

An integrated epigenetic and genetic approach to common human disease

Hans T. Bjornsson^{1,2}, M. Daniele Fallin³ and Andrew P. Feinberg²

¹Graduate Program in Human Genetics, John Hopkins School of Medicine, 720 Rutland Ave, Baltimore, MD 21205, USA

²Departments of Medicine, Molecular Biology and Genetics, and Oncology, John Hopkins School of Medicine, 720 Rutland Ave, Baltimore, MD 21205, USA

³Department of Epidemiology, John Hopkins Bloomberg School of Public Health, 615 N. Wolfe St, Baltimore, MD 21205, USA

Epigenetic information is heritable during cell division but is not contained within the DNA sequence itself. Despite increasing evidence for and interest in the role of epigenetics in human disease, particularly in cancer, virtually no epigenetic information is routinely or systematically measured at the genome level. The current population-based approach to common disease relates common DNA sequence variants to either disease status or incremental quantitative traits contributing to disease. Although this purely genetic approach is powerful and general, there is currently no conceptual framework to integrate epigenetic information. In this article, we propose an approach to common human disease that incorporates epigenetic variation into genetic studies. Epigenetic variation might also help to explain the late onset and progressive nature of most common diseases, the quantitative nature of complex traits and the role of environment in disease development, which a purely sequence-based approach might not.

Epigenetic information is stably maintained during cell division but does not involve the sequence of A, G, T and C nucleotides *per se* and has not traditionally been represented within the Human Genome Project databases, such as GenBank. The best understood epigenetic information is DNA methylation, a covalent modification of cytosine, maintained at CpG dinucleotides by DNA methyltransferase I. DNA methylation is thought to cause gene silencing but this is an oversimplified view because methylation changes can arise secondary to chromatin modification, and methylation helps to maintain boundaries between active and inactive chromatin. Chromatin modifications, generally involving post-translational modifications of the N-terminal tails of core histones in the nucleosomes, are also thought to be maintained during DNA replication. GENOMIC IMPRINTING (see Glossary) is a form of epigenetic modification that involves parent-of-origin-specific allele silencing.

All of these epigenetic forms of information are important in cancer, where DNA methylation is linked to gene activation, gene silencing and chromosomal instability. In addition, epigenetic changes are more clearly

associated with tumor progression than are specific mutations [1,2]. Recently, a common variant in imprinting was associated with colorectal cancer risk in the general population, suggesting that epigenetic variation in the population can be associated with common cancer risk [3]. (For reviews on epigenetics and cancer see Refs [1,4–6].)

Although the data linking epigenetics to cancer are strong, this is not the case for other common diseases that have a familial component [based on familial risk estimation (λ_r)] yet do not demonstrate a clear mendelian pattern of inheritance. Instead, most of the focus has been on sequence variation but direct links between variants and disease contribution have been difficult to prove.

We propose that in the etiology of common disease, an epigenetic framework can help to provide an explanation of three characteristics: (i) their age-dependence, which is not well explained by accumulated mutation; (ii) their quantitative nature; and (iii) the mechanism by which the environment might modulate genetic predisposition to disease (Table 1). There are several articles discussing the role of epigenetics in human disease [7–12], however, in this article, we focus on the potential relationship of genetic and epigenetic factors and how they might be explored experimentally.

Glossary

Genomic imprinting: a special case of stable transcriptional repression in which the relatively silenced allele is determined by the parental origin of the allele.

Epigenotype: information in a cell that is maintained through meiosis and/or mitosis but does not involve the DNA sequence itself. Epigenetics is the study of such information. Examples include: DNA methylation, a covalent modification of cytosine at CpG dinucleotides, posttranslational modifications of histone including methylation, acetylation and phosphorylation, particularly of the amino terminal tails of histones H3 and H4, and genomic imprinting (defined above). A crucial property of epigenetic marks is that they are metastable and can be reprogrammed. In our discussion we have focused mainly on DNA methylation as that is the best-understood example to date. However, we believe that the concepts are valid for all types of epigenotype. Methylation also appears to represent a more stable form of epigenetic modification compared with some of the other more dynamic modifications, which appear to have shorter half-lives.

Age-dependent degeneration of epigenetic patterns: the loss of normal epigenetic patterning. This can involve both loss and gain of epigenetic marks leading to a loss of original patterning. We use the term degeneration because its etymological root literally means to lose the original type or pattern.

Table 1. Features of complex diseases that are compatible with the CDGE^a

Complex disease phenomenon	Purely genetic model	Genetic and epigenetic model
Family association	Genetic variants inherited	Epigenetic and genetic variants inherited
Frequency	Common variants in combinations	Common genetic and epigenetic variants in combinations
Disease onset lag time	Gradual break down of homeostasis	Age-dependant epigenetic degeneration
Environmental influence	Gene–environmental interaction	Epigenetic systems are environmentally sensitive
Tissue specificity of disease	Tissue-specific transcription factors	Epigenetic patterns are tissue specific
Progressivity	Continuing breakdown of homeostasis	Epigenetic patterns erode over time
Quantitative nature of complex disease	Multiple loci	Quantitative epigenetic variation of one or more genetic loci

Abbreviation: CDGE, common disease genetic epidemiology in the context of both genetic and epigenetic variation.

An integrated approach to disease genetics and epigenetics

If the ‘genetic’ (i.e. heritable) components of common diseases can be explained via the interplay of genetic and epigenetic variation, then we should seek to better understand this relationship and how knowledge of both types of variation can help predict disease and improve understanding of disease mechanisms. To incorporate both concepts, we suggest an approach to common disease genetic epidemiology in the context of both genetic and epigenetic variation, which we have termed CDGE. The genetic determinants are no different from those that have been comprehensively formulated (i.e. there are common population variants that can then be associated with contribution to disease by large-scale population studies or within families through transmission tests for linkage disequilibrium). The rationale for inclusion of epigenetic information as potential disease determinants is its heritability during cell division in a given cell lineage, and its biological interaction with the expression of sequence variation.

The traditional model

Disease (either the probability of having a disease or values of some quantitative disease phenotype) is a function of the interaction of genes and environment (Figure 1). The environmental effects are assumed to be relatively large for many common diseases, given that concordance rates for common diseases among monozygotic (MZ) twins do not usually approach 100% and that

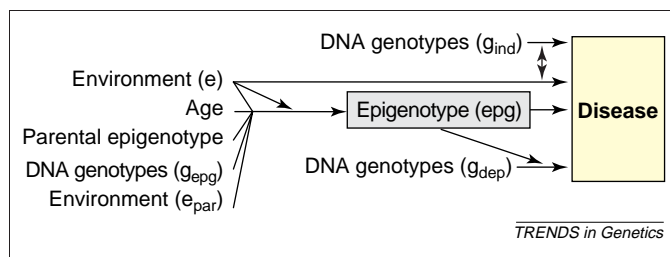


Figure 1. An integrated genetic and epigenetic approach to common disease. A schematic summary of how genetic and epigenetic (epg) factors might contribute to human disease and the factors that contribute to epigenetic variation. The sources of epg variation (genetic, environmental and stochastic) are also represented. This complicates our notation because genetic variation can refer to genes that determine epg or genes that directly affect disease phenotypes. For clarity, we have added the subscript ind to genes that affect disease *independently* of epigenetics and the subscript epg to genes that directly code for epg variation (the differences between the two are illustrated in Figure 2). g_{ind} might be epigenetically modified but the epigenetic modification does not influence disease. The relative importance of g_{ind} is inversely proportional to the degree to which common disease is epigenetically determined, which is unknown at present. The modification of gene penetrance by epigenetic context is shown by the use of arrows, which point to relationships, rather than measured values.

disease rates typically vary widely with geography and culture. For example, MZ and Dizygotic (DZ) concordance rates are 63% and 43% for Type II diabetes [13] and 67% and 20% for bipolar disorder [14], respectively, whereas for a fully penetrant dominant disorder MZ and DZ would be expected to be 100% and 50% concordant, respectively. No explicit consideration of variation at the epigenetic level has been included in these models. The closest attempts have been the inclusion of parent-of-origin effects to allow for differences in penetrance (disease risk given the genotype) according to the parental origin of the risk allele. Differences in parent-of-origin effects would suggest a role for imprinting but measured epigenetic information has not been considered directly in models of complex disease.

Epigenetic influence on disease phenotypes

Epigenetic marks could influence disease phenotypes by affecting the target gene directly, regardless of sequence variation within the gene. Alternatively, the influence of epigenetic marks on disease phenotype could be through the interaction with specific DNA-sequence variants.

Direct epigenetic effects: is cancer the tip of the iceberg?

In cancer, there is compelling evidence that epigenetic marks, such as chromatin modification, can influence phenotypes through the regulation of particular genes, without an underlying sequence variant in the gene. Alterations in methylation, imprinting and chromatin are ubiquitous in cancer. A causal role is suggested by recent data showing a gatekeeper role for altered imprinting and methylation in Wilms' tumor [15], and a common epigenetic variant appears to be associated with risk of colorectal cancer [3]. The first step in understanding the role of epigenetics in other human diseases is to investigate epigenetics in a broader disease context. We find it helpful to express our conceptual model through notation and diagrams that build on the simple concept of genes and environment in disease prediction (qualitative or quantitative). Our first notation is to denote measured epigenotypes as ‘epg’ in Figure 1 and in Eqn 1.

Epigenetic modification of disease penetrance

Although epg could influence disease directly, a more subtle, and potentially more important, source of epigenetic contribution to other common diseases is the possibility that epigenetic modification might affect the penetrance of traditional disease-causing genetic variants. In this case, a gene can have one or several disease-causing variants but the expression of these variants is epigenetically controlled

and therefore will effect the phenotype only when ‘masked’ or ‘unmasked’ by epigenetic marks; for example, the epigenetic buffering of genetic variants, which might underlie the dramatic uncovering of pre-existing genetic variants after Hsp90 mutation in *Drosophila melanogaster* [16]. A heritable altered chromatin state was observed in response to reduced activity of Hsp90, suggesting that Hsp90 acted as a capacitor for morphological evolution through epigenetic and genetic mechanisms [17]. An extreme example of epigenetic modification leading to differences in penetrance in mammals already exists. Imprinting is the best-known example, in which disease-predisposing alleles only show their effects in a particular epigenetic context, the parental origin of the variant. In imprinting, the relationship between the genetic variation and disease phenotype is dependent on EPIGENOTYPE. (Thus, we have used the subscript dep to specify this particular genetic contribution to disease in contrast to genes of direct effect on phenotype or genes that control the epg variation). Figure 2 shows the differences between g_{ind} and g_{dep} (i.e. genes that influence disease independently or are dependent on epigenetic factors, respectively). Imprinting, a special case of g_{dep} -epg interaction, has been the focus of previous studies, because it is easily identifiable by its unusual inheritance pattern. A significant

number of complex disorders have shown parent-of-origin effects in epidemiological studies including diabetes, psoriasis and Hirschsprung disease. Some complex behavioral phenotypes, including autism, schizophrenia and bipolar disease, have shown parent-of-origin effects, suggesting that imprinted genes might contribute to their etiology. A more general example of a g_{dep} class gene would be a gene affected by the CREB-binding protein, a histone acetyltransferase, which mediates the phenotype of neurodegenerative disorders caused by polyglutamine expansion [18].

Epigenetic variation might also help to explain the quantitative nature of common disease phenotypes. Although it is likely that common diseases are multigenic, even a single locus could show a gaussian distribution of phenotype based on the quantitative nature of epigenetic variation (Figure 2).

Sources of epigenetic variation

To include epigenotypes in a disease model, it is important to understand the sources of variation of epigenetic marks because these factors will have indirect effects on disease. This can also help explain the apparent heterogeneity in genetic effects on disease phenotypes as a result of age and/or environment, as described in the following sections.

Genetic factors (g_{epg}) contributing to epigenetic variation (epg)

Types of g_{epg} variation include DNA sequence variants, expression differences in chromatin remodeling genes or genes that affect or detect DNA methylation. The extreme paradigm of a loss of g_{epg} is immunodeficiency-chromosomal instability-facial anomalies syndrome (ICF syndrome), caused by mutation of the DNA methyltransferase gene *DNMT3B* [19], in which widespread methylation abnormalities cause pleiotropic changes in gene expression and chromosomal stability. Less extreme allelic variation at *DNMT3B* and/or the other methyltransferases could cause variation in g_{dep} methylation. In addition, sequence variation in the gene encoding g_{dep} could make it more or less attractive to g_{epg} -encoded proteins leading to changes in epigenotype (e.g. loss of a CpG dinucleotide in g_{dep}).

Vertically transmitted (parental) epigenetic marks contributing to epg

Considering that epigenetic variation within populations has a familial component, perhaps epigenetic marks are themselves vertically transmitted. The first example observed in humans was the mendelian transmission of methylation of a variable number of tandem repeat (VNTR) sequence [20]. A more recent example is the familial clustering of imprint-specific methylation of *H19* [21]. However, the transmission of parental epigenetic marks must be viewed somewhat theoretically, because it is unknown whether the epigenetic mark itself was vertically transmitted or if this reflects the transmission of an underlying genetic mark (g_{epg}). But data from the mouse agouti locus shows that mothers with the agouti allele are more likely to have agouti offspring rather than pseudo-agouti offspring. This is thought to be because of

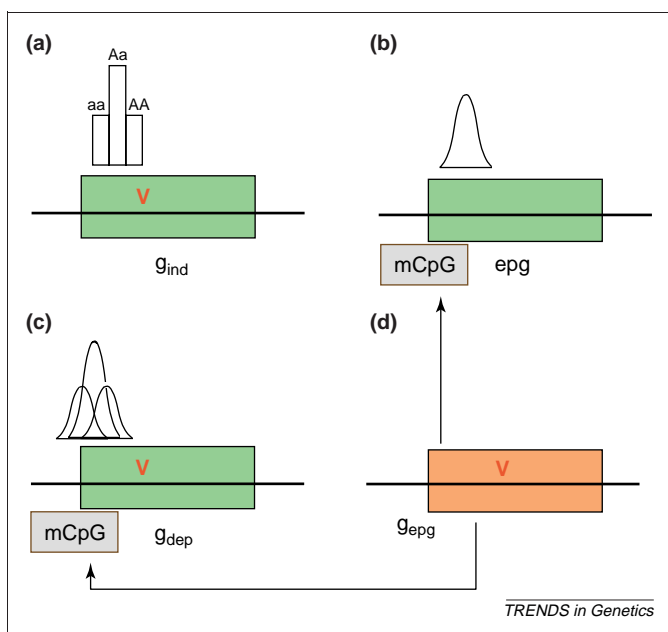


Figure 2. Interaction of genetic and epigenetic variation at a genomic level. (a) A gene (green box) can contain a sequence variant (V) that contributes to disease. Gene variants that are not influenced by epigenetic modification in their disease contribution (although epigenetics can contribute to normal function) are epigenetic-independent (g_{ind}). The distribution of phenotype of a single locus will not be gaussian in the absence of other factors. (b) Epigenetic variation can contribute to disease phenotype directly, independent of a genetic variation in the target gene (epg). Epigenetic variation itself is quantitative and thus can impart the quantitative nature to a trait, even at a single locus. (c) If the penetrance of a gene sequence variant is affected by epigenetic modification (mCpG), the gene is ‘epigenetic-dependent’ (g_{dep}). In this case, the genetic and epigenetic variation together could contribute to a gaussian distribution, even at a single locus. Note that the epigenetic modification is drawn on the gene but it could be at some distance upstream or downstream from that gene. The epigenetic modification need not be methylation, which is drawn here for convenience. (d) A genetic variant that can influence this epigenetic modification (e.g. encoding a chromatin-modifying protein) is referred to as g_{epg} , and its influence is denoted by the arrows.

the incomplete erasure of epigenetic marks and the importance of such 'transgenerational' inheritance has been reviewed recently. [12,22]

Parental environment contributing to epg

Parental environment that contributes to the epigenotype of the gamete and embryo is probably affected both by environmental exposure and by the parental genotype itself, independently of vertically transmitted genetic marks that affect epigenotype. Environmental effects that act before and after conception are important in determining epigenotypes (e.g. a mother's dietary methionine affecting the methylcytosine content of her embryo). A well-understood example is the variegated phenotype of the agouti locus in mouse that is affected by diet [23,24], mediated by the methylation of a transposon-derived repeat near the gene promoter. This balance can be shifted by providing mothers with methyl-enriching diets [23,24]. Interestingly, a quarter of human regulatory sequences originate from transposable elements [25]. The parental genotypes affecting the gametic or embryonic environment might also be relevant (e.g. those affecting methionine and homocysteine metabolism).

The contribution of aging to epg

The evidence of epigenetic change with age is mostly in the form of isolated reports and has been performed unsystematically, so more comprehensive studies are required. Nevertheless, the existing data suggest age-dependent degeneration, including erosion of global DNA methylation and hypermethylation at some sites, and age-related epigenetic silencing of gene promoters. [5,7,26,27] This erosion of normal methylation marks has been found in most tissues. [7] A recent study estimates that the frequency of epigenetic changes in mouse might be one-to-two orders of magnitude greater than the rate of somatic DNA mutation [28]. Many additional examples come from studies on the reactivation of X inactivated genes with aging [29–32].

With the idea of AGE-DEPENDENT DEGENERATION OF EPIGENETIC PATTERNS in mind, the effects of *epg* in our conceptual model should be considered time-dependent. A purely genetic model does not easily incorporate this time-dependent idea, yet increasing incidence with age is a notably common feature of complex disease epidemiology. Highly penetrant, monogenic mendelian disorders generally arise congenitally or early in childhood, whereas most complex disease incidence starts to rise in middle age and increases decade by decade thereafter (Figure 3). This age-related increase in disease can be related to genotypes and epigenotypes by considering the g_{dep} disease relationship to be time-dependent in our conceptual model of CDGE. For example, the penetrance and relative risk of a disease-associated g_{dep} -allele will increase with age under the CDGE model because of the age-related attenuation of the relevant epigenotype (Figure 4). Direct measurement of the epigenotype would explain this age relationship by including the age-related variation into the epigenotype measure, creating a more precise estimate of the g_{dep} -disease relationship in the epigenotype context.

The contribution of the environment to epigenetic variation

The environment can affect the degeneration of epigenetic marks with age and should be considered in the context of disease predisposition. For example, methylation patterns in an environment that is low in methionine, with limited methylation ability, can degenerate over time at a much faster rate than methylation decay in a non-limited environment. Several examples of environmentally mediated epigenetic effects exist. A recent study suggested that hyperhomocysteinemia in patients with uremia is associated with global hypomethylation and biallelic expression of genes that normally show monoallelic gene expression. The abnormalities could be reverted by folate treatment of the same patients [33]. Hyperhomocysteinemia is an independent risk factor for Alzheimer's disease [34], heart disease [35] and stroke [36]. Folate deficiency, a precursor for DNA methylation, predisposes to several complex disease traits including anemia [37], neural tube defects [38,39] and cancer [40]. Lack of *de novo* methylation in *Dnmt3b*-null mice also leads to neural tube defects [41], suggesting that folate might act through changes in methylation that affect gene expression patterns [41]. (For more information, see Ref. [42].) Other environmental exposures that have been shown to affect the epigenome include metallotoxins, such as Ni, As and Cd [43,44].

Taking these results into account, environment context should be incorporated into our working model. This mainly occurs at two points: (i) the direct environmental influence on the epigenetic marks; and (ii) the environmental influence on the rate of attenuation of epigenetic marks with age. Both result in an epigenotype change that can be measured directly; the inclusion of measured *epg* in an investigation of disease would help to explain the observational age and environmental heterogeneity in genetic relationships and would reduce noise.

Stochastic events contributing to epigenetic variation

Some epigenetic variation is stochastic during development and aging [28,45]. Random changes in chromatin or methylation patterns from one generation to the next, or within the lifespan of a single generation, could contribute greatly to the total variation in epigenotypes. Recent evidence supports this concept by showing rapid selection for particular epigenotypes in response to environmental pressure. Stochastic epigenetic variation might also explain the discordance of phenotype between monozygotic twins, who are genetically identical and usually raised in the same environment. For instance, the untranslated RNA *LIT1*, shows variation in methylation and imprinting, distinguishing affected and unaffected monozygotic twins with Beckwith-Wiedemann syndrome [46]. However, it is difficult to exclude the possibility of a different environment for humans, therefore, model organisms give an indication of the variability among clonal populations under the same environment. The best understood example of an epigenetically mediated stochastic event is that of X-inactivation [47] and when a recent study explored age-dependent changes in epigenetic patterns in clonal mice it demonstrates great

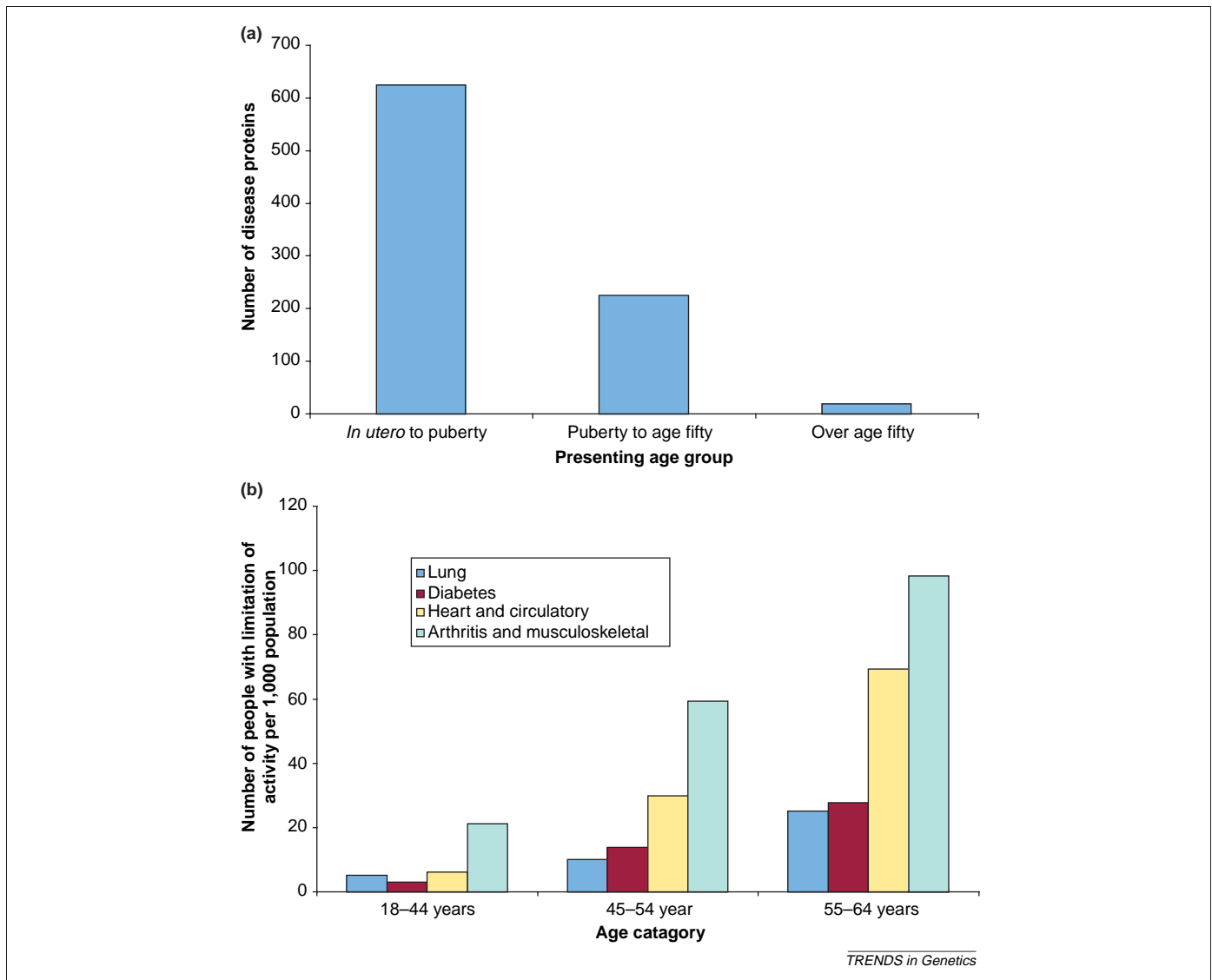


Figure 3. The age of presentation of conventional (a) mendelian disorders and (b) common complex traits causing limitation of activity among working-age adults in 1999–2001. Data for (a) are from Jimenez-Sanchez *et al.* [56], and data for (b) are from the U.S. National Center for Health Statistics (<http://www.cdc.gov/nchs/hus.htm>).

variability within individual cohorts raised in the same environment [28].

Modeling the CDGE approach to disease

We have discussed the potential for interplay between sequence variation, epigenetic variation, environment and age. We have proposed several sources of variation in the epigenetic marks themselves. The challenge, of course, is to combine these possibilities into a framework that is useful for considering these factors simultaneously to test particular hypotheses. We have summarized our overall framework in Figure 1 and as follows:

$$epg \leftarrow g_{epg} + ep g_{par} + e_{par} + age + e + e^* age \quad (\text{Eqn 1})$$

The model includes three classes of genetic variation (g_{ind} , g_{dep} and g_{epg}), at least two classes of environmental variation (e_{par} and e) and age. We should be clear that the genetic and environmental classes are represented as separate entities for ease of illustration, although this assumption will not strictly be true. For example, one could imagine the same gene having some alleles with

direct effects on disease but others that rely on an epigenetic mechanism.

CDGE is largely conjectural but is helpful for designing tests of hypotheses regarding epigenetics, for assessing the importance of including measured epigenotypes in disease-gene discovery, and for explaining the impact of aging and environment on genetic predisposition to disease. To illustrate some of these concepts, we simulated disease status in a population according to our conceptual CDGE model. For example, some genes had direct effects on disease status. These were modeled as rare alleles that were highly penetrant, regardless of age or environment. We then modeled genes with more common alleles that were highly penetrant when expressed but that were often masked by particular epigenotypes (modeled as high methylation levels). We finally modeled variation in these epigenotype values that were due to other genes and we modeled the attenuation of these methylation levels with age and environment (three environments, each with different rates of age-related attenuation). Once these populations were simulated, we examined the

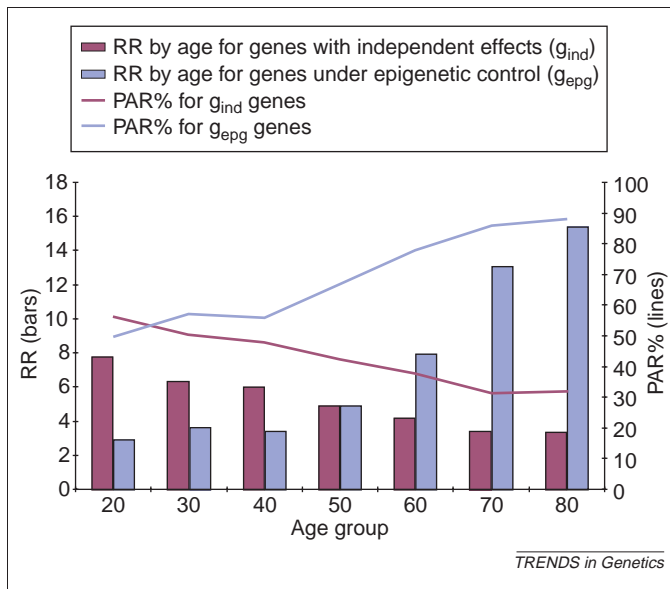


Figure 4. The relative risk (RR) and population attributable risk (PAR) by age for direct-effect genes and for those that are under epigenetic control. The results of simulations of 40 populations modeled according to Eqn 1. Simulations were used to create 40 populations containing affected and unaffected individuals for a genetic epidemiologic analysis. Age, environmental status and genotypes for three different genes were first simulated at random according to specified frequencies. Disease status was then simulated according to CDGE. One gene contained a rare allele that was highly penetrant, regardless of epigenotype, age, or environment (i.e. g_{dep}). Another gene contained a more common allele with a penetrance that depended on epigenotype value, such that lower values corresponded to higher penetrances. These epigenotype values were assigned based on a third gene, with genotypes that predicted mean epigenotype levels. These epg levels were then reduced according to age, such that higher ages had the greatest reduction of epigenotype value, and according to environment, with three levels: one with a slow rate of age-related attenuation, one with a moderate rate and one with a high rate of age-related attenuation. From these simulated populations, relative risks for g_{ind} and g_{dep} genes were estimated in cross-sectional analysis by each age decade and averaged over 40 populations (bars). Because this risk only reflects the magnitude of a genetic effect, and not the importance of that genotype with respect to all cases in the population, we also estimated population attributable risk percent (PAR%) at each decade. This reflects the proportion of cases in the population that can be explained by the particular genetic effect.

relative-risk estimates for genotypes with direct effects on disease versus genotypes in which penetrance was dependent on epigenetic context. For example, our CDGE model emphasizes that genes that act independently of epigenotype (g_{ind}) will be the most important type of DNA variation for early-onset diseases, whereas g_{dep} -type genes, which are under epigenetic control, might be of major importance in common diseases with a later onset. However, there is a decrease in relative risks by cross-sectional age group for g_{ind} genotypes and an increase in relative risks by age for g_{dep} genotypes (Figure 4).

More strikingly, the population attributable risk (PAR) by age group shifts dramatically from the majority of case burden at early ages as a result of g_{ind} genes, to late ages as a result of g_{dep} genes. The interactive effect of environment can also be seen by plotting the age-related increase in relative risk of g_{dep} genotypes according to severity of environment. Therefore, the most extreme environments, such as impaired methylation ability, would lead to a more rapid alteration of epigenetic marks and a faster increase in penetrance of the detrimental g_{dep} genotype. The current epidemiological practice of considering all

measured genotypes in the same way has reduced the ability to identify genes of the g_{dep} variety (which might be important to common diseases), and this might be one reason for the difficulty in identifying genes for such disorders. The inclusion of measured epigenotypes in such analyses would help to explain age and environment-dependent effects in a more precise way.

Practical epigenotyping

We describe how these ideas might be applied to the practical incorporation of epigenetic data along with conventionally obtained genetic data in disease-association studies in the following sections (Figure 5).

Define classes of normal variation, using high-throughput tools for epigenotyping

It will be necessary to identify and classify epigenetic marks, by the type of variation (i.e. heritable epigenetic marks that are transmitted vertically), by inter-individual variants in the population and by tissue-specific variants. This will require the development of high-throughput tools for epigenotyping. Some progress has been made towards comprehensive analysis of DNA methylation, through array-based hybridization. A European consortium has begun to perform complete methylation analysis of the MHC region of chromosome 6, using high-throughput bisulfite DNA sequencing [48]. In this manner, tissue-specific differences can be distinguished from polymorphic methylation variants and invariants (e.g. unmethylated CpG sites, can be excluded from further study). Surprisingly, more genome-scale approaches to epigenotype might be possible for allele-specific gene expression and chromatin modification than for DNA methylation because of the growing availability and accuracy of large-scale chips for hybridization. Some success has already been achieved in large-scale allele-specific gene expression measurement by hybridization of cDNA to a 'single nucleotide polymorphism (SNP)-Chip'. Similarly, chromatin immunoprecipitation to a chip is already possible at a genome level in yeast and *Drosophila* and will probably soon be practical for mammalian cells. However, it will be crucial to first identify the nature of tissue and population variation in epigenetic marks. For a given disease-based study, it might be possible to use a surrogate tissue, which is why it is important to establish which are population-based variants across tissues. For some disorders, the target tissue itself is available for analysis (e.g. lymphocytes in autoimmune disease). Given that we currently have no baseline for epigenetic variation, a useful experiment might be to determine whether epigenetic variation was less between monozygotic twins than among sibs.

Relate epigenotype variation to genotype variation in normal individuals

A potential source of epigenetic variation is genetic variation (g_{epg}). There are two ways genetic variation could affect the epigenotype, *in cis* or *in trans*. If sequence variation in the epigenetically affected gene is more susceptible to epigenetic modification, it would be the result of a *cis* effect (e.g. more CpG dinucleotides in one haplotype than in another). Such *cis* effects were

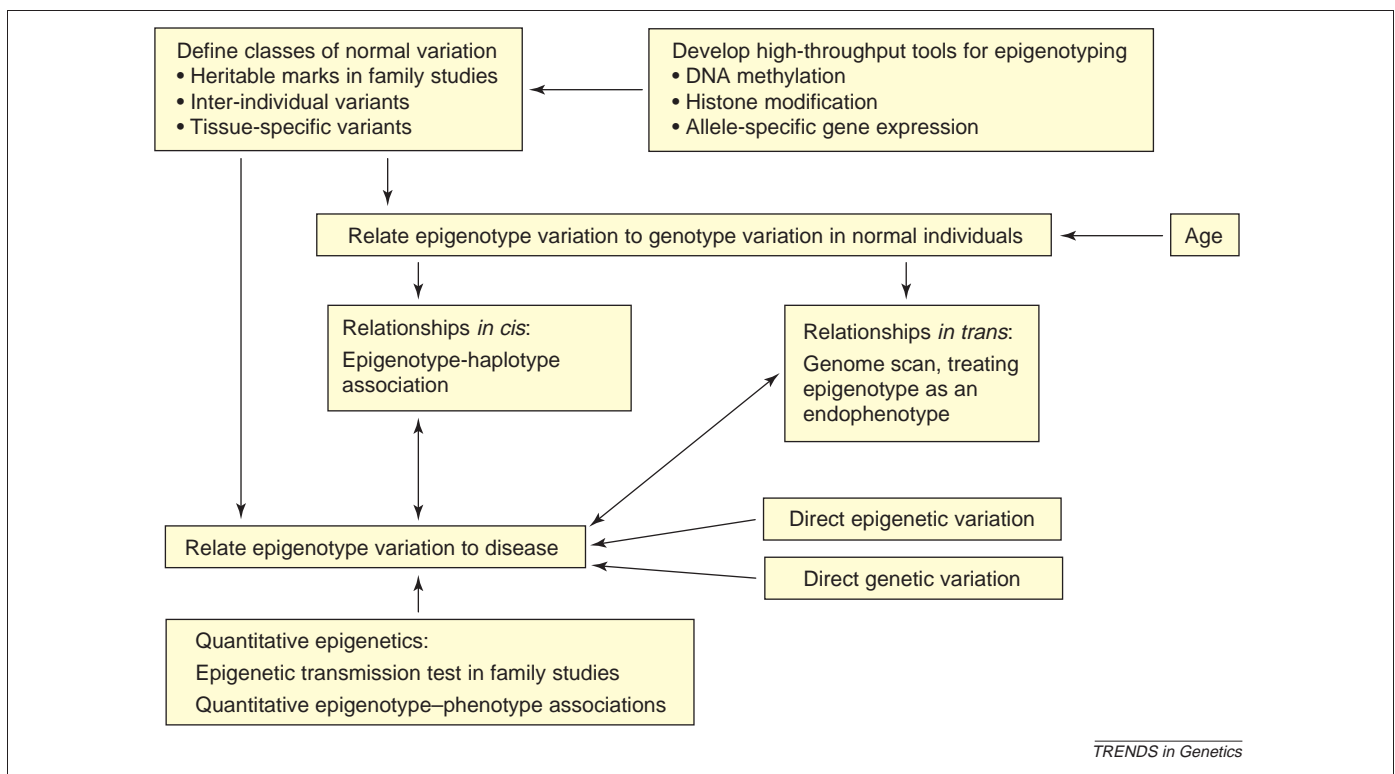


Figure 5. Practical epigenotyping. The steps needed to introduce epigenetic information into human population-based disease studies are shown. Classes of normal epigenetic variation must be defined, which will require development of high throughput tools for epigenotyping. The relationship between epigenetic and genetic variation must be examined through epigenotype-haplotype association *in cis*, and through genome scan, treating epigenotype as an endophenotype for *trans*-interactions, and their potential age-dependence. Application to complex disease must be incorporated with existing genetic tools but will also require the development of novel quantitative approaches.

suggested when individual haplotypes were compared with their corresponding epigenotype [49]. It would be worthwhile to examine the epigenome of a particular haplotype block in a population, to see if it correlated with the underlying DNA, and to see whether gene expression correlated with the haplotype block or with its epigenetic state. The detecting of *trans* effects in humans could be performed by treating the epigenotype as an endophenotype in a genome scan (i.e. measuring the relationship between the epigenetic change at one locus with sequence variants elsewhere in the genome).

Trans effects on the epigenotype could also be tested in model organisms by identifying a hypothetical epigenetic-strain-dependent effect, either using linkage to identify the epigenetic modifier, where the phenotype is the epigenetic variation of g_{dep} , or using a genetic screen. Such efforts to identify *trans*-acting regulators of gene expression have been performed in yeast [50,51] and mice [52], treating the level of gene expression as a quantitative trait. We predict that similar approaches could be used to identify the causes of epigenetic modification directly. Some modifiers of X-inactivation have already been discovered using a genetic screen [53]. It is also interesting that some *trans*-acting factors in yeast, which were originally thought to be transcription factors, have now been shown to be chromatin-modifying factors [54]. The effects of age on the epigenotype-genotype relationship could be more easily determined in a unaffected population, preferably one in which samples were obtained and banked over time.

Incorporate epigenotype variation into disease association studies

How can the epigenotype be measured? At the single gene level, the tools already exist. Cancer epigenetics has illuminated this pathway, with many studies of DNA methylation, allele-specific gene expression and, more recently, chromatin modification. In almost all cases, these studies are directed at specific genes that are implicated by genetic alterations. The simplest form of epigenetic analysis would then involve regions of interest implicated by genetic models, and these studies would also benefit from information on epigenotype-haplotype association (Figure 5). This approach is valuable because it might reveal epigenetic modification that modulates the effect of genetic variation (i.e. it might be valuable for g_{dep} genes). Furthermore, the size of the target region is not onerous because it is delimited by genetic-association- or linkage studies.

A related idea might be to examine the epigenotype at a candidate locus that has been identified in one population but fails to be confirmed in another. Perhaps, in some cases, a haplotype association reveals an overlying epigenotype or a strong association that is modified by the epigenotype. Epigenetic analysis of another population might reveal such an association even if the haplotype itself differs. However, such a gene- or region-focused approach will not identify more complex genetic-epigenetic interactions. Ideally, one would like to examine epigenetic differences at loci throughout the genome, which will require more general genome-level epigenotypic

scans. These disease studies would also benefit from the genome scan treating epigenotype as endophenotype on normal individuals described previously (Figure 5). In addition, only the direct measurement of epigenotype will reveal the direct effects of epigenotype on disease.

Novel computational approaches to epigenomics

Once experimental tools for high-throughput epigenotyping are developed, how should they be applied statistically to disease and normal populations? The answer follows from the outline given previously of the potential sources of epigenetic variation in a population. Some might be transmitted vertically in families. For example, epigenetic information at a locus could be measured when assessing disease relationships, if a given locus were epigenetically silenced in offspring that did not receive the disease-predisposing SNP. A more complex example of the use of family studies is the analysis of transmission of an epigenetic mark from parent to child. For example, it might be more important that a methylated or silenced allele, rather than a SNP, was transmitted to an affected individual. One might also need to include the parental epigenotype in such studies. Doing so would require the development of novel statistical algorithms because the transmission test for linkage disequilibrium (TDT) is essentially a test for recombination but applied to family trios, whereas epigenetic modification need not be in linkage disequilibrium with the affected locus.

We would also argue that greater attention should be paid to the quantitative nature of the assays. For example, it is common to use techniques such as single nucleotide primer extension or real-time PCR for categorizing genes as imprinted or not, even though these tests were not designed for that purpose. These techniques can be applied inappropriately, however, there is also untapped quantitative power within them that could be used in epigenotype-phenotype association studies.

Age and environment

If age and environment contribute to disease phenotype in part through epigenetic variation, the measurement of epigenotype would simplify disease identification. Currently, analysis of disease phenotypes deals with a potential age effect by stratifying for age, typically in deciles. Environmental effects are more difficult to measure and are usually limited to those suggested by earlier studies. We suggest that by measuring epigenotype, one will include within the resulting statistical models any age or environmental effects that affect disease through epigenetics. A more surprising implication of the model is that measurement of epigenotype might act as a surrogate marker for parental environment and thereby might increase the power of epidemiological studies. Such an approach might be useful because a recent epidemiological study has suggested that the diet of grandparents affects both birth weight and disease occurrence in grandchildren [55]. Current attempts to characterize disease predisposition have not taken these subtleties into account and might have therefore missed important sources of variation. We emphasize that we do not yet know to what degree epigenetics might contribute to disease generally

but the question itself has been nearly impossible to ask until now. We hope this discussion provides a useful framework for investigation.

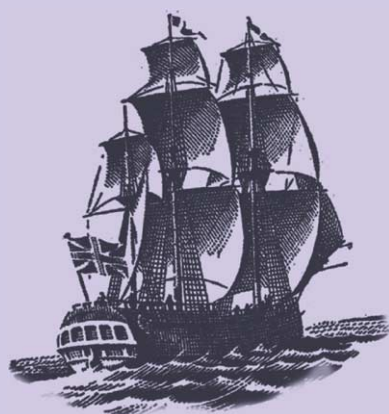
Acknowledgements

We thank Barbara Migeon, Barton Childs, David Valle, Rolf Ohlsson and Victor McKusick for their critical reading of the manuscript. This work was supported by NIH grants HG03233 and CA65145 to A.P.F., and a Fulbright Scholarship to H.B.

References

- 1 Feinberg, A.P. and Tycko, B. (2004) The history of cancer epigenetics. *Nat. Rev. Cancer* 4, 143–153
- 2 Feinberg, A.P. The epigenetics of cancer etiology. *Semin. Cancer Biol.* (in press)
- 3 Cui, H. *et al.* (2003) Loss of IGF2 imprinting: a potential marker of colorectal cancer risk. *Science* 299, 1753–1755
- 4 Jenuwein, T. and Allis, C.D. (2001) Translating the histone code. *Science* 293, 1074–1080
- 5 Herman, J.G. and Baylin, S.B. (2003) Gene silencing in cancer in association with promoter hypermethylation. *New Engl. J. Med.* 349, 2042–2054
- 6 Feinberg, A.P. *et al.* (2002) Epigenetic mechanisms in human disease. *Cancer Res.* 62, 6784–6787
- 7 Cooney, C.A. (1993) Are somatic cells inherently deficient in methylation metabolism? A proposed mechanism for DNA methylation loss, senescence and aging. *Growth Dev. Aging* 57, 261–273
- 8 Petronis, A. (2001) Human morbid genetics revisited: relevance of epigenetics. *Trends Genet.* 17, 142–146
- 9 Issa, J.P. and Baylin, S.B. (1996) Epigenetics and human disease. *Nat. Med.* 2, 281–282
- 10 Issa, J.P. (2002) Epigenetic variation and human disease. *J. Nutr.* 132, 2388S–2392S
- 11 Valle, D. (2004) Genetics, individuality, and medicine in the 21st century. *Am. J. Hum. Genet.* 74, 374–381
- 12 Whitelaw, E. and Martin, D.I. (2001) Retrotransposons as epigenetic mediators of phenotypic variation in mammals. *Nat. Genet.* 27, 361–365
- 13 Poulsen, P. *et al.* (1999) Heritability of type II (non-insulin-dependent) diabetes mellitus and abnormal glucose tolerance—a population-based twin study. *Diabetologia* 42, 139–145
- 14 Bertelsen, A. *et al.* (1977) A Danish twin study of manic-depressive disorders. *Br. J. Psychiatry* 130, 330–351
- 15 Ravenel, J.D. *et al.* (2001) Loss of imprinting of insulin-like growth factor-II (IGF2) gene in distinguishing specific biologic subtypes of Wilms tumor. *J. Natl. Cancer Inst.* 93, 1698–1703
- 16 Rutherford, S.L. and Lindquist, S. (1998) Hsp90 as a capacitor for morphological evolution. *Nature* 396, 336–342
- 17 Sollars, V. *et al.* (2003) Evidence for an epigenetic mechanism by which Hsp90 acts as a capacitor for morphological evolution. *Nat. Genet.* 33, 70–74
- 18 Taylor, J.P. *et al.* (2003) Aberrant histone acetylation, altered transcription, and retinal degeneration in a Drosophila model of polyglutamine disease are rescued by CREB-binding protein. *Genes Dev.* 17, 1463–1468
- 19 Xu, G.L. *et al.* (1999) Chromosome instability and immunodeficiency syndrome caused by mutations in a DNA methyltransferase gene. *Nature* 402, 187–191
- 20 Silva, A.J. and White, R. (1988) Inheritance of allelic blueprints for methylation patterns. *Cell* 54, 145–152
- 21 Sandovici, I. *et al.* (2003) Familial aggregation of abnormal methylation of parental alleles at the IGF2/H19 and IGF2R differentially methylated regions. *Hum. Mol. Genet.* 12, 1569–1578
- 22 Rakyán, V. and Whitelaw, E. (2003) Transgenerational epigenetic inheritance. *Curr. Biol.* 13, R6
- 23 Cooney, C.A. *et al.* (2002) Maternal methyl supplements in mice affect epigenetic variation and DNA methylation of offspring. *J. Nutr.* 132, 2393S–2400S
- 24 Waterland, R.A. and Jirtle, R.L. (2003) Transposable elements: targets for early nutritional effects on epigenetic gene regulation. *Mol. Cell. Biol.* 23, 5293–5300
- 25 Jordan, I.K. *et al.* (2003) Origin of a substantial fraction of human

- regulatory sequences from transposable elements. *Trends Genet.* 19, 68–72
- 26 Mays-Hoopers, L.L. (1989) Age-related changes in DNA methylation: do they represent continued developmental changes? *Int. Rev. Cytol.* 114, 181–220
- 27 Issa, J.P. *et al.* (1994) Methylation of the oestrogen receptor CpG island links ageing and neoplasia in human colon. *Nat. Genet.* 7, 536–540
- 28 Bennett-Baker, P.E. *et al.* (2003) Age-associated activation of epigenetically repressed genes in the mouse. *Genetics* 165, 2055–2062
- 29 Wareham, K.A. *et al.* (1987) Age related reactivation of an X-linked gene. *Nature* 327, 725–727
- 30 Brown, S. and Rastan, S. (1988) Age-related reactivation of an X-linked gene close to the inactivation centre in the mouse. *Genet. Res.* 52, 151–154
- 31 Busque, L. *et al.* (1996) Nonrandom X-inactivation patterns in normal females: lyonization ratios vary with age. *Blood* 88, 59–65
- 32 Mukherjee, A.B. and Wallace, K.C. (1990) Changes in frequency and localization of human X- and Y-chromatin bodies at interphase during *in vitro* cellular aging. *Mech. Ageing Dev.* 53, 61–71
- 33 Ingrosso, D. *et al.* (2003) Folate treatment and unbalanced methylation and changes of allelic expression induced by hyperhomocysteinemia in patients with uraemia. *Lancet* 361, 1693–1699
- 34 Seshadri, S. *et al.* (2002) Plasma homocysteine as a risk factor for dementia and Alzheimer's disease. *New Engl. J. Med.* 346, 476–483
- 35 Homocysteine Studies Collaboration, (2002) Homocysteine and risk of ischemic heart disease and stroke: a meta-analysis. *Jama* 288, 2015–2022
- 36 Li, Z. *et al.* (2003) Elevated plasma homocysteine was associated with hemorrhagic and ischemic stroke, but methylenetetrahydrofolate reductase gene C677T polymorphism was a risk factor for thrombotic stroke: a multicenter case-control study in China. *Stroke* 34, 2085–2090
- 37 Donnelly, J.G. (2001) Folic acid. *Crit. Rev. Clin. Lab. Sci.* 38, 183–223
- 38 Gos, M. Jr and Szepecht-Potocka, A. (2002) Genetic basis of neural tube defects. II. Genes correlated with folate and methionine metabolism. *J. Appl. Genet.* 43, 511–524
- 39 Copp, A.J. (1998) Prevention of neural tube defects: vitamins, enzymes and genes. *Curr. Opin. Neurol.* 11, 97–102
- 40 Pufulete, M. *et al.* (2003) Folate status, genomic DNA hypomethylation, and risk of colorectal adenoma and cancer: a case control study. *Gastroenterol.* 124, 1240–1248
- 41 Juriloff, D.M. and Harris, M.J. (2000) Mouse models for neural tube closure defects. *Hum. Mol. Genet.* 9, 993–1000
- 42 Van den Veyver, I.B. (2002) Genetic effects of methylation diets. *Annu. Rev. Nutr.* 22, 255–282
- 43 Poirier, L.A. and Vlasova, T.I. (2002) The prospective role of abnormal methyl metabolism in cadmium toxicity. *Environ. Health Perspect.* 110 (Suppl. 5), 793–795
- 44 Sutherland, J.E. and Costa, M. (2003) Epigenetics and the environment. *Ann. N.Y. Acad. Sci.* 983, 151–160
- 45 Barbot, W. *et al.* (2002) Epigenetic regulation of an IAP retrotransposon in the aging mouse: progressive demethylation and desilencing of the element by its repetitive induction. *Nucleic Acids Res.* 30, 2365–2373
- 46 Weksberg, R. *et al.* (2002) Discordant KCNQ1OT1 imprinting in sets of monozygotic twins discordant for Beckwith-Wiedemann syndrome. *Hum. Mol. Genet.* 11, 1317–1325
- 47 Migeon, B.R. (2002) X chromosome inactivation: theme and variations. *Cytogenet. Genome Res.* 99, 8–16
- 48 Novik, K.L. *et al.* (2002) Epigenomics: genome-wide study of methylation phenomena. *Curr. Issues Mol. Biol.* 4, 111–128
- 49 Murrell, A. *et al.* (2004) An association between variants in the IGF2 gene and Beckwith-Wiedemann syndrome: interaction between genotype and epigenotype. *Hum. Mol. Genet.* 13, 247–255
- 50 Brem, R.B. *et al.* (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science* 296, 752–755
- 51 Yvert, G. *et al.* (2003) Trans-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. *Nat. Genet.* 35, 57–64
- 52 Schadt, E.E. *et al.* (2003) Genetics of gene expression surveyed in maize, mouse and man. *Nature* 422, 297–302
- 53 Percec, I. *et al.* (2003) An N-ethyl-N-nitrosourea mutagenesis screen for epigenetic mutations in the mouse. *Genetics* 164, 1481–1494
- 54 Brownell, J.E. *et al.* (1996) Tetrahymena histone acetyltransferase A: a homolog to yeast Gen5p linking histone acetylation to gene activation. *Cell* 84, 843–851
- 55 Kaati, G. *et al.* (2002) Cardiovascular and diabetes mortality determined by nutrition during parents' and grandparents' slow growth period. *Eur. J. Hum. Genet.* 10, 682–688
- 56 Jimenez-Sanchez, G. *et al.* (2001) Human disease genes. *Nature* 409, 853–855



Endeavour

Coming soon in the quarterly magazine for
the history and philosophy of science:



The future of electricity in 1892 by G.J.N. Gooday
The First Personal Computer by J. November
Sherlock Holmes the Scientist by L. Snyder

Locate *Endeavour* on *ScienceDirect* (<http://www.sciencedirect.com>)