La Medicina del Lavoro

Med Lav 2017; 108, 1: 10-23
DOI: 10.23749/mdl.v108i1.6072

Epigenetics applied to epidemiology: investigating environmental factors and lifestyle influence on human health

Valeria Motta[1], Matteo Bonzini[1,2], Lotte Grevendonk[1], Simona Iodice[2], Valentina Bollati[1,2]

[1] EPIGET - Epidemiology, Epigenetics and Toxicology Lab - Department of Clinical Sciences and Community Health, Università degli Studi di Milano, Milan, Italy
[2] Fondazione IRCCS Ca' Granda - Ospedale Maggiore Policlinico, Milan, Italy

Key words: Epigenetic modifications; *in-utero* epigenetics; environmental factors; lifestyle; epidemiological study

Parole chiave: Modificazioni epigenetiche; epigenetica in-utero; fattori di rischio ambientale; stili di vita; epidemiologia

Summary

Epigenetics modifications, that include variations in DNA methylation, histone acetylation and micro RNA (miRNA) expression, co-operate together, influencing genome expression and function, in response to exogenous stimuli or exposures. Thus, epigenetic tools applied to epidemiology are useful in investigating, at the population level, the relationships between exposures to environmental, lifestyle, genetic, socioeconomic risk factors, and the epigenome, and/or specific health outcomes. But the choice of an appropriate study design and of valid epidemiological methods has a key role in determining the achievement of the study. This review summarises available evidence about the role of the most investigated epigenetic mechanisms in mediating lifestyle or environmental exposure effects on human health, considering the entire life-course, from in-utero to adulthood. Moreover, we illustrate the most important variables that should be properly considered when designing an epigenetic epidemiology study: the choice of an appropriate study design, a proper estimation of the required sample size, a correct biological sample selection, a validation strategy for epigenetics data, and an integrated exposure assessment methodology.

Riassunto

***«L'epigenetica applicata all'epidemiologia: indagine degli effetti sulla salute di ambiente e stili di vita»**. Le modificazioni epigenetiche, che comprendono alterazioni della metilazione del DNA, della composizione degli istoni e dell'espressione di micro RNA (miRNA), agiscono sinergicamente influenzando l'espressione e la funzionalità del genoma, in risposta a stimoli esterni derivanti da esposizioni ambientali. Marcatori epigenetici sono quindi sempre più utilizzati in studi epidemiologici che indagano, a livello di popolazione, la relazione tra fattori di rischio ambientali, genetici, legati agli stili di vita o socio-culturali ed effetti su specifiche malattie o sull'epigenoma nel suo complesso. La scelta di un appropriato disegno dello studio e l'uso di corretti metodi epidemiologici riveste un ruolo chiave nel determinare la qualità e il successo di questo tipo di studi. La presente revisione di letteratura riassume le evidenze disponibili circa il ruolo dei meccanismi epigenetici più studiati nel mediare gli effetti di ambiente e stili di vita sulla*

Pervenuto il 23.12.2016 - Accettato il 17.1.2016
Corrispondenza: Prof.ssa Valentina Bollati, Phd, EPIGET - Epidemiology, Epigenetics and Toxicology Lab, Department of Clinical Sciences and Community Health, Università degli Studi di Milano, Via San Barnaba 8, 20122, Milan, Italy
E-mail: valentina.bollati@unimi.it

salute umana, considerando le diverse fasi del corso della vita, dal suo sviluppo in utero all'età adulta. Si passano poi in rassegna le principali variabili metodologiche che dovrebbero essere adeguatamente considerate quando si disegni uno studio epidemiologico che utilizza l'epigenetica: la scelta di un appropriato disegno dello studio, la determinazione di una ottimale dimensione campionaria, la selezione del campione biologico più indicato, un programma di validazione delle misure epigenetiche, una valida strategia per la stima corretta dell'esposizione individuale.

New tools in epidemiology

Epidemiology investigates disease occurrence in human populations and factors associated with it. Epidemiological studies are often based on information deriving from current statistics (e.g., mortality), registries (e.g., cancer incidence) surveys, and questionnaires, and have been extremely helpful in discovering key risk factors, in identifying individuals at high risk and hence developing disease prevention measures (6, 36, 50).

In the last decades epidemiology has been progressively exploring new tools provided by the advancement of molecular sciences and tecniques, by including in the study design the collection of biological samples, which allow i) the quantification of specific molecules (e.g. DNA or RNA adducts) and early biological markers (e.g. somatic mutations), ii) the better characterization of study subjects (e.g. genomic variations in metabolic genes) and their risk stratification, and iii) the use of molecular markers to further classify diseases that are currently categorized by etiology or prognosis.

Epigenetic epidemiology has been defined as the study of the associations between epigenetic variations and risk of disease (65). Epigenetics describes several molecular mechanisms such as DNA methylation, histone modifications and miRNA expression, that operate together in a synergic manner and lead to the alteration of genome expression patterns and functions after exposure to exogenous influences. Therefore, while genetic epidemiology is focused on variation in DNA sequence, epigenetic epidemiology has to deal with the complexity of an interplay between different modifications.

Environmental epigenetics

Evaluating the role of environmental risk factors in disease occurrence is a major challenge. Environment should be understood in its broadest meaning, not just in terms of exposure to chemicals. By definition, environment includes all the physical, chemical, socio-cultural and biological factors external to an individual (figure 1). The emerging field of environmental epigenetics studies the exposure to agents that impact epigenetic mechanisms.

The epigenetic effects of a single environmental agent may vary depending on concentration and time of exposure and operate in addition to other factors such as individual genetic susceptibility, age, sex, lifestyle, exposure to other pollutants or pollutant mixtures. Epigenetic changes have been shown to reflect the effects of both acute and chronic exposures, showing responses to exogenous stimuli which are not always linear. The non-linearity of the response is mostly dependent on the individual's life stage: for example, the fetus development is considered a critical phase because of the epigenetic reprogramming that occurs during pre-implantation (53), while ageing has been shown to be strongly associated with epigenetic changes, especially in methylation levels (29). During these periods of life, characterized by a hypersusceptibility to several exposures effect, an acute, low-dose exposure to external factors might have greater effects than high-dose exposures in adulthood. Available evidences from animal models indicate that exposure to endocrine-disruptive chemicals (EDCs) during critical periods of mammalian development might not follow the dose-response relationship (27).

The identification of cause and effect relationships between exogenous stimuli, epigenetic changes, and the potential disease development might be difficult to determine, since epigenetic changes are cumulative and somatically inherited, so they can persist even in the absence of the factor that established them (18). A useful approach that can be helpful in this context might be the longitudinal study of epigenetic changes in monozygotic twins (MZT) cohorts (5). MZT share variables such as genomic polimorfisms, parental origins, age and it

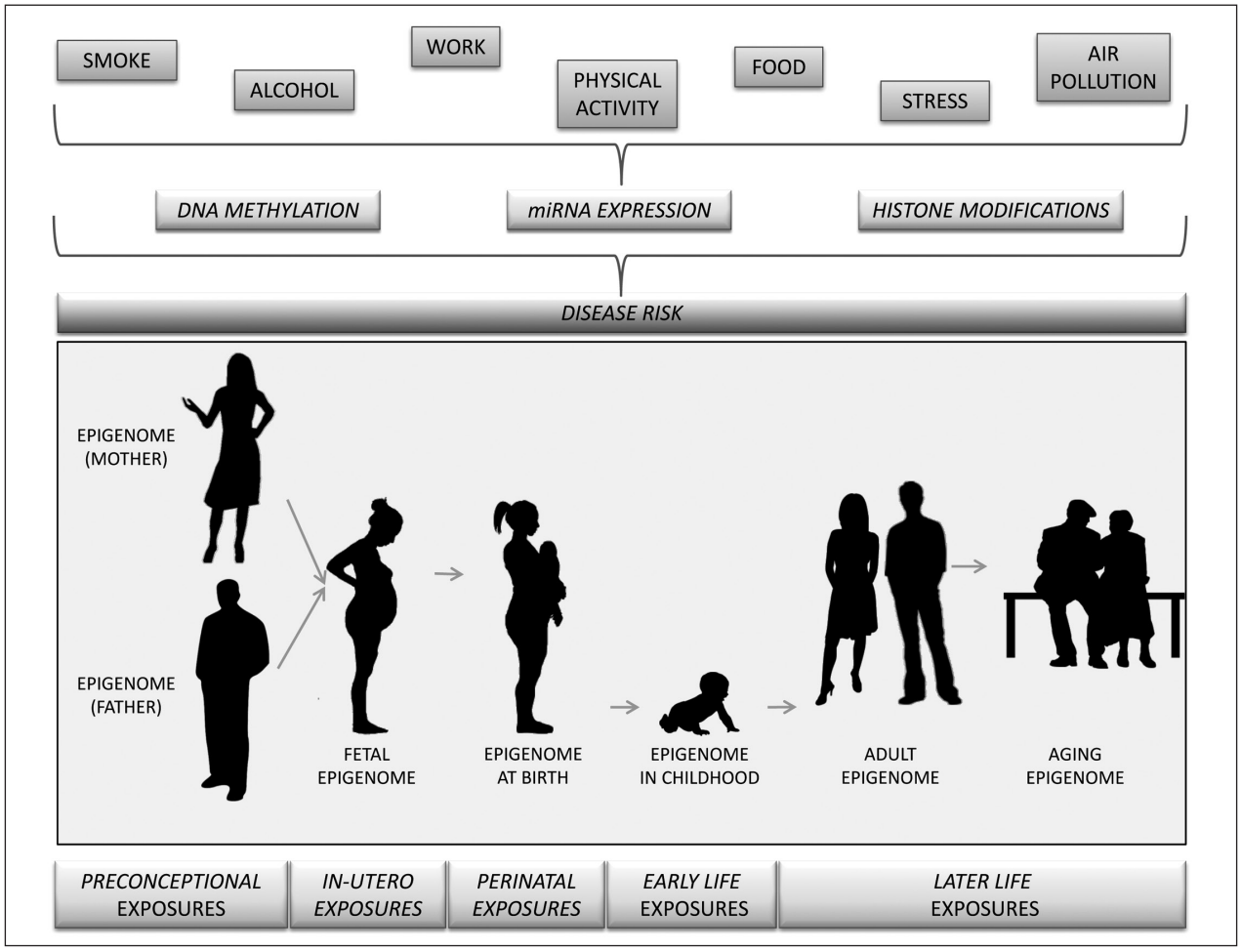

Figure 1 - Environmental factors affecting disease risk during life-course

is also possible to assume that MZT are exposed to a very similar prenatal environment. Following longitudinally MZT would highlight the environmental factors that are unique for each subject and that could drive specific epigenetic changes possibly responsible for disease development.

Current evidence demonstrates that epigenetics have a great potential to further our understanding of the molecular mechanisms of environmental and occupational exposures and to predict health-related risks due to the interplay between such exposures and individual susceptibility (9). Environmental epigenetics hold substantial potential for developing biological markers to predict which exposures would increase disease risk and which individual will be more susceptible to develop disease. In addition to this, the stability of some epigenetic marks in body fluids make them putative risk predictors in primary prevention strategies.

The environmental exposure assessment issue: the 'exposome'

In-vitro and animal studies investigate the effects of single agents and can identify with reasonable confidence which specific epigenetic alterations are associated with the exposure. These studies also allow investigating dose and time responsiveness, thanks to the possibility to control every single step of exposure.

In human studies, however, the same individual is usually exposed to several exogenous stimuli that can

influence epigenetic patterns in different ways, leading to diverse severity of development or exacerbation of a disease. For example, in a complex disease like asthma, different environmental, occupational, domestic and lifestyle factors such as air pollution, working conditions, house dust and smoking, are combined together in determining the disease status (32). Therefore in human studies, assessing multiple exposures in an accurate and reliable way is becoming an increasing need. In 2005 Wild presented the idea of studying the 'exposome' as summation of environmental exposures, 'in the broad sense of non-genetic exposures', over an individual's lifetime (67). According to this view, the 'exposome' considers three domains that have been described as general external environment, such as urban pollution and climate, specific external environment, such as smoking and diet, and internal environment including biological elements. Epigenetic mechanism are discussed among the internal environment domain.

An interesting point of view was presented in 2014 by Olden who in his commentary states that it would be intriguing to speculate that the accumulation of epigenetic changes resulting from the exposures to different environmental risk factors may be a cumulative predictive biomarker of susceptibility to chronic illnesses (42).

One major challenge for environmental epigenetics still lies in the identification of the interaction among all the exogenous agents (1). Additive, synergistic, and antagonistic effects have not yet been sufficiently explored in human epigenetics.

Effects of environmental exposures on epigenetic mechanisms

Most of the human studies conducted so far have focused on DNA methylation (45, 60) and miRNAs (37), whereas only few studies have investigated the effects of external exposures on histone modifications (74).

It is important to take into account that human exposure levels are usually orders of magnitude below the levels used in experimental studies and that therefore the observation of an effect might be inherently difficult, especially if the number of subjects is limited. Occupational exposures, which differ from ambient exposures in duration, frequency and concentration, have been given particular attention since they might allow an easier identification of the interplay between environmental factors and human host.

DNA methylation

Changes in DNA methylation have been extensively associated with exposure to air pollutants such as particulate matter (30), airborne benzene (9) and aero-dispersed heavy metals including nickel, cadmium, lead, and arsenic (4). In relation to these exposures, DNA methylation has been investigated at global level (i.e. measured as total presence of 5-methylcytosines), estimated through repetitive element (e.g. *Alu* and LINE elements) analysis, or at gene-specific level.

The analysis of *Alu* and LINE sequences allows for the amplification of a representative pool of repetitive elements and has been used as a surrogate for global DNA methylation changes (71). Decreased repetitive elements methylation, measured via bisulfite sequencing of LINE-1 and *Alu* elements, was identified in healthy individuals exposed to benzene (9), persistent organic pollutants (51), traffic related air pollution (4), lead (69), and arsenic (23). In a recent work, DNA methylation of different repetitive element subfamilies belonging to *Alus*, LINEs and HERVs, was examined in multiple groups of participants exposed to different types of airborne pollutants such as metal-rich particulate matter, benzene and elemental carbon. Interestingly, a different susceptibility of repetitive element methylation to pollutant exposure has been found, possibly explained by the sequence variation and GC-content differences between the subfamilies (11).

Specific genes found to be hypermethylated in response to environmental exposures include p15, MAGE-1, and H19 with benzene exposure (9), ACSL3 region with PAH exposure (44), and p53 and p16 with arsenic exposure (13). Changes of methylation levels in peripheral blood of human subjects have been shown after a short-term exposure to metal-rich particulate matter. The inducible nitric oxide synthase gene iNOS (also known as NOS2) promoter methylation was found to be

significantly lower in post-exposure blood samples compared with baseline, indicating a capability of the gene to rapidly respond to exogenous stimuli and a possible activation in the regulation of immune-stimulation and infection processes (60).

Besides the exposure to toxicants, there are increasing studies exploring the association between social or behavioral factors and methylation patterns (58). For example, a study of the biological effects of shiftwork in Italian male chemical plant workers showed a significant increase in methylation of TNF-α, a cytokine involved in systemic inflammation, in shift workers compared to day workers and an association between job seniority and hypomethylation of IFN-γ, a regulator of immunologic processes (10).

Methylation levels can also be altered with the absorption of dietary methyl groups that derive from foods that contain folate, choline, betaine, methionine, and serine (40) indicating that nutrition might play a role in the modification of methylation mechanisms. For example, a study on maternal diet showed different methylation patterns in candidate genes involved in metabolic disease among individuals who were pre-conceptionally exposed to famine compared with their unexposed same-sex siblings (61).

Emerging evidence indicates that epigenetic mechanisms may be involved in mediating effects of physical activity. In a recent work, physical activity was associated with higher blood methylation levels of LINE-1 elements, a class of repeated sequences known to be involved in inflammatory responses and chromosomal instability (73).

miRNA expression

The susceptibility of miRNAs to environmental exposures and the recent advances in molecular biology opened the opportunity for new approaches in population based studies. The tissue specificity of these small RNA molecules and the observed correlation with disease phenotype, suggest that circulating miRNAs are entitled to become powerful tools to serve as proxy of tissue specific miRNAs. Moreover, after the discovery that miRNAs are quite stable in biological fluids, there was an increasing interest in the possibility that changes in these miRNA levels could be used as noninvasive biomarkers for a variety of clinical settings. Circulating miRNAs have been reported in whole blood, serum, plasma, and other body fluids (26) and can be detected with high sensitivity and specificity using real-time PCR, deep sequencing, and microarray techniques.

Several epidemiologic studies suggested that circulating miRNAs could mediate the health effects of air pollution exposure. Recently Vrijens et al. identified some miRNAs associated with air pollution exposure and provided a list of putative biomarkers: miR-10b and miR-128 appeared to be differentially expressed both in vitro and in human studies in association with air pollution exposure (64).

Smoking is an exogenous factor that has been widely studied in relation to miRNAs expression patterns. Interestingly Takahashi et al. showed that smoking can alter the plasma miRNA profiles, but quitting smoking can restore the pattern, making it more similar to the miRNA profile of nonsmokers (59).

miRNA involvement in response to lifestyle habits has also been reported. miRNA levels have been observed to be altered following dietary modulation, with miRNA expression in human muscle being increased following a dietary challenge of essential amino acids (19). In a population of young healthy men, miRNA expression in circulating neutrophils resulted to be affected by brief physical exercise and strictly related to inflammatory processes (47).

Histones modifications

Histone analyses are time-consuming and problematic for large epidemiological studies. To assay for histone modifications, DNA must be extracted after cross-linking (usually by formalin treatment) to ensure that histone proteins are not removed during DNA purification. Only few studies have investigated the effects of environmental chemicals on histone modifications in human populations.

An investigation on the effects of metal-rich air particle exposure showed significantly increased levels of histone H3 with demethylation of lysine-4 (H3K4me2) and histone H3 with acetylation of lysine-9 (H3K9ac) in peripheral blood of healthy steelworkers (12). Occupational exposure to nickel

was associated with an increase in H3K4me3, a histone modification associated with transcriptional activation, and decrease in H3K9me2 in 120 healthy male subjects working in a nickel refinery in Jinchang, China (2).

Dietary compounds such as isothiocyanates, butyrate, and diallyl disulfide have been studied as modifiers of histone deacetylase (HDAC) activity (15). A study on the effects of diet was performed on healthy human subjects fed with a single serving of broccoli sprouts showing inhibition of histone deacetylase (HDAC) activity in circulating peripheral blood mononuclear cells 3-6 h after consumption, with concurrent induction of histone H3 and H4 acetylation (14).

An example of modulation of lifestyle on histones modifications in a target tissue comes from a study that found an effect of physical exercise (cycling) on increased acetylation of global histone 3 at lysine 36, a site associated with transcriptional elongation, in human skeletal biopsies (34).

In utero epigenetics

Although the epigenome is vulnerable to dysregulation throughout the life-course, it is estimated to be most susceptible during intrauterine development, when epigenetic profiles are being set. Embryogenesis includes rapid waves of epigenetic changes in which the entire genome is highly unstable and therefore most sensitive to hormonal and environmental factors (31). The first window of susceptibility is during germline cell development, when histones are modified and methylation is reprogrammed. After fertilization the entire genome undergoes demethylation followed by a re-methylation of some genes, once the embryo reaches the early blastocyst stage (22). This highly orchestrated process of de-programming and re-programming of the genome may represent a mechanism of plasticity of the organism in response to its environment as well as a mechanism through which long-term health consequences can be shaped (31).

Human epidemiological studies provided evidences that prenatal exposure to diverse environmental pollutants can alter the epigenetic programming and risk for diseases in adult life such as cancer, cardiovascular disease, obesity and even behavioral disorders including schizophrenia (22, 43). Besides the interaction between toxic and environmental exposures, investigating *in utero* epidemiology requires to consider interactions with genetic, nutritional and social factors, that can exacerbate effects. The investigated tissues should be specified in order to interpret the results of these studies and, obviously, must be as much representative of the fetus as possible. Since the placenta plays a critical role in the maternal-fetal interface, it is usual to investigate trans-placental epigenetic adaptations of key genes and pathways which may alter the *in utero* development and risks for long-term health outcomes. Alternatively, newborn cord blood can be studied to indicate how trans-placental environmental exposure can impact the composition of cells within the peripheral blood, referring to immunological effects of exposure (31, 43).

A well studied environmental toxicant is Arsenic, which has a relatively common occurrence, an interesting mechanism of action and the ability to induce epigenetic effects (31). Reported consequences of in-utero exposure include increased mortality from lung cancer and bronchiectasis in young adulthood (54). Pilsner et al. indicated an epigenetic effect in the association between maternal urinary arsenic levels and an overall increase in global 5-methylcytosine levels in infant cord blood (45). Another recent study demonstrated that an overall hypermethylation of CpG loci within CpG islands was related to increasing arsenic exposure, and indicated the possibility that even low arsenic levels induce shifts in the proportions of specific immune cell populations, specifically an increase in CD8+ T lymphocytes populations (25). A genome wide analysis associated maternal total urinary arsenic to an increased expression of 12 miRNAs in offspring cord blood that may contribute to the related immune response perturbations (48). Additional research on low-level arsenic exposure is nevertheless still needed to determine whether those epigenetic changes are associated with any adverse health effects.

Other environmental epigenetic toxicants of great interest are polycyclic aromatic hydrocarbons (PAHs) which are produced by fossil fuels burning. Parental exposure has been associated with multi-

ple adverse effects including fetal growth reduction, development delay, and behavioral disorders (43). Herbstman et al. showed that prenatal exposure to PAHs, measured by personal air monitor during the third trimester of pregnancy, was associated with lower global methylation levels measured in DNA of cord blood cells (21). Another study indicated the methylation of the ACSL3 region, a gene expressed in lung tissue, in umbilical cord DNA as a candidate biomarker of prenatal PAH exposure and a putative predictor of PAH-associated childhood asthma (44).

Besides the prenatal exposure to environmental toxicants, several maternal lifestyle factors are known to exert long term consequences for the offspring as well as epigenetic dysregulation. In several human studies prenatal exposure to maternal stressful conditions, including acute and chronic stressors, anxiety or depression, was associated with an increased risk for multiple neurobiological and behavioral problems in the offspring, during their adult life. Moreover, emerging evidence indicated that these long-lasting effects of maternal prenatal distresses are possibly mediated by DNA methylation dysregulations analyzed in cord blood and placental tissue (35, 63).

Despite increased awareness, maternal cigarette smoking during pregnancy continues to be a common habit causing low birth weight and severe health complications later in life, including asthma, cancer, obesity and type II diabetes (41).

Studies from placenta, cord blood, buccal cells and peripheral blood indicate that prenatal exposure to maternal cigarette smoking is associated with alterations in global DNA methylation patterns. In addition the probability of this epigenetic link is enforced by several specific genes including CYP1A1, AhRR, FOXP3, TSLP, IGF2, AXL, PTPRO, C11orf52, FRMD4A and BDNF, which are shown to have altered DNA methylation patterns in at least one of the above mentioned tissues (41, 57).

Other detrimental environmental factors of specific concern are the Endocrine Disrupting Compounds (EDCs) because they are widespread in the environment and because organism in early stage development is extremely sensitive to perturbation by substances with hormone-like activity (estrogenic, anti-estrogenic, and anti-androgenic) (52, 68). A wide variety of EDCs have been studied in animal models and in-vitro experiments and have been associated with epigenetic deregulation and with the capability of accumulation in early human embryos development. Among EDCs, worth mentioning bisphenol A (BPA), compound employed in making polycarbonate plastics (55), dioxins (38) and diethylstilbestrol (DES), a synthetic estrogen prescribed to prevent pregnancy complications or premature deliveries (24). There is consistent evidence that heavy metals such as cadmium, mercury, arsenic and lead may also have endocrine disrupting activity (52). A human study reported the association between prenatal cadmium exposure and cord blood DNA methylation levels (24). Interestingly, this association was sex-specific. In boys, 96% of the top 500 CpG sites indicated positive associations, whereas most correlations in girls were negative. Moreover, in girls hypermethylation was found to be associated with organ development, morphology and mineralization of bone, while in boys, changes in cell death-related genes have been showed.

Methodological tools to conduct an epigenetic epidemiology study

Study designs

In designing an epigenetic epidemiology study, an important factor that must be taken into account is that the unidirectionality of conventional genetic studies does not apply and that epigenetic changes may be an intermediate factor between exposure and disease development or may be a consequence of the disease state. Moreover, an individual's disease status may influence lifestyle and exposure to environmental factors, thus the association initially observed in the epigenome analysis may shift (49).

To account for this potential scenario, in which cause and effect are reversed - the so called "reverse causation" - it is important to determine *a priori* the temporal sequence of events to be evaluated. Several epidemiological study designs are available, each has pros and cons for epigenetic analyses, and are dependent from the research question that need to be investigated.

Cohort studies

In a cohort, groups of healthy individuals are recruited in a longitudinal study and followed over a defined period of follow-up time: weeks, months, years and potentially throughout their life. At baseline specific exposure and clinical information, as well as biospecimens, are collected from all participants. Study participants are then followed up and, at different time points after the initial study visit, additional data and biological samples are collected. As the follow-up proceeds, outcomes are recorded. Cohort studies play key roles in human epigenetics investigations: They are able to capture the dynamic nature of epigenetic changes and can contribute to the understanding of how the epigenetic mechanisms measured over time change as a results of temporal modifications in exposure. This allows any epigenetic change assessed over time to be related to any possible disease experience: long-term follow-up can help identifying important risk factors for common complex diseases, as well as follow-up of cohorts from birth or childhood can help identifying the importance of early exposures and developmental characteristics for adult health (39).

The repeated collection of exposure measurements and epigenetic markers at multiple time points throughout the study allows assessment of temporal relationships, strengthens the assessment of direction of causality, and increases the ability to detect small effects. The main disadvantage of longitudinal studies is that they are costly to investigate and maintain. Since study populations include healthy individuals, invasive sampling cannot be performed (until disease development) and only limited types of biological samples can be collected (i.e. blood, buccal cells, nasal cells, saliva, and urine) and stored. Moreover, the large numbers of participants typically involved makes biospecimen storage and laboratory analyses even more expensive.

Birth cohort studies

The longitudinal birth cohort is a well-established study design in epigenetic epidemiology that measures *in utero* and early life exposures, by recruiting parents during pregnancy or in the early postnatal period. Birth cohort studies allows for the collection of transgenerational and across-life samples such as cord blood, placenta and saliva, which are easily obtained at delivery and during early life from mothers and newborns. In longitudinal birth cohorts, epigenetic changes can be measured over time, can be related to pre-conceptional, perinatal, or early-life exposures, and may be used to assess health status, disease development and onset in children from delivery to adulthood.

Cross-sectional studies

In a cross-sectional study all factors of interest on each individual are assessed at one time point instead of long-term. Its important advantage is to be quick and less expensive and is often used to describe and assess the associations between exposure, epigenetic changes and disease. Its inability to establish the direction of causality make it difficult to analyze and explain dynamic changes in epigenome and its effect on the disease. Cross-sectional studies are primarily used when the aim is to determine the prevalence of a specific epigenetic mark, or compare the distribution of factors between well-defined groups that differ, for example, in sex, age, or smoking habits.

Case-control studies

In case-control studies, the study base includes subjects with disease (cases) and subjects free of disease (controls) from the same population. Controls are defined as individuals who would have become study cases if they had developed the disease. There are multiple types of study bases including i) a specific population, such as subjects living in the same geographic boundary, that is followed for a precise period of time, ii) a selection of cases (if cases are identified in a specific clinic, the corresponding source population is made up of all the people that would attend that clinic if they had the disease) (50), or iii) a cohort that has already been defined (Nested case-control study). Cases and controls may come from registries (such as cancer registries or demographic registries), clinics and hospitals (mono- or multi-center), or may be determined at recruit-

ment (cases admitted for first diagnosis, relapse). The main advantage of this study design is that it involves a larger number of readily available and informative cases when compared with the cohort study design. Moreover, tissue collection is more feasible, especially if the study population is based in a hospital or clinic. On the other hand, this study design is particularly susceptible to misinterpretations due to reverse causation since case and control individuals often provide biased information regarding their past exposures. In addition, biological samples are limited to a single time point when the disease is fully developed, and this leads to difficulties in distinguishing whether an epigenetic alteration is a cause or a consequence of a disease.

Nested case-control studies

It is a case-control study embedded in a cohort study. All individuals who develop the disease of interest, at any time during follow-up, are selected within a well-defined cohort study. Controls are randomly selected among those who remains free of the disease during follow-up. The risk estimate is thus calculated on a small sample of the entire cohort population. Such an approach allows to obtain more information than that already available in the cohort and makes the study less time-consuming, less costly and more efficient. Given the efficiency of the nested case-control design, it is well suited to assess rare diseases and outcomes, to assess diseases that are characterized by long induction and latent periods, and it is ideal for studies in which exposure data are difficult or expensive to obtain. Moreover, since biological samples were collected at the time of subject enrolment in the original cohort, thus precede the diagnosis of disease of interest, the epigenetic state is not influenced by disease. A major limitation of the nested case-control design is that multiple outcomes cannot be investigated because controls are selected from non-cases when an event occurs.

Intervention studies

Intervention studies investigate the effect of the modification of one or more recognized or putative risk factors on an outcome of interest. Intervention studies on epigenetic alterations in humans have been focused on the effects of lifestyle modifications, such as physical exercise (16), and dietary changes in folate and polyphenols intake (7).

Estimating the sample size

Selecting the appropriate sample size is important for all experimental studies, but is especially critical for epigenetic epidemiology studies. The complexity of these experiments and the large number of targets to be evaluated raises the risk for potential false findings. In addition, if too few subjects are enrolled in a study, that design will very likely not answer the research question that has been set forth and the risk of unreliable statistical inferences increases. The classical approach to sample size estimation involves testing a hypothesis by selecting a sample size and significance threshold that controls for the type I error rate, which is the probability of incorrectly rejecting a true null hypothesis. This approach is well suited when information from other similar experiments or prior knowledge about the reliability of the measurements are available, but this may not apply to epigenetic epidemiology studies.

Microarray experiments aim at identifying differences in gene expression among several groups. In this context statistical tests determine whether a particular gene is not differentially expressed across groups (a null hypothesis). The multiple testing problem arises because, if many hypotheses are tested simultaneously, some test statistics will be significant, even if no associations exist. Multiple test procedures are designed to control the entire set of hypotheses, to prevent study conclusions being drawn by results that could be attributed to chance alone.

A standard method to reduce error rates is the family wise error rate (FWER), which is the probability of committing at least one type I error over the course of the entire study; the Bonferroni's correction is one of the simplest type of FWER. However, FWER is known to be too conservative and may fail to identify significant differences. More recently, the false discovery rate (FDR) method proposed by Benjamini and Hochberg (1995) and Storey (2002) and variations on it have gained support (20, 56).

FDR controls for the expected proportion of rejected null hypotheses that are actually true, and has the advantage of being less conservative than FWER methods and of having more power to identify significant differences. On the other hand the probability of type I errors is higher.

The overarching goal of sample size is to enroll a number of subject that is sufficient to provide adequate statistical power to detect a meaningful effect. An effect size can be expressed in various forms, depending on the nature of the outcome. This should be defined as fold changes between differentially expressed genes, mean differences between cases and controls, or odds ratios for binary responses. To determine the sample size a measure of the variability of the effect is neededе (28, 72) that should be evaluated: this is usually derived from similar previous experiments. The larger is the identified variability, the larger is the sample sizes required. Further, to be conservative, variability from genes that display the greatest variation should be selected. If sample size is limited, then variability may not be estimated reliably (17). For complex multivariate models, sample calculations may be difficult and may require ad hoc methods. Even if sample size planning procedures for nonstandard analyses have not generally been developed, a general principle of sample size planning is that sample size can be planned in any situation with an a priori Monte Carlo simulation study. This implies implementing the particular statistical technique, and repeating it a large number of times with different sample sizes until the minimum sample size is found where the particular goal is accomplished. This requires knowledge of the distributional form and population parameters, comparable to traditional analytic methods of sample size planning (33).

Biological sample selection

Since epigenetic marks are highly tissue-specific, the study of target tissues would be ideal but this may not always be possible, as prospective studies can only collect non-invasive biospecimens. Determining which biospecimen to collect is not trivial, and for some diseases, the simplest approach is to collect samples that are most proximal to the target tissue. For example, when studying bladder cancer, cells in urine sediments could be collected and when studying leukemia the best surrogate biospecimen are given by peripheral white blood cells. For many diseases, such as brain diseases, this approach is not applicable, and three alternative approaches have been proposed. The "embryo layer" approach involves collecting cells derived from the same embryo layer as the target tissue (for example, brain and buccal cells both derive from neuroectoderma). This method is well suited for studying epigenetic changes induced *in utero*. The "uniform effect" approach suggests that, even if the baseline levels of epigenetic marks vary between tissues, all tissues may be equally impacted by exposure. But, it should be considered that different tissues also have different levels of exposure due to the distinct distribution of toxicants throughout the body. The "highest dose, first target" approach involves collecting biospecimens from the first target of exposure because the exposure effect will presumably be greatest in those biospecimens (for example, nasal mucosae would be collected when studying inhalable pollutants). Nevertheless, most epidemiological studies are based on existing cohorts, in which the only available biological samples are blood or buffy coat, buccal cells, or urine. Furthermore, archived biospecimens are often a collection of different cell types, for example, blood is a heterogeneous tissue and epigenetic levels may vary as a result of variable blood cell compositions between samples. For blood, normalizing samples using blood cell counts can partially solve this problem.

Validation of epigenomic data

Given the multifaceted nature of information produced by epigenetic epidemioly investigations, it is challenging to separate robust signals from noise.

The discovery-only study is a single study without a validation step. Because a large number of associations are investigated simultaneously, discovery-only studies often produce false findings. To overcome this limitation, two study design strategies have been proposed. In the split-sample design, the study population is divided into two groups. One group is analyzed using an epigenome-wide approach, and

the second group is analyzed using candidate target analyses. In the single-study cross-validation design and subsets of the study population are assayed multiple times to evaluate result reproducibility.

In the two-stages design, where stage one is discovery and stage two is replication, the investigation is conducted as two independent studies to ensure the validity of the findings. The choice of a proper study design and validation of results are important to minimize false positive and false negative findings. Methods to validate results may be technical (on the same study subjects however using an alternative technique) or biological (on different study subjects).

The importance of data collection for individual's characterization

Given the increasing need of measuring the individual's 'exposome', it is necessary to try to collect as much data as possible to cover all exposures that might influence the onset or the development of a disease.

Many attempts have been made, aimed at evaluating all personal interaction with the environment, not only chemical and physical agents present in the air and water, but also everything that is part of lifestyle (food, alcohol, physical exercise, smoking, stress, etc.). Research groups are working to develop blood-based tests to trace any kind of substance that could give an index of personal exposure, some other groups are taking into consideration the importance of gathering mobility information and are suggesting the use of GPS or mobile phone tracking tools to account for the individual's time-location configurations (70).

While waiting for technology to give us more advanced tools, it is possible to consider some aspects of the study design that can help to collect individual exposure data properly.

Fisrst it might be useful to integrate different exposure data from air quality monitoring stations, dispersion models, land use regression models and indoor measurements. The gold standard would be the use of portable and lightweight personal sampler that record time and air pollution levels (62), although for large epidemiological studies the cost of these sampler might be an issue. Questionnaires have sometimes been considered as a weak point of the epidemiological studies because they are mostly self-reporting tools and the information they contain could be inaccurate and misleading. However questionnaires can easily gather information on workplace, home, smoking status and diet helping the implementation of information. As long as the questionnaire is validated and standardized it is possible to gather all the information that can be useful in modeling the data analysis.

Epigenetic data, as well as genetic, molecular and biomarkers data, need to be integrated too. Epigenomic approaches can generate huge amount of data per experiment that need to be interpreted per se and in association with other variables. Bioinformatic software are usually available and can support data cleaning, data visualization, preprocessing and some basic statistical analysis. Although some skills belonging to bioinformatics overlap with data management and partly with statistics, there are aspects of data analysis (such as the role of confounding variables and effect modificators) that cannot be dealt with bioinformatics softwares only. Statistical modeling is then mandatory for data integration and analysis.

Conclusions

Exposure to several toxicants has been consistently associated to alterations at the epigenetic level, including DNA methylation, miRNA expression and histone modifications. In addition, diverse prenatal environmental factors have been correlated with epigenetic programming of the individual, as well as with the risk of developing diseases later in life.

There is an increasing awareness that epidemiology studies should take into account that people are simultaneously exposed to a multiplicity of risk factors (toxicants, social and lifestyle factors). Integration of epigenetic data and environmental factors within the framework of epidemiological studies, may contribute to the evaluation of the external influences over an individual's lifetime, and provide a better understanding of the molecular pathways leading to disease outcomes.

No potential conflict of interest relevant to this article was reported

References

1. Alegria-Torres JA, Baccarelli A, Bollati V: Epigenetics and lifestyle. Epigenomics 2011; 3: 267-277
2. Arita A, Shamy MY, Chervona Y, et al: The effect of exposure to carcinogenic metals on histone tail modifications and gene expression in human subjects. J Trace Elem Med Biol 2012; 26: 174-178
3. Baccarelli A, Bollati V: Epigenetics and environmental chemicals. Curr Opin Pediatr 2009; 21: 243-251
4. Baccarelli A, Wright RO, Bollati V, et al: Rapid DNA methylation changes after exposure to traffic particles. Am J Respir Crit Care Med 2009; 179: 572-578
5. Bell JT, Spector TD: A twin approach to unraveling epigenetics. Trends Genet 2011; 27: 116-125
6. Bertazzi PA, Pesatori AC, Landi MT, Consonni D: Occupational epidemiology and new challenges in occupational medicine. Med Lav 1999; 90: 445-459
7. Bishop KS, Ferguson LR: The interaction between epigenetics, nutrition and the development of cancer. Nutrients 2015; 7: 922-947
8. Bollati V, Baccarelli A: Environmental epigenetics. Heredity (Edinb) 2010; 105-12
9. Bollati V, Baccarelli A, Hou L, et al: Changes in DNA methylation patterns in subjects exposed to low-dose benzene. Cancer Res 2007; 67: 876-880
10. Bollati V, Baccarelli A, Sartori S, et al: Epigenetic effects of shiftwork on blood DNA methylation. Chronobiol Int 2010; 27: 1093-1104
11. Byun HM, Motta V, Panni T, et al: Evolutionary age of repetitive element subfamilies and sensitivity of DNA methylation to airborne pollutants. Part Fibre Toxicol 2013; 10: 28
12. Cantone L, Nordio F, Hou L, et al: Inhalable metal-rich air particles and histone H3K4 dimethylation and H3K9 acetylation in a cross-sectional study of steel workers. Environ Health Perspect 2011; 119: 964-969
13. Chanda S, Dasgupta UB, Guhamazumder D, et al: DNA hypermethylation of promoter of gene p53 and p16 in arsenic-exposed people with and without malignancy. Toxicol Sci, 2006; 89: 431-437
14. Dashwood RH, Ho E: Dietary histone deacetylase inhibitors: from cells to mice to man. Semin Cancer Biol 2007; 17: 363-369
15. Delage B, Dashwood RH: Dietary manipulation of histone structure and function. Annu Rev Nutr 2008; 28: 347-366
16. Denham J, O'Brien BJ, Harvey JT, Charchar FJ: Genome-wide sperm DNA methylation changes after 3 months of exercise training in humans. Epigenomics 2015; 1-15
17. Dobbin K, Simon R: Sample size determination in microarray experiments for class comparison and prognostic classification. Biostatistics 2005; 6: 27-38
18. Dolinoy DC: The agouti mouse model: an epigenetic biosensor for nutritional and environmental alterations on the fetal epigenome. Nutr Rev 2008; S7-11
19. Drummond MJ, Glynn EL, Fry CS, et al: Essential amino acids increase microRNA-499, -208b, and -23a and downregulate myostatin and myocyte enhancer factor 2C mRNA expression in human skeletal muscle. J Nutr 2009; 139: 2279-2284
20. Efron B, Tibshirani R: Empirical bayes methods and false discovery rates for microarrays. Genet Epidemiol 2002; 23: 70-86
21. Herbstman JB, Tang D, Zhu D, et al: Prenatal exposure to polycyclic aromatic hydrocarbons, benzo[a]pyrene-DNA adducts, and genomic DNA methylation in cord blood. Environ Health Perspect 2012; 120: 733-738
22. Jirtle RL, Skinner MK: Environmental epigenomics and disease susceptibility. Nat Rev Genet 2007; 253-262
23. Kile ML, Baccarelli A, Hoffman E, et al: Prenatal arsenic exposure and DNA methylation in maternal and umbilical cord blood leukocytes. Environ Health Perspect 2012; 120: 1061-1066
24. Kippler M, Engström K, Mlakar SJ, et al: Sex-specific effects of early life cadmium exposure on DNA methylation and implications for birth weight. Epigenetics 2013; 8: 494-503
25. Koestler DC, Avissar-Whiting M, Houseman EA, et al: Differential DNA methylation in umbilical cord blood of infants exposed to low levels of arsenic in utero. Environ Health Perspect 2013; 121: 971-977
26. Krutzfeldt J, Poy MN, and Stoffel M, Strategies to determine the biological function of microRNAs. Nat Genet 2006; 38: S14-19
27. Liang Q, Gao X, Chen Y, et al: Cellular mechanism of the nonmonotonic dose response of bisphenol A in rat cardiac myocytes. Environ Health Perspect 2014; 122: 601-608
28. Liu P, Hwang JT: Quick calculation for sample size while controlling false discovery rate with application to microarray analysis. Bioinformatics, 2007; 23: 739-746
29. Madrigano J, Baccarelli A, Mittleman MA, et al: Aging and epigenetics: longitudinal changes in gene-specific DNA methylation. Epigenetics 2012; 7: 63-70
30. Madrigano J, Baccarelli A, Mittleman MA, et al: Prolonged exposure to particulate pollution, genes associated with glutathione pathways, and DNA methylation in a cohort of older men. Environ Health Perspect 2011; 119: 977-982

31. Marsit CJ: Influence of environmental exposure on human epigenetic regulation. J Exp Biol 2015; 218: 71-79
32. Martinez FD: Gene-environment interactions in asthma: with apologies to William of Ockham. Proc Am Thorac Soc 2007; 4: 26-31
33. Maxwell SE, Kelley K, Rausch JR: Sample size planning for statistical power and accuracy in parameter estimation. Annu Rev Psychol 2008; 59: 537-563
34. McGee SL, Fairlie E, Garnham AP, Hargreaves M: Exercise-induced histone modifications in human skeletal muscle. J Physiol 2009; 587: 5951-5958
35. Monk C, Spicer J, Champagne FA: Linking prenatal maternal adversity to developmental outcomes in infants: the role of epigenetic pathways. Dev Psychopathol 2012; 24: 1361-1376
36. Morabia A: A history of epidemiologic methods and concepts Basel (CH), Birkauser Verlag, 2004
37. Motta V, Angelici L, Nordio F, et al: Integrative Analysis of miRNA and inflammatory gene expression after acute particulate matter exposure. Toxicol Sci 2013; 132: 307-316
38. Newbold RR: Lessons learned from perinatal exposure to diethylstilbestrol.Toxicol Appl Pharmacol 2004; 199: 142-150
39. Ng JW, Barrett LM, Wong A, et al: The role of longitudinal cohort studies in epigenetic epidemiology: challenges and opportunities. Genome Biol 2012; 13: 246
40. Niculescu MD, Zeisel SH: Diet, methyl donors and DNA methylation: interactions between dietary folate, methionine and choline. J Nutr 2002; 132: 2333S-2335S
41. Nielsen CH, Larsen A, Nielsen AL: DNA methylation alterations in response to prenatal exposure of maternal cigarette smoking: A persistent epigenetic impact on health from maternal lifestyle? Arch Toxicol 2016; 90: 231-245
42. Olden K, Lin YS, Gruber D, Sonawane B: Epigenome: biosensor of cumulative exposure to chemical and nonchemical stressors related to environmental justice. Am J Public Health 2014; 104: 1816-1821
43. Perera F, Herbstman J: Prenatal environmental exposures, epigenetics, and disease. Reprod Toxicol 2011; 31: 363-373
44. Perera F, Tang WY, Herbstman J, et al: Relation of DNA methylation of 5'-CpG island of ACSL3 to transplacental exposure to airborne polycyclic aromatic hydrocarbons and childhood asthma. PLoS One 2009; 4: e4488
45. Pilsner JR, Hall MN, Liu X, et al: Influence of prenatal arsenic exposure and newborn sex on global methylation of cord blood DNA. PLoS One 2012; 7: e37147
46. Pilsner JR, Liu X, Ahsan H, et al: Folate deficiency, hyperhomocysteinemia, low urinary creatinine, and hypomethylation of leukocyte DNA are risk factors for arsenic-induced skin lesions. Environ Health Perspect 2009; 117: 254-260
47. Radom-Aizik S, Zaldivar F Jr, Oliver S, et al: Evidence for microRNA involvement in exercise-associated neutrophil gene expression changes. J Appl Physiol (1985), 2010; 109: 252-261
48. Rager JE, Bailey KA, Smeester L, et al: Prenatal arsenic exposure and the epigenome: altered microRNAs associated with innate and adaptive immune signaling in newborn cord blood. Environ Mol Mutagen 2014; 55: 196-208
49. Relton CL, Davey Smith G: Is epidemiology ready for epigenetics? Int J Epidemiol 2012; 41: 5-9
50. Rothman KJ: Epidemiology An Introduction. 2nd Edition 2012: Oxford University Press
51. Rusiecki JA, Baccarelli A, Bollati V, et al: Global DNA hypomethylation is associated with high serum-persistent organic pollutants in Greenlandic Inuit. Environ Health Perspect 2008; 116: 1547-1552
52. Schug TT, Janesick A, Blumberg B, Heindel JJ: Endocrine disrupting chemicals and disease susceptibility. J Steroid Biochem Mol Biol 2011; 127: 204-215
53. Shi L, Wu J: Epigenetic regulation in mammalian preimplantation embryo development. Reprod Biol Endocrinol 2009; 7: 59
54. Smith AH, Marshall G, Yuan Y, et al: Increased mortality from lung cancer and bronchiectasis in young adults after exposure to arsenic in utero and in early childhood. Environ Health Perspect 2006; 114: 1293-1296
55. Somm E, Stouder C, Paoloni-Giacobino A: Effect of developmental dioxin exposure on methylation and expression of specific imprinted genes in mice. Reprod Toxicol 2013; 35: 150-155
56. Storey JD, Tibshirani R: Statistical significance for genomewide studies. Proc Natl Acad Sci USA 2003; 100: 9440-9445
57. Suter MA, Anders AM, Aagaard KM: Maternal smoking as a model for environmental epigenetic changes affecting birthweight and fetal programming. Mol Hum Reprod 2013; 19: 1-6
58. Szyf M: The early life social environment and DNA methylation: DNA methylation mediating the long-term impact of social environments early in life. Epigenetics 2011; 6: 971-978
59. Takahashi K, Yokota S, Tatsumi N, et al: Cigarette smoking substantially alters plasma microRNA profiles in healthy subjects. Toxicol Appl Pharmacol 2013; 272: 154-160
60. Tarantini L, Bonzini M, Apostoli P, et al: Effects of particulate matter on genomic DNA methylation content and iNOS promoter methylation. Environ Health Perspect 2009; 117: 217-222

61. Tobi EW, Lumey LH, Talens RP, et al: DNA methylation differences after exposure to prenatal famine are common and timing- and sex-specific. Hum Mol Genet 2009; 18: 4046-4053
62. Tsai CJ, Liu CN, Hung SM, et al: Novel active personal nanoparticle sampler for the exposure assessment of nanoparticles in workplaces. Environ Sci Technol 2012; 46: 4546-4552
63. Vaiserman A: Early-life Exposure to Endocrine Disrupting Chemicals and Later-life Health Outcomes: An Epigenetic Bridge? In Aging Dis 2014: United States. p. 419-429
64. Vrijens K, Bollati V, Nawrot TS: MicroRNAs as Potential Signatures of Environmental Exposure or Effect: A Systematic Review. Environ Health Perspect 2015; 123: 399-411
65. Waterland RA, Michels KB: Epigenetic epidemiology of the developmental origins hypothesis. Annu Rev Nutr 2007; 27: 363-388
66. Wild CP: Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. Cancer Epidemiol Biomarkers Prev 2005; 14: 1847-1850
67. Wild CP: The exposome: from concept to utility. Int J Epidemiol 2012; 41: 24-32
68. Wong RL, Walker CL: Molecular pathways: environmental estrogens activate nongenomic signaling to developmentally reprogram the epigenome. Clin Cancer Res 2013; 19: 3732-3737
69. Wright RO, Schwartz J, Wright RJ, et al: Biomarkers of lead exposure and DNA methylation within retrotransposons. Environ Health Perspect 2010; 118: 790-795
70. Wu J, Jiang C, Liu Z, et al: Performances of different global positioning system devices for time-location tracking in air pollution epidemiological studies. Environ Health Insights 2010; 4: 93-108
71. Yang AS, Estecio MR, Doshi K, et al: A simple method for estimating global DNA methylation using bisulfite PCR of repetitive DNA elements. Nucleic Acids Res 2004; 32: e38
72. Yeung SH, Liu P, Del Bueno N, et al: Integrated sample cleanup-capillary electrophoresis microchip for high-performance short tandem repeat genetic analysis. Anal Chem 2009; 81: 210-217
73. Zhang FF, Cardarelli R, Carroll J, et al: Physical activity and global genomic DNA methylation in a cancer-free population. Epigenetics 2011; 6: 293-299
74. Zhou X, Sun H, Ellen TP, et al: Arsenite alters global histone H3 methylation. Carcinogenesis 2008; 29: 1831-1836

Acknowledgements: *Prof. Bollati, Dr. Iodice and Dr. Grevendonk received support from the EU Programme "Ideas" (ERC-2011-StG 282413 to V. Bollati).*