Int J Mol Epidemiol Genet 2010;1(2):124-133
www.ijmeg.org /IJMEG912004

Review Article

Towards a more functional concept of causality in cancer research

Eiliv Lund, Vanessa Dumeaux

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

Received December 21, 2009, accepted March 2, 2010, available online: March 15, 2010

Abstract: Advances in molecular technologies challenge the different concepts of causality in biology, epidemiology and multistage mathematical models. The lack of integration of the different aspects of causality into a common framework could postpone our attempts to build a human causal model of carcinogenesis. We present here some aspects of differences in methodology, terminology and traditions between the scientific disciplines and propose a research strategy using functional analyses of the transcriptome and epigenetics to illuminate causality in complex biological systems. Overcoming the challenges of biological material collection suitable for such analyses into a prospective design, this could give unique opportunities for verification of mechanistic information from basic biological research in a human model system. The ultimate goal is to obtain a dynamic causal description of the different carcinogenesis stages. The success of this novel approach depends on the biological relationship between the gene expression of the somatic driver mutations or co-expressed genes in tumours and the gene expressions mirrored in peripheral blood along the different stages of carcinogenesis. The use of gene expression profiles and epigenetics could produce a functional concept of causality to explain the human multistage carcinogenic process.

Keywords: Causality, transcriptomic, epigenetics, carcinogenesis, epidemiology, multistage models

Introduction

The description of the genome [1,2] has been followed by an extraordinary technology-driven research at the borders between traditional epidemiology and basic research investigating single nucleotide polymorphism (SNPs) in relation to disease risk. Similar innovations are ongoing for the analyses of the cancer genome, transcriptome, proteome, and metabolome [3]. This development has fostered new multidisciplinary and interdisciplinary fields of research under headings like systems biology [4] or systems epidemiology [5]. In this context, the different definitions and traditions for handling the question of causality have been well described and discussed [6-8], but so far no common concept have been proposed.

Here, we will present three major approaches to causality from basic biological research, epidemiological research, and mathematics and argue for an exploration of pathways analysis as a functional tool for assessing causality in interdisciplinary cancer research.

Concept of causality in basic genetic research

Basic genetic research is mechanistic oriented and mostly based on studies with experimental design using animal models or cell lines [9]. Within this tradition, the rather reductionist definition of causality [10,11] corresponds to the effects related to one change in the experiment. Also, the concept of causality in cancer research has also been promoted in a more biological sense as the effects of mutations in oncogenes, suppressor genes and microRNA [12]. Hence, we now comprehend that cancer arises from successive genetic changes by which a number of cellular processes are altered. This paradigm of genetic causality has been challenged by highly complicated and multilevel functions described in systems biology [13]. It is also currently understood that gene analysis by itself provides an incomplete picture. Due to alterna-

Table 1. Causality criteria proposed by Hill [19]

(1) Strength – of the observed statistical association
(2) Consistency – repeated observations by different persons, in different places and time
(3) Specificity – the association is limited to specific exposures or particular sites and types of disease
(4) Temporality – the order of exposure before disease
(5) Biological gradient – a dose-response curve
(6) Plausibility – existing biological knowledge
(7) Coherence – should not conflict with known facts of natural history and biology
(8) Experiment – experimental or semi-experimental evidence
(9) Analogy – judged by analogy

tive splicing of both mRNA and proteins, combined with more than 100 unique post-translational modifications, one gene can give rise to multiple protein species. In tumour biology, mutations are divided in driver and passenger mutations [14]. A driver mutation is a somatic mutation in the tumour that is causally linked to carcinogenesis. It is not required for maintenance of the invasive cancer, but it must have been functional sometimes during cancer development. Passenger mutation will have no impact on the carcinogenic process since they have not conferred clonal growth advantages being somatic mutations without functional consequences. Driver mutations could occur in at least five or six genes, while the numbers of passenger mutations could be substantially higher. The presence of somatic mutations is partly the reason for the new cancer taxonomy based on gene expression profiles [15]. These molecular profiles based on hierarchical modelling could improve etiological research by classifying tumours according to gene expression patterns that could also be linked to specific exposures. Animal models have been constructed to show the carcinogenic process in murine models, e.g. for gastric cancer [16]. However, animal models have been shown to be too simplistic compared to human carcinogenesis. For example, in rats only two mutations are sufficient for creating a tumour cell from a normal cell, but these finding have not been generalized to human cell lines [17]. Cell lines show the same lack of direct generalisation because of changes partly due to laboratory manipulations in order to make the cell lines immortal [18]. In spite of an overwhelming literature on *in vitro* experimental pathway research, neither the exact functions nor the succession of mutations necessary for the multistage model have been sufficiently described for implementation in *in vivo* research.

Finally, the introduction of high-throughput analyses of DNA, RNA, proteins and metabolites in basic biological studies has fostered the need for advanced statistics and data mining making the interpretation of the results closer to the criteria used in biostatistics and epidemiology. This development has reinforced the research discipline named systems biology and the closely related methodologies of bioinformatics.

Concept of causality in epidemiology

Epidemiology is mostly an observational science with few experiments. The nature of causality in studies of human health and diseases has been discussed over several hundred years. The criteria used today are most often referred to Hill [19], but philosophers like Hume previously noted several of the important aspects of causality [20]. A major discussion of criteria for determining causality was brought up by the early works related to smoking and lung cancer. At that time, epidemiologists still used to think of causality in mono causal terms by the postulates of Koch [21]. The causality criteria by Hill, see **Table 1**, were partly rules for judging the time frame, statistical associations, relationship to earlier works, ecological data, analogy and experiments. Of his nine criteria, some are today obsolete like analogy, and others are often better specified like specificity of both exposure and disease. For a more dynamic concept of causality the most important criteria of Hill is plausibility. Hill himself had a very cautious view on this aspect since he very clearly saw that the criteria would depend upon the biological knowledge of the day. Presently, the discussion of biological plausibility draws heavily on the knowledge from basic research with inherent problems of generalization to human conditions. As noted earlier, most information originates from animal and *in vitro* experiments and could

be too simplistic for the interpretation of gene function in the real life situation with multiple exposures and the flexibility of the pathway network to adjust for different living conditions.

Also, whole genome scans in gene-environment studies have given birth to another set of criteria [22]. Some of these are related to the validity of the study design and the statistical problem of mass testing of significance. These criteria have always been part of the scientific process, but traditionally they were discussed as part of the study design, before the discussion of causality. The concern over lack of validity and comparability between studies now foster initiatives for more systematic judgement about the validity of questionnaire information, biomarkers and outcomes [23].

Mathematical modelling of the carcinogenic process

There has been longstanding and wide interest for mathematical modelling of the carcinogenic process. Different models have been applied on incidence figures [24,25]. These models mostly describe a two or multiple step process. Each step has been interpreted as a mutation. These mutations changes the metabolic pathways, but the models do not say anything about the nature of each step, only that several steps seems to be necessary to describe the incidence curves for most cancer sites. As written by P. Armstrong more than 20 years ago: "Until and unless we obtain direct evidence about the presence and nature of intermediate stages, any statistical theory is likely to remain largely unfalsifiable" [24]. The mathematical models have been important for the concepts of initiation (first mutation), promotion (clonal growth) and progression (final mutation) [26]. They leave little possibility for a judgement of the effect linked to any allele variant or to changes in gene expression or protein synthesis as a consequence of lifestyle [27]. The dependence of the cancer incidence on molecular processes can still not be quantified and the lack of details on how cancer evolves keeps attempts to link relevant biological processes to risk patterns at a fairly simple level.

Improving concepts of causality through transcriptome analyses

Due to the traditional mono-disciplinary research the three different concepts of causality in cancer research have been poorly integrated, but recent research headed under the name of gene-environment interaction studies has for a long time been forcefully pushed forward [28]. These gene-environment studies combine epidemiology with basic genetics. They are mostly based on a simple design with exposures measured once at start of follow-up and analysed together with variants of single nucleotide polymorphisms, SNPs, also considered as exposures in the same statistical design. Several of the large initiatives also use case-control samples [e.g. 29]. So far this approach has given little information on the multistage process of cancer. High-throughput SNPs assays have not revealed a strong association to tumour development. In a previous genome-wide association study of lung cancer as an example, only one single locus was associated with an increased risk of lung cancer (RR<2) after a search of around 30,000 genes [29]. In contrast, the risk for lung cancer for heavy smokers in the same study was found to be in the order of ten or more. In addition, the identification of one or more putative SNPs will presumably not inform us about which stage in carcinogenesis could be affected. This lack of information makes results less useful for public health strategies since effective prevention depends on the ability to intervene on the last steps of the multistage process [30]. In similar analyses, no major SNP in the oestrogen metabolism was identified as a risk factor for breast cancer [31]. Hence, some research is ongoing in order to improve the analyses of the missing heritability of complex disease [32].

Human biological processes are the outcome of complex interactions between lifestyle factors, environment, and genes, so there might be strong reasons for studying gene expression or protein changes *in vivo*. Also, the genome and the proteome can provide a dynamic reflection of both the intrinsic genetic programme of the cell and the impact of its immediate environment. At least three different approaches are currently applied in order to obtain a better understanding of causality in relation to the multistage carcinogenesis. First, several animal and cell models have successfully imitated the multistage aspect of cancer development but a major problem remains, as noted previously, the generalization from *in vitro* till *in vivo* models. Second, a huge effort is under way to describe

Table 2. A short overview of current *in vivo* knowledge of gene expression in blood and tumour tissue relevant for a more functional concept of causality

Tumour tissue
Gene profiling of tumour tissue have added a new taxonomy for some cancer sites like breast cancer [15]
Large cohort studies of cancer patients have been constructed for clinical research investigating cancer prediction and prognosis [37]
Peripheral blood
Expression patterns in peripheral blood cells have been found to reflect exposures like radiation [38], metal fumes [39], smoking [40], dioxin [41], hormone therapy[42], and benzene [43] which could be confounders in gene expression analyses related to disease status.
There is *in vitro* evidence for a dietary regulation of microRNA expression in cancer cells [44,45]
One recent study gives a reference library for gene expression in blood cell subtypes, Haematlas, which identifies key genes with roles in blood cell function [46]
Information bridging the blood – tumour gap, necessary for a transcriptional model of carcinogenesis
The peripheral blood transcriptome may dynamically reflect system wide biology which could be use as a potential diagnostic tool [47-55]. Last year, the first commercial diagnostic tests based on gene expression in blood were launched. The ColonSentry™ is an RT-PCR based assay of 7 genes for colorectal cancer screening [50] and the BCtect™ is also an RT-PCR based assay using 96 genes to detect breast cancer at an early stage of the disease [48], even though the sensitivity and specificity so far is not excellent.
A relationship between gene expression in blood and adipose tissue for traits related to clinical obesity has been observed with a weaker relationship in blood [54]. A significant genetic component to gene expression traits has been demonstrated for genes related to obesity.

the different cancer genomes (e.g. the International Cancer Genome Consortium, ICGC) [14]. Somatic mutations can be considered as consequences of the archaeology of the cells with the use of the mutations prevalence as a measure of the driver or passenger status, but still verifications by other methods will be necessary. The third approach is to build new prospective studies with a more complicated design in order to follow the changes in transcriptome and proteome of peripheral blood and target tumour tissue over time. Obviously, these studies should overcome the problems of chance or statistical power by being large enough (50.000 – 1.000.000), be representative of defined populations, with information on major lifestyle factors from questionnaires and with biobanks for the analyses of the genome, transcriptome, proteome, and metabolome. A major limitation of such approach so far has been lack of application of standard epidemiological methods [33]. Several proposals exist under headings such as population laboratory [34] or the globolomic design [5, 35, 36].

Biological premises for a functional causality concept

The idea of building large prospective studies to explore the transcriptome and the proteome in cancer epidemiology as a basis for a functional concept of causality depends on some assumptions about gene expression and regulation in peripheral blood and tumour tissues. Taking the example of the transcriptome, peripheral blood gene expression may dynamically reflect system wide biology and possibly pre-invasive stages of cancer (**Table 2**). Recognition of the role of molecular changes in carcinogenesis demands a new generation of molecular biomarkers of exposure in order to account for confounding signals and reinforce the investigation of biological plausibility for these associations [7, 41].

Somatic mutations and their associated patterns in gene expression should be accessible through collection of tumour tissue. However, for most sites of cancer only the last invasive stage is accessible for tissue biopsies. Tissue

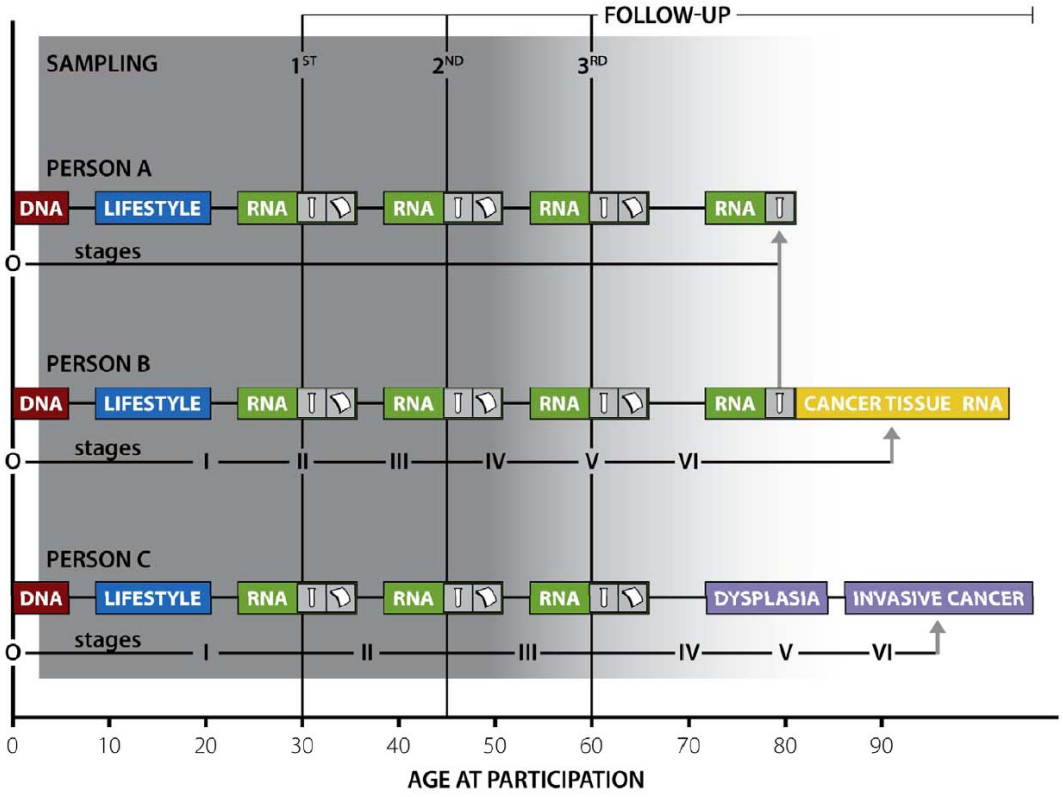

Figure 1. A schematic representation of the relationship between the design of a globolomic study with three repeated samples of blood and eventually one of tumor tissue sample, and the multistage model, for three persons with different life histories or stages in the model. The dark shaded area shows the optimal design with repeated measurements, the light one the restricted design with only one measurement.

samples from pre-malignant stages are only taken out regularly from the cervix and to some extent from the breast. Of note, tumour evolution for example in the breast does not seem to follow a simple linear progression model but rather appears to involve a more complex process characterized by periods of expansion followed by contraction concomitant with simultaneous evolution of multiple independent tumour regions – a scenario likely responsible for the great mutational heterogeneity witnessed in this disease [57, 58]. It has been noted by several investigators that tumour heterogeneity (i.e. expansion) is greatest at transition phase (i.e., from ductal carcinoma *in situ* to invasive ductal carcinoma, or from invasive ductal carcinoma to metastatic boundaries) and less so following these phases (i.e. contraction) presumably owing to alternating periods of mutational diversity followed by clonal selection.

Obviously, the success of a prospective design with genomic, transcriptomic, and proteomic options will depend heavily on the collected biological specimen and new technologies. mRNA is particularly sensitive to degradation by abundant and ubiquitous RNAses and several techniques to overcome this challenge are used or under investigation [60, 61]. The questions of technologies are utterly more important for studies of the transcriptome in peripheral blood than in tumour tissue due to lower yield of RNA and less highly expressed genes. Moderne technology has enabled studies of microRNA from serum (62). Similarly, robust technology exists for the analysis of epigenetics, mainly as DNA methylation, and to a lesser extent for histone modifications [63, 64]. Finally, the profiling of proteins is difficult in numerous other ways. Apart from the technological challenge presented by the range of protein concentrations, proteins have properties arising from their folded structures, so generic methods are difficult to design and apply, and the analysis and

significance of post-translational modifications provide a major challenge, both in normal and disease conditions [65].

New designs

The major challenge of a more functional approach to causality using a surrogate tissue like peripheral blood as proposed here will be to investigate whether there is specific communication between the tumour tissue till peripheral blood by the means of mRNA, microRNA or other molecular mechanisms through the different stages of initiation, promotion and progression. Obviously one would expect such signals to be weak, maybe a few transcripts or proteins changes, pointing to the need of high quality biological specimens and the use of even more improved novel technologies (e.g. deep sequencing, protein microarrays).

We propose a new analytical strategy as extensions of the common prospective design. The optimal design for a functional study would be to sample blood suitable for gene expression analyses regularly during follow-up to facilitate analyses of possible changes in gene expression within the same persons over time and compare with tumour tissue expression. **Figure 1** illustrates a globolomic design with multiple (e.g. three) collections of peripheral blood suitable for gene expression, protein and metabolites analyses throughout the lifespan. In addition, tumour tissue from cases can be collected at the time of diagnosis. Schematically, person A remains healthy and presents no somatic driver mutation. Person B contracts a cancer some years after the third blood sample - the proposed numbers of six somatic mutations have taken place. At time of diagnosis tumour tissue sample is taken and buffered for expression profiling together with another peripheral blood sample. Concurrently, person A is randomly drawn as a control and also asked for blood samples. This nested case-control design should give correct estimates of differences in gene expression related to the carcinogenic process between diseased and healthy individuals and with the possibilities of adjustment for potential confounders. The comparison could be done from the first collected blood samples until cancer diagnosis of cancer. In the same design, collection of tumour tissue would enable a comparison between driver mutations and their expression in tumour tissue and blood profiles over time. There exist quite a few large prospective studies with repeated blood samples suitable to investigate genome and proteome; however in the discussion of pro et cons for new prospective biobanks gene expression analyses are mostly neglected [65-67]. Prospective study including biological specimen suitable for expression analyses in peripheral blood and tumour tissue has only recently been implemented [36].

A restricted analytical design could be implemented even if only one blood sample suitable for expression analyses was collected before diagnosis. The restricted design, light shaded area in **Figure 1**, would start its follow-up at the time shown for the third blood collection. Person A and B can be compared, but only for the one blood sample taken a few years before. In addition, some participants will at time of blood sampling have neoplasm at different pre-invasive stages, illustrated with person C at stage III. Through the follow-up period, this neoplasm develops into invasive cancer and is diagnosed. The comparison between the stored blood samples for this person C with person A could give information on the possible gene expression of the somatic mutation in stage III. Also, if the two stage clonal expansion model [68] is correct the success rate of this approach would improve due to fewer mutations and consequently critical carcinogenesic stages.

Concluding remarks

The lack of information on the exact nature of the multistage model is one of several obstacles for an integrated concept of causality in current translational cancer research. For the moment we have no commonly accepted carcinogenic model with a description of the number of mutations or their succession. It has been proposed that six essential alterations in cell physiology dictate malignant growth [69]. Another explanation for the lack of understanding of the carcinogenic model could be the effects of resistance genes that may stop incipient cancerous foci [70]. Alternatively, since gene expression and protein expression differ along the time scale modelling biochemistry at the metabolic level could be another solution [71].

Since we lack almost completely knowledge about the transcriptome or proteome in peripheral blood and its relationship to the carcino-

genic process in tumour tissues we propose that future prospective studies should include high quality biological material for gene expression analysis, preferably with repeated samples through several decades. In this way the fragmented and mono-disciplinary concept of causality used today could be replaced by a more functional and dynamic view on the carcinogenic process integrating multiple perspectives. The term systems epidemiology has been introduced to cover such a new scientific discipline [5]. Ongoing projects should in a few years more clearly demonstrate any important effects of this design on the concept of causality – the core term for most epidemiologists and basic researchers.

Aknowledgements

This work was supported by Grant: ERC-2008-AdG 232997-TICE.

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

References

[1] Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blöcker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ; International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature 2001; 409: 860-921.

[2] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng

ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigó R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X. The sequence of the human genome. Science 2001; 291: 1304-1351.

[3] Vineis P, Perera F. Molecular epidemiology and biomarkers in etiologic cancer research: the new in light of the old. Cancer Epi Biomarker Prev 2007; 16: 1954-65.

[4] Institute of Systems and Synthetic Biology at Imperial College London. What is systems biology? www3.imperial.ac.uk/systemsbiology (accessed July,22 2009).

[5] Lund E, Dumeaux V. Systems epidemiology in cancer. Cancer Epi Biomarker Prev 2008; 17: 2954-7.

[6] Green LW. Public health asks for systems science: to advance our evidence-based practice, can you help us get more practice-based evidence? Am J Public Health 2006; 96: 406-9.

[7] Wild CP. Environmental exposure measurements in cancer epidemiology. Mutagenesis 2009; 24: 117-25

[8] Carbone M, Klein G, Gruber J, Wong M. Modern criteria to establish human cancer etiology. Cancer Res 2004; 64: 5518-24.

[9] Zhao JJ, Roberts TM, Hahn WC. Functional genetics and experimental models of human cancer. Trends Mol Med 2004; 10: 344-50.

[10] Lagiou P, Adami HO, Trichopoulos D. Causality in cancer epidemiology. Eur J Epi 2005; 20: 565-74.

[11] Lazebnik Y. Can a biologist fix a radio? Cancer Cell 2002; 2: 179-82.

[12] Croce CM. Oncogenes and cancer. N Eng J Med 2008; 358: 502-1.

[13] Noble D. Genes and causation. Phil Trans R Soc 2008; 266: 3001-15.

[14] Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature 2009; 458: 719-24.

[15] Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lønning PE, Børresen-Dale AL, Brown PO, Botstein D. Molecular portraits of human breast tumours. Nature 2000; 406: 747-752.

[16] Heeg S, Doebele M, Werder A, Opitz OG. In vitro transformation models. Modelling human cancer. Cell Cycle 2006; 6: 630-4.

[17] Sun B, Chen M, Hawks CL, Pereira-Smith OM, Hornsby PJ. The minimal set of genetic alterations required for conversion of primary human fibroblasts to cancer cells in the subrenal capsule array. Neoplasia 2005; 7: 585-93.

[18] Beerenwinkel N, Antal T, Dingli D, Traulsen A, Kinzler KW, Velculescu VE, Vogelstein B, Nowak MA. Genetic progression and the waiting time to cancer. PLoS Comp Biol 2007; 3: 2239-46.

[19] Hill AB. The environment and disease: association or causation? Proc Royal Soc Med 1965; 58: 295-300

[20] Morabia A. On the origin of Hill's causal criteria. Epidemiology 1991; 2: 367-9.

[21] Walker L. Koch's postulates and infectious proteins. Acta Neuropathol 2006; 112: -1–4.

[22] Ioannides JPA, Bofetta P, Little J, Little J, O'Brien TR, Uitterlinden AG, Vineis P, Balding DJ, Chokkalingam A, Dolan SM, Flanders WD, Higgins JPT, Mccarthy MI, Mcdermott DH, Page GP, Rebbeck TR, Seminara D, Khoury MJ. Assessment of cumulative evidence on genetic associations: interim guidelines. Int J Epi 2008; 37: 120-32.

[23] Thelle DS. STROBE and STREGA: instruments for improving transparency and quality of reporting scientific results. Eur J Epidemiol 2009; 24: 7-8.

[24] Armitage P. Multistage models of carcinogenesis. Env Health Perspective 1985; 63: 195-201.

[25] Hornsby C, Page KM, Tomlinson IPM. What can we learn from the population incidence of cancer? Armitage and Doll revisited. Lancet Oncology 2007; 8: 1030-8

[26] Colditz GA, Rosner BA. What can be learnt from models of incidence rates? Breast Cancer Res 2006; 8: 208.

[27] Meza R, Hazelton WD, Colditz GA, Moolgavkar SH. Analysis of lung cancer incidence in the nurses' health and health professionals' follow-up studies using a multistage model. Cancer Causes Control 2008; 19: 317-28.

[28] Webb PM, Merritt MA, Boyle GM, Green AC. Microarrays and epidemiology: not the beginning of the end but the end of the beginning. Cancer Epi Biomarker Prev 2007; 16: 637-8.

[29] Hung RJ, McKay JD, Gaborieau V, Boffetta P,

Hashibe M, Zaridze D, Mukeria A, Szeszenia-Dabrowska N, Lissowska J, Rudnai P, Fabianova E, Mates D, Bencko V, Foretova L, Janout V, Chen C, Goodman G, Field JK, Liloglou T, Xinarianos G, Cassidy A, McLaughlin J, Liu G, Narod S, Krokan HE, Skorpen F, Elvestad MB, Hveem K, Vatten L, Linseisen J, Clavel-Chapelon F, Vineis P, Bueno-de-Mesquita HB, Lund E, Martinez C, Bingham S, Rasmuson T, Hainaut P, Riboli E, Ahrens W, Benhamou S, Lagiou P, Trichopoulos D, Holcátová I, Merletti F, Kjaerheim K, Agudo A, Macfarlane G, Talamini R, Simonato L, Lowry R, Conway DI, Znaor A, Healy C, Zelenika D, Boland A, Delepine M, Foglio M, Lechner D, Matsuda F, Blanche H, Gut I, Heath S, Lathrop M, Brennan P. A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. Nature 2008; 452: 633-7

[30] Day NE, Brown CC. Multistage models and primary prevention of cancer. J Natl Cancer Inst 1980; 64: 977-89.

[31] Haiman CA, Dossus L, Setiawan VW, Stram DO, Dunning AM, Thomas G, Thun MJ, Albanes D, Altshuler D, Ardanaz E, Boeing H, Buring J, Burtt N, Calle EE, Chanock S, Clavel-Chapelon F, Colditz GA, Cox DG, Feigelson HS, Hankinson SE, Hayes RB, Henderson BE, Hirschhorn JN, Hoover R, Hunter DJ, Kaaks R, Kolonel LN, Le Marchand L, Lenner P, Lund E, Panico S, Peeters PH, Pike MC, Riboli E, Tjonneland A, Travis R, Trichopoulos D, Wacholder S, Ziegler RG. Genetic variation at the CYP19A1 locus predicting circulating oestrogen levels but not breast cancer risk in postmenopausal women. Cancer Res 2007; 67: 1-5.

[32] Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM. Finding the missing heritability of complex diseases. Nature 2009; 461: 747-753.

[33] Potter JD. Epidemiology, cancer genetics and microarrays: making correct inferences, using appropriate designs. Trends Genetics 2003; 19: 690-95.

[34] Potter JD. Epidemiology informing clinical practice; from bills of mortality to population laboratories. Nat Clin Pract Oncol 2005; 2: 625-34.

[35] Dumeaux V, Børresen-Dale AL, Frantzen JO, Kumle M, Kristensen VN, Lund E. Cohort profile: The Norwegian Women and Cancer study – NOWAC – Kvinner og kreft. Int J Epidemio 2008; 37: 36-41.

[36] 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.

[37] Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, Sweet-Cordero A, Ebert BL, Mak RH, Ferrando AA, Downing JR, Jacks T, Horvitz HR, Golub TR. MicroRNA expression profiles classify human cancers. Nature 2005; 435: 834-8.

[38] Amundson SA, Do KT, Shahab S, Bittner M, Meltzer P, Trent J, Fornace AJ Jr. Identification of potential mRNA biomarkers in peripheral blood lymphocytes for human exposure to ionizing radiation. Radiat Res 2000; 154: 342-346.

[39] Wang Z, Neuburg D, Li C, Su L, Kim JY, Chen JC, Christiani DC. Global gene expression profiling in whole-blood samples from individuals exposed to metal fumes. Environ Health Perspect 2005; 113: 233-41.

[40] Lampe JW, Stepaniants SB, Mao M, Radich JP, Dai H, Linsley PS, Friend SH, Potter JD. Signature of environmental exposures using peripheral leukocyte gene expression: tobacco smoke. Cancer Epidemiol Biomarkers Prev 2004; 13: 445-53.

[41] Dumeaux V, Olsen SK, Paulssen RH, Børresen-Dale AL, Lund E. Deciphering blood gene expression variation – The postgenome NOWAC study. PLoS Genetics 2009 (in press).

[42] Terasaka S, Aita Y, Inoue A, Hayashi S, Nishigaki M, Aoyagi K, Sasaki H, Wada-Kiyama Y, Sakuma Y, Akaba S, Tanaka J, Sone H, Yonemoto J, Tanji M, Kiyama R. Using a customized DNA microarray for expression profiling of the estrogen-responsive genes to evaluate estrogen activity among natural estrogens and industrial chemicals. Environ Health Perspect 2004; 112: 773-81.

[43] Dumeaux V, Johansen J, Børresen-Dale AL, Lund E. Gene expression profiling of whole-blood samples from women exposed to hormone replacement therapy. Mol Cancer Ther 2006; 5: 868-876.

[44] Forrest MS, Lan Q, Hubbard AE, Zhang L, Vermeulen R, Zhao X, Li G, Wu YY, Shen M, Yin S, Chanock SJ, Rothman N, Smith MT. Discovery of novel biomarkers by microarray analysis of peripheral blood mononuclear cell gene expression in benzene-exposed workers. Environ Health Perspect 2005; 113: 801-7.

[45] Davis CD and Ross SA. Evidence for dietary regulation of microRNA expression in cancer cells. Nutrition Rev 2008; 66: 477-82.

[46] Konstantinidou V, Khymenets O, Covas MI, de la Torre R, Muñoz-Aguayo D, Anglada R, Farré M, Fito M. Time course of changes in the expression of insulin sensitivity-related genes after an acute load of virgin olive oil. OMICS 2009; 13: 431-438.

[47] Watkins NA, Gusnanto A, de Bono B, De S, Miranda-Saavedra D, Hardie DL, Angenent WG, Attwood AP, Ellis PD, Erber W, Foad NS, Garner SF, Isacke CM, Jolley J, Koch K, Macaulay IC, Morley SL, Rendon A, Rice KM, Taylor N, Thijssen-Timmer DC, Tijssen MR, van der Schoot CE, Wernisch L, Winzer T, Dudbridge F, Buckley CD,

Langford CF, Teichmann S, Göttgens B, Ouwehand WH; Bloodomics Consortium. A HaemAtlas: characterizing gene expression in differentiated human blood cells. Blood 2009; 113: 1-9.
[48] Mohr S and Liew CC. The peripheral-blood transcriptome: new insights into disease and risk assessment. Trends Mol Med 2007; 13: 422-32.
[49] Aarøe J, Lindahl T, Dumeaux V, Sæbø S, Tobin D, Hagen N, Skaane P, Lönneborg A, Sharma P, Børresen-Dale AL. Gene expression profiling of peripheral blood cells for early detection of breast cancer. Breast Cancer Research 2009; 12: R7.
[50] Han M, Liew CT, Zhang HW, Chao S, Zheng R, Yip KT, Song ZY, Li HM, Geng XP, Zhu LX, Lin JJ, Marshall KW, Liew CC. Novel blood-based, five-gene biomarker set for the detection of colorectal cancer. Clin Cancer Res 2008; 14: 455-60.
[51] Solmi R, Ugolini G, Rosati G, Zanotti S, Lauriola M, Montroni I, del Governatore M, Caira A, Taffurelli M, Santini D, Coppola D, Guidotti L, Carinci P, Strippoli P. Microarray-based identification and RT-PCR test screening for epithelial-specific mRNAs in peripheral blood of patients with colon cancer. BMC Cancer 2006; 6: 250.
[52] Burczynski ME, Peterson RL, Twine NC, Zuberek KA, Brodeur BJ, Casciotti L, Maganti V, Reddy PS, Strahs A, Immermann F, Spinelli W, Schwertschlag U, Slager AM, Cotreau MM, Dorner AJ Molecular classification of Crohn's disease and ulcerative colitis patients using transcriptional profiles in peripheral blood mononuclear cells. J Mol Diagn 2006; 8: 51-61
[53] Osman I, Bajorin DF, Sun TT, Zhong H, Douglas D, Scattergood J, Zheng R, Han M, Marshall KW, Liew CC Novel blood biomarkers of human urinary bladder cancer. Clin Cancer Res 2006; 12; 3374-3380
[54] Li Y, Elashoff D, Oh M, Sinha U, St John MA, Zhou X, Abemayor E, Wong DT. Serum circulating human mRNA profiling and its utility for oral cancer detection. J Clin Oncol 24(11), 1754-1760 (2006
[55] Emilsson V, Thorleifsson G, Zhang B, Leonardson AS, Zink F, Zhu J, Carlson S, Helgason A, Walters GB, Gunnarsdottir S, Mouy M, Steinthorsdottir V, Eiriksdottir GH, Bjornsdottir g, Reynisdottir I, Gudbjartson D, Helgadottir A, Jonasdottir A, Jonasdottir A, Styrkarsdottir S, Magnusson KP, Stefansson H, Fossdal R, Kristjansson K, Gislason HG, Stefansson T, Leifsson BG, Thorsteinsdottir U, lamb JR, Gulscher JR, Reitman ML, Kong A, Schadt EE, Stefanson K. Genetics of gene expression and its effect on disease. Nature 2008; 452: 423-8.
[56] Liew CC, Ma J, Tang HC, Zheng R, Dempsey AA. The peripheral blood transcriptome dynamically reflects system wide biology: a potential diagnostic tool. J Lab Clin Med 2006; 147: 126-132.
[57] Kuukasjärvi T, Tanner M, Pennanen S, Karhu R, Kallioniemi OP, Isola J. Genetic changes in intraductal breast cancer detected by comparative genomic hybridization. Am J Path 1997; 150: 1465-71.
[58] Fujii H, Marsh C, Cairns P, Sidransky D, Gabrielson E. Genetic divergence in the clonal evolution of breast cancer. Cancer Res 1996; 56: 1493-96.
[59] Amoli MM, Carthy D, Platt H, Ollier WE. EBV Immortalization of human B lymphocytes separated from small volumes of cryo-preserved whole blood. Int J Epidemiol 2008; 37: i41-45.
[60] Salway F, Day PJR, Ollier WER, Peakman TC. Levels of 5' RNA tags in plasma and buffy coat from EDTA blood increase with time. Int J Epi 2008; 37: i11-15.
[61] Lodes MJ, Caraballo M, Suciu D, Munro S, kumar A, Anderson B. Detection of cancer with serum miRNAs on an oligonucleotide microarray. Plos One 4(7): e6229.doi:10.1371/journal.pone.0006229.
[62] Vineis P, Khan AE, Vlaanderen J, vermeulen R. The impact of new research tecknologies on our understanding of environmental causes of disease: the concept of clinical vulnerability. Env Health 2009; 8: 1-10.
[63] Hanash SM, Pitteri SJ, Faca VM. Mining the plasma proteome for cancer biomarkers. Nature 2008; 452: 571-57
[64] Collins FS, Manolio TA. Necessary but not sufficient. Nature 2007; 445: 259.
[65] Willett WC, Blot WJ, Colditz GA, Folsom AR, Henderson BE, Stampfer MJ. Merging and emerging cohorts : not worth the wait. Nature 2007: 445; 257-258.
[66] Colditz GA, Sellers TA, Trapido E. Epidemiology – identifying the causes and preventability of cancer? Nature Rev 2006; 75-83.
[67] Hazelton WD, Clements MS, Moolgavkar SH. Multistage carcinogenesis and lung cancer mortality in three cohorts. CEBP 2005; 14; 1171-81.
[68] Hanahan D, Weinberg RA. The hallmarks of cancer. Cell 2000: 100; 57-70.
[69] Klein G. Toward a genetics of cancer resistance. PNAS 2009; 106: 859-863.
[70] Nicholson JK, Holmes E, Lindon JC, Wilson ID. The challenges of modelling mammalian biocomplexity. Nature Biotech 2004; 22; 1268-74.