Int J Mol Epidemiol Genet 2012;3(2):107-114
www.ijmeg.org /ISSN1948-1756/IJMEG1111008

Review Article

Transcriptional output in a prospective design conditionally on follow-up and exposure: the multistage model of cancer

Eiliv Lund, Sandra Plancade

Institute of Community Medicine, University of Tromsø, 9037 Tromsø, Norway

Received November 25, 2011; accepted April 21, 2012; Epub May 10, 2012; Published May 30, 2012

Abstract: Transcriptomics as the analysis of mRNA and microRNA could be implemented in prospective studies both in peripheral blood and tissues. Its application in cancer epidemiology could provide a new understanding of the functional changes underlying the multistage model of carcinogenesis, as well as the relationship between these changes and exposure to carcinogens. Transcriptomics is not merely another –omics technology for risk assessment in traditional prospective studies. Instead, this novel approach has the potential to estimate the distribution of gene expression conditionally on different exposures, and to study the length of the different stages of carcinogenesis. If it proves to be a valid approach, transcriptomics could be an opportunity to make meaningful advances in our understanding of the carcinogenic process.

Keywords: Carcinogenesis, systems epidemiology, transcriptomics, prospective study, latent variable, multistage model

Introduction

There are now new opportunities to explore the dynamics of the multistage model of carcinogenesis in prospective cohort studies through functional studies of transcriptomics (the analysis of gene expression through mRNA and non-coding microRNA in blood and tissue) [1, 2]. Indeed, functional studies of transcriptomics could be used to identify genes that are responsible for the different stages of carcinogenesis through changes in their transcriptional output, In epidemiology such functional studies based on transcriptomics will operate with the underlying assumption or general hypothesis that the carcinogenic process can be detected in peripheral blood either due to communication from the cancer cells in the target tissue or simply as by-products of the intracellular process [3]. This view is gaining momentum both from *in-vitro* and *in-vivo* research, and the recent description of exosomes is just one confirmation [4]. The use of transcriptomics in tumour tissue is already well established. The application of transcriptomics in prospective studies could be considered in parallel to proposed *in-vitro* post-genome-wide association studies (GWAS) [5].

So far, GWAS performed in large epidemiological cohorts have identified more than 200, mostly new, common low-penetrance cancer-risk loci (single nucleotide polymorphisms, SNPs), with predicted odds ratios of less than 2. Thus the hypothesis emerged that trait-associated alleles exert their effects by influencing transcriptional output. Therefore it has been recommended that post-GWAS research focus on the functional characterisation of cancer-risk loci through *in-vitro* experiments. To complement this approach, we previously proposed to explore the potential of functional analyses of transcriptomics as part of systems epidemiology [6], through an extended globolomic design of the prospective study [7].

The aim of this article is to explore the use of transcriptional outputs, applied to a prospective -based design, to estimate the length of the different stages of carcinogenesis, and the related changes in the expression of genes that

play a role in human carcinogenesis, conditionally on exposure to carcinogens. This novel analytical strategy differs from current survival analyses of GWAS or gene-environment studies. Indeed, GWAS and gene-environment studies are used to estimate the relative risk of gene profiles based on the Cox model, which assumes proportional hazard rates throughout a given follow-up time, outcomes with non-parametric distribution, and no relationship to any biological model.

Testable models of carcinogenesis

There is no single model of carcinogenesis in basic research or epidemiology that can be tested directly in a prospective study design. Consequently, the analytical strategy commonly used for GWAS studies, i.e., a gene-specific or pathway analysis, and an agnostic search through all genes [8] taking into account the false discovery rate [9], could also be applied to the search for functional changes related to carcinogenesis.

In basic research several models of carcinogenesis have been proposed. Important concepts include the hallmarks of cancer that cover important mutations or functional changes related to major areas of cell-cycle control [10]. However, although in basic research one can test single genes or pathways, there is no information on time dependency, the number of necessary mutations and functional changes, or their sequence, even if some functions could be assumed to occur at later stages in the carcinogenic process. The driver and passenger mutation hypothesis depends on the observation of a large amount of mutations in tumour tissue [11], which could be a consequence of some driver mutations, i.e., mutations necessary for the carcinogenic process, with other passenger mutations simply following along. However, so far no mutations have been classified as either driver or passenger mutations. Another model of carcinogenesis is the division of the carcinogenic process into initiator effects, usually considered to be mutations, and promoter effects, which are related to later stages of carcinogenesis and are more functional in nature.

In cancer epidemiology the estimation of the carcinogenic multistage model is more than 50 years old and has at least five different variants [12]. The usefulness of these mathematical models is hampered by a lack of biological and observational data, which is needed to estimate the parameters in the model [13]. Currently, estimations are still based on non-identifiable solutions of the multistage model equations [14] by arbitrarily assigning a set value to one parameter in the mathematical model. Originally the multistage model was based on incidence figures for different cancer sites, and consisted of five or six stages [15]. Later work showed that a two-stage model could perform better [16]. The two-stage model assumes a first stage of mutations with clonal growth of tumour cells until the second stage occurs. However, the importance of estimating the last stage of carcinogenesis has been put forward [17].

For this reason, we proposed an exposure-driven functional model for the analysis of the carcinogenic process and the multistage model of carcinogenesis in cancer epidemiology [18]. The assumption in our model is that exposure to carcinogens is the major driving force of the carcinogenic process, that cessation of exposures could stop, or even reverse this process, and that the stages usually covers decades of human life. The different stages of carcinogenesis can be due to both mutations and functional changes in oncogenes or tumour suppressor genes. The hypothesis underlying the proposed post-GWAS functional research is that trait-associated alleles exert their effects by influencing transcriptional output [5]. This corresponds to our hypothesis that different stages of carcinogenesis can be identified by measuring the expression of specific genes that are affected by exposure to carcinogens. Different exposures might also lead to variations in the metabolic pathways, and result in an exposure-specific multistage model. This has been shown for exposures like radiation [19], viruses [20] or bacteria [21].

Limitations and opportunities in the study design of a human model of carcinogenesis

In order to understand and describe the carcinogenic process or the multistage model, the primary interest is to first estimate the functional changes in single genes or pathways that are related to different stages of carcinogenesis, as well as the length of these stages. Given the nature of carcinogenesis, the epidemiological design with the most potential would be a prospective study. 1) The prospective study de-

sign is the only epidemiological study design that can take time dependency into account by looking at the carcinogenic process during decades of follow-up. 2) Biological material suitable for high quality gene expression analyses should be collected from peripheral blood before cancer diagnosis and at time of diagnosis as described for the globolomic design. One problem specific to functional analyses of transcriptomicsare length of follow-up. Indeed, as high quality mRNA for whole genome expression analyses requires samples buffered for the effect of Rnase, most biobank material meeting this criterion would typically not be older than a decade (coinciding with the availability of the necessary commercial products), which is generally not sufficient for functional studies of cancer development. Repeated measurements would be necessary both for the –omics analyses and for questionnaire information on lifestyle factors. The transcriptome is changing over time and potentially mirroring the carcinogenic process. It is obvious [22] that repeated sampling is most important. For practical purposes the intervals would be of several years. The cumulation of the exposures could be based on repeated questionnaires or markers of long term exposures. 3) Prospectively collected biological material should be analysed according to a nested case-control design within a cohort, as the classic case-control design is vulnerable to systematic errors in exposure measurements which will be impossible to compensate through statistical techniques. In a cross-sectional study of lifestyle factors and gene expression, different batch effects during laboratory testing dominated the overall gene expression profiles and could seriously compromised the results [23]. A closely matched case-control study design that keeps the case-control pairs together throughout all laboratory analyses will reduce batch effects since the major measure of the gene expression to be used in the statistical analyses will be the differences in gene expression between case and control within each pair. In this way, even if the cases-control pairs are handled differently at some point during laboratory analyses due to different batches over time, the potential differences due to systematic errors between cases and controls will remain small. 4) The prospective study design gives what is usually considered to be non-differential biased estimates of relative risk. While obtaining blood samples from cases is generally fairly easy, finding valid controls that represent the source population could be a major problem. However, this problem can be eliminated by using a nested case-control design in existing cohorts, as cases and controls will originate from the same population. 5) The use of buffered whole blood gives a mixture of mRNA from different subpopulations of peripheral blood cells. This mixture could confuse the transcriptional profiles. There is some ongoing work trying to identify each cell subtype by unique expression patterns [24, 25]. While mRNA suitable for whole genome analysis is not detectable in plasma, microRNA can be extracted both from whole, buffered blood and from plasma or serum in standard biobank material. The potential differences between these two sources of microRNA are not well studied. 6) The proposed carcinogenic model should also take into account the potential of protective factors like dietary factors, but unfortunately these are not that well characterized. One such example would be high parity and decreased risk of breast cancer. 7) The statistical power of sample size necessary for the search of gene expression patterns related to the carcinogenesis in peripheral blood is still hard to calculate. As proposed by Potter the number of persons to be involved should be hundred of thousands or a million [26], in the same range as ongoing new biobank studies in England and Sweden.

Functional anlyses and the multistage model

Here below is a brief description of the functional design and analytical challenges of transcriptomics in a prospective study.

Cancer cases should be identified through linkages to cancer registers, hospital systems or through feedback from study participants. Once identified, random controls from the cohort should be selected as per a nested case-control study design. Stored biological samples for each case-control pair should then be processed together through extraction, hybridisation and scanning, in order to reduce technical noise and batch effects. The researcher will then receive a quantitative measure of the gene expressions and the differences for each case-control pair either from micro-array or deep sequencing. Biological samples used for transcriptomics should be of high quality and collected according to the specific analyses to be done. Most microRNA testing can be performed on good existing biobank material. The major limitation

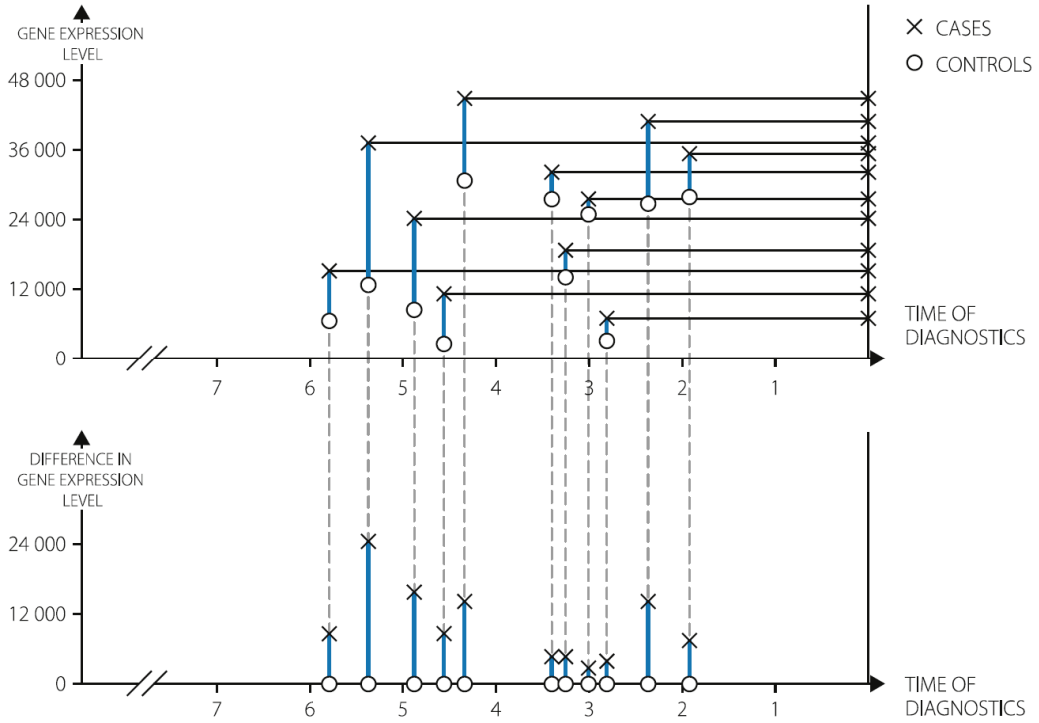

Figure 1. Hypothetical gene expression levels (Illumina Hu-6 chip) and differences in expression between cases and controls for single genes taken from case-control pairs with different follow-up time, illustrating a potential distribution of time in years for the start of the last stage.

is the quality of mRNA in non-buffered samples of blood and tissue. Epidemiological studies in which blood samples were rapidly frozen in liquid nitrogen could have high quality mRNA depending on the actual procedures. The challenge will then be to find the most sensitive analyses of the differences in gene expression for all 25000 genes.

This approach faces at least four major problems. 1) Time dependency should ideally be controlled through repeated measurements for all participants. This would allow previous samples taken from decades ago from cases at the time of diagnosis, to be compared to controls without the disease at the time of case diagnosis. Unfortunately, there are only a few prospective studies that can offer this possibility. A restricted, but more realistic strategy would be to use time-dependent information from case-control pairs with blood samples collected at different follow-up times or times prior to diagnosis (cases) or matching date (controls) (**Figure 1**). For each case-control pair, the gene expression, and the differences in gene expression, for single genes can be estimated from stored blood samples at a given point of follow-up, upper panel **Figure 1**. The distribution of the differences in gene expression during follow-up for genes differentially expressed by cases and controls could create an observed time distribution. This observed distribution would consist of both time before diagnosis, and a quantitative measure for each gene. This could be used to estimate the time distribution of the unknown last stage of carcinogenesis, lower panel **Figure 1**. 2) In the model proposed here, exposures are considered to be the major driving force of the carcinogenic process. This implies that all analyses of mutations or functional changes should be related to information on potential carcinogens, as in gene-environment studies. Since each exposure could have a specific gene expression profile, the overall gene expression profile in a given blood sample could be due to the unique mixture of lifestyle factors for that person. Finding genes differentially expressed by cases and controls due to a mutation or functional change would be difficult with little, or no exposure information, as the overall gene expression profile would consist of many overlapping exposure-related profiles, which could be

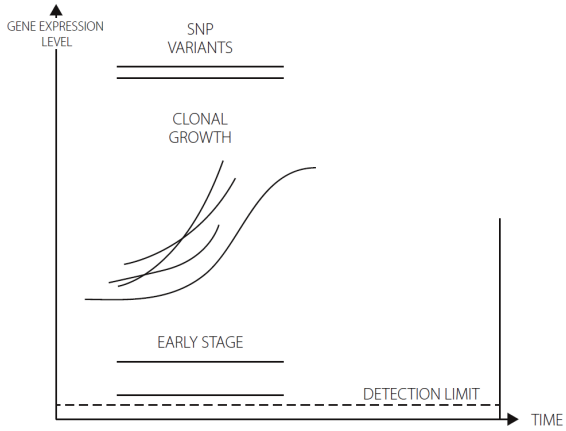

Figure 2. Time-dependant gene expression during follow-up, showing potential distributions; somatic mutations or functional changes (upper panel), SNPs (middle panel) or early stage (lower panel).

seen as a lack of controlling for confounders. Spontaneous or endogenous mutations due to the cellular processes would complicate the analyses. 3) An important issue is the distribution of gene expression over time for each single gene (**Figure 2**). A new somatic mutation or functional change might be followed by a change in gene expression. An increased risk related to germ line mutations given as SNPs has been proposed to translate into a constant difference in gene expression over time [5] (upper panel **Figure 2**). A plausible assumption would be that changes in gene expression following a clonal growth model should yield an exponential, or maybe a sigmoid distribution during follow-up (middle panel **Figure 2**). The same pattern of constant differences could be the effect of an early stage of carcinogenesis (lower panel **Figure 2**). Estimation of functions describing the single gene expression pattern during follow-up would help to better understand the mechanisms behind stage changes. Currently, the expression of any single gene has an unknown distribution over time. Nevertheless, the values observed could be used to estimate the length of the last stage of carcinogenesis using mathematical models. 4) The possibility that the strength of the gene expression signals related to mutations or functional changes in different stages of carcinogenesis could be weaker than those related to housekeeping or other exposures raises some statistical problems. Another aspect of this problem is the lack of knowledge about when in carcinogenesis one could expect to find traces of it in blood. Studies of the last stage could seem most promising. The use of standard statistical testing and false discovery rate, focusing on the lowest p-values or most significant associations, might remove the weaker signals related to carcinogenesis due to less significant p-values. Therefore, weaker associations between exposures, gene expressions and disease should be held up against existing biological knowledge. The difficulty here lies in the fact that current biological knowledge has been obtained mostly from *in-vitro* experimental systems and not from *in-vivo* observations of humans, and with little knowledge about the effects of different exposures.

Challenges in mathematical modelling

A new approach to the statistical analysis of prospective studies will be important to the goal of estimating the distribution of the length of the last stage of carcinogenesis, and identifying the genes involved. Observations should be left-censored due to the short follow-up time in relation to the length of the carcinogenic process. The challenge will be to estimate the distribution of the length of the last stage as a latent variable conditionally on changes in exposures. The most realistic approach could be to start by defining the length of the last stage of carcinogenesis as the time elapsed from diagnosis backwards until the change in gene expression, as changes that lead to invasive cancer may be relatively close in time to cancer diagnosis. The analytical challenge will be to separate changes in gene expression due to exposure to carcinogens, from gene expression signals related to the carcinogenic process caused conditionally on these same exposures. Of major importance is the possibility for cumulation exposures over time potentially by repeated collection of questionnaires and blood.

For example, let us assume that an exposure A changes the gene expression of a single gene. For a given gene affected by exposure A, there are four exposure relationships possible in a matched, nested case-control study design: both cases and controls exposed, only cases exposed, only controls exposed, or neither cases nor controls exposed (**Table 1**). Gene expression is usually given as a quantitative, continuous variable. For simplicity it is divided into yes or no for cases and controls in **Table 1**. The

Table 1. Differences in gene expression between cases and controls in a nested case-control design, one gene expressed only for exposure A, and one gene related to carcinogenesis in cases conditionally on exposure A.

	Gene exposure				Gene carcinogenesis			
	Exposure				Exposure			
Cases	yes	yes	no	no	yes	yes	no	no
level of gene expression	a	c	e	g	i	k	m	o
Controls	yes	no	yes	no	yes	no	yes	no
level of gene expression	b	d	f	h	j	l	n	p
Differences in gene expression	a-b	c-d	e-f	g-h	i-j	k-l	m-n	o-p
Null hypotheses	+(-)	0	-(+)	0	+(-)	+(-)	0	0

differences in gene expression should be zero for exposed cases versus exposed controls. For exposed cases and unexposed controls, the difference should be positive if the mutation or functional change increases gene expression, and negative if it decreases expression. If the cases are not exposed and the controls are, the difference in gene expression would be inverted. And lastly, if both cases and controls are unexposed the difference should be zero. The result should be 0, +, -, 0 as a combination score, given that the follow-up time is shorter than the length of the last stage of carcinogenesis. If the follow-up time is longer than the last stage, the combination should be 0,0,0,0. For a gene related to carcinogenesis in cases conditionally on exposure A, the pattern of differences in gene expression would be different. The gene should only be expressed in exposed cases, giving a combination of +, +, 0, 0. Knowledge about the length of the last stage – potentially dependent on the exposures – should improve the sensitivity in detecting genes involved in the last. It is obvious that this analytical strategy will depend on the quality of the exposure information. Different mathematical models [18] could be used for the search of relationships like this under the hypothesis that the carcinogenesis is dependent on exposure to carcinogens and expressed through specific changes due to mutations or functional changes.

The sensitivity of the matched nested case-control design is highly dependent on the lack of batch effects and the use of the latest microarray technology. As an example, the Illumina Hu-6 chip has on average 60 replicates or repeated measurements of each gene. The distribution of the differences in gene expression for 100 random genes run for 150 case-control pairs is shown in **Figure 3**. The figure shows the mean gene expression (whole line) with 95% confidence intervals for the matched design (bold lines). For the non-matched case-control design the 95% confidence lines (broken lines) are much wider. The improved sensitivity of the matched design is obvious.

This novel epidemiological approach has the possibility to provide further information on the functional changes that follow exposure to carcinogens, and offers the opportunity to advance

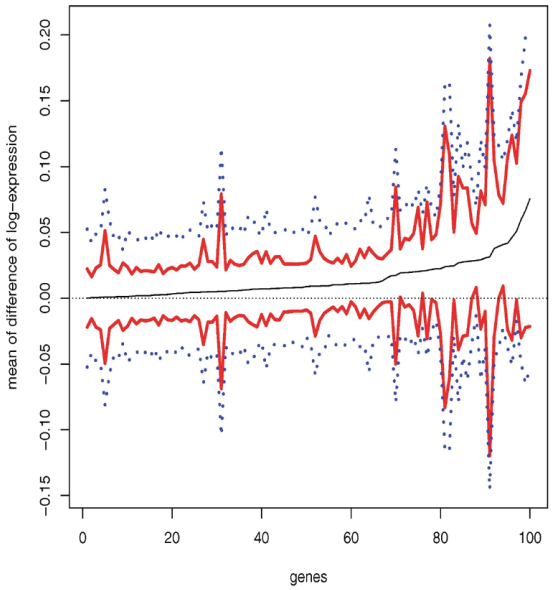

Figure 3. The distribution of differences in gene expression with 95% confidence intervals for 100 random genes measured in peripheral blood given for a matched and a non-matched case-control design, the mean difference in gene expression given as a whole line, 150 pairs from the Norwegian Women and Cancer postgenome cohort [3, 7] order by increasing differences, (Illumina Hu6 micro-array).

our understanding of the carcinogenic process. It could give us pre-diagnostic tests that could be used in connection with other diagnostics like mammography. This would depend on common pathways for different exposures.

Conclusion

Several researchers have claimed that the current views on the carcinogenic process are approaching paradigm instability [27-29]. The two different post-GWAS approaches of basic genetic research and epidemiology could mutually improve the understanding of the mechanistic or functional aspects of multistage carcinogenesis.

Acknowledgement

This study was supported by Grant ERC-2008-AdG 232997-TICE "Transcriptomics in cancer epidemiology"

Address correspondence to: Dr. Eiliv Lund, Institute of Community Medicine, University of Tromsø, 9037 Tromsø, Norway Tel: +47 91144064; Fax: +47 77644831; E-mail: eiliv.lund@uit.no

References

[1] Croce CM. Oncogenes and cancer. N Engl J Med 2008; 358: 502-511.
[2] Farazi TA, Spitzer JI, Morozow P, Tuschl T. miRNA in human cancer. J Pathol 2011; 223; 102-115.
[3] Dumeaux V, Børresen-Dale AL, Frantzen JO, Kumle M, Kristensen VN, Lund E. Gene expression analyses in breast cancer epidemiology: the Norwegian Women and Cancer postgenome cohort study. Breast Cancer Res 2008; 10: R13.
[4] Brase JC, Wuttig D, Kuner R, Sultmann H. Serum microRNAs as non-invasive biomarkers for cancer. Molecular Cancer 2010; 9; 306.
[5] Freedman ML, Monteiro AN, Gayther SA, Coetzee GA, Risch A, Plass C, Casey G, De Biasi M, Carlson C, Duggan D, James M, Liu P, Tichelaar JW, Vikis HG, You M, Mills IG. Principles for the post-GWAS functional characterization of cancer risk loci. Nature Genetics 2011; 43; 513-518.
[6] Lund E, Dumeaux V. Systems epidemiology. CEBP 2008; 17: 2954-2957.
[7] Lund E, Dumeaux V, Braaten T, Hjartåker, Engeset D, Skeie G, Kumle M. Cohort profile: The Norwegian women and cancer study - NOWAC - Kvinner og kreft. International Journal of Epidemiology 2008; 37: 36-41.
[8] Spitz MR, Bondy ML. The evolving discipline of molecular epidemiology of cancer. Carcinogenesis 2010; 31; 127-134.
[9] Reiner A, Yekutieli D, Benjamini Y. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics 2003; 19; 368-375.
[10] Hanahan D, Weinberg RA. Hallmarks of cancer: The next generation. Cell 2011; 144: 646-674.
[11] Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature 2009; 458: 719-724.
[12] Vineis P, Schatzkin A, Potter JD. Models of carcinogenesis: An overview. Carcinogenesis 2010; 31: 1703-1709.
[13] Armitage P. Multistage models of carcinogenesis. Env Health Perspective 1985; 63: 195-201.
[14] Cox LA, Huber WA. Symmetry, identifiability, and prediction uncertainties in multistage clonal expansion (MSCE) models of carcinogenesis. Risk Analysis 2007; 6: 1441-1453.
[15] Armitage P, Doll R. The age distribution of cancer and a multi-stage theory of carcinogenesis. Br J Cancer 1954; 8: 1-12.
[16] Moolgavkar SH, Day NE, Stevens RG. Two-stage model for carcinogenesis: epidemiology of breast cancer in females. JNCI 1980; 65: 559-569.
[17] Day NE, Brown CC. Multistage models and primary prevention of cancer. J Natl Cancer Inst 1980; 64: 977-989.
[18] Lund E, Dumeaux V. Towards a more functional concept of causality in cancer research. Int J Mol Epidemiol Genet 2010; 1: 124-133.
[19] Imaoka T, Nishimura M, Iizuka D, Daino K, Takabatake T, Okamoto M, Kakinuma S, Shimida Y. Radiation-induced mammary carcinogenesis in rodent models: What's different from chemical carcinogenesis? J Radiat Res 2009; 50: 281-293.
[20] Lizano M, Berumen J, Garcia-Carranca A. HPV-related carcinogenesis: Basic concepts, viral types and variants. Arch Med Res 2009; 40: 428-434.
[21] Hernandez LG, van Steeg H, Luitjen M, van Benthem J. Mechanisms of non-genotoxic carcinogens and importance of a weight of evidence approach. Mut Res 2009; 682: 94-109.
[22] Vineis P, Chadeau-Hyam M. Integrating biomarkers into molecular epidemiological studies. Curr Opin Oncol 2011; 23: 100-105.
[23] Dumeaux V, Olsen SK, Paulssen RH, Børresen-Dale AL, Lund E. Deciphering blood gene expression variation – The postgenome NOWAC study. PLoS Genetics 2010; 6: e1000873.
[24] Allantaz F, Cheng DT, Bergauer T, Ravindran P, Rossier MF, Ebeling M, Badi L, Reis B, Bitter H, D'Asaro M, Chiappe A, Sridhar S, Pacheco GD, Burczynski ME, Hochstrasser D, Vonderscher J, Matthes T. Expression profiling of human immune cell subsets identifies miRNA-mRNA regulatory relationships correlated with cell type specific expression. Plos ONE 2012; 7:

e29979.
[25] Birnbaum KD, Kussell E. measuring cell identity in noisy biological systems. Nucl Acids res 2011; 39: 9037-9107.
[26] Potter JD. Epidemiology informing clinical practice: from bills of mortality to population laboratories. Nat Clin Pract Oncol 2005; 2: 625-34.
[27] Baker SG, Cappuccio A, Potter JD. Research on early-stage carcinogenesis: Are we approaching paradigm instability? J Clin Onc 2010; 28: 3215-3218.
[28] Weinstein B, Joe A. Oncogene addiction. Cancer Res 2008; 68: 3077-3080.
[29] Felsher DW. Oncogene addiction versus oncogene amnesia: Perhaps more than just a bad habit. Cancer Res 2008; 68: 3081-3086.