CHAPTER 27.

Future perspectives on molecular epidemiology

Martyn T. Smith, Pierre Hainaut, Frederica Perera, Paul A. Schulte, Paolo Boffetta, Stephen J. Chanock, and Nathaniel Rothman

The current status of molecular epidemiology

As witnessed in the previous chapters, since the development of molecular epidemiology in the early 1980s (1), the field has advanced so that large-scale, in-depth studies have been performed or are ongoing. These studies have not only monitored the external environment and ascertained clinical disease status, but have also collected data on biomarkers of exposure, biologically effective dose, preclinical effects, and susceptibility within population studies. Links have been drawn between various environmental and nutritional factors and diseases as diverse as childhood asthma, cardiovascular disease, cancer, developmental disorders, obesity and metabolic disorders. In some cases, educational or regulatory interventions have been mounted as a result of these studies.

As described throughout this book, non-genetic environmental factors, broadly defined to include diet, lifestyle, infections, stress, ionizing radiation, and chemical pollutants in the air, water, food supply and workplace, are important contributors to chronic disease. Adverse gene–environment interactions (GxE) probably influence most chronic diseases, including neurological disorders and cancer. The genetic (G) contribution to different diseases varies, but several lines of evidence, including classic studies of migrant populations in which the genetics remain essentially the same but the incidence of disease changes because of the new environment, clearly show that non-genetic factors have high attributable risks (2). For some diseases, incidence rates increase or decrease dramatically within the first or second generation of immigrants, with disease patterns becoming more similar to the adoptive country and less similar to the country of origin. This highlights the fact that environmental factors (E) can contribute to a large portion of at least some chronic diseases (3,4).

Genomic tools arising from the Human Genome Project combined with bioinformatics have allowed scientists to begin to examine the

genetic component of many chronic diseases. Initially, variations in candidate genes were examined in great detail, most notably in xenobiotic metabolizing and DNA repair genes (5–7). More recently, genome-wide association studies (GWAS) have been increasing in number and scope and have provided important insights into the roles that particular genes, gene regions, regulatory elements, and other parts of the genome with function yet to be defined play in disease development (8). Thus, a key focus of most current molecular epidemiology studies is on the genome and genetic variation. The reason for this focus on genetics, even though the environment may be more critical, is simply that we have extremely precise, accurate, and global tools to examine the genome, either measured as external factors or biologically as reflected by the "exposome". Such tools are not available to examine the environment. At the same time, by examining the G component of GxE we may find clues as to where to look for E factors (9,10), although success in this regard is very limited to date.

The most productive approach to assessing the environmental contribution to disease may be to examine environmental exposures agnostically (11). Unfortunately, compared to genomics, the tools for assessment of exposures, based upon measurements of chemicals in air, water, food and the human body, have undergone a more gradual evolution in the past 30 years and have not experienced the same exponential gains. This is due to both lack of technological progress in the tools available for exposure assessment, as well as the more challenging task of obtaining data on, or estimates of, non-fixed exogenous and endogenous individual exposures. These exposures can vary day-to-day as well as over time, as individuals age and secular changes occur in a given population. The use of questionnaires has been the core approach for assessing exposure in studies of chronic diseases in the general population that arise, in part, from exposure patterns present over many years. This approach relies on self-reports, which can be imprecise and inaccurate. However, they have been successfully used to identify consistent patterns of chronic disease risk for several exposures such as tobacco, alcohol, obesity, components of the reproductive history, air pollution, and some aspects of diet (e.g. intake of cruciferous vegetables). Also, the increased ability to obtain objective occupational and residential histories from study subjects, linked by sophisticated methods to comprehensive exposure databases, has allowed advances in identifying associations between certain chemical exposures and disease risk. At the same time, methods to measure chemicals in biologic samples have steadily evolved to measure a wider array of compounds in smaller amounts of samples. Nevertheless, these advances are not comparable to the quantum leap that has occurred in genomics.

The Human Genome Project and at least 20 years of investment in genetics are very helpful to molecular epidemiologists in understanding the genetic determinants of diseases, but we remain much more limited when it comes to quantifying human exposures. This disparity in current knowledge between genetic contributions and environmental exposures was recognized by Wild, who defined the exposome, representing all environmental exposures and lifestyle factors from conception onwards, as a quantity of critical interest to disease etiology (12). If we expect to have any success at identifying the effects of G, E and GxE on chronic diseases, we must develop 21st century tools to measure exposure levels in large human populations (11). That is, we need to quantify the exposome, a topic we will return to later.

Many lifestyle factors such as exercise levels, dietary choices and stress levels also contribute to the environmental component of disease, but are hard to quantify retrospectively and prospectively. Modern tools to capture, store and use information about physical activity, diet and stress levels are needed for epidemiological studies. Such tools are being developed. For example, it soon may be possible to perform population-scale, longitudinal measurement of physical activity using common cell phones that include internal accelerometers and low-power wireless communication capabilities (13). Dietary assessment methods suitable for use in large epidemiologic studies (e.g. dietary recall, food diaries and food frequency questionnaires) are subject to considerable inaccuracy. More accurate methods (e.g. metabolic ward studies and doubly-labelled water) are prohibitively costly and/or labour-intensive for use in population-based studies. Several research groups are assessing methods that use cell phones to capture both voice recordings and photographs of dietary intake in real time that, along with computerized analysis, may revolutionise nutritional epidemiology studies (14).

Accumulating evidence is also consistent with the role of psychosocial stress in moderating the effects of genetic and other environmental factors on health outcomes. Further advances in this area will require the development

of standardized, psychometrically sound instruments for quantifying exposures to psychosocial stress. Again, progress is being made in this area using, for example, colorimetric test strips to rapidly detect and quantify salivary α-amylase, a biomarker of the body's adrenergic stress response (15,16). The measurement of telomere length is another stress biomarker that is gaining acceptance (17,18).

The potential contributions of genomics to molecular epidemiology

Genomics is the study of all of the nucleotide sequences, including structural genes, regulatory sequences, and noncoding DNA segments, in the chromosomes of an organism. Because of the tools provided by the Human Genome Project, the current focus of many molecular epidemiological studies is on genomic variation. GWAS and, most recently, examinations of copy number variation have revealed many surprising insights. Results from these studies show that many common causal variants, each of small, additive effect, probably contribute to complex disease risk.

As of 2010, GWAS had identified over 750 regions in the genome strongly associated with more than 125 traits and diseases (http://www.genome.gov/gwastudies) (19). In chronic complex diseases, such as type 2 diabetes and Crohn's disease, over 40 genetic regions have been associated with each disease. For certain heritable traits such as height, recent studies have identified several hundred regions, each of which contributes to their heredity (20). In cancer alone there are over 135 regions associated with 21 cancers (21). However, the early GWAS have not sufficiently explained the heredity of any given common disease. This is not surprising since GWAS interrogate only common variants, which represent only a proportion of genetic variation in the human genome. For example, despite the addition of 10 positively-associated SNPs, the performance of breast cancer risk models only improved modestly; the area under the curve of the receiver operating curve increased from 58% to a mere 61.8% (22). Thus far, GWAS have been most successful in identifying regions that harbour genetic variants that are directly associated with risk for a complex disease, such as cancer or inflammatory bowel disease. For the latter, GWAS have pointed towards a region on chromosome 2q37.1 and identified a novel mechanism of autophagy previously not well described in the pathogenesis of inflammatory bowel disease, specially as it relates to the genes in this pathway (23). Fine mapping of regions together with functional work is required to elucidate the biological underpinnings of the direct association of common variants with complex diseases such as cancer (24). Certainly, the advent of new technologies, in conjunction with computational resources, will enable investigators to use next generation sequencing to explore the contribution of uncommon and rare variants in the near future.

The potential contributions of molecular epidemiology in the near future

In the near future, there will also likely be a maturing of omic applications and the incorporation of systems biology into molecular epidemiology, which will produce what some have called systems epidemiology (25). Studies of epigenetic changes are already coming to the forefront of molecular epidemiology, and studies of changes in DNA methylation, histone modifications and microRNA (miRNA) expression, both in cells and body fluids, have recently been published (26–28). ChIP-on-Chip (chromatin immunoprecipitation with microarray analysis) and ChIP-seq (chromatin immunoprecipitation with sequencing) will help in understanding epigenetic effects on gene–protein interactions. Advances in mass spectrometry will soon make it possible to measure post-translation modifications of proteins such as histones in small volumes of biological sample, adding to our repertoire of epigenetic changes that can be studied in human populations.

Advances in mass spectrometry and in laboratory-on-a-chip devices that use nanotechnology may also soon permit us to profile all the major protein and DNA adducts in humans using adductomics. This will allow for the examination of multiple biomarkers in very small sample volumes, such as a few microlitres of serum, a drop of blood, or a dried blood spot.

These tools are expected to have great application in molecular epidemiology studies in the near future. There are emerging opportunities to apply these technologies in molecular epidemiologic studies with banked biological samples, including cross-sectional, case–control, and, in particular, prospective cohort studies, to study a wide range of diseases.

This should advance the ability of molecular epidemiology to more broadly explore exposure–disease relationships, to study effect modifiers, and to obtain insight into the fundamental underlying pathogenesis of these conditions. Further, beyond providing etiologic insights, it is expected that molecular epidemiology will be

newly positioned to make important contributions to translating these findings into primary, secondary and tertiary prevention strategies. This would begin with broad public health practices that could include removal or substantial reduction of exposure to hazardous environmental compounds, making available healthier food in schools and better education on lifestyle risk factors.

At the same time, molecular epidemiology is likely to play an important role in the upcoming revolution in personalized medicine (29). At present, identifying individual genetic risk is at the forefront of this personalization of health care. But given the limited role for genetics in comparison to the environment in causing disease, the focus must eventually shift to include individual environmental risk factors, again broadly defined to include toxic exposures, lifestyle, diet, drugs, etc. This could help bring about not only lifestyle modification to prevent disease and improve drug treatment, but it could also help individuals gain an understanding of their prior and current chemical exposures and other risk factors, leading to personalized risk assessment. Molecular epidemiologists may be able to identify not only broad subgroups of the population with a higher probability of developing disease given genetic and other risk factors, but also move to further develop predictive models that can be applied to individuals by preventive and clinical medicine practitioners. An example of this is the Gail Model for predicting breast cancer, which is based on all known risk factors including *BRCA1* and *BRCA2* mutations (30). Genetics is now poised to augment this model and provide even greater sensitivity and specificity, but as mentioned previously, success to date using GWAS data is limited.

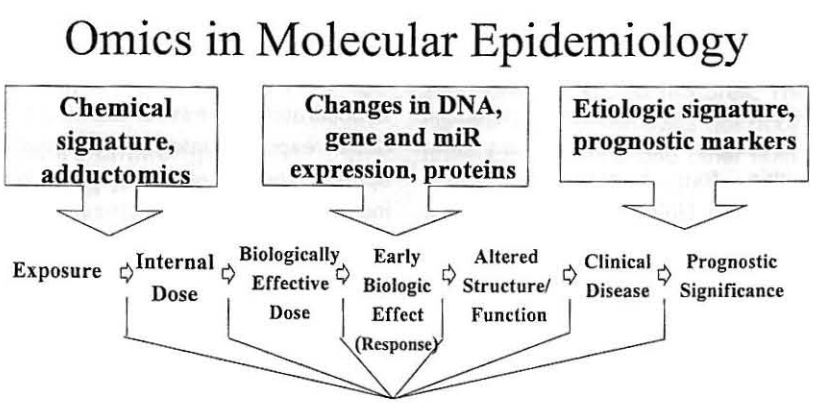
Additional profoundly important steps taking place in molecular epidemiology are the increased size of studies and the formation of dozens of international consortia, including those that focus on specific diseases as well as those that are based on study design (e.g. various cohort consortia). There are now a large number of prospective cohort studies in North America, Europe, Asia and Australia that have enrolled or are continuing to enrol several million study subjects. These cohort studies have millions of samples of DNA, serum, blood cells, and other biological material stored at low temperatures. Some studies are tracking individuals *in utero* through adolescence, providing an opportunity to assess the earliest determinants of disease. These samples are precious as well as numerous. Efficient, high-throughput methods that work on minute amounts of sample are needed to analyse nested case–control or case–cohort studies carried out within them. The combination of new nanotechnology/laboratory-on-a-chip methodologies with large prospective cohort studies holds great promise for new research findings. At the same time, there will still be a need for focused, hypothesis-testing studies carried out within these cohorts, in addition to the application of discovery technologies. Such studies can often be carried out on smaller sample sizes, as they generally do not need to contend with the low prior hypothesis/multiple testing problem.

In addition, there is still a need for focused, cross-sectional studies of well-defined populations with particular exposure or lifestyle patterns of interest, such as new exposures recently introduced into the environment, high or low levels of exposures, etc. (e.g. (31,32)). These studies can use extensive and complicated protocols as sample sizes are generally small (a few hundred subjects). Often they can have very detailed assessment of the target exposure, evaluate potential confounders, modifiers, and other contributors to endpoints under study in greater detail (e.g. nutritional, genetic, psychosocial factors) (33), and arrange for samples to be transported and processed very quickly, allowing specialty assays to be carried out. They can incorporate both state-of-the-art omics platforms as well as in-depth hypothesis-testing (34).

Future use of omic technologies in molecular epidemiology

The field of molecular epidemiology is entering an exciting new phase in which the innovative tools of omics, such as microarrays and metabolic and peptide profiling, are being applied along with novel laboratory-on-a-chip microdevices that can act as biosensors of everything from glucose levels to protein adducts (Figure 27.1). The term omics has come to mean any field of study in biology in which the totality of something is studied, beginning with genomics which surveys across the genome. The tools of genomics, developed as a consequence of the Human Genome Project, include microarrays allowing the examination of gene variation and expression and high-throughput sequencing. The latter is now being used not only to sequence DNA but the RNA transcriptome to give a more complete picture of gene and siRNA expression. Transcriptomics is the study of all forms of RNA that are transcribed from the DNA and includes mRNA and miRNA expression.

Figure 27.1



Recent and near-future contributions of transcriptomics

Distinctive blood transcriptional profiles have been demonstrated for over 35 human diseases (35). As more data become available on global gene expression in the blood of humans following exposure, it will become easier to identify molecular mechanisms by which environmental chemicals promote/cause human disease. Initiatives such as the Comparative Toxicogenomics Database (http://ctd.mdibl.org/) (36) have been developed towards this goal.

A broad array of environmental exposures including pharmaceuticals, pesticides, air pollutants, industrial chemicals, heavy metals, hormones, nutrition and behaviour can change gene expression through several gene regulatory mechanisms (37). The potential of toxicogenomics in the discovery of biomarkers of complex environmental exposures was illustrated by a study in which gene expression profiling of leukocytes was shown to distinguish individuals exposed to cigarette smoke (CS) from unexposed individuals (38). An association between CS-induced gene expression and DNA adduct formation was later shown in a study of monozygotic twin pairs (39). The impact of air pollution on children at the transcriptional level in blood cells was investigated by comparing children from urban and rural regions of the Czech Republic (40). Several genes were differentially expressed and a correlation with micronuclei frequencies was shown. Further, the effects on children and adults at the transcriptional level differed (41). A small study of children in New York City found that a gene-specific methylation change in umbilical cord white blood cell DNA was associated both with prenatal exposure to PAH air pollutants and with reported asthma in the children by age 5 years (42).

More recently, many groups have begun profiling miRNA expression. A role for miRNAs in mediating the response to environmental exposures has been demonstrated by a study showing that smoking induces gene expression changes in the human airway epithelium (43) with some genes modulated by miRNA (44). Expression profiling analyses have revealed characteristic miRNA signatures in certain human cancers (45) and other diseases. The study of miRNA in molecular epidemiology will likely explode in the near future as new tools become available and the biology is better understood.

Applications of proteomics

While toxicogenomics studies using global transcriptional analysis have enormous potential, the transcriptome does not always reflect the functional proteome. Further, proteins may be subject to post-translational modification and translocation. However, proteomics, the analysis of the total protein output encoded by the genome using techniques such as mass spectrometry and antibody arrays, is more challenging and less amenable to application in a high-throughput capacity due to differences in protein properties, location and abundance. Recently, a multilaboratory study has attempted to dispel some of the notions of the irreproducibility of mass spectrometry-based proteomics by pinpointing where the methodological problems are and where challenges remain (46). By addressing these methodological issues researchers hope to bring proteomics to the forefront of biomarker research.

Applications of metabolomics

Metabolomics is defined as the study of metabolic profiles in easily collected biological samples such as urine, saliva or plasma. The metabolome is highly variable and time-dependent, and consists of several thousand chemical structures. Since it is sensitive to

age, gut microbial composition, and lifestyle, metabolomics is ideal for the characterization of dietary and therapeutic interventions, metabolism and metabolism-related disorders (47). While successfully established in the screening of inborn errors in neonates, metabolomics is being increasingly applied to several diseases. For many years specific metabolites have been measured in body fluids to diagnose particular diseases such as diabetes, by measurement of glucose, and vascular diseases, by determination of cholesterol. Metabolomics, with its ever-increasing coverage of endogenous compounds and its high-throughput capacity, now provides a much more comprehensive assessment of health status and can be used in the identification, qualification, and development of biomarkers.

An important challenge in metabolomics is the acquisition of qualitative and quantitative information concerning the metabolites that occur under normal circumstances to be able to detect perturbations in the complement of metabolites as a result of changes in environmental factors. Technologies that rely on UPLC-MS/MS, FT-ICR-MS, Orbitrap, and asymmetric waveform ion mobility analysers are emerging as dominant analytical methods for metabolomic studies because of the accuracy, high throughput and coverage (>1000 unique metabolites) that can be achieved (48). However, even though these methods provide accurate mass values that may reduce the number of potential molecular formulas down to a few candidates, further development is needed to provide complete structural information. The exchange of chemical and analytical information must be encouraged for metabolomics to expand.

Importantly for epidemiologists, metabolomics is relatively easy to apply in large-scale human studies. For example, a large-scale exploratory analytical approach investigated metabolic phenotype variation across and within four human populations using 1H NMR spectroscopy (49). Metabolites discriminating across populations were then linked to data for individuals on blood pressure. Spectra were analysed from two 24-hour urine specimens for each of 4630 participants from the INTERMAP epidemiological study, which involved 17 population samples in China, Japan, the United Kingdom, and the USA. It was shown that urinary metabolite excretion patterns for East Asian and western population samples, with contrasting diets, diet-related major risk factors, and coronary heart disease/stroke rates, were significantly differentiated ($P < 10(-16)$). Among discriminatory metabolites, four were quantified and showed associations with blood pressure.

Potential impact of molecular epidemiology on public health and regulatory policy

A bioinformatics database could be built of the human response to different chemical exposures and associated chronic diseases. This database may well be useful in many ways for risk assessment. For example, by comparing the molecular effects of newly tested chemicals to those of established carcinogens, we could identify potential carcinogens (hazard identification) and establish modes of action by studying the effects of the same chemicals in experimental animals and on human cells *in vitro*. This would allow for better prediction of human carcinogenicity and assessment of carcinogenic mechanisms (50). Given the sensitivity of omic analyses, low-dose adverse effects can also be observed and distinguished from high-dose phenomena, if exposure is accurately assessed, allowing for dose–response data from molecular epidemiology studies to be incorporated into risk assessments.

For additional public health impact, molecular epidemiology must continue to expand its contributions to surveillance, mechanistic research, efficacy trials, translational research and health policy. We must assemble and communicate information to decision-makers, medical and health professionals, and the public. If molecular epidemiology is to make a major impact on population health, it must be preventive and have a global as well as a local focus. A life-course approach is also important in establishing the earliest causes of diseases both in children and adults. We must expand our horizons to develop affordable population-wide tools for combating common diseases.

Serving as the linking hub for laboratory and population, problem and solution, molecular epidemiology can help translate research to practice. To do this, there will be a need to continue current trends in the discipline and establish new ones. Continuation of the trend towards large-scale consortia and biobanks, use of bioinformatics, and attention to individual and collective ethical issues will serve to move the field forward, as will in-depth hypothesis-driven studies of at-risk populations. Powerful impacts will be achieved by incorporating epigenetic and biological systems theory in research and by expanding skill sets and professional knowledge to complement translation research and risk communication and to foster public health perspectives. A broad population-wide vision for

using biological markers is required to leverage the power of molecular scale insight to give beneficial macro-scale impacts on public health.

Future challenges: Dealing with complexity and lack of resources

A major challenge to many of the novel approaches described above is the size and complexity of the data generated. Currently, it is a major biostatistical undertaking to analyse terabytes of data, and the emerging results require extensive further analysis by bioinformatics. Efforts must be made to simplify the analysis and reduce the data. New statistical approaches and computer programs are urgently needed to assist in the analysis.

Exposure assessment must also be able to address low-level exposure to complex mixtures. The current cost of analysis for most chemicals in blood and other fluids is prohibitive if one wishes to assess multiple compounds. New analytical chemical approaches are needed to assess the thousands of chemicals and their metabolites to which we are exposed.

One method to overcome resource difficulties may be to pool samples. Recently, this approach has been used with considerable success in GWAS and in studies of the plasma proteome (51,52).

Conclusion

Molecular epidemiology is poised to make ever-greater contributions to understanding the genetic and environmental causes of human disease. Both agnostic and hypothesis-driven approaches to both categories of risk factors could lead to leaps in our understanding. Investment in new methods and approaches will be needed, however. Strong links between population scientists, bench scientists, bioinformaticians and engineers must also be forged if progress is to be made.

Disclaimer: The findings and conclusions in this chapter are those of the author and do not necessarily represent the views of the Centers for Disease Control and Prevention.

References

1. Perera FP, Weinstein IB (1982). Molecular epidemiology and carcinogen-DNA adduct detection: new approaches to studies of human cancer causation. *J Chronic Dis*, 35:581–600.doi:10.1016/0021-9681(82)90078-9 PMID:6282919

2. Willett WC (2002). Balancing life-style and genomics research for disease prevention. *Science*, 296:695–698.doi:10.1126/science.1071055 PMID:11976443

3. Muir CS (1996). Epidemiology of cancer in ethnic groups. *Br J Cancer Suppl*, 29:S12–S16. PMID:8782793

4. Sasco AJ (2003). Breast cancer and the environment. *Horm Res*, 60 Suppl 3;50.doi:10.1159/000074500 PMID.14671396

5. García-Closas M, Malats N, Real FX *et al.* (2007). Large-scale evaluation of candidate genes identifies associations between VEGF polymorphisms and bladder cancer risk. *PLoS Genet*, 3:e29.doi:10.1371/journal.pgen.0030029 PMID:17319747

6. Lan Q, Zhang L, Shen M *et al.* (2009). Large-scale evaluation of candidate genes identifies associations between DNA repair and genomic maintenance and development of benzene hematotoxicity. *Carcinogenesis*, 30:50–58.doi:10.1093/carcin/bgn249 PMID:18978339

7. Wang SS, Purdue MP, Cerhan JR *et al.* (2009). Common gene variants in the tumor necrosis factor (TNF) and TNF receptor superfamilies and NF-kB transcription factors and non-Hodgkin lymphoma risk. *PLoS One*, 4:e5360.doi:10.1371/journal.pone.0005360 PMID:19390683

8. Khoury MJ, Bertram L, Boffetta P *et al.* (2009). Genome-wide association studies, field synopses, and the development of the knowledge base on genetic variation and human diseases. *Am J Epidemiol*, 170:269–279.doi:10.1093/aje/kwp119 PMID:19498075

9. Rothman N, Wacholder S, Caporaso NE *et al.* (2001). The use of common genetic polymorphisms to enhance the epidemiologic study of environmental carcinogens. *Biochim Biophys Acta*, 1471:C1–C10. PMID:11342183

10. Vineis P, Marinelli D, Autrup H *et al.* (2001). Current smoking, occupation, N-acetyltransferase-2 and bladder cancer: a pooled analysis of genotype-based studies. *Cancer Epidemiol Biomarkers Prev*, 10:1249–1252. PMID:11751441

11. Rappaport SM, Smith MT (2010). Epidemiology. Environment and disease risks. *Science*, 330:460–461.doi:10.1126/science.1192603 PMID:20966241

12. Wild CP (2005). Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiol Biomarkers Prev*, 14:1847–1850.doi:10.1158/1055-9965.EPI-05-0456 PMID:16103423

13. Bexelius C, Löf M, Sandin S *et al.* (2010). Measures of physical activity using cell phones: validation using criterion methods. *J Med Internet Res*, 12:e2. PMID:20118036

14. Arab L, Winter A (2010). Automated camera-phone experience with the frequency of imaging necessary to capture diet. *J Am Diet Assoc*, 110:1238–1241.doi:10.1016/j.jada.2010.05.010 PMID:20656101

15. Groer M, Murphy R, Bunnell W *et al.* (2010). Salivary measures of stress and immunity in police officers engaged in simulated critical incident scenarios. *J Occup Environ Med*, 52:595–602.doi:10.1097/JOM.0b013e3181e129da PMID:20523239

16. Klein LC, Bennett JM, Whetzel CA *et al.* (2010). Caffeine and stress alter salivary alpha-amylase activity in young men. *Hum Psychopharmacol*, 25:359–367.doi:10.1002/hup.1126 PMID:20589924

17. Epel ES, Blackburn EH, Lin J *et al.* (2004). Accelerated telomere shortening in response to life stress. *Proc Natl Acad Sci USA*, 101: 17312–17315.doi:10.1073/pnas.0407162101 PMID:15574496

18. Puterman E, Lin J, Blackburn E *et al.* (2010). The power of exercise: buffering the effect of chronic stress on telomere length. *PLoS One*, 5:e10837.doi:10.1371/journal.pone .0010837 PMID:20520771

19. Hindorff LA, Sethupathy P, Junkins HA *et al.* (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA*, 106:9362–9367.doi:10.1073/ pnas.0903103106 PMID:19474294

20. Lango Allen H, Estrada K, Lettre G *et al.* (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature*, 467:832–838. doi:10.1038/nature09410 PMID:20881960

21. Chung CC, Magalhaes WCS, Gonzalez-Bosquet J, Chanock SJ (2010). Genome-wide association studies in cancer–current and future directions. *Carcinogenesis*, 31:111–120. doi:10.1093/carcin/bgp273 PMID:19906782

22. Wacholder S, Hartge P, Prentice R *et al.* (2010). Performance of common genetic variants in breast-cancer risk models. *N Engl J Med*, 362:986–993.doi:10.1056/NEJMoa09 07727 PMID:20237344

23. Rioux JD, Xavier RJ, Taylor KD *et al.* (2007). Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nat Genet*, 39:596–604.doi:10. 1038/ng2032 PMID:17435756

24. Donnelly P (2008) Progress and challenges in genome-wide association studies in humans. *Nature*, 456:728–731.doi: 10.1038/nature07631 PMID:19079049

25. Lund E, Dumeaux V (2008). Systems epidemiology in cancer. *Cancer Epidemiol Biomarkers Prev*, 17:2954–2957.doi:10.1158/ 1055-9965.EPI-08-0519 PMID:18990736

26. Perera F, Herbstman J (2011). Prenatal environmental exposures, epigenetics, and disease. *Reprod Toxicol*, 31:363–373 doi: 10.1016/j.reprotox.2010.12.055 PMID:21256208

27. Smeester L, Rager JE, Bailey KA *et al.* (2011). Epigenetic changes in individuals with arsenicosis. *Chem Res Toxicol*, 24:165–167. doi:10.1021/tx1004419 PMID:21291286

28. Ren X, McHale CM, Skibola CF *et al.* (2011). An emerging role for epigenetic dysregulation in arsenic toxicity and carcinogenesis. *Environ Health Perspect*, 119:11–19.doi:10.1289/ ehp.1002114 PMID:20682481

29. Aebersold R, Auffray C, Baney E *et al.* (2009). Report on EU-USA workshop: how systems biology can advance cancer research (27 October 2008). *Mol Oncol*, 3:9–17.doi:10.1016/j.molonc.2008.11.003 PMID:19383362

30. Gail MH (2009). Value of adding single-nucleotide polymorphism genotypes to a breast cancer risk model. *J Natl Cancer Inst*, 101:959–963.doi:10.1093/jnci/djp130 PMID: 19535781

31. Lan Q, Zhang L, Li G *et al.* (2004). Hematotoxicity in workers exposed to low levels of benzene. *Science*, 306:1774–1776. doi:10.1126/science.1102443 PMID:15576619

32. Turner PC, Collinson AC, Cheung YB *et al.* (2007). Aflatoxin exposure in utero causes growth faltering in Gambian infants. *Int J Epidemiol*, 36:1119–1125.doi:10.1093/ije/ dym122 PMID:17576701

33. Vineis P, Perera F (2007). Molecular epidemiology and biomarkers in etiologic cancer research: the new in light of the old. *Cancer Epidemiol Biomarkers Prev*, 16:1954–1965.doi:10.1158/1055-9965.EPI-07-0457 PMID:17932342

34. Zhang L, McHale CM, Rothman N *et al.* (2010). Systems biology of human benzene exposure. *Chem Biol Interact*, 184:86–93.doi:10.1016/j.cbi.20 09.12.011 PMID:20026094

35. Mohr S, Liew CC (2007). The peripheral-blood transcriptome: new insights into disease and risk assessment. *Trends Mol Med*, 13:422–432.doi:10.1016/j.molmed.2007. 08.003 PMID:17919976

36. Mattingly CJ, Rosenstein MC, Colby GT *et al.* (2006). The Comparative Toxicogenomics Database (CTD): a resource for comparative toxicological studies. *J Exp Zool A Comp Exp Biol*, 305:689–692 doi:10.1002/jez.a.307 PMID:16902965

37. Edwards TM, Myers JP (2008). Environmental exposures and gene regulation in disease etiology. *Cien Saude Colet*, 13:269–281.doi:10.1590/S1413-81232008000100030 PMID:18813540

38. Lampe JW, Stepaniants SB, Mao M *et al.* (2004). Signatures of environmental exposures using peripheral leukocyte gene expression: tobacco smoke. *Cancer Epidemiol Biomarkers Prev*, 13:445–453. PMID:15006922

39. van Leeuwen DM, Gottschalk RW, van Herwijnen MH *et al.* (2005). Differential gene expression in human peripheral blood mononuclear cells induced by cigarette smoke and its constituents. *Toxicol Sci*, 86:200–210. doi:10.1093/toxsci/kfi168 PMID:15829617

40. van Leeuwen DM, van Herwijnen MH, Pedersen M *et al.* (2006). Genome-wide differential gene expression in children exposed to air pollution in the Czech Republic. *Mutat Res*, 600:12–22. PMID:16814814

41. van Leeuwen DM, Pedersen M, Hendriksen PJ *et al.* (2008). Genomic analysis suggests higher susceptibility of children to air pollution. *Carcinogenesis*, 29:977–983. doi:10.1093/carcin/bgn065 PMID:18332047

42. Perera F, Tang WY, Herbstman J *et al.* (2009). Relation of DNA methylation of 5'-CpG island of ACSL3 to transplacental exposure to airborne polycyclic aromatic hydrocarbons and childhood asthma. *PLoS One*, 4:e44 88.doi:10.1371/journal.pone.0004488 PMID: 19221603

43. Sridhar S, Schembri F, Zeskind J *et al.* (2008). Smoking-induced gene expression changes in the bronchial airway are reflected in nasal and buccal epithelium. *BMC Genomics*, 9:259.doi:10.1186/1471-2164-9-259 PMID:18513428

44. Schembri F, Sridhar S, Perdomo C *et al.* (2009). MicroRNAs as modulators of smoking-induced gene expression changes in human airway epithelium. *Proc Natl Acad Sci USA*, 106:2319–2324.doi:10.1073/pnas.0806 383106 PMID:19168627

45. Calin GA, Croce CM (2006). MicroRNA signatures in human cancers. *Nat Rev Cancer*, 6:857–866.doi:10.1038/nrc1997PMID:170609 45

46. Aebersold R (2009). A stress test for mass spectrometry-based proteomics. *Nat Methods*, 6:423–430 doi:10.1038/nmeth.f.25 5. PMID:19448641

47. Biclash M, Ebbels TM, Chan Q *et al.* (2010). Opening up the "Black Box": metabolic phenotyping and metabolome-wide association studies in epidemiology. *J Clin Epidemiol*, 63:970–979.doi:10.1016/j. jclinepi.2009.10.001 PMID:20056386

48. Want EJ, Wilson ID, Gika H *et al.* (2010). Global metabolic profiling procedures for urine using UPLC-MS. *Nat Protoc*, 5:1005–1018. doi:10.1038/nprot.2010.50 PMID:20448546

49. Holmes E, Loo RL, Stamler J *et al.* (2008). Human metabolic phenotype diversity and its association with diet and blood pressure. *Nature*, 453:396–400.doi:10.1038/ nature06882 PMID:18425110

50. Guyton KZ, Kyle AD, Aubrecht J *et al.* (2009). Improving prediction of chemical carcinogenicity by considering multiple mechanisms and applying toxicogenomic approaches. *Mutat Res*, 681:230–240. doi:10.1016/j.mrrev.2008.10.001 PMID:19010444

51. Skibola CF, Bracci PM, Halperin E *et al.* (2009). Genetic variants at 6p21.33 are associated with susceptibility to follicular lymphoma. *Nat Genet*, 41:873–875.doi:10.10 38/ng.419 PMID:19620980

52. Prentice RL, Paczesny S, Aragaki A *et al.* (2010). Novel proteins associated with risk for coronary heart disease or stroke among postmenopausal women identified by in-depth plasma proteome profiling. *Genome Med*, 2:48.doi:10.1186/gm169 PMID:20667078