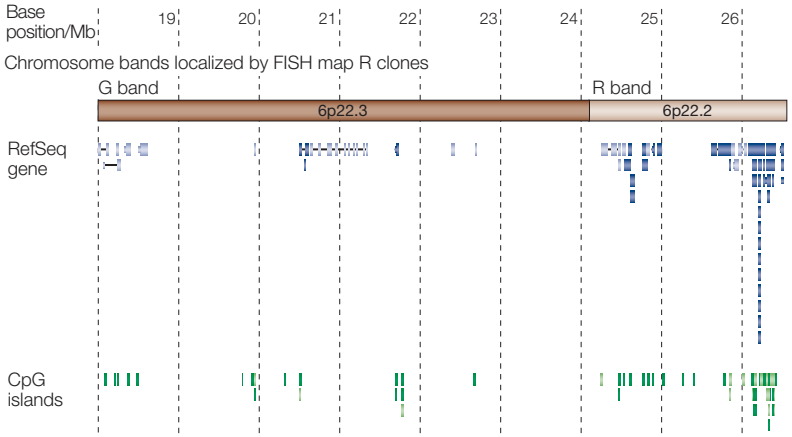
BISULPHITE SEQUENCING
A technique that is used to identify methylcytosines that depends on the relative resistance of the conversion of methylcytosine to uracil compared with cytosine. PCR amplification and sequencing of the DNA following conversion shows a thymine where a cytosine was located, whereas persistence of a cytosine reflects its methylation in the starting DNA sample.

MALDI MASS SPECTROMETRY
Matrix-assisted laser desorption/ionization mass spectroscopy is based on the co-crystallization of a test compound with an ultraviolet-light-absorbing matrix, which allows ionization using laser excitation to determine the mass of the test compound.

L1 LINES
The currently active long interspersed nuclear element in the eutherian genome. These elements are capable of retrotransposition but lack the long terminal repeats that characterize retroviruses.

MIR AND ALU SINES
Short interspersed nuclear elements, of which the Alu type is currently active in primates, whereas the MIR (mammalian interspersed repeat) type became extinct since eutherians diverged from marsupials.

a



b

Properties	(A+T) rich	(G+C) rich	Refs
Ultrastructural features			
Cytogenetic band type	G (Giemsa) band	R (reverse) band	80
DNA sequence features			
Genes	Fewer	More	71
Intron size	Longer	Smaller	81
CpG islands	Fewer	More	72
L1 LINEs	More	Fewer	73
Alu SINEs	Fewer	More	81
Functional correlations			
DNA replication	Later	Earlier	75
Meiotic recombination	Lower rate	Higher rate	77
Transcriptional activity	Lower rate	Higher rate	74
Methylcytosine content	Lower	Higher	76

Figure 2 | **Correlations between sequence features and functional characteristics in the mammalian genome.** At cytogenetic resolution, the genome is heterogeneously organized, recognizable by banding patterns that are detectable by light microscopy, and correlated with both DNA sequence and functional features. The 6p22.2-6p22.3 G to R band transition is shown in panel **a**. Note the increased gene and CpG island densities that are apparent in the R band (right). Data from the UCSC Genome Browser (hg16, July 2003 freeze). Additional properties of the G and R band sequences are shown in the table in panel **b**. As discussed in the text, the tendency of certain DNA sequence features to co-localize in the genome is apparent at this level of resolution, and correlates with the functional outcomes described. These correlations, which exist for unknown reasons, lead to problems with the analysis and interpretation of epigenomics studies.

number of sequence features analysed. The striking outcome was that it seems that there is a constraint on SINE accumulation in the 100-kb flanking imprinted promoters[64] — not only ALU SINES, which populated the human genome mostly since our ancestral divergence from rodents[65], but also MIR SINEs, which became extinct in the eutherian genome, following divergence from marsupials[66].

The same study showed that maternally-expressed imprinted genes tend to lie in the more (G+C)-rich compartment of the genome, whereas paternally-expressed genes segregate to the complementary L1-LINE-rich compartment[64]. With small sample sizes for each group, this is not a robust observation, but raises the question as to whether the male and female germlines treat these genomic compartments differentially. Although we used UNIVARIATE ANALYSIS in our study, MULTIVARIATE TECHNIQUES subsequently confirmed the paucity of SINEs in imprinted regions in the human genome, and also for some (but not all) classes of SINEs in the mouse genome[67]. The tendency of paternally- and maternally-expressed imprinted genes to segregate to genomic regions with different sequence characteristics was confirmed by the same investigators[67], although the sample sizes of these subgroups of imprinted genes are extremely limited.

An analysis of autosomal loci that are monoallelically expressed, but randomly with respect to parental origin (therefore not imprinted), showed that L1 LINEs and not CpG islands nor SINEs accumulated in the 200 kb that flank these genes[68]. Based on these findings, the authors proposed criteria to predict other monoallelically-expressed genes in the mouse and human genomes.

All of the above studies point to transposable element frequencies as a variable that is predictive of epigenetic regulation. This indicates that these supposedly neutral genomic parasites might have a direct influence on the epigenetic outcome , or it could reflect a separate influence on transposable element accumulation that also influences the epigenetic outcome. The study of epigenetic regulation is therefore converging with the understanding of the influences on transposable element accumulation on a genome-wide scale.

Isochores and transposable elements

Two approaches originally led to the recognition that transposable elements are heterogeneously distributed in the mammalian genome: the study of sequence isochores and molecular cytogenetics. Isochore analysis is founded on studies of base composition in plant[69] and animal[70] genomes, and defines regions of hundreds of kilobases in terms of their (G+C) content. Regions of similar base composition on this scale are referred to as isochores. The initial observation that the genome is compositionally heterogeneous was followed by studies that found correlations between base composition and other sequence features, such as the presence of increased numbers of genes[71], CpG dinucleotides and islands[72], and transposable elements such as Alu SINEs in regions of high (G+C) content[72]. Notably, the same regions are characterized by decreased numbers of L1 LINE transposable elements[73]. Functionally, (G+C) content positively correlates with transcriptional levels[74], earlier replication timing[75], greater overall cytosine methylation levels[76] and increased meiotic recombination frequencies[77]. Molecular cytogenetic studies, in which fluorescently labelled sequence features were hybridized to metaphase chromosomes, confirmed these findings. These sequence features include transposable elements such as Alu SINEs[51,52,78] and L1 LINEs[51,52], CpG island libraries[79], and samples of DNA from different isochores[80]. It is apparent that, at the

UNIVARIATE ANALYSIS
Analysis of functions of one variable.

MULTIVARIABLE ANALYSIS
Analysis of functions of several variables.

cytogenetic level of resolution, different sequence features segregate to distinct genomic regions (FIG. 2). The use of isochore samples in these *in situ* hybridization experiments links base composition (and all of the correlated features described above) with genomic function and with genomic organization on the scale of hundreds of kilobases. Transposable elements of specific types accumulate preferentially within certain genomic 'compartments' of similar sequence features. Whether the constraints on the accumulation of transposable elements reflected by higher-order patterning genome-wide are mechanistically related to the unusual accumulation of transposable elements in epigenetically-distinctive regions remains to be seen.

So, what is known about transposable element accumulation? Transposable elements have managed to populate ~45% of the human genome[81], but surprisingly little is known about how their further accumulation is controlled. They can certainly be targeted by cytosine methylation[82,83] and silenced as a consequence, but during germ-cell development — a crucial time for their continued accumulation — this methylation protection fails for Alu SINEs[84], whereas the L1 LINE protein products are produced during spermatogenesis[85]. The potential for further accumulation exists in every generation, but for some reason this potential is obviously not being realized.

The L1 LINEs and Alu SINEs are the more abundant transposable elements in the human genome and are heterogeneously distributed in the human genome with respect to (G+C) content[81]. This distribution could be a result of non-random insertion or specific exclusion of these elements from some genomic regions. Genomic distribution of the L1 LINE endonuclease cleavage site[86] does not by itself explain non-random accumulation of these retrotransposed sequences. Furthermore, as L1 LINEs and Alu SINEs seem to use the identical enzymatic mediators for retrotransposition[87] but end up segregating to different regions within the genome, the process must be more complex than insertion bias alone. Further weakening the argument that these insertion events are non-random is the observation that young Alu SINEs are, in fact, randomly distributed with respect to (G+C) content[81]. Non-random accumulation therefore seems to be due to post-insertion influences (FIG. 3).

The post-insertional influences that are usually invoked are non-random amplification[88] or deletion[89] of these transposable elements. Whereas Alu SINE density is higher in genomic regions that have undergone segmental duplications[88], they only account for ~5% of the genome[88]. Non-random deletion is a more plausible mechanism, and is currently, by exclusion, the more likely mechanism. It is not clear why non-random deletion should occur. Given the data on SINE accumulation in regions that flank promoters, might it be possible that transposable element accumulation in imprinted regions (and possibly elsewhere in the genome) deregulates nearby genes? If so, insertion of transposable elements into some sequences would not be tolerated by selection, leading to selective deletions. Dysregulation of imprinted genes by SINEs is a tenable hypothesis, given that SINEs are a major target of cytosine methylation in the genome[82,83] and that SINEs have been found to induce methylation in flanking sequences[90,91]. SINEs therefore constitute candidate sequences for epigenomic studies, especially for epigenetic phenomena that involve cytosine methylation. We are compelled to consider the possibility that transposable elements are not neutral but actively influence gene expression *in cis*. Consequently, their accumulation would influence fitness and the likelihood of the maintenance of the genotype in the population.

In suggesting that epigenetic mechanisms might underlie the process of non-random transposable element accumulation, we are not only linking epigenomics studies with insights into gene regulation but also with aspects of genome architecture. However, we have to also consider the possibility that a DNA sequence feature does not cause or direct the epigenetic outcome, but instead is caused by the epigenetic process.

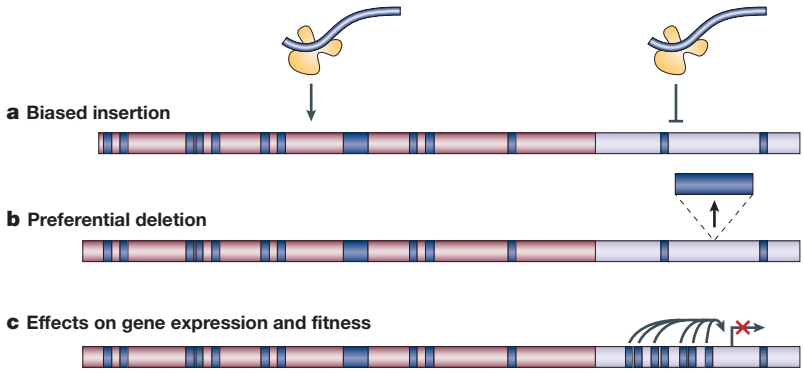


Figure 3 | **Models to explain the non-random distribution of transposable elements in the mammalian genome.** Transposable elements accumulate heterogeneously in the mammalian genome for reasons that are unknown. We suggest three possible mechanisms that might be occurring. **a** | Transposable elements insert into certain genomic environments (represented by lighter shading of the target DNA) less readily. As younger retro-elements in the human genome are randomly dispersed throughout the genome[81], this is an unlikely general mechanism, although the exclusion of young and old Alu SINEs from imprinted regions[64] indicates that it might be influential at certain loci. **b** | A biased tendency to deletion from certain regions is a possibility that is very difficult to test bioinformatically. **c** | If transposable elements are non-neutral and influence gene expression when they accumulate to excess in the vicinity of a gene, the effect might be to reduce organismal fitness and loss of that genotype from the population. The emerging correlations between transposable element accumulation and epigenetic regulation lend support to a *cis*-effect model.

Correlation and causation

Correlating epigenetic outcomes with genomic sequence features can yield important hypotheses and significant findings. But statistical analyses of epigenomic data are challenging (BOX 1), partly due to the complexity of the data sets themselves, involving, for example, internal correlations and sample size. Further problems arise when a statistically significant correlation is found, at which time the problem becomes how to determine the manner in which the descriptive (genomic) variable is biologically related to the functional (epigenetic) outcome. If the goal is to use the result predictively (as opposed to descriptively) — to identify further loci that might undergo similar epigenetic regulation based on discriminatory sequence features — then some of the

problems are circumvented. For example, new imprinted genes can be predicted in this way. There are only several dozen known imprinted genes in humans and mice[64], whereas hundreds are believed to exist[92]. A test to predict new imprinted genes would be very valuable, especially given the increasing number of human diseases with parent-of-origin effects on their inheritance[93–97] and the technical difficulty inherent in proving that a given gene is imprinted[98]. To be able to rank genes in a region of interest by likelihood of imprinting, based on similarities to known imprinted genes, should make the prediction much more efficient.

There is also, of course, the descriptive aspect to epigenomics — the identification of sequence characteristics that could be involved in the mechanism or reflect the evolution of the epigenetic process. The authors (ourselves included) of epigenomics studies of monoallelically-expressed genes described earlier made an implicit assumption that the correlated DNA sequence feature is directly involved in the functional outcome. For example, the L1 LINE enrichment on the mammalian X chromosome was postulated to directly mediate the spreading of X-chromosome inactivation[54]. But L1 LINE enrichment occurs in the genome in regions that are also distinctive for other reasons (FIG. 2). The L1 LINE enrichment might be statistically significant on the X chromosome but it might have nothing to do with X-chromosome inactivation, and accumulate for independent reasons or accumulate in the same regions as the actual mechanistic mediator of X inactivation. An enrichment of L1 LINEs could therefore be a valid statistical result, but on a biological level merely indicative of a separate sequence that is the actual mechanistic mediator of inactivation. In the interpretation of even the most stringent epigenomic studies, the issue of a separate, correlated but untested variable has to be considered.

Another issue is that of cause and effect. As an example, consider meiotic recombination and L1 LINE accumulation. Cytogenetic studies in human and mouse gametes have long recognized that in GIEMSA (G) BANDS, which are enriched in L1 LINEs, meiotic recombination is less frequent[99], a finding confirmed using bioinformatic approaches[51]. A study that mapped full-length L1 LINEs in the human genome found their relative preservation only in regions of low recombination on chromosome 21 and on the sex chromosomes, prompting the authors to suggest that they were being lost through meiotic recombination by purifying selection[89]. Meiotic recombination might avoid L1-LINE-enriched regions, or deplete L1 LINE content in the regions that it does target. Only in the former case would L1 LINEs be influential in directing meiotic recombination, but a bioinformatics study would show the negative correlation in either case.

An epigenomic study therefore has to be carefully analysed at every stage, paying attention to the possibility that the correlated DNA sequence feature might have no direct influence on the outcome, but might physically co-segregate in the genome with the mechanistically important sequence feature that was not analysed. In addition, the functional 'outcome' might not be caused by the correlated DNA sequence, but might instead be causing the DNA sequence to accumulate in these regions. Finally, epigenetic effects on genome composition could be subject to evolutionary selection. When a descriptive epigenomic study has been accomplished and significant

GIEMSA (G) BANDS
The chromosomal bands that are resistant to protease treatment (relative to reverse (R) bands), allowing them to stain more darkly with Giemsa stain. In chromosomal ideograms, the G bands are indicated by black/grey regions, the R bands by white regions.

HIERARCHICAL CLUSTERING
An unsupervised clustering technique. Each data point initially forms a separate cluster and then clusters are merged sequentially based on similarity, reducing the number of clusters at each step until only one cluster is left.

K-MEANS CLUSTERING
An unsupervised clustering technique. Data points are partitioned into a predetermined number of non-hierarchical clusters based on similarity.

LOGISTIC REGRESSION
A statistical model that is used when the outcome is binary in nature. Relates the log odds of *Pr*(event) to a linear combination of predictor variables.

TREE-BASED CART MODELS
A statistical tool that is used for identifying structure in data that uses binary recursive partitioning to obtain a tree classifier.

DISCRIMINANT FUNCTION ANALYSIS
A statistical method that is used to determine which variables and function best maximize the distance between two groups. Similar to logistic regression computationally, but generally less flexible in its assumptions.

Box 1 | Statistics and epigenomics: analytical techniques in sequence analysis

Epigenetic studies often involve bioinformatic and statistical analysis and can be generally subdivided into three broad categories: class discovery, comparison and prediction. Class discovery is often called an 'unsupervised' analysis as it is performed without reference to the outcome or group labels, such as gene type or imprinting status. Cluster analysis is commonly used in discovery. Allen[68], for example, used cluster analysis to examine heterogeneity in monoallelic genes with respect to sequence features. Feltus[102] used hierarchical clustering to assess similarity between cell lines that overexpress *DNMT1* on the basis of methylation profiles of the clones. There are many forms of clustering, including HIERARCHICAL and K-MEANS, but the primary goal of each is the separation of observations into distinct groups, the members of which are as similar to each other as possible while maximizing the differences between clusters. Central to this method is the notion of similarity. Similarity is often measured by Euclidean distance (connecting observations with a straight line in Euclidean space) or correlations (angles between the observation vectors). The choice of similarity measure (algorithm that is used) and the number of resulting clusters are subjective, and sensitivity of final clusters to each should be examined.

Comparison is a supervised form of analysis in which features of known gene types or classes are compared. Typically, each feature is examined in a univariate way using t-tests or a non-parametric equivalent. Greally[64] compared several sequence features of imprinted compared with non-imprinted genes. Bailey[55] compared the mean repeat contents of X-inactivated genes with those that escape inactivation. Permutation and simulation allow *p* values to be generated when standard large-sample approximation is invalid. Comparison studies are descriptive in nature — the emphasis is on understanding relevant features of the data and on obtaining an idea of the magnitude and direction of the effect.

Prediction is similar to comparison, but its main goal is to predict group membership in a new set of genes or observations. The complexity of epigenetic events requires the use of a multivariable model to generate accurate predictions. Model selection methods are often required to avoid overfitting the data. The most common models used for prediction are LOGISTIC REGRESSION MODELS, TREE-BASED (CART) MODELS and DISCRIMINANT FUNCTION ANALYSIS. Ke[103] uses a discriminant function to identify genes as candidates for imprinting. A set of important sequence features was identified through the use of information theoretic approaches and model-based prediction scores were implemented across the genome, identifying genes with the highest likelihood of imprinting.

BOOTSTRAP METHODS
Computer-intensive methods for statistical analysis. Treats the observed sample as the population and resamples from this population.

Box 2 | Statistics and epigenomics: a model-based approach

The challenge of statistical modelling in epigenomic studies is to use the large amount of genomic data and to identify the few sequence features that best predict the epigenetic outcome. Variables offer varying levels of discriminatory ability with respect to outcome; therefore, some effort has to be made both before and during the analysis stage to reduce the set of predictors to those that are informative and might be properly characterized in the model. To find this reduced set, many statistical approaches should be considered.

One of the most popular and best-understood strategies is variable subset selection (VSS). Variables are removed from the model if they have relatively little predictive ability. Univariate screening is a common form of subset selection in which the relevance of each variable is judged by its univariate performance in the model. Given that there are usually complicated relationships among sequence characteristics (such as the correlations shown in FIG. 2), the best univariate predictors might not be the best set of predictors collectively.

A useful tool for model selection is the 'all subsets' method that is available in most statistical packages. All potential models are fitted, resulting in a set of candidate models differing in complexity (number of variables in the model). Model information might be assessed by Akaike's Information Criteria (AIC)[104], which balances model fit with model complexity. Candidate models might then be further examined for biological validity and a representative model selected for predictive purposes.

In descriptive models, the goals of model selection are different. Removing non-informative or highly correlated variables yields a clearer picture of the variables that are important with respect to the outcome. Interpretation of direction and magnitude of effect for descriptive purposes is based on parameter estimates and requires each variable in the model to contribute independent information. BOOTSTRAP METHODS provide information with respect to the stability of each variable in the model. Variables that are consistently selected in model selection procedures over many bootstrap samples are likely to be stable descriptors of outcome.

correlations have been identified, these caveats have to be considered when interpreting the data.

Conclusions

We are entering a period during which epigenomic research will generate copious amounts of data, both from DNA sequence mining and from direct molecular assays. Although progress is being made using whole-genome cytosine methylation and ChIP assays, even more ambitious aims should be established. Currently, it is not at all easy to test the methylation state of every CpG dinucleotide in the genome, especially when they are located in repetitive DNA or in sites other than those recognized by methylation-sensitive restriction enzymes. We need to aim to develop techniques that can be used by individual investigators to determine the methylation of every cytosine at CpG dinucleotides (or CpNpG trinucleotides[100]) in the genome if we want to define the full extent of the methylome or epigenome.

The intriguing observations regarding transposable element accumulation in regions of distinctive epigenetic regulation strongly indicate that the possible *cis* effects of these elements need to be explored further. The development of mouse models of active, tagged retroelements[101] might be the most powerful approach to this issue. If threshold effects of groups of transposable elements rather than of individual elements are functionally more critical, the experiments to demonstrate *cis* effects of transposable elements will be very difficult to design and interpret.

The most important immediate issue in epigenomics will not involve data generation but data analysis (BOX 2). The number of ways in which a locus can be described in terms of its local and flanking DNA sequence characteristics is potentially immense. Add to this the increasing number of histone-tail modifications and transcription factors that can be immunoprecipitated from chromatin, the ability to quantify the amount of cytosine methylation at many loci and the ability to reproduce these molecular assays in a plethora of cell types, and the scope of the problem of data management and analysis becomes apparent. We have pointed out how correlations between descriptive variables complicate data analysis and how a variable subset selection approach might be the best way of addressing these analytical issues. The goal remains worthwhile; identification of DNA sequences that determine where epigenetic processes are targeted is central to our understanding of diverse phenomena that range from position effects to genome evolution and to neoplasia.

1. Zweiger, G. & Scott, R. W. From expressed sequence tags to 'epigenomics': an understanding of disease processes. *Curr. Opin. Biotechnol.* **8**, 684–687 (1997).
2. Beck, S., Olek, A. & Walter, J. From genomics to epigenomics: a loftier view of life. *Nature Biotechnol.* **17**, 1144 (1999).
3. Holliday, R. Epigenomics. *Nature Biotechnol.* **18**, 243 (2000).
4. Novik, K. L. *et al.* Epigenomics: genome-wide study of methylation phenomena. *Curr. Issues Mol. Biol.* **4**, 111–128 (2002).
Introduces the Human Epigenome Project within the context of a review of cytosine methylation and its role in human disease.
5. Plass, C. Cancer epigenomics. *Hum. Mol. Genet.* **11**, 2479–2488 (2002).
6. Reik, W., Santos, F. & Dean, W. Mammalian epigenomics: reprogramming the genome for development and therapy. *Theriogenology* **59**, 21–32 (2003).
7. Bird, A. P. CpG-rich islands and the function of DNA methylation. *Nature* **321**, 209–213 (1986).
8. Gardiner-Garden, M. & Frommer, M. CpG islands in vertebrate genomes. *J. Mol. Biol.* **196**, 261–282 (1987).
9. Sinsheimer, R. L. The action of pancreatic deoxyribonuclease. II. Isomeric dinucleotides. *J. Biol. Chem.* **215**, 579–583 (1955).
10. Hark, A. T. *et al.* CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus. *Nature* **405**, 486–489 (2000).
11. Kim, J., Kollhoff, A., Bergmann, A. & Stubbs, L. Methylation-sensitive binding of transcription factor YY1 to an insulator sequence within the paternally expressed imprinted gene, Peg3. *Hum. Mol. Genet.* **12**, 233–245 (2003).
12. Bird, A. P. & Wolffe, A. P. Methylation-induced repression: belts, braces, and chromatin. *Cell* **99**, 451–454 (1999).
13. Sims, R. J., Nishioka, K. & Reinberg, D. Histone lysine methylation: a signature for chromatin function. *Trends Genet.* **19**, 629–639 (2003).
14. Roth, S. Y., Denu, J. M. & Allis, C. D. Histone acetyltransferases. *Annu. Rev. Biochem.* **70**, 81–120 (2001).

15. Thomson, S., Clayton, A. L. & Mahadevan, L. C. Independent dynamic regulation of histone phosphorylation and acetylation during immediate-early gene induction. *Mol. Cell* **8**, 1231–1241 (2001).
16. Zhang, Y. Transcriptional regulation by histone ubiquitination and deubiquitination. *Genes Dev.* **17**, 2733–2740 (2003).
17. Jenuwein, T. & Allis, C. D. Translating the histone code. *Science* **293**, 1074–1080 (2001).
18. Feinberg, A. P. Methylation meets genomics. *Nature Genet.* **27**, 9–10 (2001).
19. Singer, J., Roberts-Ems, J. & Riggs, A. D. Methylation of mouse liver DNA studied by means of the restriction enzymes msp I and hpa II. *Science* **203**, 1019–1021 (1979).
20. Bird, A. P. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res.* **8**, 1499–1504 (1980).
21. Gruenbaum, Y., Stein, R., Cedar, H. & Razin, A. Methylation of CpG sequences in eukaryotic DNA. *FEBS Lett.* **124**, 67–71 (1981).
22. Bestor, T. H., Hellewell, S. B. & Ingram, V. M. Differentiation of two mouse cell lines is associated with hypomethylation of their genomes. *Mol. Cell. Biol.* **4**, 1800–1806 (1984).
23. Duncan, B. K. & Miller, J. H. Mutagenic deamination of cytosine residues in DNA. *Nature* **287**, 560–561 (1980).
24. Ioshikhes, I. P. & Zhang, M. Q. Large-scale human promoter mapping using CpG islands. *Nature Genet.* **26**, 61–63 (2000).
25. Ponger, L., Duret, L. & Mouchiroud, D. Determinants of CpG islands: expression in early embryo and isochore structure. *Genome Res.* **11**, 1854–1860 (2001).
Whereas CpG islands at promoters are usually associated with 'housekeeping' gene functions, these authors found an even more striking association with expression during early embryogenesis.
26. Takai, D. & Jones, P. A. Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proc. Natl Acad. Sci. USA* **99**, 3740–3745 (2002).
27. Yoder, J. A., Walsh, C. P. & Bestor, T. H. Cytosine methylation and the ecology of intragenomic parasites. *Trends Genet.* **13**, 335–340 (1997).
An influential review that linked epigenetic regulation with transposable element biology, setting the stage for subsequent exploratory bioinformatic studies.
28. Li, E., Beard, C. & Jaenisch, R. Role for DNA methylation in genomic imprinting. *Nature* **366**, 362–365 (1993).
29. Pfeifer, G. P., Tanguay, R. L., Steigerwald, S. D. & Riggs, A. D. *In vivo* footprint and methylation analysis by PCR-aided genomic sequencing: comparison of active and inactive X chromosomal DNA at the CpG island and promoter of human PGK-1. *Genes Dev.* **4**, 1277–1287 (1990).
30. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).
31. Shiota, K. *et al.* Epigenetic marks by DNA methylation specific to stem, germ and somatic cells in mice. *Genes Cells* **7**, 961–969 (2002).
32. Kremenskoy, M. *et al.* Genome-wide analysis of DNA methylation status of CpG islands in embryoid bodies, teratomas, and fetuses. *Biochem. Biophys. Res. Commun.* **311**, 884–890 (2003).
33. Baylin, S. B. *et al.* Aberrant patterns of DNA methylation, chromatin formation and gene expression in cancer. *Hum. Mol. Genet.* **10**, 687–692 (2001).
34. Lengauer, C., Kinzler, K. W. & Vogelstein, B. Genetic instabilities in human cancers. *Nature* **396**, 643–649 (1998).
35. Gaudet, F. *et al.* Induction of tumors in mice by genomic hypomethylation. *Science* **300**, 489–492 (2003).
This report of increased chromosomal instability in mice with a hypomorphic *Dnmt1* allele demonstrated increased tumorigenesis and chromosomal instability, and prompted an interesting correspondence in the journal.
36. Shibata, H. *et al.* Genetic mapping and systematic screening of mouse endogenously imprinted loci detected with restriction landmark genome scanning method (RLGS). *Mamm. Genome* **5**, 797–800 (1994).
37. Plass, C. *et al.* Identification of Grf1 on mouse chromosome 9 as an imprinted gene by RLGS-M. *Nature Genet.* **14**, 106–109 (1996).
38. Yan, P. S. *et al.* Applications of CpG island microarrays for high-throughput analysis of DNA methylation. *J. Nutr.* **132**, S2430–S2434 (2002).
39. Frigola, J., Ribas, M., Risques, R. A. & Peinado, M. A. Methylome profiling of cancer cells by amplification of inter-methylated sites (AIMS). *Nucleic Acids Res.* **30**, e28 (2002).
40. Chen, C. M. *et al.* Methylation target array for rapid analysis of CpG island hypermethylation in multiple tissue genomes. *Am. J. Pathol.* **163**, 37–45 (2003).
41. Huang, T. H., Perry, M. R. & Laux, D. E. Methylation profiling of CpG islands in human breast cancer cells. *Hum. Mol. Genet.* **8**, 459–470 (1999).
42. Yan, P. S. *et al.* CpG island arrays: an application toward deciphering epigenetic signatures of breast cancer. *Clin. Cancer Res.* **6**, 1432–1438 (2000).
43. Day, J. K. *et al.* Genistein alters methylation patterns in mice. *J. Nutr.* **132**, S2419–S2423 (2002).
44. Bradbury, J. Human epigenome project-up and running. *PLoS Biol.* **1**, E82 (2003).
45. Martone, R. *et al.* Distribution of NF-kappaB-binding sites across human chromosome 22. *Proc. Natl Acad. Sci. USA* **100**, 12247–12252 (2003).
46. Cawley, S. *et al.* Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell* **116**, 499–509 (2004).
47. Weinmann, A. S., Yan, P. S., Oberley, M. J., Huang, T. H. & Farnham, P. J. Isolating human transcription factor targets by coupling chromatin immunoprecipitation and CpG island microarray analysis. *Genes Dev.* **16**, 235–244 (2002).
48. Wells, J., Yan, P. S., Cechvala, M., Huang, T. & Farnham, P. J. Identification of novel pRb binding sites using CpG microarrays suggests that E2F recruits pRb to specific genomic sites during S phase. *Oncogene* **22**, 1445–1460 (2003).
49. Li, Z. *et al.* A global transcriptional regulatory role for c-Myc in Burkitt"s lymphoma cells. *Proc. Natl Acad. Sci. USA* **100**, 8164–8169 (2003).
50. Buck, M. J. & Lieb, J. D. ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. *Genomics* **83**, 349–360 (2004).
An excellent examination of the technical challenges and solutions for whole-genome chromatin immunoprecipitation experiments.
51. Korenberg, J. R. & Rykowski, M. C. Human genome organization: Alu, lines, and the molecular structure of metaphase chromosome bands. *Cell* **53**, 391–400 (1988).
52. Boyle, A., Ballard, G. & Ward, D. Differential distribution of long and short interspersed element sequences in the mouse genome: chomosome karyotyping by fluorescence *in situ* hybridisation. *Proc. Natl Acad. Sci. USA* **87**, 7757–7761 (1990).
53. Rastan, S. & Brown, S. D. The search for the mouse X-chromosome inactivation centre. *Genet. Res.* **56**, 99–106 (1990).
54. Lyon, M. F. X-chromosome inactivation: a repeat hypothesis. *Cytogenet. Cell Genet.* **80**, 133–137 (1998).
55. Bailey, J. A., Carrel, L., Chakravarti, A. & Eichler, E. E. Molecular evidence for a relationship between LINE-1 elements and X chromosome inactivation: the Lyon repeat hypothesis. *Proc. Natl Acad. Sci. USA* **97**, 6634–6639 (2000).
The first major application of genome sequencing data to address whether a DNA sequence feature is correlated with an epigenetic outcome.
56. Ke, X. & Collins, A. CpG islands in human X-inactivation. *Ann. Hum. Genet.* **67**, 242–249 (2003).
57. Saitoh, S. & Wada, T. Parent-of-origin specific histone acetylation and reactivation of a key imprinted gene locus in Prader-Willi syndrome. *Am. J. Hum. Genet.* **66**, 1958–1962 (2000).
58. Xin, Z., Allis, C. D. & Wagstaff, J. Parent-specific complementary patterns of histone H3 lysine 9 and H3 lysine 4 methylation at the Prader-Willi syndrome imprinting center. *Am. J. Hum. Genet.* **69**, 1389–1394 (2001).
59. Schweizer, J., Zynger, D. & Francke, U. *In vivo* nuclease hypersensitivity studies reveal multiple sites of parental origin-dependent differential chromatin conformation in the 150 kb SNRPN transcription unit. *Hum. Mol. Genet.* **8**, 555–566 (1999).
60. Hurst, L. D., McVean, G. & Moore, T. Imprinted genes have few and small introns. *Nature Genet.* **12**, 234–237 (1996).
61. Alami, R. *et al.* β-globin YAC transgenes exhibit uniform expression levels but position effect variegation in mice. *Hum. Mol. Genet.* **9**, 631–636 (2000).
62. Feng, Y. Q., Lorincz, M. C., Fiering, S., Greally, J. M. & Bouhassira, E. E. Position effects are influenced by the orientation of a transgene with respect to flanking chromatin. *Mol. Cell. Biol.* **21**, 298–309 (2001).
63. Martin, D. I. & Whitelaw, E. The vagaries of variegating transgenes. *Bioessays* **18**, 919–923 (1996).
64. Greally, J. M. Short interspersed transposable elements (SINEs) are excluded from imprinted regions in the human genome. *Proc. Natl Acad. Sci. USA* **99**, 327–332 (2002).
65. Jurka, J. & Smith, T. A fundamental division in the Alu family of repeated sequences. *Proc. Natl Acad. Sci. USA* **85**, 4775–4778 (1988).
66. Smit, A. F. & Riggs, A. D. MIRs are classic, tRNA-derived SINEs that amplified before the mammalian radiation. *Nucleic Acids Res.* **23**, 98–102 (1995).
67. Ke, X., Thomas, S. N., Robinson, D. O. & Collins, A. The distinguishing sequence characteristics of mouse imprinted genes. *Mamm. Genome* **13**, 639–645 (2002).
68. Allen, E. *et al.* High concentrations of long interspersed nuclear element sequence distinguish monoallelically expressed genes. *Proc. Natl Acad. Sci. USA* **100**, 9940–9945 (2003).
69. Salinas, J., Matassi, G., Montero, L. M. & Bernardi, G. Compositional compartmentalization and compositional patterns in the nuclear genomes of plants. *Nucleic Acids Res.* **16**, 4269–4285 (1988).
70. Bernardi, G., Mouchiroud, D. & Gautier, C. Compositional patterns in vertebrate genomes: conservation and change in evolution. *J. Mol. Evol.* **28**, 7–18 (1988).
71. Federico, C., Andreozzi, L., Saccone, S. & Bernardi, G. Gene density in the Giemsa bands of human chromosomes. *Chromosome Res.* **8**, 737–746 (2000).
72. Jabbari, K. & Bernardi, G. CpG doublets, CpG islands and Alu repeats in long human DNA sequences from different isochore families. *Gene* **224**, 123–127 (1998).
73. Pavlicek, A. *et al.* Similar integration but different stability of Alus and LINEs in the human genome. *Gene* **276**, 39–45 (2001).
74. Arhondakis, S., Auletta, F., Torelli, G. & D'Onofrio, G. Base composition and expression level of human genes. *Gene* **325**, 165–169 (2004).
75. Smith, Z. E. & Higgs, D. R. The pattern of replication at a human telomeric region (16p13.3): its relationship to chromosome structure and gene expression. *Hum. Mol. Genet.* **8**, 1373–1386 (1999).
76. Caccio, S. *et al.* Methylation patterns in the isochores of vertebrate genomes. *Gene* **205**, 119–124 (1997).
77. Kong, A. *et al.* A high-resolution recombination map of the human genome. *Nature Genet.* **31**, 241–247 (2002).
78. Manuelidis, L. & Ward, D. C. Chromosomal and nuclear distribution of the HindIII 1.9-kb human DNA repeat segment. *Chromosoma* **91**, 28–38 (1984).
79. Craig, J. M. & Bickmore, W. A. The distribution of CpG islands in mammalian chromosomes. *Nature Genet.* **7**, 376–382 (1994).
80. Saccone, S. *et al.* Correlations between isochores and chromosomal bands in the human genome. *Proc. Natl Acad. Sci. USA* **90**, 11929–11933 (1993).
81. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
This landmark description of the human genome sequence is also the most comprehensive description of genomic heterogeneity and DNA sequence feature correlations.
82. Hellmann-Blumberg, U., Hintz, M. F., Gatewood, J. M. & Schmid, C. W. Developmental differences in methylation of human Alu repeats. *Mol. Cell. Biol.* **13**, 4523–4530 (1993).
83. Kochanek, S., Renz, D. & Doerfler, W. Transcriptional silencing of human Alu sequences and inhibition of protein binding in the box B regulatory elements by 5′-CG-3′ methylation. *FEBS Lett.* **360**, 115–120 (1995).
84. Rubin, C. M., VandeVoort, C. A., Teplitz, R. L. & Schmid, C. W. Alu repeated DNAs are differentially methylated in primate germ cells. *Nucleic Acids Res.* **22**, 5121–5127 (1994).
85. Ergun, S. *et al.* Cell type-specific expression of LINE-1 ORF1 and ORF2 in fetal and adult human tissues. *J. Biol. Chem.* 31 Mar 2004 [epub ahead of print].
86. Feng, Q., Moran, J. V., Kazazian, H. H. Jr & Boeke, J. D. Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. *Cell* **87**, 905–916 (1996).
87. Okada, N., Hamada, M., Ogiwara, I. & Ohshima, K. SINEs and LINEs share common 3′ sequences: a review. *Gene* **205**, 229–243 (1997).
Based on sequence homology at the 3′ ends of co-evolving LINE and SINE elements, the authors suggest that this is the site of association of the retroelement's transcript with the LINE's reverse transcriptase/endonuclease product. Interestingly, human L1 LINEs and Alu SINEs are an exception to this 3′ homology observation.
88. Jurka, J., Kohany, O., Pavlicek, A., Kapitonov, V. V. & Jurka, M. V. Duplication, coclustering, and selection of human Alu retrotransposons. *Proc. Natl Acad. Sci. USA* **101**, 1268–1272 (2004).
89. Boissinot, S., Entezam, A. & Furano, A. V. Selection against deleterious line-1-containing loci in the human lineage. *Mol. Biol. Evol.* **18**, 926–935 (2001).
The finding of a negative correlation between meiotic recombination frequency and full-length L1 LINE abundance in the human genome prompted the authors to consider a causal relationship in which the primary influence is meiotic recombination and the L1 LINE abundance the outcome, implying a role for meiotic recombination in genome evolution.
90. Hasse, A. & Schulz, W. A. Enhancement of reporter gene *de novo* methylation by DNA fragments from the alpha-fetoprotein control region. *J. Biol. Chem.* **269**, 1821–1826 (1994).
91. Yates, P. A., Burman, R. W., Mummaneni, P., Krussel, S. & Turker, M. S. Tandem B1 elements located in a mouse methylation center provide a target for *de novo* DNA methylation. *J. Biol. Chem.* **274**, 36357–36361 (1999).
92. Reik, W. & Walter, J. Genomic imprinting: parental influence on the genome. *Nature Rev. Genet.* **2**, 21–32 (2001).
93. Strauch, K., Bogdanow, M., Fimmers, R., Baur, M. P. & Wienker, T. F. Linkage analysis of asthma and atopy including models with genomic imprinting. *Genet. Epidemiol.* **21** (Suppl. 1), S204–S209 (2001).

94. Green, J. *et al.* Impact of gender and parent of origin on the phenotypic expression of hereditary nonpolyposis colorectal cancer in a large Newfoundland kindred with a common MSH2 mutation. *Dis. Colon Rectum* **45**, 1223–1232 (2002).
95. Karason, A. *et al.* A susceptibility gene for psoriatic arthritis maps to chromosome 16q: evidence for imprinting. *Am. J. Hum. Genet.* **72**, 125–131 (2003).
This study reveals linkage of psoriatic arthropathy to a locus on chromosome 16q21 only when the model for linkage analysis assumes paternal transmission of the disease allele. Linkage studies that do not test this model will fail to reveal imprinted gene effects.
96. McInnis, M. G. *et al.* Genome-wide scan of bipolar disorder in 65 pedigrees: supportive evidence for linkage at 8q24, 18q22, 4q32, 2p12, and 13q12. *Mol. Psychiatry* **8**, 288–298 (2003).
97. Pezzolesi, M. G. *et al.* Examination of candidate chromosomal regions for type 2 diabetes reveals a susceptibility locus on human chromosome 8p23.1. *Diabetes* **53**, 486–491 (2004).
98. Suda, T. *et al.* Use of real-time RT-PCR for the detection of allelic expression of an imprinted gene. *Int. J. Mol. Med.* **12**, 243–246 (2003).
99. Ashley, T. G-band position effects on meiotic synapsis and crossing over. *Genetics* **118**, 307–317 (1988).
100. Clark, S. J., Harrison, J. & Frommer, M. CpNpG methylation in mammalian cells. *Nature Genet.* **10**, 20–27 (1995).
101. Ostertag, E. M. *et al.* A mouse model of human L1 retrotransposition. *Nature Genet.* **32**, 655–660 (2002).
102. Feltus, F. A., Lee, E. K., Costello, J. F., Plass, C. & Vertino, P. M. Predicting aberrant CpG island methylation. *Proc. Natl Acad. Sci. USA* **100**, 12253–12258 (2003).
103. Ke, X., Thomas, N. S., Robinson, D. O. & Collins, A. A novel approach for identifying candidate imprinted genes through sequence analysis of imprinted and control genes. *Hum. Genet.* **111**, 511–520 (2002).
A genuinely multivariate approach to the analysis of DNA sequence features that discriminates imprinted from non-imprinted genes.
104. Akaike, H. in *Second International Symposium on Information Theory* 267–281 (1973).

Acknowledgements
Dedicated with affection to F. Ruddle, on the occasion of his retirement.

Competing interests statement
The authors declare that they have no competing financial interests.

Online links

FURTHER INFORMATION
Genetic Information Research Institute (Girinst): http://www.girinst.org/
Human Epigenome Project: http://www.epigenome.org/
UCSC Genome Browser: http://genome.ucsc.edu/

SUPPLEMENTARY INFORMATION
See online article: S1 (box)
Access to this links box is available online.